



2021 Keystone Technical Report: Algebra I, Biology, and Literature

Provided by Data Recognition Corporation

This document has been formatted to be ADA compliant.

TABLE OF CONTENTS

- Preface: An Overview of the Assessments7**
 - The Keystone Exams from 2008 to Present7
 - Assessment Activities Occurring from 2010 to Present7
- Chapter One: Background of the Keystone Exams14**
 - Assessment History in Pennsylvania14
 - The Keystone Exams14
- Chapter Two: Test Development Overview of the Keystone Exams.18**
 - Keystone Blueprint/Assessment Anchors and Eligible Content.18
 - High-Level Test Design Considerations20
 - Online Testing Design Considerations20
 - Algebra I21
 - Biology23
 - Literature.24
 - Literature Passages25
- Chapter Three: Item and Test Development Processes27**
 - General Keystone Test Development Processes27
 - General Test Definition28
 - Algebra I Test Definitions28
 - Biology Test Definitions30
 - Literature Test Definitions32
 - Item Development Considerations34
 - Item and Test Development Cycle36
 - General Item and Test Development Process39
- Chapter Four: Universal Design Procedures Applied to the Keystone Exams Test Development Process .44**
 - Universal Design44
 - Elements of Universally Designed Assessments44
 - Guidelines for Universally Designed Items45
 - Item Development.46
 - Item Format47
 - Assessment Accommodations48
- Chapter Five: Embedded Field Test.49**
 - Field Test Overview.49
 - Classical Item Analysis50
 - Criteria for Identifying Items53
 - Review of Items with Data57

Chapter Six: Operational Forms Construction for 2021 Administrations	58
Final Selection of Items and Keystone Forms Construction	58
Special Forms Used with the Operational 2020 Keystone Exams	59
Chapter Seven: Test Administration Procedures	61
Sections, Sessions, Timing, and Layout of the Keystone Exams	61
Sections and Sessions	61
Timing	62
Layout	63
Shipping, Packaging, and Delivery of Materials	64
Chapter Eight: Processing and Scoring	67
Receipt of Materials	67
Scanning of Materials	68
Materials Storage	70
Online Testing	70
Scoring Multiple-Choice Items	72
Handscoring Open-Ended Items	72
Rangefinding	72
Rater Recruitment/Qualifications	73
Leadership Recruitment/Qualifications	73
Training	73
Handscoring Process	75
Handscoring Validity Process	75
Quality Control	76
Chapter Nine: Description of Data Sources	81
Student Filtering Criteria	81
Key Verification Data	82
Pre-Equating Verification	82
Final Data	82
Spiraling of Forms	82
Chapter Ten: Summary Demographic and Accommodation Data for Spring 2021 Keystone Exams	85
Assessed Students	85
Reasons for Student Non-Assessment	86
Demographic Characteristics of Students Receiving Test Scores	88
Test Accommodations Provided	97
Glossary of Accommodation Terms	111

Chapter Eleven: Classical Item Statistics	114
Item-Level Statistics	114
Item Difficulty	114
Item Discrimination	115
Scatter Plots of Item Discrimination and Difficulty	116
Observations and Interpretations	119
Chapter Twelve: Rasch Item Calibration	121
Description of the Rasch Model	121
Checking Rasch Assumptions	122
Rasch Item Statistics	126
Chapter Thirteen: Standard Setting	132
Standard Setting and Performance Level Descriptors	132
Development Overview for the Performance Level Descriptors	132
Performance Level Descriptors Meeting 1	133
Performance Level Descriptors Meeting 2	135
Standard Setting	137
Chapter Fourteen: Scaling	151
Raw Scores to Rasch Ability Estimates	151
Rasch Ability Estimates to Scaled Scores	152
Raw-to-Scaled-Score Tables	153
Chapter Fifteen: Equating	154
Pre- vs. Post-Equating	154
Equating Design for Keystone Exams	155
Scale Linking	155
Pre-Equating Verification	156
Performance Level Classification	159
Scale Stability and Maintenance	160
Test Characteristic Curves and Logit Plots	160
Chapter Sixteen: Scores and Score Reports	166
Scoring	166
Description of Total-Test Scores	166
Highest Total Test Scaled Score to Date	167
Description of Module Scores	168
Appropriate Score Use	169
Cautions for Score Use	169
Report Development	170
Reports	170

Chapter Seventeen: Operational Test Statistics	175
Performance Level Statistics	175
Scaled Scores	177
Raw Scores	177
Chapter Eighteen: Reliability	184
Reliability Indices	185
Coefficient Alpha	185
Further Interpretations	186
Standard Error of Measurement	189
Results and Observations	190
Rasch Conditional Standard Errors of Measurement	191
Results and Observations	192
Reliability of Performance Level Classification Decisions	195
Rater Agreement	197
Chapter Nineteen: Validity	201
Purposes and Intended Uses of the Keystone Exams	201
Evidence Based on Test Content	201
Evidence Based on Response Process	203
Evidence Based on Internal Structure	203
Evidence Based on Relationships with Other Variables	207
Evidence Based on Consequences of Tests	208
Evidence Related to the Use of Rasch Model	209
Validity Evidence Summary	209
Appendix A: Understanding Depth of Knowledge and Cognitive Complexity	210
Algebra I Depth of Knowledge	211
Biology Depth of Knowledge	213
Literature Depth of Knowledge	216
Appendix B: General Scoring Guidelines	218
Algebra I	218
Biology	219
Literature	220
Appendix C: Item and Test Development Process for the Keystone Exams	221
Appendix D: Item and data Review Card examples	224
Item Review Card Example	224
Data Review Card Example	225
Appendix E: Item Rating Sheet	227
Appendix F: Tally Sheets	231
Appendix G: Keystone Exams Module Layout Plans	249

Appendix H: Mean Raw Scores by Form.	250
Algebra I: Spring	251
Biology: Spring	252
Literature: Spring	253
Appendix I: Demographic and Accommodation Data	254
Winter	254
Summer	276
Appendix J: Item Statistics.	300
Multiple-Choice Items.	301
Constructed-Response Items.	337
Appendix K: Raw-To-Scale Score Conversion Tables	340
Winter	341
Spring	347
Summer	353
Appendix L: Pre-Equating Verification Results.	359
Winter	359
Spring	375
Appendix M: Reliabilities	390
Appendix N: Opportunity to Learn Survey	400
Summary.	400
Results by Subject and Mode	401
Results by Instructional Mode	404
Appendix O: Examining Student Performance in Spring 2019 and Spring 2021 Using Propensity Score Matching	407
Summary.	407
Introduction.	407
Methods	409
Analyses	409
Results	411
Discussion	415
Supplemental Information	416
Group Comparison on Covariates Before Propensity Score Matching	419
Propensity Score Matching Analyses by Subgroups	421
References.	431

PREFACE: AN OVERVIEW OF THE ASSESSMENTS

THE KEYSTONE EXAMS FROM 2008 TO PRESENT

COMPREHENSIVE GRADUATION COMPETENCY ASSESSMENT PROGRAM

In 2008, the Commonwealth of Pennsylvania initiated a comprehensive graduation competency assessment program. The goals of this program include the following:

- To provide a system that is aligned, focused, standards-based, accurate, universally applicable, and publicly accessible
- To develop, produce, distribute, administer (both online and in paper-and-pencil), collect, score, analyze, track, and report results of graduation competency assessments for ten high school-level content areas: Algebra I, Algebra II, Biology, Chemistry, Civics and Government, English Composition, Geometry, Literature, U.S. History, and World History, with each area or course comprised of modules containing unique content
- To provide graduation competency testing opportunities for students three times each school year—spring, summer, and fall—with students permitted to retake modules until proficiency is achieved on each module
- To report graduation competency results under accelerated timelines
- To ensure validity and reliability of the assessment systems through technically sound test development and psychometric practices, detailed statistical analyses and research studies, and well-documented processes and quality procedures

The Keystone Exams, as the graduation competency assessments are named, are just one component of Pennsylvania's system of high school graduation requirements. Keystone Exams are designed to help school districts guide students toward meeting state standards—standards aligned with expectations for success in college and the workplace. In order to receive a diploma, students are also required to meet local district credit and attendance requirements and to complete a culminating project, along with any additional district requirements.

For graduating classes, students are to demonstrate successful completion of secondary-level course work in Algebra I, Biology, and Literature, in which the Keystone Exam served as the final course exam. Students' Keystone Exam scores count for at least one-third of the final course grades.

Based upon Chapter 4 regulations, each Keystone Exam is designed in modules that reflect distinct, related academic content common to the traditional progression of course work. Students who do not score Proficient or above on a Keystone Exam module may choose to complete a project-based assessment for that module based upon other specific requirements.

ASSESSMENT ACTIVITIES OCCURRING FROM 2010 TO PRESENT

The first assessment activities took place in the 2010–2011 school year. Prior to November 2010, there were no Keystone Exams assessment events. The table below outlines the field tests and operational exams administered during the 2010–11 school year.

Following the development of Assessment Anchors and Eligible Content, exams were developed for the initial field test in 2010 and were subsequently administered as operational exams in 2011. Additional exams, which were based on the Assessment Anchors and Eligible Content developed in 2009 and 2010, were developed for the initial field test in 2011. Detailed information about the operational exam activities that occurred during the 2010–2011 school year is in the *Keystone Exams Spring 2011 Algebra I, Biology, and Literature Technical Report*.

Field Test and Operational Exams during the 2015–16 School Year

Exam	Assessment Activity	Date
Algebra I	Initial Stand-Alone Field Test	Fall 2010 (November)
Algebra I	Inaugural Operational Exam Administration	Spring 2011 (May)
Algebra II	Initial Stand-Alone Field Test	Spring 2011 (May)
Biology	Initial Stand-Alone Field Test	Fall 2010 (November)
Biology	Inaugural Operational Exam Administration	Spring 2011 (May)
English Composition	Initial Stand-Alone Field Test	Spring 2011 (May)
Geometry	Initial Stand-Alone Field Test	Spring 2011 (May)
Literature	Initial Stand-Alone Field Test	Fall 2010 (November)
Literature	Inaugural Operational Exam Administration	Spring 2011 (May)

Following a one-year program hiatus in 2012, the field test items embedded in the Spring 2011 operational forms were used to construct the forms for the next four administrations (spring, summer, winter, and possible breach). The table below outlines exams administered during the 2012–13 school year. Detailed information about the operational exam activities that occurred during the 2012–2013 school year is in the *Keystone Exams Spring 2013 Algebra I, Biology, and Literature Technical Report*.

Operational Exams during the 2012–13 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2012/2013 (December–January)
Algebra I	Operational Exam Administration	Spring 2013 (May)
Algebra I	Operational Retest Exam Administration	Summer 2013 (August)
Biology	Operational Retest Exam Administration	Winter 2012/2013 (December–January)
Biology	Operational Exam Administration	Spring 2013 (May)
Biology	Operational Retest Exam Administration	Summer 2013 (August)
Literature	Operational Retest Exam Administration	Winter 2012/2013 (December–January)
Literature	Operational Exam Administration	Spring 2013 (May)
Literature	Operational Retest Exam Administration	Summer 2013 (August)

Some of the field test items embedded in the Spring 2013 operational forms were used to construct the forms for the next four administrations (spring, summer, winter, and possible breach). The core items on the 2013–2014 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2013–14 school year. Detailed information about the operational exam activities that occurred during the 2013–2014 school year is in the *Keystone Exams Spring 2014 Algebra I, Biology, and Literature Technical Report*.

Operational Exams during the 2013–14 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2013/2014 (December–January)
Algebra I	Operational Exam Administration	Spring 2014 (May)
Algebra I	Operational Retest Exam Administration	Summer 2014 (August)
Biology	Operational Retest Exam Administration	Winter 2013/2014 (December–January)
Biology	Operational Exam Administration	Spring 2014 (May)
Biology	Operational Retest Exam Administration	Summer 2014 (August)
Literature	Operational Retest Exam Administration	Winter 2013/2014 (December–January)
Literature	Operational Exam Administration	Spring 2014 (May)
Literature	Operational Retest Exam Administration	Summer 2014 (August)

Some of the field test items embedded in the Spring 2014 operational forms were used to construct the forms for the next year’s administrations (spring, summer, winter). The core items on the 2014–2015 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2014–15 school year.

Operational Exams during the 2014–15 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2014/2015 (December–January)
Algebra I	Operational Exam Administration	Spring 2015 (May)
Algebra I	Operational Retest Exam Administration	Summer 2015 (August)
Biology	Operational Retest Exam Administration	Winter 2014/2015 (December–January)
Biology	Operational Exam Administration	Spring 2015 (May)
Biology	Operational Retest Exam Administration	Summer 2015 (August)
Literature	Operational Retest Exam Administration	Winter 2014/2015 (December–January)
Literature	Operational Exam Administration	Spring 2015 (May)
Literature	Operational Retest Exam Administration	Summer 2015 (August)

Some of the field test items embedded in the Spring 2015 operational forms were used to construct the forms for the next year’s administrations (spring, summer, winter). The core items on the 2015–2016 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2015–16 school year.

Operational Exams during the 2015–16 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2015/2016 (December–January)
Algebra I	Operational Exam Administration	Spring 2016 (May)
Algebra I	Operational Retest Exam Administration	Summer 2016 (August)
Biology	Operational Retest Exam Administration	Winter 2015/2016 (December–January)
Biology	Operational Exam Administration	Spring 2016 (May)
Biology	Operational Retest Exam Administration	Summer 2016 (August)
Literature	Operational Retest Exam Administration	Winter 2015/2016 (December–January)
Literature	Operational Exam Administration	Spring 2016 (May)
Literature	Operational Retest Exam Administration	Summer 2016 (August)

Some of the field test items embedded in the Spring 2016 operational forms were used to construct the forms for the next year’s administrations (spring, summer, winter). The core items on the 2016–2017 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2016-17 school year.

Operational Exams during the 2016–17 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2016/2017 (December–January)
Algebra I	Operational Exam Administration	Spring 2017 (May)
Algebra I	Operational Retest Exam Administration	Summer 2017 (July–August)
Biology	Operational Retest Exam Administration	Winter 2016/2017 (December–January)
Biology	Operational Exam Administration	Spring 2017 (May)
Biology	Operational Retest Exam Administration	Summer 2017 (July–August)
Literature	Operational Retest Exam Administration	Winter 2016/2017 (December–January)
Literature	Operational Exam Administration	Spring 2017 (May)
Literature	Operational Retest Exam Administration	Summer 2017 (July–August)

Some of the field test items embedded in the spring 2017 operational forms were used to construct the forms for the next year’s administrations (spring, summer, winter). The core items on the 2017–2018 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2017–2018 school year.

Operational Exams during the 2017–18 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2017/2018 (December–January)
Algebra I	Operational Exam Administration	Spring 2018 (May)
Algebra I	Operational Retest Exam Administration	Summer 2018 (July–August)
Biology	Operational Retest Exam Administration	Winter 2017/2018 (December–January)
Biology	Operational Exam Administration	Spring 2018 (May)
Biology	Operational Retest Exam Administration	Summer 2018 (July–August)
Literature	Operational Retest Exam Administration	Winter 2017/2018 (December–January)
Literature	Operational Exam Administration	Spring 2018 (May)
Literature	Operational Retest Exam Administration	Summer 2018 (July–August)

Some of the field test items embedded in the spring 2018 operational forms were used to construct the forms for the next year’s administrations (spring, summer, winter). The core items on the 2018–2019 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2018–2019 school year.

Operational Exams during the 2018–19 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2018/2019 (December–January)
Algebra I	Operational Exam Administration	Spring 2019 (May)
Algebra I	Operational Retest Exam Administration	Summer 2019 (July–August)
Biology	Operational Retest Exam Administration	Winter 2018/2019 (December–January)
Biology	Operational Exam Administration	Spring 2019 (May)
Biology	Operational Retest Exam Administration	Summer 2019 (July–August)
Literature	Operational Retest Exam Administration	Winter 2018/2019 (December–January)
Literature	Operational Exam Administration	Spring 2019 (May)
Literature	Operational Retest Exam Administration	Summer 2019 (July–August)

Some of the field test items embedded in the spring 2019 operational forms were used to construct the forms for the next year’s administrations (summer, winter). The core items on the 2019–2020 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2019–2020 school year.

The spring administration of the Keystone exams was cancelled due to Coronavirus (COVID-19) mitigation efforts. On March 27, 2020, the U.S. Department of Education (USDE) approved Pennsylvania’s request to waive federal assessment requirements for the 2019–20 school year, along with accountability and certain reporting requirements based on data derived from the 2019–20 school year.

The waiver’s coverage of assessment requirements applies to the cohort of 2019–2020 test takers who were scheduled to take one or more Keystone Exams in the spring of 2020 (Spring 2019–2020 Cohort). Accordingly, the Federal government is not requiring any student enrolled in a Keystone Exam trigger course (Algebra I, Biology, English Literature) during the spring of the 2019–2020 school year, regardless of their current grade level or expected graduation date, to take the associated Keystone Exam(s) once schools reopen and federal assessment requirements resume. Because Keystone Exams scores are “banked” for accountability purposes and applied in grade 11, this waiver will affect students enrolled in grades other than grades 11 and 12, including as low as grade 6.

Students who took the Keystone Exams during the 2019–2020 school year prior to the pandemic (i.e., summer 2019 or winter 2020) may use those results to satisfy their individual Act 158 requirements and for any local purposes. However, because 2019–2020 school year cohort results will not be comparable with prior or future years results, these results will not be factored into future accountability determinations such as cyclical Comprehensive Support and Improvement (CSI) and Additional Targeted Support and Improvement determinations (A-TSI) (scores of students who took the Keystone Exams in the summer of 2019 or the winter of 2019–2020 as a re-tester will be attributed for the purposes of federal accountability and reporting.)

For this reason, these partial results were not reported through Future Ready in fall 2020.

Operational Exams during the 2019–20 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2019/2020 (December–January)
Algebra I	Operational Exam Administration	Cancelled due to Coronavirus (COVID-19) mitigation efforts.
Algebra I	Operational Retest Exam Administration	Autumn 2020 (September–October) *previously a summer administration
Biology	Operational Retest Exam Administration	Winter 2019/2020 (December–January)
Biology	Operational Exam Administration	Cancelled due to Coronavirus (COVID-19) mitigation efforts.
Biology	Operational Retest Exam Administration	Autumn 2020 (September–October) *previously a summer administration
Literature	Operational Retest Exam Administration	Winter 2019/2020 (December–January)
Literature	Operational Exam Administration	Cancelled due to Coronavirus (COVID-19) mitigation efforts.
Literature	Operational Retest Exam Administration	Autumn 2020 (September–October) *previously a summer administration

The core items on the 2019–2020 forms also consisted of items that appeared on the core forms of the administrations two years prior. More information on these core-to-core overlap items can be found in Chapter Three. The table below outlines exams administered during the 2020–2021 school year.

Operational Exams during the 2020–21 School Year

Exam	Assessment Activity	Date
Algebra I	Operational Retest Exam Administration	Winter 2020/2021 (December–January)
Algebra I	Operational Exam Administration	Spring 2021 (May–September)
Biology	Operational Retest Exam Administration	Winter 2020/2021 (December–January)
Biology	Operational Exam Administration	Spring 2021 (May–September)
Literature	Operational Retest Exam Administration	Winter 2020/2021 (December–January)
Literature	Operational Exam Administration	Spring 2021 (May–September)

The core items on the 2020–2021 forms also consisted of items that appeared on the core forms of prior administrations. More information on these core-to-core overlap items can be found in Chapter Three.

CHAPTER ONE: BACKGROUND OF THE KEYSTONE EXAMS

This brief overview of the Pennsylvania Keystone Exams summarizes the history of the program’s development process, intent and purpose, and recent changes.

ASSESSMENT HISTORY IN PENNSYLVANIA

Pennsylvania’s involvement in statewide assessment actually began in the 1969–1970 school year with a purely school-based assessment known as *Educational Quality Assessment (EQA)*, which continued through the 1987–1988 school year. A state-mandated student competency testing program called *Testing for Essential Learning and Literacy Skills (TELLS)* also operated from the school years of 1984–1985 through 1990–1991. Also in 1990, the state initiated an on-demand writing assessment.

The Pennsylvania System of School Assessment (PSSA) program was instituted in 1992 as a school evaluation model with reporting at the school level only. The PSSA initially measured performance in the content areas of mathematics and reading at grades 5, 8, and 11, and in writing at grades 6 and 9. Starting in 1994, as part of Chapter 5 regulations, the PSSA added student-level reports. In 1999, as part of Chapter 4 regulations, the State Board of Education adopted the Pennsylvania Academic Standards for mathematics and for reading, writing, speaking, and listening. Proficiency levels for Advanced, Proficient, Basic, and Below Basic were defined in 2000. In 2001 and 2004, the reading and mathematics assessments underwent various content enhancements to improve alignment to the 1999 Academic Standards. Grade 11 was added to the writing assessment in 2001. Then, in 2004–2005, the PSSA Assessment Anchors and Eligible Content were developed to clarify content structure and improve articulation between assessment and instruction. In addition, in 2005, the grade 6 and 9 writing assessments were moved to grades 5 and 8. By 2006, the operational mathematics and reading assessments incorporated grades 3 through 8 and 11. In 2007, the PSSA and the PSSA Assessment Anchors and Eligible Content underwent additional content enhancements. In 2008, science was added to the PSSA as an operational assessment. Starting with the 2013 field test, PSSA began a multiyear transition to a new set of standards called the Pennsylvania Core Standards. Detailed information about the operational exam activities that occurred during the 2013–2014 school year is in the *2014 PSSA Technical Report*.

THE KEYSTONE EXAMS

In 2008, the Commonwealth of Pennsylvania initiated a comprehensive graduation competency assessment program. As a key piece of this program, the Keystone Exams are designed to assess proficiency in various subject areas, including Algebra I, Algebra II, Biology, Chemistry, Civics and Government, English Composition, Geometry, Literature, U.S. History, and World History. The Keystone Exams are just one component of Pennsylvania’s high school graduation requirements. Students must also earn state-specified credits, fulfill the state’s service learning and attendance requirements, and complete any additional local school system requirements to receive a Pennsylvania high school diploma.

The stated goals of the Keystone program are to

- provide for a system that is aligned, focused, standards-based, accurate, universally applicable, and publicly accessible.
- develop, produce, distribute, administer (both online and in paper-and-pencil), collect, score, analyze, track, and report results of graduation competency assessments for ten high school-level content areas: Algebra I, Algebra II, Biology, Chemistry, Civics and Government, English Composition, Geometry, Literature, U.S. History, and World History, with each area or course composed of modules containing unique content.
- provide graduation competency testing opportunities for students three times each school year—spring, summer, and fall—with students permitted to retake modules until proficiency is achieved in each module.
- report graduation competency results under accelerated timelines.
- ensure validity and reliability of the assessment systems through technically sound test development and psychometric practices, detailed statistical analyses and research studies, and well-documented processes and quality procedures.

GRADUATION REQUIREMENTS AND THE KEYSTONE EXAMS

Based upon Chapter 4 regulations, each Keystone Exam is designed in modules that reflect distinct, related academic content common to the traditional progression of coursework. Students who do not score Proficient or above on a Keystone Exam module may choose to complete a project-based assessment for that module based on the requirements detailed below.

If a student is unable to meet the requirements in § 4.24(b)(1)(iv)(A) (relating to high school graduation requirements) after two attempts on a Keystone Exam, the student may supplement a Keystone Exam score with satisfactory completion of a project-based assessment. Points earned through satisfactory performance on one or more project modules related to the Keystone Exam module or modules that the student did not pass shall be added to the student's highest Keystone Exam score.

A student may qualify to participate in one or more project-based assessments if the student has met all of the following conditions:

1. The student has taken the course.
2. The student was unsuccessful in achieving a score of Proficient on the Keystone Exam after at least two attempts.
3. The student has met the district's attendance requirements for the course.
4. The student has participated in a satisfactory manner in supplemental instructional services under § 4.24(i).

KEYSTONE ASSESSMENT ANCHORS AND ELIGIBLE CONTENT

In 2009, the state initiated development of test designs and test blueprints for the Keystone Exams based on Pennsylvania Keystone Course Standards. Committees of Pennsylvania educators met in 2009, 2010, and 2011 to write, review, and approve Assessment Anchors and Eligible Content statements and sample exam items. To provide initial focus, each test blueprint committee was presented with materials specific to the exam in question, including a basic blueprint structure, the Pennsylvania State Standards, and draft Eligible Content statements based on the standards. The results from the initial committee work were evaluated by national, state, and local subject experts, and following revisions, they were ultimately validated by another committee of Pennsylvania educators. Following committee approval, the Keystone Assessment Anchors and Eligible Content statements for literacy, mathematics, and science were approved by the State Board of Education in September 2010.

- Mathematics
 - The first committee meetings took place in April 2009, where initial drafts of the test blueprints were developed for Algebra I, Algebra II, and Geometry.
 - A follow-up committee meeting for the three mathematics exams was held in August 2009.
- Literacy
 - The first committee meetings took place in April 2009, where initial drafts of the test blueprints were developed for English Composition and Literature.
 - A follow-up committee meeting for the two literacy exams was held in November 2009.
- Science
 - The first committee meetings took place in October 2009, where the initial draft of the test blueprint was developed for Biology.
 - A follow-up committee meeting for Biology was held in January 2010.
 - In addition, in January 2010, the initial draft of the test blueprint was developed for Chemistry.
 - Chemistry was part of a follow-up committee meeting held in late January 2010.

- Social Studies
 - The first committee meetings took place in November 2010, where initial drafts of the test blueprints were developed for Civics and Government, U.S. History, and World History.
 - A follow-up committee meeting for the Civics and Government exam was held in October 2011.
 - A follow-up committee meeting for U.S. History and World History remains unscheduled pending further decisions about the future of these Keystone exams.

WAVE IMPLEMENTATION OF THE EXAMS

The implementation plan for the Keystone Exams envisioned the ten Keystone Exams becoming operational through a series of waves. The initial wave included Algebra I, Biology, and Literature. These first three exams were field tested in fall 2010 and reached operational status with the spring 2011 administration. The second wave included Algebra II, English Composition, and Geometry; these were field tested in spring 2011. English Composition is scheduled to reach operational status at a future date. Civics and Government is projected to reach operational status following English Composition. The implementation of the five remaining courses, Algebra II, Geometry, Chemistry, U.S. History, and World History, is currently unscheduled. The Pennsylvania Department of Education continues to evaluate the implementation schedule. Table 1–1 reflects the implementation plans as of September 2021.

Table 1–1. Keystone Exams Wave Implementation Plan

Wave	Exam(s)	Initial Field Test	First Operational
1	Algebra I, Biology, Literature	Fall 2010	Spring 2011
2	English Composition	Spring 2011	TBD
2	Algebra II, Geometry	Spring 2011	Not Scheduled
3	Civics and Government	TBD	Not Scheduled
4	Chemistry, U.S. History, World History	TBD	Not Scheduled

MODE OF DELIVERY FOR THE EXAMS

One key feature of the Keystone Exams is the dual mode of delivery of the testing materials that is available to districts. In addition to the traditional paper-and-pencil format, the Keystone Exams are also available in a computer-based online format using test-delivery software.

While exam materials are still available in the traditional format (two pieces of exam materials—a test book and a separate answer book [or, in the case of English Composition, a single test/answer book]), districts are given the option to administer the exams using computer-based online testing software instead of the paper-and-pencil format.

For more information about how the online exams were developed in concert with the traditional paper-and-pencil format, see Chapter Three.

MULTIPLE TESTING OPPORTUNITIES

Another key feature of the Keystone Exams is the multiple testing opportunities provided to students. Main administrations in both spring and winter provide options for students completing course work at various times of the year and accommodate both traditional and block scheduling. In addition, a summer retest opportunity is also available. More information about the spring, winter, and summer administrations can be found in Chapter Seven.

PERFORMANCE LEVELS FOR THE KEYSTONE EXAMS

The State Board approved a set of criteria defining Advanced, Proficient, Basic, and Below Basic levels of performance for the Keystone Exams. More information about these Performance Level Descriptors (PLDs) is found in Chapter Thirteen.

OPERATIONAL TEST DESIGN INFORMATION

The test definition of each of the operational Keystone Exams, including information about exam-specific test designs, test blueprints, test layouts, item types, and other exam elements, is detailed in Chapter Three.

CHAPTER TWO: TEST DEVELOPMENT OVERVIEW OF THE KEYSTONE EXAMS

KEYSTONE BLUEPRINT/ASSESSMENT ANCHORS AND ELIGIBLE CONTENT

The Keystone Test Blueprints—known as the Keystone Exams Assessment Anchors and Eligible Content—are based on Pennsylvania Keystone Course Standards and the Pennsylvania Core Standards. Prior to the development of the Assessment Anchors, multiple groups of Pennsylvania educators convened to create a set of standards for each of the Keystone Exams. Derived from a review of existing standards, these Enhanced Standards (Course Standards) focus on what students need to know and be able to do in order to be ready for college and career.

Although the Keystone Course Standards indicate what students should know and be able to do, Assessment Anchors are designed to indicate the parts of the Keystone Course Standards (Instructional Standards) that will be assessed on the Keystone Exams. Based on recommendations from Pennsylvania educators, the Assessment Anchors were designed as a tool to improve the articulation of curricular, instructional, and assessment practices. The Assessment Anchors clarify what is expected and focus the content of the standards into what is assessable on a large-scale exam. The Assessment Anchor documents also serve to communicate Eligible Content—the range of knowledge and skills from which the Keystone Exams are designed.

The Keystone Exams Assessment Anchors and Eligible Content have been designed to hold together, or anchor, the state assessment system and curricular/instructional practices in schools by following these design parameters:

- **Clear:** The Assessment Anchors are easy to read and user friendly; they clearly detail which standards are assessed on the Keystone Exams.
- **Focused:** The Assessment Anchors identify a core set of standards that can be reasonably assessed on a large-scale assessment; this keeps educators from having to guess which standards are critical.
- **Rigorous:** The Assessment Anchors support the rigor of the state standards by assessing higher order and reasoning skills.
- **Manageable:** The Assessment Anchors define the standards in a way that can be easily incorporated into a course to prepare students for success.

The Assessment Anchors and Eligible Content are organized into cohesive blueprints, each structured with a common labeling system. This framework is organized by increasing levels of detail: first, Module (Reporting Category); second, Assessment Anchor; third, Anchor Descriptor; and fourth, Eligible Content statement. The common format of this outline is followed across the Keystone Exams.

A description of each level in the labeling system for the Keystone Exams is as follows:

- **Module:** The Assessment Anchors are organized into two thematic modules for each of the Keystone Exams, and these modules serve as the Reporting Categories for the Keystone Exams. The module title appears at the top of each page in the Assessment Anchor document. The module level is also important because the Keystone Exams are built using a module format, with each of the Keystone Exams divided into two equally sized test modules. Each module is made up of two or more Assessment Anchors.
- **Assessment Anchor:** The Assessment Anchor appears in the shaded bar across the top of each Assessment Anchor table in the Assessment Anchor document. The Assessment Anchors represent categories of subject matter that anchor the content of the Keystone Exams. Each Assessment Anchor is part of a module and has one or more Anchor Descriptors unified under it.
- **Anchor Descriptor:** Below each Assessment Anchor in the Assessment Anchor document is a specific Anchor Descriptor. The Anchor Descriptor level details the scope of content covered by the Assessment Anchor. Each Anchor Descriptor is part of an Assessment Anchor and has one or more Eligible Content unified under it.

- **Eligible Content:** The column to the right of the Anchor Descriptor in the Assessment Anchor document contains the Eligible Content statements. The Eligible Content is the most specific description of the content that is assessed in the Keystone Exams. This level is considered the assessment limit. It helps educators identify the range of content covered on the Keystone Exams.
- **Enhanced Standard:** In the column to the right of each Eligible Content statement is a code representing one or more Enhanced Standards that correlate to the Eligible Content statement. Some Eligible Content statements include annotations that clarify the scope of an Eligible Content.
- **Notes:** There are three types of notes included in the Assessment Anchor document.
 - “e.g.” (“for example”)—sample approach, but not a limit to the Eligible Content
 - “i.e.” (“that is”)—specific limit to the Eligible Content
 - “Note”—content exclusions or definable range of the Eligible Content

The Assessment Anchor’s coding is read like an outline. The coding includes the Subject (Exam), Reporting Category/Module, Assessment Anchor, Anchor Descriptor, and Eligible Content. Each exam has two modules. Each module has two or more Assessment Anchors. Each of the Assessment Anchors has one or more Anchor Descriptors, and each Anchor Descriptor has at least one Eligible Content (generally more than one). The Assessment Anchors form the basis of the test design for the exams undergoing test development. In turn, this hierarchy is the basis for organizing the total module and exam scores (based on the core [common] portions).

Table 2–1. Sample Keystone Assessment Anchor Coding

Sample Code	Subject (Exam)	Reporting Category (Module)	Assessment Anchor (AA)	Anchor Descriptor (AD)	Eligible Content (EC)
A1.1.1.2.1	A1 Algebra I	1 Operations and Linear Equations & Inequalities	1 Linear Equations	2 Write, solve, and/or graph linear equations using various methods.	1 Write, solve, and/or apply a linear equation (including problem situations).
BIO.A.2.1.1	BIO Biology	A Cells and Cell Processes	2 The Chemical Basis for Life	1 Describe how the unique properties of water support life on Earth	1 Describe the unique properties of water and how these properties support life on Earth (e.g., freezing point, high specific heat, cohesion).
L.F.2.4.1	L Literature	F Fiction	2 Analyzing and Interpreting Literature—Fiction	4 Use appropriate strategies to interpret and analyze the universal significance of literary fiction.	1 Interpret and analyze works from a variety of genres for literary, historical, and/or cultural significance.

The complete set of Assessment Anchors and Eligible Content aligned to the Pennsylvania Academic Standards can be referenced at PDE’s website: www.education.pa.gov.

HIGH-LEVEL TEST DESIGN CONSIDERATIONS

The Keystone Exams employ two types of test items (questions): multiple-choice and constructed-response. These item types assess different levels of knowledge and provide different information about achievement. Psychometrically, multiple-choice items are very useful and efficient tools for collecting information about a student's academic achievement. Constructed-response performance tasks generally generate fewer scorable points than multiple-choice items generate in the same amount of testing time; however, they provide tasks that are more realistic and sample eligible content that best lends itself to this item type. Furthermore, well-constructed scoring guides have made it possible to include constructed-response tasks in large-scale assessments, and trained scorers apply the scoring guides to efficiently score large numbers of student responses in a highly reliable way. The design of the Keystone Exams attempts to achieve a reasonable balance between the two item types.

Table 2–2. Keystone Exams High-Level Design Considerations

Exam	MC as Percentage of Core Points	CR as Percentage of Core Points	Number of Points per MC	Number of Points per CR	Number of Modules	Number of Assessment Anchors	Number of Eligible Content
Algebra I	60	40	1	4	2	6	33
Biology	73	27	1	3	2	8	38
Literature	65	35	1	3	2	4	56

DEPTH OF KNOWLEDGE

The goal of each Keystone Exam is for each item to be of sufficient rigor, or Webb's Depth of Knowledge (DOK) Level 3. Webb's DOK was created by Norman Webb of the Wisconsin Center for Education Research. Webb's definition of depth of knowledge is the degree or complexity of knowledge that the content curriculum standards and expectations require. Therefore, when reviewing items for depth of knowledge, the item is reviewed to determine whether it is as demanding cognitively as what the actual content curriculum standard expects. In the case of the Pennsylvania Keystone items, the item meets the criterion if the DOK of the item is in alignment with the DOK of the Assessment Anchor as defined by the Eligible Content. Webb's DOK includes four levels, from the lowest (basic recall) level to the highest (extended thinking) level.

In some specific cases, DOK level 2 was allowed when the cognitive intent of an Eligible Content was level 2. For more information on DOK, see Chapter Three and Appendix A.

ONLINE TESTING DESIGN CONSIDERATIONS

The Keystone Exams were designed from the beginning to provide a dual mode of test delivery, using traditional paper-and-pencil forms and using computer-based online forms. The computer-based online testing environment (called INSIGHT) is designed to provide a testing experience that mirrors the elements of traditional paper-and-pencil-based test delivery. This includes not only standard ancillary testing materials available in or with the printed forms, like formula sheets, periodic tables, scoring guidelines, and response spaces, but also analogs of the mechanical elements of response generation not necessarily associated with a computer-screen interface. These elements include line guides, rulers, screen highlighters, magnifiers, equation-building software, online calculators and graphing tools, and keyboard shortcuts.

Consideration of other components of online testing—like item layout, passage layout, font, screen resolution, navigation tools, and other interface mechanisms—all played a role in the overall design constraints, with some considerations having a more meaningful impact on specific exams. For more information on how the online test design impacted the overall test design considerations, see the sections below under each exam.

Online testing also provides an opportunity to utilize software to generate scores for student responses. In cases where responses to questions invoke numerical strings or equations, online responses can be scored through the use of lookup tables. Lookup tables are automated scoring rubrics that contain common correct and incorrect responses. When a response does not match a record in the lookup table, a human scorer is used to adjudicate the score. Operational autoscoring was only used for the Algebra I Exam; see below for more information on its use in Algebra I. For more information on scoring, see Chapter Eight.

ALGEBRA I

The Keystone Algebra I Exam has two reporting categories: Module 1, Operations and Linear Equations & Inequalities, and Module 2, Linear Functions and Data Organizations. Both modules include three Assessment Anchors. Module 1 has 18 Eligible Content, and Module 2 has 15 Eligible Content. Each module corresponds to specific content aligned to statements and specifications included in the course-specific Assessment Anchor documents. The Algebra I content included in the Keystone Algebra I multiple-choice items aligns with the Assessment Anchors and Eligible Content statements. The process skills, directives, and action statements also specifically align with the Assessment Anchors and Eligible Content statements. The content included in Algebra I constructed-response items aligns with content included in the Eligible Content statements. The process skills, directives, and action statements included in the performance demands of the Algebra I constructed-response items align with specifications included in the Assessment Anchor statements, the Anchor Descriptor statements, and/or the Eligible Content statements. In other words, the verbs or action statements used in the constructed-response items or stems can come from the Eligible Content, Anchor Descriptor, or Assessment Anchor statements.

ALGEBRA I ONLINE CONSIDERATIONS

Students taking the computer-based online delivery of the Algebra I exam are provided with online versions of several common tools typically available to a student taking a traditional paper-and-pencil exam. Each student has access to the following online tools: a standard four-function calculator, a scientific calculator, a graphing tool (similar, but not identical to, a graphing calculator), a ruler (available in metric and English units), a highlighter, a line guide, a magnifier, a sticky note generator, and a cross-off tool. In addition, an equation builder—which allows students to generate complex equations not normally possible with a standard keyboard—is also made available with all constructed-response items. Also, if the constructed-response item requests that the student draw, label, or otherwise change a graph, special graph-drawing tools are provided for on-screen graph generation. The Algebra I general scoring guideline and formula sheets are also available to students.

Layout of both the multiple-choice and constructed-response items is optimized for minimal screen manipulation (minimal scrolling required to see graphics or text that extend beyond the visible working space on the computer screen), and exam items are scrutinized carefully in both print and online versions for continuity and accuracy.

ALGEBRA I MULTIPLE-CHOICE ITEMS

Sixty percent of the possible points on the Algebra I Exam are derived from multiple-choice items. This item type is especially efficient for measuring a broad range of content. Each multiple-choice item has four response options, only one of which is correct. The student is awarded one point for choosing the correct response. Distractors typically represent incorrect concepts, incorrect logic, incorrect application of an algorithm, or computational errors.

Algebra I multiple-choice items are intended to take about one and a half minutes of response time per item. They are used to assess a variety of skill levels, including problem solving. Algebra I items involving application emphasize the requirement to carry out some mathematical process to find an answer rather than simply recalling information from memory.

ALGEBRA I CONSTRUCTED-RESPONSE ITEMS

Constructed-response items (tasks) require that students read a problem description and develop an appropriate solution. Algebra I constructed-response items are designed to take about ten minutes of response time per item. Most of the constructed-response items have several components in the overall task that may enable students to enter or begin the problem at different places. In some items, each successive component is designed to assess progressively more difficult skills or higher knowledge levels. Certain components may ask students to explain their reasoning for applying particular operations or for arriving at certain conclusions. The types of tasks utilized do not necessarily require computations. Students may also be asked to perform such tasks as constructing a graph, shading some portion of a figure, or listing object combinations that meet specified criteria.

Constructed-response tasks are especially useful for measuring students' problem-solving skills in Algebra. They offer the opportunity to present real-life situations that necessitate that the students solve problems using mathematics abilities learned in the classroom. Students must read the task carefully, identify the necessary information, devise a method of solution, perform the calculations, enter the solution directly in the answer document, and, when required, offer an explanation. This provides insight into the students' mathematical knowledge, abilities, and reasoning processes.

The constructed-response Algebra items are scored on a 0–4 point scale using an item-specific scoring guideline. The item-specific scoring guideline outlines the requirements for each score point. Item-specific scoring guidelines are based on the Algebra I General Description of Scoring Guidelines. The general guidelines describe a hierarchy of responses, which represent the five score levels, see Appendix B.

The Algebra I Keystone Exam includes two types of constructed-response items: Scaffolded Constructed-Response Items (SCR) and Extended Constructed-Response Items (ECR). Both types are scored on the same 0–4 point scale using the same Algebra I General Description of Scoring Guidelines as the base. SCR items are constructed to generally elicit four distinct responses (a response may contain more than one answer blank), and each response has the potential to earn a discrete number of score points (generally just one [1] score point per response). In turn, the four distinct responses are generally organized into four sections, with each labeled as a “Part” within an SCR. The next table shows a generic (nonauthentic) illustration of the application of this concept.

Table 2–3. Generic Example [Nonauthentic] Showing Concept of Four Distinct Responses

Stem	Part A	Part B	Part C	Part D
Presents a numerical distribution	In the answer spaces, write the list of numbers from least to greatest.	Write the mean in an answer blank.	Write the median in an answer blank.	Write the mode in an answer blank.
4 points	1 distinct point even though students enter more than one number	1 distinct point with one distinct entry	1 distinct point with one distinct entry	1 distinct point with one distinct entry

- SCR items do not require narrative, explanation, or “show all your work” responses.
- Most SCR item responses lend themselves to automatic scoring; however, **not all items can be automatically scored exclusively with the use of lookup tables.** The full application of Assessment Anchors and Eligible Content sometimes requires item construction that is incompatible with lookup tables.

In familiar and probably the most descriptive terms, Algebra I ECR items—in form, format, and scoring provisions—adhere to the philosophy of PSSA OE item format. Like SCR items, development is based on the item qualities that best measure the skills and concepts with which the item aligns.

- ECR items intentionally elicit narrative, explanation of reasoning, “explain why . . .”, and/or “show your work” responses.
- In contrast to SCR items, in which DOK level 3 cognitive engagement is inferred from student responses, ECR items (through explanations and recorded work) often provide direct evidence of DOK level 3 engagement. This aspect of ECR items is intentionally included during development. Following initial development, the ECR item will be approved by PDE as accepted by the review committee, or PDE and DRC will collaborate in amending the item.

BIOLOGY

The Keystone Biology Exam has two reporting categories: Module 1[A], Cells and Cell Processes; and Module 2[B], Continuity and Unity of Life. Both modules have four Assessment Anchors. Module A has 16 Eligible Content, and Module B has 22 Eligible Content. Each module corresponds to specific content aligned to statements and specifications included in the course-specific assessment anchor documents. The Biology content included in the Keystone Biology multiple-choice items aligns with the Assessment Anchors and Eligible Content statements. The process skills, directives, and action statements also specifically align with the Assessment Anchors and Eligible Content statements. The content included in Biology constructed-response items aligns with content included in the Eligible Content statements. The process skills, directives, and action statements included in the performance demands of the Biology constructed-response items align with specifications included in the Assessment Anchor statements, the Anchor Descriptor statements, and/or the Eligible Content statements. In other words, the verbs or action statements used in the constructed-response items or stems can come from the Eligible Content, Anchor Descriptor, or Assessment Anchor statements.

BIOLOGY ONLINE CONSIDERATIONS

Students taking the computer-based online delivery of the Biology Exam are provided with online versions of several common tools typically available to a student taking a traditional paper-and-pencil exam. Each student has access to the following online tools: a highlighter, a line guide, a magnifier, a sticky note generator, and a cross-off tool. The Biology general scoring guideline and a periodic table are also provided to students.

Layout of both the multiple-choice and constructed-response items is optimized for minimal screen manipulation (minimal scrolling to see graphics or text that extend beyond the visible working space on the computer screen), and exam items are scrutinized carefully in both print and online versions for continuity and accuracy.

BIOLOGY MULTIPLE-CHOICE ITEMS

Seventy-three percent of the possible points on the Biology Exam are derived from multiple-choice items. Multiple-choice items are especially efficient for measuring a broad range of content. Each multiple-choice item has four response options, only one of which is correct. The student is awarded one point for choosing the correct response. Distractors typically represent incorrect concepts, incorrect logic, or incorrect application of a biological principle.

Biology multiple-choice items are intended to take about one and a quarter minutes of response time per item. They are used to assess a variety of skill levels, including the application of Biology content. Biology items involving application emphasize the requirement to utilize science content to find an answer rather than simply recalling information from memory.

BIOLOGY CONSTRUCTED-RESPONSE ITEMS

Constructed-response items (tasks) require students to read a description of a Biology problem and to develop an appropriate solution. Biology constructed-response items are designed to take about eight minutes of response time per item. Constructed-response tasks are especially useful for measuring students' skills in biology. These tasks may present real-life situations that require students to solve problems using science abilities learned in the classroom. Students must read a task carefully, identify the necessary information, devise a method of solution, enter the solution directly into the answer document, and when required, offer an explanation. This provides insight into students' science knowledge, abilities, and reasoning processes.

The constructed-response science items are scored on a 0–3 point scale with an item-specific scoring guideline, and each task is carefully constructed with a scoring guideline reflecting the task requirements. The general guidelines describe a hierarchy of responses, which represents the four score levels. Each item-specific scoring guideline outlines the requirements at each score point, and each item-specific scoring guideline is based on the Biology General Description of Scoring Guidelines, see Appendix B.

LITERATURE

The Keystone Literature Exam has two reporting categories: Module 1, Fiction; and Module 2, Nonfiction. Both modules have two Assessment Anchors. Module 1 has 25 Eligible Content, and Module 2 has 33 Eligible Content. The Literature Exam employs two types of test items, multiple-choice and constructed-response, and the content included aligns with content included in the Eligible Content statements. The items are designed to measure students' comprehension of the content contained in the literature passages. Each module corresponds to specific content aligned to statements and specifications included in the course-specific Assessment Anchor documents. The Literature content included in the Keystone Literature multiple-choice items aligns with the Assessment Anchors and Eligible Content statements. The process skills, directives, and action statements also specifically align with the Assessment Anchors and Eligible Content statements. The content included in Literature constructed-response items aligns with content included in the Eligible Content statements. The process skills, directives, and action statements included in the performance demands of the Literature constructed-response items align with specifications included in the Assessment Anchor statements, the Anchor Descriptor statements, and/or the Eligible Content statements. In other words, the verbs or action statements used in the constructed-response items or stems can come from the Eligible Content, Anchor Descriptor, or Assessment Anchor statements.

LITERATURE ONLINE CONSIDERATIONS

Students taking the computer-based online delivery of the Literature Exam are provided with online versions of several common tools typically available to a student taking a traditional paper-and-pencil exam. Each student has access to the following online tools: a highlighter, a line guide, a magnifier, a sticky note generator, and a cross-off tool. The Literature general scoring guideline is also provided to students.

Layout of passages, multiple-choice items, and constructed-response items is optimized for minimal screen manipulation (minimal scrolling to see text and graphics that extend beyond the visible working space on the computer screen), and exam items are scrutinized carefully in both print and online versions for continuity and accuracy. In addition, the amount of space devoted to the passage compared to the amount of space devoted to the exam questions was also optimized.

LITERATURE MULTIPLE-CHOICE ITEMS

Sixty-five percent of the possible points on the Literature Exam are derived from multiple-choice items. Literature multiple-choice items are intended to take about one minute of response time per item. They are designed to measure how well students comprehend the overall meaning of a passage or make basic inferences about it. At times, asking students to choose a preferred answer is the best way to determine whether they have gleaned certain information from a story. Such information may include central idea, setting, or main events and their sequence.

Each Literature multiple-choice item has four response options, only one of which is correct. The student is awarded one point for choosing the correct response. Distractors typically represent some kind of misinterpretation, predisposition, unsound reasoning, or casual reading.

LITERATURE CONSTRUCTED-RESPONSE ITEMS

Constructed-response items (tasks) are designed to address comprehension of text in ways that multiple-choice items cannot. Literature constructed-response items are designed to take about five minutes of response time per item. A short written response allows students to prepare an answer and summarize using supporting details or examples derived from the text.

The Literature constructed-response items are scored on a 0–3 point scale using an item-specific scoring guideline. Each task is text-dependent and is carefully constructed with the scoring guideline reflecting the task requirements. All item-specific scoring guidelines are based on the Literature General Description of Scoring Guidelines. The general guidelines describe a hierarchy of responses, which represent the four score levels, see Appendix B.

LITERATURE PASSAGES

One of the key requirements of the Keystone Literature Exam is that students should be able to read and comprehend both literature and informational texts of sufficient text complexity and quality as required by the Assessment Anchors and Eligible Content. For example, the Literature Keystone Assessment Anchors and Eligible Content require students to engage with appropriately complex literary fiction, literary nonfiction, and informational works. Passage genres include, but are not limited to, the following: stories; excerpts from novels, biographies, and autobiographies; letters; dramas; poems; myths from diverse cultures and different time periods; texts in history/social studies, science, and other disciplines; seminal U.S. documents; the classics of American, British, and world literature; and current articles and editorials.

TEXT COMPLEXITY

Text complexity involves three components: matching reader to text and task, qualitative evaluation of the text, and quantitative evaluation of the text.

MATCHING READER TO TEXT AND TASK

A number of factors are taken into consideration when deciding whether a passage will be placed in the pool for possible use on the Keystone Literature Exam. The factors include, but are not limited to, the following:

- Are the conceptual load, vocabulary, syntactic patterns, sentence length, and clarity appropriate for the grade level?
- Does the passage stand the test of time as an example of literary fiction, literary nonfiction, and/or informational text, and is it judged by the committee of Pennsylvania educators as having sufficient quality?
- Is the passage “rich” enough to generate a variety of items?
- Do the passages represent a range of reading levels appropriate to the grade level?
- Do the passages lend themselves well to measuring the Keystone Assessment Anchors and Eligible Content, including text structures and elements?
- Are the passages free of issues of bias, fairness, and/or sensitivity?
- Does the pool of passages represent diversity in the areas of gender, culture, ethnicity, urban/rural status, socioeconomic status, physical differences, and age?

QUALITATIVE EVALUATION OF THE TEXT

Evaluating the text complexity of a passage is essentially a judgmental process by individuals familiar with the classroom context and what is linguistically appropriate at a given grade level. All Keystone passages to be included in the pool of passages for possible use on the Keystone Literature Exam are reviewed and approved by PDE and the Pennsylvania Reading Content Committee (a committee of Pennsylvania educators). The passages are reviewed by Pennsylvania educators to judge whether each passage meets the criteria outlined above. All potential passages are also reviewed by the Pennsylvania Bias, Fairness, and Sensitivity Committee.

QUANTITATIVE EVALUATION OF THE TEXT

Each readability program uses different methods to determine the readability for a particular passage (e.g., syllables, sentence length, number of words, vocabulary lists). Each readability formula is designed for a particular grade range of materials. When using the various readability formulas, a wide range of readability levels may be identified for a particular passage. Some readability formulas are better suited to a particular grade level. If a particular formula being used is outside of the intended range, then the results may be unreliable.

Readability of the Keystone Literature Exam passages has been determined using several of the most widely accepted readability formulas. These formulas are not used in a rigid way, but rather more informally to provide for several “snapshots” of a passage. The readability formulas used for the passages that appear on the Keystone Literature Exam are the Dale-Chall Formula, the Flesch Grade Level Formula, and the Fry Graph.

CHAPTER THREE: ITEM AND TEST DEVELOPMENT PROCESSES

GENERAL KEYSTONE TEST DEVELOPMENT PROCESSES

The 2021 Keystone Exams continued to use the core-to-core biennial overlap. Approximately 30% to 50% of the operational points in each module overlap with items used operationally 2 years prior. The 2021 Keystone Exam cores were made up of items that had appeared on the Spring 2018, Summer 2018, and/or Winter 2018/2019 cores. The remainder of the operational 2021 exams were made up of items that were field tested on the Spring 2019 Keystone Exams embedded field test administration. Table 3–1 is a graphic representation of the basic process flow and overlap of the development cycles.

Table 3–1. General Development and Usage Cycle of the Algebra I, Biology, and Literature Keystone Exams

Admin Year	Events Occurring in 2016	Events Occurring in 2017	Events Occurring in 2018	Events Occurring in 2019	Events Occurring in 2020	Events Occurring in 2021*
2014–2015		Biennial Core-to-Core Overlap (2015 core included as a portion of the 2017 core)				
2015–2016	Spring 2016 Oper. & Embedded FT; Data Review of Spring 2016 FT; Summer 2016 Admin		Biennial Core-to-Core Overlap (2016 core included as a portion of the 2018 core)			
2016–2017	Winter 2016/17 Admin; New Item Dev. for Spring 2017 FT	Spring 2017 Oper. & Embedded FT; Data Review of Spring 2017 FT; Summer 2017 Admin		Biennial Core-to-Core Overlap (2017 core included as a portion of the 2019 core)		
2017–2018		Winter 2017/18 Admin; New Item Dev. for Spring 2018 FT	Spring 2018 Oper. & Embedded FT; Data Review of Spring 2018 FT; Summer 2018 Admin		Biennial Core-to-Core Overlap (2018 core included as a portion of the 2020 core)	
2018–2019			Winter 2018/19 Admin; New Item Dev. for Spring 2019 FT	Spring 2019 Oper. & Embedded FT; Data Review of Spring 2019 FT; Summer 2019 Admin		
2019–2020				Winter 2019/20 Admin; New Item Dev. for Spring 2021 FT	Summer 2020 Admin	

Table 3–1 (continued). General Development and Usage Cycle of the Algebra I, Biology, and Literature Keystone Exams

Admin Year	Events Occurring in 2016	Events Occurring in 2017	Events Occurring in 2018	Events Occurring in 2019	Events Occurring in 2020	Events Occurring in 2021*
2020–2021*					Winter 2020/21 Admin	Spring 2021 Oper. & Embedded FT; Data Review of Spring 2021 FT; Summer 2021 Admin

*Projected/scheduled tasks and activities

GENERAL TEST DEFINITION

The plan for the Keystone Exam was developed through the collaborative efforts of the Pennsylvania Department of Education (PDE) and Data Recognition Corporation (DRC). The exams are presented online or in two printed testing materials, a test book and a separate answer book. The test book contains multiple-choice (MC) items. The answer book contains scannable pages for multiple-choice responses, constructed-response (CR) items with response spaces, and demographic data collection areas. All MC items are worth 1 point. Algebra I CR items receive a maximum of 4 points (on a scale of 0–4), and all Biology and Literature CR items receive a maximum of 3 points (on a scale of 0–3).

CORE-TO-CORE OVERLAP ITEMS

The operational items consist of a set of core items taken by all students. Starting in 2014 these core items included core-to-core overlapping items, which are items that also appeared on the core form of the administration two years before. The overlap connects the spring and summer administrations of year x and the winter administration of year $x+1$, with the year $x+2$ spring and summer and year $x+3$ winter administrations. The first biennial core-to-core overlap from the spring 2011 and winter 2011–2012 core was scheduled to begin with the spring 2013 administration. However, when the program was placed on hiatus during the 2011–2012 school year, the overlap was moved to the spring 2014 administration.

ALGEBRA I TEST DEFINITIONS

- The Spring 2021 Algebra I Keystone Exam was composed of 20 forms. All of the forms contained operational core items identical for all students and sets of generally unique items. The following two tables display the design for Algebra I for forms 1 through 20. The column entries for these tables denote the following:
 - Number of unique core MC items
 - Number of unique core CR items
 - Number of embedded MC field test items
 - Number of embedded CR field test items
 - Total number of MC and CR items in the form

Table 3–2. Algebra I Test Plan (Spring 2021) per Operational Form

Module	Core per Form MC Items	Core per Form CR Items	Field Test per Form MC Items	Field Test per Form CR Items	Total per Form Core & FT MC Items	Total per Form Core & FT CR Items
1	18	3	5	1	23	4
2	18	3	5	1	23	4
Total	36	6	10	2	46	8

Table 3–3. Algebra I Test Plan (Spring 2021) per 20 Operational Forms

Module	Core per 20 Forms MC Items	Core per 20 Forms CR Items	Field Test per 20 Forms MC Items	Field Test per 20 Forms CR Items	Total per 20 Forms Core & FT MC Items	Total per 20 Forms Core & FT CR Items
1	18	3	100	20	118	23
2	18	3	100	20	118	23
Total	36	6	200	40	236	46

The operational (core) portions of the Spring 2021, Summer 2021, and the Winter 2020/2021 administrations came from the same sources. Therefore 30% to 50% of the 2020/2021 Winter, Spring and Summer cores overlap with the Spring 2018, Summer 2018, and Winter 18/19 cores. The remaining core items that appeared on the 2020/2021 forms were field tested on prior administrations. Although each spring administration includes embedded field test items, the summer, winter, and breach forms do not include any embedded field test items due to the lower *n*-counts for these administrations. However, summer, winter, and breach forms still include the same number of items that appear in the spring administration. Instead of field test items, the slots in these exams are filled by placeholder (PH) items. Table 3–4 displays the design for the Algebra I Summer, Winter, and Breach operational forms.

Table 3–4. Algebra I Test Plan (2021 Summer, Winter, and Breach) per Operational Form

Module	Core per Form MC Items	Core per Form CR Items	Placeholders per Form MC Items	Placeholders per Form CR Items	Total per Form Core & PH MC Items	Total per Form Core & PH CR Items	Number of Forms Master Core	Number of Forms Scrambles
1	18	3	5	1	23	4	1	3
2	18	3	5	1	23	4	1	3
Total	36	6	10	2	46	8	1	3

Since an individual student’s score is based solely on the operational (or core) items, the total number of operational points is 60 for Algebra I. The total score is obtained by combining the points from the core MC (1 point each) and core CR (up to 4 points each) portions of the exam as follows:

Table 3–5. Algebra I Core Points

Category	Module 1 MC Items	Module 1 CR Items	Module 2 MC Items	Module 2 CR Items	Total MC Items	Total CR Items
Total Points	30 (50%)		30 (50%)		60 (100%)	
Core Items	18	3	18	3	36	6
Core Points	18	12	18	12	36	24

The Algebra I Exam results will be reported in two categories based on the two modules of the Algebra I Exam. The code letters for these Assessment Anchor categories are

1. Operations and Linear Equations & Inequalities
2. Linear Functions and Data Organization

The distribution of Algebra I items into these two categories is shown in the following table.

Table 3–6. Algebra I Module and Anchor Distribution

Algebra I Module	Raw Points	Module Weight	Number of Anchors	Number of Eligible Content
1	30	50%	3	18
2	30	50%	3	15

The reporting categories are further subdivided for specificity and Eligible Content (limits). Each subdivision is coded by adding an additional character to the framework of the labeling system. These subdivisions are called Assessment Anchors and Eligible Content. More information about Assessment Anchors and Eligible Content is in Chapter Two.

For more information concerning the process used to convert the Algebra I operational test plan into forms (i.e., form construction), see Chapter Six.

For more information concerning the test sessions, timing, and layout for the Algebra I operational exam, see Chapter Seven.

BIOLOGY TEST DEFINITIONS

The Spring 2021 Biology Keystone Exam was composed of 20 forms. All of the forms contained operational core items identical for all students and sets of generally unique items. The following two tables display the design for Biology for forms 1 through 20. The column entries for these tables denote the following:

- Number of unique core MC items
- Number of unique core CR items
- Number of embedded MC field test items
- Number of embedded CR field test items
- Total number of MC and CR items in the form

Table 3–7. Biology Test Plan (Spring 2021) per Operational Form

Module	Core per Form MC Items	Core per Form CR Items	Field Test per Form MC Items	Field Test per Form CR Items	Total per Form Core & FT MC Items	Total per Form Core & FT CR Items
1	24	3	8	1	32	4
2	24	3	8	1	32	4
Total	48	6	16	2	64	8

Table 3–8. Biology Test Plan (Spring 2021) per 20 Operational Forms

Module	Core per 20 Forms MC Items	Core per 20 Forms CR Items	Field Test per 20 Forms MC Items	Field Test per 20 Forms CR Items	Total per 20 Forms Core & FT MC Items	Total per 20 Forms Core & FT CR Items
1	24	3	160	20	184	23
2	24	3	160	20	184	23
Total	48	6	320	40	368	46

The operational (core) portions of the Spring 2021, Summer 2021, and the Winter 2020/2021 administrations came from the same sources. Therefore 30% to 50% of the 2020/2021 Winter, Spring and Summer cores overlap with the Spring 2018, Summer 2018, and Winter 18/19 cores. The remaining core items that appeared on the 2020/2021 forms were field tested on prior administrations. Although each spring administration includes embedded field test items, the summer, winter, and breach forms do not include any embedded field test items due to the lower *n*-counts for these administrations. However, summer, winter, and breach forms still include the same number of items that appear in the spring administration. Instead of field test items, the slots in these exams are filled by placeholder (PH) items. Table 3–9 displays the design for the Biology Summer, Winter, and Breach operational forms.

Table 3–9. Biology Test Plan (2021 Summer, Winter, and Breach) per Operational Form

Module	Core per Form MC Items	Core per Form CR Items	Placeholders per Form MC Items	Placeholders per Form CR Items	Total per Form Core & PH MC Items	Total per Form Core & PH CR Items	Number of Forms Master Core	Number of Forms Scrambles
1	24	3	8	1	32	4	1	3
2	24	3	8	1	32	4	1	3
Total	48	6	16	2	64	8	1	3

Since an individual student’s score is based solely on the operational (or core) items, the total number of operational points is 66 for Biology. The total score is obtained by combining the points from the core MC (1 point each) and core CR (up to 3 points each) portions of the exam as follows:

Table 3–10. Biology Core Points

Category	Module 1 MC Items	Module 1 CR Items	Module 2 MC Items	Module 2 CR Items	Total MC Items	Total CR Items
Total Points	33 (50%)		33 (50%)		66 (100%)	
Core Items	24	3	24	3	48	6
Core Points	24	9	24	9	48	18

The Biology Exam results will be reported in two categories based on the two modules of the Biology Exam.

1. Cells and Cell Processes
2. Continuity and Unity of Life

The distribution of Biology items into these two categories is shown in the following table.

Table 3–11. Biology Module and Anchor Distribution

Biology Module	Raw Points	Module Weight	Number of Anchors	Number of Eligible Content
1	33	50%	4	16
2	33	50%	4	22

The reporting categories are further subdivided for specificity and Eligible Content (limits). Each subdivision is coded by adding an additional character to the framework of the labeling system. These subdivisions are called Assessment Anchors and Eligible Content. More information about Assessment Anchors and Eligible Content is in Chapter Two.

For more information concerning the process used to convert the Biology operational test plan into forms (i.e., form construction), see Chapter Six.

For more information concerning the test sessions, timing, and layout for the Biology operational exam, see Chapter Seven.

LITERATURE TEST DEFINITIONS

The Spring 2021 Literature Keystone Exam was composed of 20 forms. All of the forms contained operational core items identical for all students and sets of generally unique items. The following two tables display the design for Literature for forms 1 through 20. The column entries for these tables denote the following:

- Number of unique core passages
- Number of unique core MC items
- Number of unique core CR items
- Number of embedded field test passages
- Number of embedded MC field test items
- Number of embedded CR field test items
- Total number of passages, MC items, and CR items in the form

Table 3–12. Literature Test Plan (Spring 2021) per Operational Form

Module	Core per Form Passages	Core per Form MC Items	Core per Form CR Items	Field Test per Form Passages	Field Test per Form MC Items	Field Test per Form CR Items	Total per Form Passages	Total per Form Core & FT MC Items	Total per Form Core & FT CR Items
1	2	17	*3	1	6	1	3	23	4
2	2	17	*3	1	6	1	3	23	4
Total	4	34	6	2	12	2	6	46	8

*For each module, one core passage has two CRs and one core passage has one CR.

Table 3–13. Literature Test Plan (Spring 2021) per 20 Operational Forms

Module	Core per 20 Forms Passages	Core per 20 Forms MC Items	Core per 20 Forms CR Items	Field Test per 20 Forms Passages	Field Test per 20 Forms MC Items	Field Test per 20 Forms CR Items	Total per 20 Forms Passages	Total per 20 Forms Core & FT MC Items	Total per 20 Forms Core & FT CR Items
1	2	17	*3	12	120	20	10	137	23
2	2	17	*3	12	120	20	10	137	23
Total	4	34	6	24	240	40	20	274	46

*For each module, one core passage has two CRs and one core passage has one CR.

The operational (core) portions of the Spring 2021, Summer 2021, and the Winter 2020/2021 administrations came from the same sources. Therefore 30% to 50% of the 2020/2021 Winter, Spring and Summer cores overlap with the Spring 2018, Summer 2018, and Winter 18/19 cores. The remaining core items that appeared on the 2020/2021 forms were field tested on prior administrations. Although each spring administration includes embedded field test items, the summer, winter, and breach forms do not include any embedded field test items due to the lower *n*-counts for these administrations. However, summer, winter, and breach forms still include the same number of items that appear in the spring administration. Instead of field test items, the slots in these exams are filled by placeholder (PH) items. Table 3–14 displays the design for the Literature Summer, Winter, and Breach operational forms.

Table 3–14. Literature Test Plan (2021 Summer, Winter, and Breach) per Operational Form

Module	Core per Form MC Items	Core per Form CR Items	Placeholders per Form MC Items	Placeholders per Form CR Items	Total per Form Core & PH MC Items	Total per Form Core & PH CR Items	Number of Forms Master Core	Number of Forms Scrambles
1	2	17	*3	1	6	1	1	3
2	2	17	*3	1	6	1	1	3
Total	4	34	6	2	12	2	1	3

*For each module, one core passage has two CRs and one core passage has one CR.

Since an individual student’s score is based solely on the operational (or core) items, the total number of operational points is 52 for Literature. The total score is obtained by combining the points from the core MC (1 point each) and core CR (up to 3 points each) portions of the exam as follows:

Table 3–15. Literature Core Points

Category	Module 1 Passages	Module 1 MC Items	Module 1 CR Items	Module 2 Passages	Module 2 MC Items	Module 2 CR Items	Total Passages	Total MC Items	Total CR Items
Total Points	26 (50%)			26 (50%)			52 (50%)		
Core Items	2	17	3	2	17	3	4	34	6
Core Points	NA	17	9	NA	17	9	NA	34	18

The Literature Exam results will be reported in two broad categories based on the two modules of the Literature Exam.

1. Fiction Literature
2. Nonfiction Literature

The distribution of Literature items into these two categories is shown in the following table.

Table 3–16. Literature Module and Anchor Distribution

Literature Module	Raw Points	Module Weight	Number of Anchors	Number of Eligible Content
1	26	50%	2	25
2	26	50%	2	31

The reporting categories are further subdivided for specificity and Eligible Content (limits). Each subdivision is coded by adding an additional character to the framework of the labeling system. These subdivisions are called Assessment Anchors and Eligible Content. More information about Assessment Anchors and Eligible Content is in Chapter Two.

For more information concerning the process used to convert the Literature operational test plan into forms (i.e., form construction), see Chapter Six.

For more information concerning the test sessions, timing, and layout for the Literature operational exam, see Chapter Seven.

ITEM DEVELOPMENT CONSIDERATIONS

Alignment to the Keystone Assessment Anchors and Eligible Content, course-level appropriateness (as specified by PDE), depth of knowledge (DOK), item/task level of complexity, estimated difficulty level, relevancy of context, rationale for distractors, style, accuracy, and correct terminology were major considerations in the item development process. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) and *Universal Design* (Thompson, Johnstone, & Thurlow, 2002) guided the development process. In addition, *Fairness in Testing: Training Manual for Issues of Bias, Fairness, and Sensitivity* (DRC, 2010) was used for developing items. All items were reviewed for fairness by bias and sensitivity committees and for content by Pennsylvania educators and field specialists.

BIAS, FAIRNESS, AND SENSITIVITY OVERVIEW

At every stage of the item and test development process, DRC employs procedures that are designed to ensure that items and tests meet Standard 7.4 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999):

Standard 7.4: Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

To meet Standard 7.4, DRC uses a series of internal quality steps. DRC provides specific training for test developers, item writers, and reviewers on how to write, review, revise, and edit items related to issues of bias, fairness, and sensitivity (as well as based on technical quality). Training also includes an awareness of and sensitivity to issues of cultural diversity. In addition to providing *internal* training in reviewing items in order to eliminate potential bias, DRC also provides *external* training to the review panels of minority experts, teachers, and other stakeholders.

DRC's guidelines for bias, fairness, and sensitivity include instruction concerning how to eliminate language, symbols, words, phrases, and content that might be considered offensive by members of racial, ethnic, gender, or other groups. Areas of bias that are specifically targeted include, but are not limited to, stereotyping, gender, region/geography, ethnic/cultural, socioeconomic/class, religion, experience, and biases against a particular age group (ageism) or persons with disabilities. DRC catalogues topics that should be avoided and maintains balance in gender and ethnic emphasis within the pool of available items and passages.

See the sections below in this chapter for more information about the Bias, Fairness, and Sensitivity Review meetings conducted for the Keystone Exams.

UNIVERSAL DESIGN OVERVIEW

The principles of universal design were incorporated throughout the item development process to allow participation of the widest possible range of students in the Keystone Exams. The following checklist was used as a guideline:

- Items measure what they are intended to measure.
- Items respect the diversity of the assessment population.
- Items have a clear format for text.
- Stimuli and items have clear pictures and graphics.
- Items have concise and readable text.
- Items allow changes to other formats, such as Braille, without changing meaning or difficulty.
- The arrangement of the items on the test has an overall appearance that is clean and well organized.

A more extensive description of the application of the principles of universal design is provided in Chapter Four.

DEPTH-OF-KNOWLEDGE OVERVIEW

An important element in statewide graduation exams is the alignment between the overall assessment system and the state’s standards. A methodology developed by Norman Webb (1999, 2006) offers a comprehensive model that can be applied to a wide variety of contexts. With regard to the alignment between standards statements and the assessment instruments, Webb’s criteria include five categories, one of which deals with content. Within the content category is a useful set of levels for evaluating DOK. According to Webb (1999), “depth-of-knowledge consistency between standards and assessments indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards” (Webb, 1999, pp. 7–8). The four levels of cognitive complexity (i.e., DOK) are as follows:

- Level 1: Recall
- Level 2: Application of Skill/Concept
- Level 3: Strategic Thinking
- Level 4: Extended Thinking

DOK levels were incorporated into the item writing and review process, and items were coded with respect to the level they represented. The DOK level for MC and CR items are at Level 3, Level 2, or Level 1 depending on the cognitive intent of an Eligible Content. DOK Level 4 items are not included on the Keystone Exams. For more information on DOK (and a comparison of DOK to Bloom’s Taxonomy), see Appendix A.

PASSAGE READABILITY OVERVIEW

Evaluating the readability of a passage is essentially a judgment by individuals familiar with the classroom context and what is linguistically appropriate (PDE recommends that the Literature Keystone Exam be administered at grade 10). Although various readability indices were computed and reviewed, it is recognized that such methods measure different aspects of readability and are often fraught with particular interpretive liabilities. Thus, the commonly available readability formulas were not used in a rigid way, but more informally to provide for several snapshots of a passage that senior test development staff considered, along with experience-based judgments in guiding the passage-selection process. In addition, passages were reviewed by committees of Pennsylvania educators who evaluated each passage for readability and grade-level appropriateness. For more information on Literature passages, see Chapter Two and the literature passage-selection process described below.

TEST ITEM READABILITY OVERVIEW

Careful attention was given to the readability of the items to make certain that the assessment focus of the item did not shift based on the difficulty of reading the item. Subject/course areas such as Algebra I or Biology contain many content-specific vocabulary terms. As a result, readability formulas were not used. However, wherever it was practicable and reasonable, every effort was made to keep the vocabulary at or one level below the course level for non-Literature exams. There was a conscious effort made to ensure that each question was evaluating a student’s ability to build toward mastery of the course standards rather than evaluating the student’s reading ability. Resources used to verify the vocabulary level were the *EDL Core Vocabularies* and the *Children’s Writer’s Word Book*.

In addition, every test question is brought before several different committees composed of Pennsylvania educators who are course-level/grade-level experts in the content field in question. They review each question from the perspective of the students they teach, determine the validity of the vocabulary used, and work to minimize the level of reading required.

Vocabulary was also addressed at the Bias, Fairness, and Sensitivity Review, although the focus was on how certain words or phrases may represent possible sources of bias or issues of fairness or sensitivity. See the sections that follow in this chapter for more information about the Bias, Fairness, and Sensitivity Review meetings conducted for the Keystone Exams.

ITEM AND TEST DEVELOPMENT CYCLE

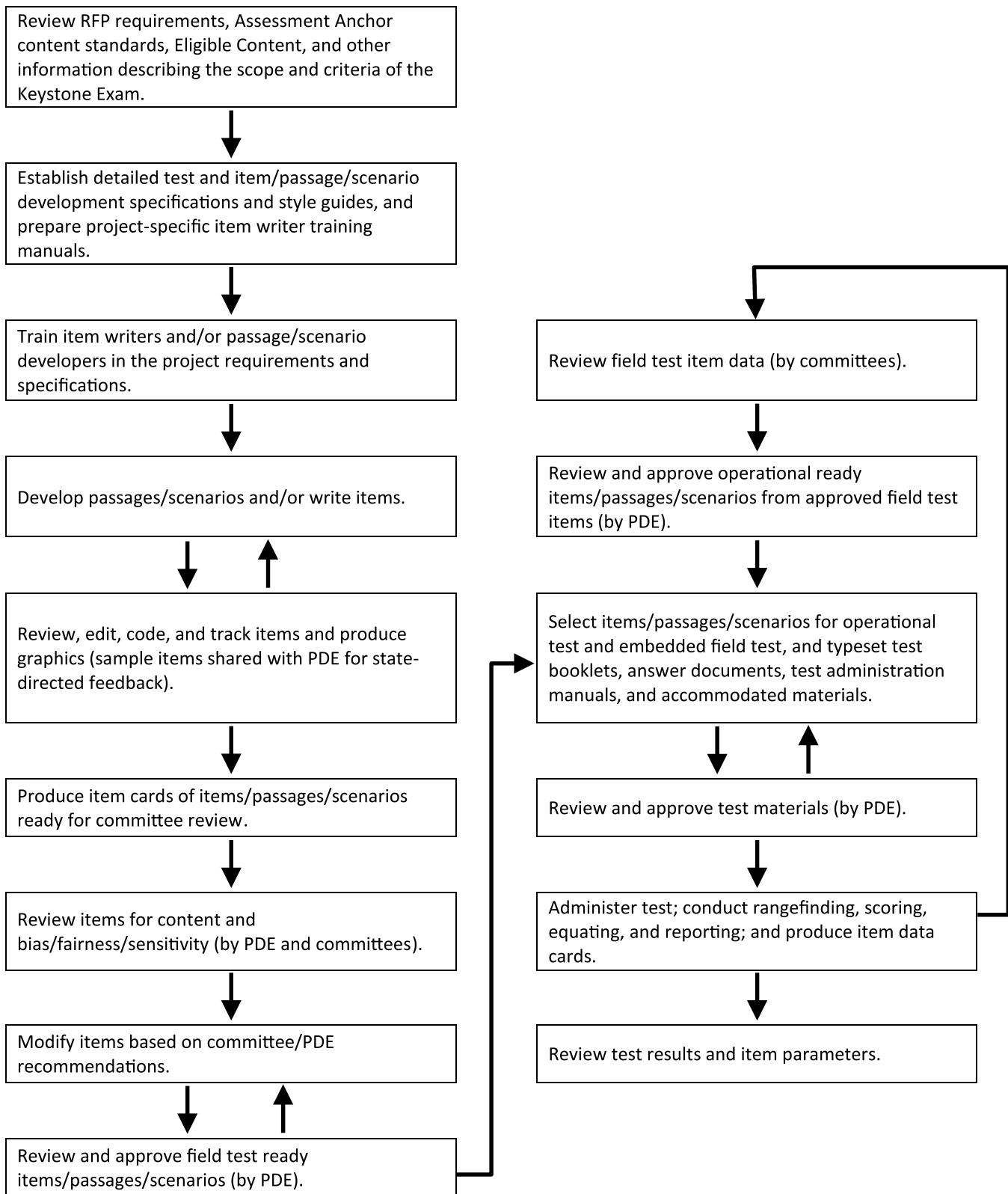
The item development process for items followed a logical cycle and time line, which are outlined in the table and figure on the next pages. On the front end of the schedule, tasks were generally completed with the goal of presenting field test candidate items to committees of Pennsylvania educators. On the back end of the schedule, all tasks led to the field test data review and operational test construction. This process represents a typical life cycle for an embedded Keystone field test event, not a stand-alone field test event or an accelerated development cycle.

The process flowchart, also shown below, illustrates the interrelationship among the steps in the primary cycle that occurs in a normal process of development (i.e., when the items for field testing are primarily from new development, as opposed to being selected from an existing item bank). In addition, a detailed process table describing the item and test development processes also appears in Appendix C.

Table 3–17. General Item and Test Development Life Cycle for Spring Keystone Administrations

Cycle	Steps in Development Life Cycle	Time Line	Approximate Window	
Primary	Development Planning	Summer/Fall	Month 1–4	Jul–Oct
Primary	Literature Passage Selection	Summer/Fall	Month 1–6	Jul–Dec
Primary	Item Writer Training	Fall	Month 5	Nov
Primary	Initial Item Authoring	Fall/Winter	Month 5–9	Nov–Mar
Primary	Internal Reviews and PDE Reviews	Fall to Spring	Month 6–12	Dec–Jun
Primary	Bias, Fairness, and Sensitivity Review	Summer	Month 13	Jul
Primary	New Item Content Committee Review (PA Educators)	Summer	Month 13	Jul
Primary	Post-Review Resolution and Cleanup	Summer	Month 13–14	Jul–Aug
Primary	Build Field Test Forms	Summer/Fall	Month 15–16	Sep–Oct
Primary	Internal Form Reviews and PDE Reviews	Summer/Fall	Month 15–18	Sep–Dec
Primary	Final Form and Printer Proof Approvals	Fall/Winter	Month 18–19	Dec–Jan
Primary	Ancillary and Accommodated Form Development	Fall/Winter	Month 18–20	Dec–Feb
Primary	Form Printing, Spiraling, Packaging, and Shipping	Winter/Spring	Month 19–22	Jan–Apr
Primary	Field Test Administration	Spring	Month 23	May
Primary	Material/Data Processing, Rangefinding, and Scoring	Spring/Summer	Month 23–26	May–Aug
Primary	Field Test Item Data Review (PA Educators)	Summer	Month 27	Sept
Primary	Select Operational Items	Summer/Fall	Month 27–28	Sep–Oct
Primary	Build Operational Forms	Fall	Month 28–29	Oct–Nov
Primary	Internal Form Reviews and PDE Reviews	Fall	Month 29–30	Nov–Dec
Primary	Final Form and Printer Proof Approvals	Fall/Winter	Month 30–31	Dec–Jan
Primary	Ancillary and Accommodated Form Development	Fall/Winter	Month 31–33	Jan–Mar
Primary	Form Printing, Spiraling, Packaging, and Shipping	Winter/Spring	Month 31–33	Jan–Mar
Primary	Operational Test Administration	Spring	Month 35	May
Primary	Material/Data Processing and Scoring	Spring/Summer	Month 35–36	May–Jun
Primary	Score Reporting	Summer	Month 35–39	May–Sep
Secondary	Select Biennial Core-to-Core Overlap Items (Operational Items)	Summer/Fall	Month 51–52	Sep–Oct
Secondary	Build Operational Forms	Fall/Winter	Month 52–53	Oct–Nov
Secondary	Internal Form Reviews and PDE Reviews	Winter	Month 53–54	Nov–Dec
Secondary	Final Form and Printer Proof Approvals	Winter	Month 54–55	Dec–Jan
Secondary	Ancillary and Accommodated Form Development	Winter/Spring	Month 55–57	Jan–Mar
Secondary	Form Printing, Spiraling, Packaging, and Shipping	Winter/Spring	Month 56–58	Feb–Apr
Secondary	Second Operational Test Administration	Spring	Month 59	May
Secondary	Material/Data Processing and Scoring	Spring/Summer	Month 59–60	May–Jun
Secondary	Score Reporting	Summer	Month 59–63	May–Sep
Tertiary	Release Core-to-Core Overlap Items in Samplers	Fall	Month 63	Sep

Figure 3–1. DRC Item and Test Development Primary Cycle



GENERAL ITEM AND TEST DEVELOPMENT PROCESS

This section describes the processes which lead up to an operational exam. These processes were used to develop the entire pool of items that appeared in the field test and operational administrations.

ITEM DEVELOPMENT PLANNING MEETING

Prior to the start of any item development work, DRC's test development staff meets with PDE's assessment office to discuss the test development plans for the next administration, including the test blueprint, the field test plan (including development counts), procedures, timelines, etc. With a complete development cycle lasting about three years (from item authoring through field test, data review, and operational usage), the initial planning begins well in advance of the anticipated administration.

ITEM WRITER TRAINING

Item writers were selected and trained for the subject areas of Algebra I, Biology, and Literature. Qualified writers—either hired independently by the testing vendor, DRC, or through subcontractors like Victory—were college graduates with teaching experience and a demonstrated base of knowledge in the content area. Many of these writers were content assessment specialists and curriculum specialists. The writers were trained individually and had previous experience in writing MC and CR items. Prior to developing items for the Keystone Exams, the cadre of item writers was trained with regard to the following areas:

- Keystone Assessment Anchors and Eligible Content
- Webb's levels of cognitive complexity, DOK
- Subject-specific general scoring guidelines
- Specific and general guidelines for item writing
- Bias, fairness, and sensitivity guidelines
- Principles of universal design
- Item quality technical style guidelines
- Reference information
- Sample items

LITERATURE PASSAGE SELECTION

The task of searching for passages was conducted by DRC professionals with classroom experience in reading/language arts. These professionals also underwent specialized training (provided by DRC) in the characteristics of acceptable passages. Guidelines for passage selection included appropriate length, text structure, density, and vocabulary. A judgment was also made about whether the reading level required by a particular passage was at the independent level—that is, where the average student should be able to read 90 percent of words in the text independently. Passage finders were given the task of searching for a specified number of passages for each genre. Generally, they looked for at least twice as many passages as were needed. Passages acquired were either authentic (permissioned), in that they were culled from published materials, or commissioned by experienced authors. See Chapter Two for more information on the types of passages used on the Literature Keystone Exam.

For permissioned passages, approval to reprint was secured from the publishers. Passages underwent an internal review by several test development content editors to judge their merit with regard to the following criteria:

- Passages have interest value for students.
- Passages are appropriate in terms of vocabulary and language characteristics.
- Passages are free of bias, fairness, and sensitivity issues.
- Passages represent different cultures.
- Passages are from a variety of sources.
- Passages are able to stand the test of time.

- Passages are sufficiently rich to generate a variety of MC and CR items.
- Passages are complete with all necessary permissions documentation.
- Passages avoid dated subject matter unless a relevant historical context is provided.
- Passages should not require students to have extensive background knowledge in a certain discipline or area to understand a text.

Once through the internal review process, the passages deemed potentially acceptable were reviewed by the Reading Content Committee and the Bias, Fairness, and Sensitivity Committee for final approval.

ITEM AUTHORIZING AND TRACKING

Initially, items are generated with software-prepared Keystone Exams Item Cards, which allow for preliminary sorting and reviewing. Although very similar, the Keystone Exams Item Card for Multiple-Choice Items differs from the Keystone Exams Item Card for Constructed-Response Items in that the former has a location at the bottom of the card for comments regarding the distractors. Blank examples of these two cards are shown in Appendix D. In both instances, a column against the right margin includes codes to identify the subject area, grade, content categories, passage information (in the case of reading), item type, DOK (cognitive complexity), estimated difficulty, answer key (for MC items), and calculator use (for mathematics items).

All items undergoing field testing were entered into the DRC Item Development and Educational Assessment System (IDEAS), which is a comprehensive, secure, online item banking system. It accommodates item writing, item viewing and reviewing, and item tracking and versioning. IDEAS manages the transition of an item from its developmental stage to its approval for use in a test form (for both print and online delivery). The system supports item history records that include item usage within a form, item-level notes, content categories and subcategories, item statistics from both classical and Rasch item analyses, and classifications derived from analyses of differential item functioning (DIF). A sample IDEAS Item Card is presented in Appendix D.

INTERNAL REVIEWS AND PDE REVIEWS

To ensure that the items produced were sufficient in number and adequately distributed across subcategories and levels of difficulty, item writers were informed of the required quantities of items. As items were written, an item authoring card was completed. It contained information about the item, such as subject, content category, and subcategories. Based on the item writer's classroom teaching experience, his/her knowledge of the content area curriculum, and the cognitive demands required by the item, estimates were recorded for level of cognitive complexity and difficulty level. Items were written to provide for a range of difficulty and for cognitive complexity focused on DOK Level 3.

As part of the item construction process, each item was reviewed by content specialists and editors at DRC. Content specialists and editors evaluated each item to make sure that it measured the intended Eligible Content and/or Assessment Anchor. They also assessed each item to make certain that it was appropriate for the intended grade and that it provided and cued only one correct answer (MC items only). In addition, the difficulty level, DOK, graphics, language demand, and distractors were also evaluated. Other elements considered in this process included, but were not limited to, universal design, bias, source of challenge, grammar/punctuation, and Keystone style.

Following this internal process, items were reviewed by content specialists at PDE, who then consulted with DRC about any general issues or concerns (e.g., style, format, interpretation of Assessment Anchors and Eligible Content) and about edits to specific items. Following PDE's review, the items were prepared for the content review meetings conducted with Pennsylvania educators.

ITEM CONTENT REVIEWS

Prior to the 2021 field testing, all newly developed test items were submitted to content committees for review. The content committees consisted of Pennsylvania educators from school districts throughout the Commonwealth of Pennsylvania, some with postsecondary university affiliations. The primary responsibility of the content committee was to evaluate items with regard to quality and content classification, including grade-level (course) appropriateness, estimated difficulty, DOK, and source of challenge. They also suggested revisions and made recommendations for reclassification of items. In some cases when an item was deleted, the committee suggested a replacement item and/or reviewed a suggested replacement item provided by the facilitators. The committee also reviewed the items for adherence to the principles of universal design, including language demand and issues of bias, fairness, and sensitivity.

With source of challenge, items were identified where the cognitive demand was focused on an unintended content, concept, or skill (Webb, 2002). Source of challenge may be a contributing factor if the reason that an answer could be given results from a cultural bias, an inappropriate reading level, a flawed graphic in an item, or if an item requires specialized, noncontent-related knowledge to answer. Source of challenge could result in a student who has mastered the intended content or skill answering the item incorrectly or a student who has not mastered the intended content or skill answering the item correctly. Committee members were asked to note any items with a source of challenge and to suggest revisions to remove the source of challenge.

The content review meetings were held on August 4–7, 2020, for Algebra I, Biology, and Literature. Committee members were approved by PDE, and PDE-approved invitations were sent to them by DRC. PDE also selected internal staff members for attendance. The meeting commenced with a welcome by PDE and DRC. This was followed by an overview of the test development process by DRC. PDE, along with DRC, also provided training on the procedures and forms to be used for item content review.

DRC content assessment specialists facilitated the reviews and were assisted by representatives of PDE. Committee members, grouped by exam, worked through and reviewed the items for quality and content, as well as for the following categories:

- Assessment Anchor alignment
- Content limits
- Grade-level (course-level) appropriateness
- Difficulty level
- DOK
- Appropriate source of challenge
- Correct answer
- Quality of distractors
- Graphics in regards to appropriateness
- Appropriate language demand
- Freedom from bias

The members then came to consensus and assigned a status to each item: Approved, Accepted with Revision, or Rejected. All comments were recorded, and a master rating sheet was completed. Committee facilitators recorded the committee consensus on the Item Review Rating Sheet. A sample form and rating criteria may be found in Appendix E.

Security was addressed by adhering to a strict set of procedures. Items in binders were distributed for committee review by number and signed for by each member on a daily basis. All attendees, with the exception of PDE staff, were required to sign a confidentiality agreement. All materials not in use at any time were stored in a locked room. Secure materials that did not need to be retained after the meetings were deposited in secure barrels, the contents of which were shredded.

BIAS, FAIRNESS, AND SENSITIVITY REVIEWS

Prior to field testing, all newly developed test items were also submitted to a Bias, Fairness, and Sensitivity Committee for review. This review took place from July 27–31, 2020, for Algebra I, Biology, and Literature. The committee’s primary responsibility was to evaluate items with regard to bias, fairness, and sensitivity issues. They also made recommendations for changes or deletion of items in order to remove the potential for issues of bias, fairness, and/or sensitivity. Included in the review were proposed reading passages. An expert, multiethnic committee composed of men and women was trained by a DRC test development lead to review items for bias, fairness, and sensitivity issues. Training materials included a manual developed by DRC (DRC, 2017). Members of the committee also had expertise with special needs students and English Learners (EL). PDE staff members were also trained and participated in the review. All items were read by a cross-section of committee members. Each member noted bias, fairness, and/or sensitivity comments on tracking sheets and on the item, if needed, for clarification. Committee members individually categorized any concerns as related to ageism, disability, ethnicity/culture, gender, region, religion, socioeconomics, or stereotypes. These categories were the framework through which recommendations for modification or rejection of items occurred during the subsequent committee consensus process. The committee discussed each of the issues as a group and came to consensus as to which decisions should represent the view of the committee. All consensus comments were then compiled, and the suggested actions on these items were recorded and submitted to PDE. This review followed the same security procedures as outlined above. The table below shows the gender and race/ethnicity of the members of the bias committee who reviewed the Keystone items and passages for bias, fairness, and sensitivity.

Table 3–18. Demographic Composition of the 2019 Keystone Bias, Fairness, and Sensitivity Committee

Member #	Gender	Race/Ethnicity	Background
1.	Male	Asian American	National Consultant (Retired Educator)
2.	Male	Asian American	Educator
3.	Female	Asian American	Building Administrator
4.	Male	Caucasian American	National Consultant (Educator)
5.	Female	Caucasian American	Educator (SPED expertise)
6.	Female	Caucasian American	Educator
7.	Female	Caucasian American	National Consultant (Educator)
8.	Male	Caucasian American	University Professor
9.	Female	African American	Educator (SPED expertise)
10.	Female	African American	Retired Educator
11.	Female	Latino	National Consultant (Community Leader, Disability Rights Activist)
Totals	7 Females, 4 Males	1 Latina, 3 Asian Americans, 5 Caucasian Americans, 2 African Americans	

The results from the 2019 Bias, Fairness, and Sensitivity Committee reviews are summarized in the next set of tables.

Table 3–19A. Number of Items—Bias, Fairness, and Sensitivity Committee Review: Algebra I

Date	Total Reviewed	Accepted As Is	Accepted with Revision	Rejected
July-August 2019	270 Items	269 Items	1 Item	0 Items

Table 3–19B. Number of Items—Bias, Fairness, and Sensitivity Committee Review: Biology

Date	Total Reviewed	Accepted As Is	Accepted with Revision	Rejected
July-August 2019	16 Scenarios, 410 Items	16 Scenarios, 410 Items	0 Scenarios, 0 Items	0 Scenarios, 0 Items

Table 3–19L. Number of Items—Bias, Fairness, and Sensitivity Committee Review: Literature

Date	Total Reviewed	Accepted As Is	Accepted with Revision	Rejected
July-August 2019	22 Passages, 343 Items	22 Passages, 343 Items	0 Passages, 0 Items	0 Passage, 0 Items

Table 3–19T. Number of Items—Bias, Fairness, and Sensitivity Committee Review: Total

Date	Total Reviewed	Accepted As Is	Accepted with Revision	Rejected
July-August 2019	16 Scenarios, 22 Passages, 1,023 Items	16 Scenarios, 22 Passages, 1,022 Items	0 Scenarios, 0 Passages, 1 Item	0 Scenarios, 0 Passages, 0 Items

CHAPTER FOUR: UNIVERSAL DESIGN PROCEDURES APPLIED TO THE KEYSTONE EXAMS TEST DEVELOPMENT PROCESS

UNIVERSAL DESIGN

Universally designed assessments allow participation of the widest possible range of students and contribute to valid inferences about participating students. Principles of Universal Design are based on the premise that each child in school is a part of the population to be tested and that testing results should not be affected by disability, gender, race, or English language ability (Thompson, Johnstone, & Thurlow, 2002). At every stage of the item and test development process, procedures were employed to ensure that items and subsequent tests (in both print and online delivery methods) were designed and developed using the elements of universally designed assessments established by the National Center on Educational Outcomes (NCEO).

Federal legislation addresses the need for universally designed assessments. The No Child Left Behind Act (Elementary and Secondary Education Act) requires that each state must “provide for the participation in [statewide] assessments of all students” [Section 1111(b)(3)(C)(ix)(I)]. Both Title I and IDEA regulations call for universally designed assessments that are accessible and valid for all students, including English Learners and students with disabilities. The benefits of universally designed assessments not only apply to these groups of students, but to all individuals with wide-ranging characteristics.

DRC’s test development team was trained in the elements of Universal Design as they relate to developing large-scale statewide assessments. Team leaders were trained directly by NCEO, and other team members were subsequently trained by team leaders. Committees involved in content review included some members who were familiar with the unique needs of students with disabilities and English Learners. Likewise, some members of the Bias, Fairness, and Sensitivity Committee were conversant with these issues. What follows are the Universal Design guidelines that were followed during all stages of the item development process for the Keystone Exams.

ELEMENTS OF UNIVERSALLY DESIGNED ASSESSMENTS

After a review of research relevant to the assessment development process and the Principles of Universal Design (Connell et al., 1997), NCEO has produced seven elements of Universal Design as they apply to assessments (Thompson, Johnstone, & Thurlow, 2002). These elements served to guide item development for the Keystone Exams.

- **Inclusive Assessment Population**

The target population includes students attending Commonwealth schools who participate in one or more of the graduation competency exams.

- **Precisely Defined Constructs**

An important function of well-designed assessments is that the assessments actually measure what they are intended to measure. The Keystone Exams Assessment Anchor Content Standards (Assessment Anchors) provided clear descriptions of the constructs to be measured on each of the exams. Universally designed assessments must remove all non-construct-oriented cognitive, sensory, emotional, and physical barriers.

- **Accessible, Nonbiased Items**

DRC conducted both internal and external reviews of items and test specifications to ensure that they did not create barriers due to lack of sensitivity to disability, culture, or other subgroups. Items and test specifications were developed by a team who understood the varied characteristics of items that might create difficulties for any group of students. Accessibility is incorporated as a primary dimension of test specifications, so accessibility was woven into the fabric of the test rather than being added after the fact.

- **Amenable to Accommodations**

Even though items on universally designed assessments are accessible for most students, there are some students who continue to need accommodations. This essential element of a universally designed assessment requires that the exam is compatible with accommodations and a variety of widely used adaptive equipment and assistive technology (see also the section on Assessment Accommodations later in this chapter).

- **Simple, Clear, and Intuitive Instructions and Procedures**

Assessment instructions should be easy to understand regardless of a student’s experience, knowledge, language skills, or current concentration level. Questions that are posed using complex language can invalidate the test if students cannot understand how they are expected to respond to a question. To meet this guideline, directions and questions were prepared in simple, clear, and understandable language that underwent multiple reviews.

- **Maximum Readability and Comprehensibility**

A variety of guidelines exist to ensure the maximum readability and comprehensibility of a test. These features go beyond what is measured by readability formulas. Readability and comprehensibility are affected by many factors, including student background, sentence difficulty, and text organization. All of these features were considered as item text was developed.

Plain language is a concept now being highlighted in research on assessments. Plain language has been defined as language that is straightforward and concise. The following strategies for editing text to produce plain language were used during the editing process of the new Keystone Exam items:

- Reduction of excessive length
- Use of common words
- Avoidance of ambiguous words
- Avoidance of irregularly spelled words
- Avoidance of proper names
- Avoidance of inconsistent naming and graphic conventions
- Avoidance of unclear signals about directing attention

- **Maximum Legibility**

Legibility is the physical appearance of text, the way that the shapes of letters and numbers enable people to read text easily. Bias can result when tests contain physical features that interfere with a student’s focus on or understanding of the constructs that test items are assessing. A style guide was developed and utilized that included dimensions of style consistent with Universal Design.

GUIDELINES FOR UNIVERSALLY DESIGNED ITEMS

All test items written and reviewed adhered closely to the following guidelines for Universal Design. Item writers and reviewers used a checklist during the item development process to ensure that each aspect was followed. For more information on the checklist, see the Universal Design Overview section in Chapter Three of this report.

1. **Items measure what they are intended to measure.** Item writing training included making certain that writers and reviewers had a clear understanding of Pennsylvania’s Academic Standards and the Keystone Assessment Anchors. During all phases of test development, items were presented with content-standard information to ensure that each item reflected the intended Assessment Anchor. Careful consideration of the content standards was important in determining which skills involved in responding to an item were extraneous and which were relevant. With certain types of items, an additional skill was necessary, such as the Algebra I test, which requires the student to read.
2. **Items respect the diversity of the assessment population.** To develop items that avoid content that might unfairly advantage or disadvantage any student subgroup, item writers, test developers, and reviewers were trained to write and review items to avoid issues of bias, fairness, and sensitivity. Training also included an awareness of and sensitivity to issues of cultural and regional diversity.

3. **Items have a clear format for text.** Decisions about how items were presented to students must allow for maximum readability for all students. Appropriate fonts and point sizes were employed with minimal use of italics, which are far less legible and are read considerably more slowly than standard typeface. Captions, keys, and legends were at least a 12-point size, while footnotes and sentence numbers used a 10-point font.¹ Legibility was enhanced by sufficient spacing between letters, words, and lines. Blank space was used around paragraphs and between columns and staggered right margins.
4. **Stimuli and items have clear pictures and graphics.** When pictures and graphics were used, they were designed to provide essential information in a clear and uncluttered manner. Illustrations were placed directly next to the information to which they referred, and labels were used when possible. Sufficient contrast between the background and text, with minimal use of shading, increased readability for students with visual impairments. Color was not used to convey important information.
5. **Items have concise and readable text.** Linguistic demands of stimuli and items can interfere with a student's ability to demonstrate knowledge of the construct being assessed. During item writing and review, the following guidelines were used:
 - Simple, clear, commonly used words were used whenever possible.
 - Extraneous text was omitted.
 - Vocabulary and sentence complexity were appropriate for the grade level being assessed.
 - Technical terms and abbreviations were used only if they were related to the content being measured.
 - Definitions and examples were clear and understandable.
 - Idioms were avoided unless idiomatic speech was being assessed.
 - Questions to be answered were clearly identifiable.
6. **Items allow changes to format without changing meaning or difficulty.** A Braille version was available for each operational exam. Attention was given to using items that allow for Braille. Specific accommodations were permitted, such as signing to a student, the use of oral presentation under specified conditions, and the use of various assistive technologies. A Spanish version of the Algebra I and Biology exams was available for use by English Learners who would benefit from this accommodation and who were in U.S. schools for less than three years.
7. **The test has an overall appearance that is clean and organized.** Information was organized in a left-right, top-bottom format. Images, pictures, and text that may not be necessary (e.g., sidebars, overlays, callout boxes, shading, visual crowding caused by excess information) and that could be potentially distracting to students were avoided. Also avoided were purely decorative features that did not serve a purpose.

ITEM DEVELOPMENT

DRC worked closely with the Pennsylvania Department of Education to ensure that the Keystone Exams complied with nationally recognized principles of Universal Design. The implementation of accommodations on large-scale statewide assessments for students with disabilities was supported in the development of the Keystone Exams. In addition to the principles of Universal Design as described in the Pennsylvania Technical Report, DRC applied to each exam the standards for test accessibility as described in *Tests Access: Making Tests Accessible for Students with Visual Impairments—A Guide for Test Publishers, Test Developers, and State Assessment Personnel* (Allman, 2004).

¹ While font size follows specific requirements during online setup of an exam, the screen resolution used at the local level can impact whether the effective font size is visible to the student.

To this end, DRC embraces the following precepts:

- Test directions are worded to allow for alternate responses to constructed-response items.
- During item and bias reviews, committee members are made aware of the Principles of Universal Design and of issues that may adversely affect students with disabilities. The goal is to make certain that the Keystone Exams are bias free for all students. With the goal of ensuring that the Keystone Exams are accessible to the widest range of diverse student populations, PDE instructs DRC to limit item types that are difficult to format in Braille and that may become distorted when published in large print. DRC is instructed to limit the following on the Keystone Exams.
 - Algebra I: Complicated tessellations; charts or graphs that extend beyond one page
 - Literature: Graphics and illustrations that are not germane to the content presented
 - All exams: Unnecessary boxes and framing of text, unless enclosing the text provides necessary context for the student; use of italics (limited to only when it is absolutely necessary, such as with variables)

ITEM FORMAT

For all Keystone Exams (both online and print), DRC formats the items to maximize accessibility for all students by using text that is in an easily readable size and font style. DRC limits shading, graphics, charts, and the number of items per page so that there is sufficient white space on each page. Whenever possible, DRC ensures that graphics, pictures, diagrams, charts, and tables are positioned on the page with the associated test items. DRC uses high contrast for text and background when possible to convey pertinent information. Tests are published on dull-finish paper to avoid the glare encountered on glossy paper. DRC pays close attention to the binding of the exam books to ensure that they lie flat for two-page viewing and ease of reading and handling.

DRC ensures consistency across Keystone Exams by following these Principles of Universal Design:

- High contrast and clarity is used to convey detailed information.
- Typically, shading is avoided; when necessary for content purposes, 10-percent screens are used as the standard.
- Overlaid print on diagrams, charts, and graphs is avoided.
- Charts, graphs, diagrams, and tables are clearly labeled with titles and with short descriptions when applicable.
- Only relevant information is included in diagrams, pictures, and graphics.
- Symbols used in keys and legends are meaningful and provide reasonable representations of the topics they depict.
- Pictures that require physical measurement are true to size.

ASSESSMENT ACCOMMODATIONS

While universally designed assessments provide for participation of the widest range of students, many students require accommodations in order to participate in the regular assessment. Clearly, the intent of providing accommodations for students is to make certain that students are not unfairly disadvantaged during testing and that the accommodations used during instruction, if appropriate, are made available as students take the test. The literature related to assessment accommodations is still evolving and often focuses on state policies regulating accommodations rather than on providing empirical data that supports the reliability and validity of the use of accommodations. On a yearly basis, the Pennsylvania Department of Education examines accommodations policies and current research to ensure that valid, acceptable accommodations are available for students. Accommodations manuals for Pennsylvania assessments titled *Accommodations Guidelines* and *Accommodations Guidelines for English Learners* were developed for use with the Keystone Exams. The PDE guideline manuals can be accessed by going to www.education.pa.gov.

In addition, Spanish-language versions, translated from the original English versions were made available for both the Algebra I Exam and the Biology Exam. The Spanish-translation editions of the exams are discussed in Chapter Six.

CHAPTER FIVE: EMBEDDED FIELD TEST

FIELD TEST OVERVIEW

Every Keystone spring administration, field-test items are embedded on operational forms. The main purposes of field testing items prior to future operational use are (a) to calculate item statistics, (b) to determine if items meet the criteria with respect to statistical properties for future operational use, and (c) to obtain item parameters for pre-equating purposes. In comparison to stand-alone field tests, embedded field tests (EFT) allow for more accurate item statistics and item parameters by alleviating concerns of whether students may perceive differences between field-test and operational items. The EFT approach allows item parameters to be used for future pre-equating purposes and is based on the assumption that students should be equally motivated to take the operational and field-test items because they are not aware of which items are field-test items. To minimize item context and item position effects (i.e., lack of motivation and fatigue), field-test items were interspersed within the operational sections. With this design, students have a smaller chance of knowing the field-test item positions.

Keystone test forms contained common operational items that were identical on all forms along with embedded field-test items. In most instances, the field-test items were unique to a given form; however, there were instances in which an embedded field-test item appeared on more than one form. More information on the field-test designs for all subjects can be found in the content-specific portions of Chapter Three. In general, the field-test items for each year represent a portion of the following year's operational forms (Spring, Summer, and Winter). In other words, items from the field-test in Spring 2018 were selected for the Spring 2019, Summer 2019, and Winter 2019–2020 operational forms. This chapter presents information about the 2019 EFT, including classical item analyses, Differential Item Functioning (DIF) analyses, identification of items for data review, and outcomes from data review.

Due to circumstances surrounding the global Covid-19 pandemic beginning in March 2020 and lasting throughout the 2020–2021 school year, several processes were modified. Several of the administration conditions in 2021 were adjusted to allow for additional testing (e.g., elongated testing windows, waiving of accountability), which ultimately changed the administration conditions of the Keystone exams. The Pennsylvania Technical Advisory Committee (TAC) provided guidance and direction for processes that were modified. With respect to field-test processes, the 2021 field-test items will not be eligible for future forms construction given that Keystone is predicated on pre-equating; therefore, item statistics for 2021 FT items are not presented in this chapter. The information provided in this chapter discusses the typical procedures for Keystone, with footnotes that outline any changes for the 2021 processes.

The number of field-test positions on spring forms varies by subject. Table 5–1 displays the number of field-test positions by subject and module. It is important to note that the number of item positions and the number of field-test items on a spring administration may differ slightly.

Table 5–1. Number of FT Item Positions by Item Type (MC, CR), Content Area and Module

Content Area	Module	MC Item Positions per Form	CR Item Positions per Form	Passages per Form	MC Item Positions (Total)*	CR Item Positions (Total)*	Passages (Total)*
Algebra I	1	5	1	0	100	20	0
Algebra I	2	5	1	0	100	20	0
Algebra I	Total	10	2	0	200	40	0
Biology	1	8	1	0	160	20	0
Biology	2	8	1	0	160	20	0
Biology	Total	16	2	0	320	40	0
Literature	1	6	1	1	120	20	10
Literature	2	6	1	1	120	20	10
Literature	Total	12	2	2	240	40	20

*Total FT Item Positions across 20 forms.

Note. Each FT passage is repeated across 2 forms with different sets of FT items to yield the highest volume of eligible items for subsequent forms selection.

CLASSICAL ITEM ANALYSIS

Classical item analyses of field-test items are conducted in order to assess the items' quality and to identify items for data review. Specifically, item difficulty (also referred to as p -values, which represent the proportion of students correctly answering the item) and item-total correlations (the relationship between whether a student answers an item correctly and that student's total test score) are estimated for the item, for each option for MC items, and for each score point for multi-point items.

ITEM DIFFICULTY

At the most general level, an item's difficulty is indicated by its mean score in some specified group (e.g., content area).

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students (n). For MC items, student scores are represented by 0's and 1's (0 = wrong, 1 = right). With 0/1 scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. So this is also the *proportion correct* for the item, or as it is better known, the p -value. In theory, p -values can range from 0.00¹ to 1.00 on the proportion-correct scale. For example, if an item has a p -value of 0.89, it means 89 percent of students answered the item correctly. Additionally, this value might suggest that the item is relatively easy or that the students who attempted the item are relatively high achievers. In other words, item difficulty and student ability are somewhat confounded.

For CR items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score (e.g., four points in the case of Algebra I CR items, three points in the case of Biology and Literature CR items). Sometimes a *pseudo p*-value is provided for a CR item by dividing the mean item score by the maximum possible item score.

¹ For MC items with four response options, pure random guessing would lead to an expected p -value of 0.25.

The minimum and maximum extremes of the difficulty scale are virtually never seen in applied practice. However, understanding what those values are helps illustrate that relatively lower values correspond to more difficult items and that relatively higher values correspond to easier items. (Because of this, some assert that this index would be better referred to as the item's *easiness*.)

Item difficulty is an important consideration for the Keystone Exams because of the various student achievement levels (Below Basic, Basic, Proficient, and Advanced). Items that are either very hard or very easy provide little information about student differences in achievement. However, an item answered correctly by a high percentage of students would suggest that the knowledge or skill the item measures has been mastered by most students. Conversely, an item answered correctly by a low percentage of students would suggest that few students have mastered the knowledge or skill the item measures. So on a criteria-referenced test like the Keystone Exams, a test development goal is to include a wide range of item difficulties.

ITEM DISCRIMINATION

At the most general level, item discrimination² indicates an item's ability to differentiate between high and low achievers. It is expected that students with high ability (i.e., those who perform well on the Keystone Exams overall) would be more likely to answer any given item correctly, while students with low ability (i.e., those who perform poorly on the Keystone Exams overall) would be more likely to answer the same item incorrectly. For the Keystone Exams, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate discrimination. As commonly practiced, Data Recognition Corporation (DRC) removes the item score from the total score so that the resulting correlations will not be spuriously high. The correlation coefficient can range from negative 1.0 to positive 1.0. If higher-scoring students tend to answer the item correctly while lower-scoring students do not answer the item correctly, the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., well above zero), indicating that the item is good at discriminating between higher-scoring and lower-scoring students.

Item-total correlation for each option is another indicator of an item's ability to differentiate between high and low achievers. It is expected that students with high ability would be less likely to choose a distractor, while students with low ability would be more likely to choose a distractor. In other words, the item-total correlations for the distractors are expected to be negative.

In summary, the correlation will be positive in value when the mean test score of the students answering the item correctly is higher than the mean test score of the students answering the item incorrectly.³ In other words, students who did well on the total test tended to do well on the item as well. However, an interaction can exist between item discrimination and item difficulty. Items answered correctly (or incorrectly) by a large proportion of examinees (i.e., items with extreme p -values) can have reduced power to discriminate and thus can have lower correlations.

CLASSICAL ITEM ANALYSIS RESULTS

Table 5–2 provides the summary statistics for the difficulty and discrimination for the 2019 field-test items with respect to each content area and item type (see Chapter Eleven for summary statistics for operational items). There is a range of p -values across all content areas, where mean p -values were between 0.52 to 0.67 for MC items and 0.23 to 0.53 for CR items. The mean item-total correlations were between 0.34 to 0.35 for MC items and 0.61 to 0.66 for CR items.

² As noted earlier, the discrimination index for dichotomous MC items is typically referred to as the *point-biserial correlation coefficient*. For CR items, the term *item-test correlation* is sometimes used.

³ It is legitimate to view the point-biserial correlation as a standardized mean. A positive value indicates that students who chose that response had a higher mean score than the average score; a negative value indicates that students who chose that response had a lower mean score than the average score.

Table 5–2. Summary Statistics of Difficulty and Discrimination by Content Area and Item Type

Content Area	Item Type	N	Mean <i>p</i> -value	Min <i>p</i> -val.	Median <i>p</i> -value	Max <i>p</i> -value	Mean I-T Corr.	Min I-T Corr.	Median I-T Corr.	Max I-T Corr.
Algebra I	CR	37	0.23	0.05	0.25	0.41	0.61	0.42	0.63	0.73
Algebra I	MC	200	0.52	0.12	0.51	0.85	0.34	-0.10	0.36	0.56
Biology	CR	40	0.46	0.26	0.45	0.68	0.63	0.27	0.65	0.71
Biology	MC	320	0.53	0.16	0.52	0.88	0.35	-0.07	0.37	0.58
Literature	CR	40	0.53	0.45	0.54	0.61	0.66	0.54	0.66	0.71
Literature	MC	240	0.67	0.20	0.69	0.95	0.35	-0.05	0.37	0.55

Note. I-T Corr. is the item-test score correlation.

DIFFERENTIAL ITEM FUNCTIONING

Differential item functioning (DIF) occurs when examinees who have the same ability level but belong to different groups do not have the same probability of answering an item correctly. When the probability differs, it is important for content experts to review items for any potential *item bias*. It is important to note that, as a statistical concept, DIF is different from item bias. DIF detects a difference in performance after controlling for student ability, whereas bias is a content issue that can arise in situations where something other than the intended construct of measurement affects the probability of a correct response for a particular group. For example, bias is present when an item presents negative group stereotypes that draw the attention of the examinee, uses non-construct relevant language that is more familiar to one subpopulation than to another, or is presented in a non-construct relevant format that disadvantages certain learning styles. While the source of item bias can be plain to trained judges, DIF may have no clear cause. In such cases, something other than bias, including construct relevant content, may be explaining the differential performance on the item. Flagging items with DIF provides the opportunity for reviewers to assess and correct potential bias, but DIF does not necessarily mean that bias is present.

DIF DETECTION PROCEDURES

For MC items, the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959) for detecting DIF is a commonly used technique. It does not depend on the application or the fit of any specific measurement model. However, it does have significant philosophical overlap with the Rasch model since it uses a test's total score for the analysis.

The MH procedure as implemented by DRC contrasts a focal group with a reference group. While it makes no practical difference in the analysis which group is defined as the focal group, the group most likely to be disadvantaged by a biased measurement is typically defined as the focal group. In these analyses, the focal group was female for gender-based DIF, black and Hispanic for ethnicity-based DIF, and the computer-based-test (CBT) group for the test administration mode-based DIF; reference groups were male, white, and the paper-and-pencil test (PPT) group for the gender, ethnicity, and mode-based DIFs, respectively. The MH statistic for each item is computed from a contingency table. The table has two groups (focal and reference) and two outcomes (right or wrong). The ability groups are defined by the total score distribution for the total examinee populations.

The basic MH statistic is a single degree of freedom chi-square that compares the observed number in each cell to the expected number. The expected counts are computed to ensure that the analysis is not confounded with differences in the achievement level of the two groups.

For CR items, a comparable statistic is computed based on the standardized mean difference (SMD) (Dorans, Schmitt, & Bleistein, 1992), computed as the differences in mean scores for the focal and reference groups if both groups had the same score distribution.

To assist the review committees in interpreting the analyses, all items are assigned a severity code (A, B, or C) based on the magnitude of the MH statistic and a direction (minus or plus) based on the direction of the MH statistic. Items classified as A+ or A- have little or no statistical indication of DIF. Items classified as B+ or B- have some indication of DIF but may be judged to be acceptable for future use. Items classified as C+ or C- have strong evidence of DIF and should be reviewed and possibly rejected from the eligible item pool. A plus sign indicates that the item favors the focal group, and a minus sign indicates that the item favors the reference group.

LIMITATIONS OF STATISTICAL DETECTION

No statistical procedure should be used as a substitute for rigorous, hands-on reviews by content and bias specialists. The statistical results can help organize reviews so the effort is concentrated on the most problematic cases. Further, no items should be automatically rejected simply because a statistical method flagged them or automatically accepted because the statistical method did not flag them.

Statistical detection of DIF is not an exact science. There have been a variety of methods proposed for detecting DIF, but no single statistic can be considered either necessary or sufficient. Different methods are more or less successful and detect DIF at different rates. No analysis can guarantee that a test is free of bias, but thoughtful item development and field-test analysis can prevent potentially biased items from unfairly impacting student scores.

A fundamental shortcoming of all statistical methods used in DIF evaluation is that they are all intrinsic to the test being evaluated. If a test is unbiased overall but contains one or two DIF items, any method can identify DIF. However, because all current methods use total test performance as the measure on which to control for group abilities, a test with all DIF items will not be able to separate DIF effects from true differences in achievement on the test.

CRITERIA FOR IDENTIFYING ITEMS

As previously discussed, all field-test items were analyzed statistically using conventional item analysis methods. For MC items, classical item statistics included the corrected point-biserial correlation (Pt. Bis.) for the correct and incorrect responses (distractors), the percentage correct (p -value), and the percentage incorrect. For constructed-response (CR) items, the statistical indices included the item-test correlation, the point-biserial correlation for each score point, the percentage of responses in each score point, and the percentage of non-scorable responses.

In general, more capable students are expected to respond correctly to easy items and less capable students are expected to respond incorrectly to difficult items. If either of these situations does not occur, the item in question will be reviewed by DRC test development staff and committees of Pennsylvania educators to determine the nature of the problem and the characteristics of the students affected. The primary ways of detecting such conditions are through the point-biserial correlation coefficient for dichotomously scored MC items and the item-total correlation for polytomously scored CR items. In each case the statistic will be positive if the total test mean score is higher for the students who respond correctly to MC items (or attain a higher CR item score) and negative when the reverse is true.

The following set of criteria was used to identify items for additional review.

For an MC item to be flagged, the item needed to meet any of the following criteria:

1. Percentage Correct (p -value) less than 0.3 or greater than 0.9
2. Item-total correlation for the correct response less than 0.25
3. Point-biserial correlation for any incorrect response greater than 0.0
4. Percentage responding to any incorrect responses greater than the percentage correct
5. Gender, ethnic, or mode DIF code of C+ or C-

For a CR item to be flagged, the item needed to meet any of the following criteria:

1. Percentage Correct (p -value) less than 0.3 or greater than 0.9
2. Item-total correlation for the correct response less than 0.25
3. Score proportion less than 0.05
4. Gender, ethnic, or mode DIF code of C+ or C-

The intent of the above criteria is to flag items for review. For this purpose, the preference is to over-identify rather than under-identify items to ensure they are appropriate for future use. Any of the criteria would cause the item to be reviewed by content experts, but there are many reasons the experts may want to keep an item in spite of the statistics.

RESULTS AND OBSERVATIONS

Each form has a set of operational items along with embedded FT items. Test forms are spiraled to ensure that a representative sample of students answer each set of FT items (see Chapter Nine for additional information about spiraling). For MC items, all data is used for classical item analyses and DIF analyses. In contrast, a random sample of responses are selected to be scored for FT analyses for CR items.

This section focuses on reporting the number (N) and percentage (%) of items flagged by different criteria (see Tables 5–3 to 5–5). For the DIF analysis, the number and percentage of items were provided not only for the C- and C+ bias codes, which were used as the criteria to flag items, but also for the bias codes A-, A+, B-, and B+.

Table 5–3. Summary of Items Flagged by the Classical Item Statistics

Item Type	Flagging Criterion	Alg. I Total N	Alg. I N	Alg. I %	Bio. Total N	Bio. N	Bio. %	Lit. Total N	Lit. N	Lit. %
MC	1 (Low P -value)	200	17	8.5	320	20	6.3	240	4	1.7
MC	2 (Low Point-Biserial Correlation)	200	43	21.5	320	67	20.9	240	47	19.6
MC	3 (Non-negative distractor correlation)	200	37	18.5	320	64	20.0	240	29	12.1
MC	4 (High distractor proportion)	200	31	15.5	320	27	8.4	240	11	4.6
CR	1 (Low P -value)	37	37	100.0	40	40	100.0	40	40	100.0
CR	2 (Low Point-Biserial Correlation)	37	35	94.6	40	32	80.0	40	40	100.0
CR	3 (Low score-point proportion)	37	10	27.0	40	0	0.0	40	0	0.0

Note. Refer to Criteria for Identifying Items.

Table 5–4. DIF Summary – MC Items

Reference Group	Focal Group	Bias Code	Alg. I (Total N=200) N	Alg. I (Total N=200) %	Biology (Total N=320) N	Biology (Total N=320) %	Literature (Total N=240) N	Literature (Total N=240) %
Male	Female	A-	94	47.0	156	48.8	111	46.3
Male	Female	A+	99	49.5	160	50.0	122	50.8
Male	Female	B-	4	2.0	3	0.9	5	2.1
Male	Female	B+	2	1.0	1	0.3	2	0.8
Male	Female	C-	1	0.5	0	0.0	0	0.0
Male	Female	C+	0	0.0	0	0.0	0	0.0
White	Black	A-	155	77.5	240	75.0	167	69.6
White	Black	A+	33	16.5	76	23.8	58	24.2
White	Black	B-	12	6.0	4	1.3	14	5.8
White	Black	B+	0	0.0	0	0.0	0	0.0
White	Black	C-	0	0.0	0	0.0	1	0.4
White	Black	C+	0	0.0	0	0.0	0	0.0
White	Hispanic	A-	143	71.5	256	80.0	175	72.9
White	Hispanic	A+	56	28.0	63	19.7	44	18.3
White	Hispanic	B-	1	0.5	1	0.3	19	7.9
White	Hispanic	B+	0	0.0	0	0.0	1	0.4
White	Hispanic	C-	0	0.0	0	0.0	1	0.4
White	Hispanic	C+	0	0.0	0	0.0	0	0.0
PPT	CBT	A-	56	28.0	125	39.1	124	51.7
PPT	CBT	A+	139	69.5	194	60.6	110	45.8
PPT	CBT	B-	3	1.5	0	0.0	2	0.8
PPT	CBT	B+	2	1.0	1	0.3	4	1.7
PPT	CBT	C-	0	0.0	0	0.0	0	0.0
PPT	CBT	C+	0	0.0	0	0.0	0	0.0

Table 5–5. DIF Summary – CR Items

Reference Group	Focal Group	Bias Code	Alg. I	Alg. I	Biology	Biology	Literature	Literature
			(Total N=37) N	(Total N=37) %	(Total N=40) N	(Total N=40) %	(Total N=40) N	(Total N=40) %
Male	Female	A-	11	29.7	16	40.0	0	0.0
Male	Female	A+	24	64.9	21	52.5	6	15.0
Male	Female	B-	1	2.7	0	0.0	0	0.0
Male	Female	B+	1	2.7	3	7.5	28	70.0
Male	Female	C-	0	0.0	0	0.0	0	0.0
Male	Female	C+	0	0.0	0	0.0	6	15.0
White	Black	A-	24	64.9	25	62.5	20	50.0
White	Black	A+	7	18.9	5	12.5	20	50.0
White	Black	B-	6	16.2	7	17.5	0	0.0
White	Black	B+	0	0.0	0	0.0	0	0.0
White	Black	C-	0	0.0	3	7.5	0	0.0
White	Black	C+	0	0.0	0	0.0	0	0.0
White	Hispanic	A-	30	81.1	34	85.0	21	52.5
White	Hispanic	A+	5	13.5	2	5.0	15	37.5
White	Hispanic	B-	2	5.4	4	10.0	3	7.5
White	Hispanic	B+	0	0.0	0	0.0	1	2.5
White	Hispanic	C-	0	0.0	0	0.0	0	0.0
White	Hispanic	C+	0	0.0	0	0.0	0	0.0
PPT	CBT	A-	28	75.7	8	20.0	28	70.0
PPT	CBT	A+	7	18.9	31	77.5	5	12.5
PPT	CBT	B-	2	5.4	0	0.0	6	15.0
PPT	CBT	B+	0	0.0	1	2.5	0	0.0
PPT	CBT	C-	0	0.0	0	0.0	1	2.5
PPT	CBT	C+	0	0.0	0	0.0	0	0.0

* Algebra I has 3 unscored CR items

REVIEW OF ITEMS WITH DATA

In the preceding section, it was stated that test development content-area specialists used certain statistics from classical item analyses and DIF analyses of the 2019 field test to identify items for review by Pennsylvania educators. Items not identified for this review had good statistical characteristics and, consequently, were entered into the eligible pool for future item selection. DRC content-area test development specialists and DRC psychometric specialists identified the remaining items for further review by a committee of Pennsylvania educators. The intent was to capture all items that needed a closer review; thus, the criteria employed tended to over-identify rather than under-identify items.

The review of the relevant items was conducted by more than 30 Pennsylvania educators (teachers and PDE staff) broken out into exam-based committees. The review took place on September 4 and 5, 2019. In these sessions, committee members were first trained by a representative from DRC’s psychometrics staff with regard to the statistical indices used in item evaluation. This training was followed by a discussion with examples concerning reasons an item might be retained regardless of the statistics. The committee review process involved a brief exploration of possible reasons for the statistical profile of an item (e.g., possible bias, grade appropriateness, instructional issues) and a decision regarding acceptance. DRC content-area test development specialists facilitated the review of the items. Each committee reviewed the pool of flagged EFT items and made recommendations on each item. The results of the committee reviews are shown in the table below.

Table 5–6. Data Review Committee Results

Exam	Module	Total No. of FT Items	No. Reviewed Items	% of Reviewed Items	No. of Items Rejected by Committee	% of Items Rejected by Committee	No. of Items Classified as Rejected*	% of Items Classified as Rejected*
Algebra I	1	120	45	37.5%	17	14.2%	18	15.0%
Algebra I	2	120	47	39.2%	18	15.0%	20	16.7%
Biology	1	180	41	22.8%	18	10.0%	18	10.0%
Biology	2	180	30	16.7%	14	7.8%	14	7.8%
Literature	1	140	25	17.9%	4	2.9%	4	2.9%
Literature	2	140	21	15.0%	2	1.4%	2	1.4%
	Total	880	209	23.8%	73	8.3%	76	8.6%

* Items Classified as “Rejected” from Spring 2019 EFT (all sources: Data Review Committee, PDE, and DRC)

CHAPTER SIX: OPERATIONAL FORMS CONSTRUCTION FOR 2021 ADMINISTRATIONS

FINAL SELECTION OF ITEMS AND KEYSTONE FORMS CONSTRUCTION

Approximately 50% of the items that made up the Spring 2021, Summer 2021, and Winter 2021/2022 operational forms emerged from the Spring 2019 embedded field test. The remaining operational (core) items were part of the biennial core-to-core overlap. For more information about the core-to-core overlapping items, please see Chapter Three. Prior to being placed on the operational tests, these items had undergone multiple reviews, including the following:

- Reviews by Data Recognition Corporation (DRC) content-area test development specialists and curriculum specialists to ensure that all items were properly aligned with content standards
- Formal bias, fairness, and sensitivity review by the Bias, Fairness, and Sensitivity Committee, which consisted of a multiethnic group of men and women having expertise with special-needs students and English Learners (EL)
- Formal review by the content committees consisting of Pennsylvania educators, including teachers as well as district personnel
- Pennsylvania Department of Education (PDE) review
- Item data review by members of the PDE subject-area teacher committees

The item and bias reviews are detailed in Chapter Three. The results of the data review are summarized in Chapter Five.

The end product of the above process was an item status designation for each field test item. All items having an item status code of Accepted/Operational Ready were candidates to be selected for the 2020 Keystone Exams. To have an item status code of Accepted/Operational Ready meant that the item met the following criteria:

- Appropriately aligned with its designated Keystone Assessment Anchor Content Standard (Assessment Anchor) and subclassifications
- Acceptable in terms of bias/fairness/sensitivity issues, including differential item functioning (for gender and ethnicity)
- Acceptable in terms of psychometric standards, including a special review of flagged items

Next, all relevant information regarding the acceptable items, including associated graphics, was entered into the item banking system known as IDEAS (Item Development and Education Assessment System). From IDEAS and other database sources, Microsoft Excel files were created for each exam. These files contained all relevant content codes and statistical characteristics. IDEAS also created an item card displaying each acceptable item, any associated graphic, and all relevant exam codes and item statistics for use by the subject-area test development specialists and psychometric services staff.

DRC test development specialists reviewed the test design blueprint, including the number of items per strand for each content-area test. Psychometricians provided content-area test development specialists with an overview of the psychometric guidelines for forms construction.

Senior DRC content-area test development specialists reviewed all items in the operational pool to make an initial selection (pull) for common (core) positions according to test blueprint requirements and psychometric guidelines. Changes to items were not encouraged since alterations could affect how an item would perform in subsequent testing.

For these common items, this meant that the combination of multiple-choice (MC) and constructed-response (CR) items would yield the appropriate range of points while tapping an appropriate variety of the Assessment Anchors and related Eligible Content within each Reporting Category (module). Items selected in the first round were examined with regard to how well they fit together as a set. Of particular concern were the following:

- One item providing cues as to the correct answer to another item
- Context redundancy (e.g., mathematics items with a sports context)
- Presence of clang (distractors not unique from one another)
- Diversity of names and artwork for gender and ethnicity

A core-building software tool known as PerForm was used in concert with performance data and metadata from IDEAS to aid in the organization and communication of the pulled data. PerForm automatically tabulates the statistical characteristics of the proposed core, updating instantly whenever item swaps were performed. Using PerForm, the first round of items was then evaluated for statistical features such as an acceptable point-biserial correlation and whether correct answers were distributed equally—that is, whether approximately 25 percent of correct answers appeared in each of the four possible positions (A, B, C, or D). Selected items that were deemed psychometrically less advantageous in contrast to the overall psychometric characteristics of the core resulted in a search by the senior reviewer for suitable replacements. At this point, the second round of items was analyzed. If necessary, this iterative process between content-based selections and statistical properties continued in an effort to reach the best possible balance.

Once the recommendations were finalized for the core items, they were submitted to PDE for review. Department staff provided feedback, which could be in the form of approval or recommendations for replacement of certain items. Any item replacement was accomplished by the collective effort of the test development specialists, psychometricians, and PDE staff until final PDE approval was given. See Appendix F for the Keystone Exams Tally Sheets.

Following final approval by PDE, test development specialists developed print and online forms based on the approved core and approved embedded field test items. Both modes of delivery were built using IDEAS. Highly skilled test development specialists and editors used specialized checklists to verify accuracy of layout and formatting in both modes of delivery. Following final approval to print, the documents were prepared for the printing presses.

SPECIAL FORMS USED WITH THE OPERATIONAL 2020 KEYSTONE EXAMS

SPANISH TRANSLATION

Starting with the operational exams in spring 2011, school personnel had the option of allowing Spanish-speaking students who had been enrolled in schools in the United States for less than three years to respond to a Spanish version of the Keystone Exams for Algebra I and Biology. The original translation of the items and the *Directions for Administration Manual* was initiated by Language Services Associates and completed/verified by Exact Communications. These companies use translators with varying cultural and regional backgrounds to create the Spanish versions. The translations were then reviewed and verified by DRC's internal Spanish group. As part of the internal review, a Spanish style guide is maintained to document Spanish word choice from administration-to-administration and across exams within an administration.

Following PDE's approval of the translation, the translated text was typeset into print delivery forms. The test book is constructed with a side-by-side format with the English text and Spanish-translated text on facing pages. The Spanish-translated text is on the left-hand side followed by the original English text on the right-hand (facing) side. Each CR item covered either two or four pages in the answer book, depending on the length of the original English-language item. In the case of four-page open-ended items, the first set of facing pages of an item was presented in Spanish. The second set of facing pages of an item was presented in the original English.

Those students using this accommodated version are permitted to write their answers on either the English language pages or on the translated Spanish language pages. Their answers can be written in English, Spanish, or a combination of both Spanish and English because all pages are evaluated and scored, and the highest possible scores from those combinations are recorded for the students.

On a yearly basis, the PDE examines accommodations policies and current research to ensure that valid, acceptable accommodations are available for students. Accommodations manuals for Pennsylvania assessments titled *Accommodations Guidelines* and *Accommodations Guidelines for English Learners (ELs)* were developed for use with the Keystone Exams. The PDE guideline manuals can be accessed by going to www.education.pa.gov. For more information about the general on-screen testing aids available to students taking the online mode of delivery, see Chapter Two.

AUDIO

For students requiring an auditory presentation accommodation, a text-to-speech synthesizer is available to students taking the Algebra I and/or Biology Exams using the online mode of test delivery. For each operational exam, one form was selected for the creation of the audio version. Special scripts are crafted, writing out each item, distractor, graphic, and directions to utilize the rich, synthesized voice features while accounting for specific nuances of the intended sounds. The resulting audio information is provided to students receiving the accommodation. Since additional software is required to generate the vocalization from the scripted text and since headphones are required to minimize disruptions within a computer lab setting, local school personnel generally must preplan to use the audio version in order to ensure that the student has a properly equipped computer and a proper setting.

BRILLE, LARGE PRINT, AND VIDEO SIGN LANGUAGE

Students were able to respond to test materials that were available in Braille, large print, or Video Sign Language. At each grade level assessed, one form was selected for the creation of these accommodations.

The large print edition is a replication of the standard print form; 8.5×11 standard form is enlarged to an 11×17 page format to achieve a font size of approximately 18-point. A side-by-side verification is completed between the standard print and large print forms to ensure that the integrity of all formatting and graphics is maintained on the large print forms.

For Braille production, the final selected form is delivered to American Printing House for the Blind (APH) via APH's secure website. APH ensures that all tests are translated correctly and accurately by using a translator and a validator. After all Braille booklets are printed, APH conducts a quality assurance step to ensure all items are bound in order and directions are included. All Braille booklets are shipped from APH to DRC via UPS.

DRC applies a security barcode to each large print and Braille booklet for purposes of shipping, distributing, and collecting the materials. This security barcode is used with DRC's Operations Materials Management System (Ops MMS).

School personnel were directed to transcribe all student answers (SR and CR) into scannable answer documents exactly as the student responded. No alterations or corrections of student work were permitted, and the transcribed answer document had to have the same form designation as the Braille and large print version.

DRC utilizes Victory Productions for the production of Sign Language Videos. The items are passed to Victory Productions via a secure ftp site. Two to three different interpreters are used to interpret and validate the translations during video recording. After the interpretations are recorded and returned to DRC via a secure ftp site, DRC loads these videos in the online test engine. When school personnel assign the specific sign language accommodation, the student will be able to play each video next to the item.

CHAPTER SEVEN: TEST ADMINISTRATION PROCEDURES

SECTIONS, SESSIONS, TIMING, AND LAYOUT OF THE KEYSTONE EXAMS

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

The design for most Keystone Exams utilizes separate test books and answer books. An answer book is used to respond to the multiple-choice (MC), evidence-based selected-response (EBSR), and constructed-response (CR) items and to collect demographic information. The MC items and all stimulus text are placed within the test book. One exam uses a single consumable book. When a single scannable answer book is utilized, the contents of the answer book and the test book are combined into one integrated book. The table below identifies the exam material format for each 2020 Keystone Exam.

Table 7–1. Book Type by Exam

Exam	Test Book	Answer Book	Single Consumable Book
Algebra I	✓	✓	
Biology	✓	✓	
Literature	✓	✓	

Generally, a separate test book and answer book are used to separate the MC items and the CR items. For passage-based exams, like Literature, the separate exam materials allow the students to reference stimulus materials at the same time that a response to a CR item is composed. In addition, since all student responses must be scanned for scoring and storage purposes, a separate answer book limits the volume of data that must be stored.

SECTIONS AND SESSIONS

Each operational Keystone Exam is organized around two equally sized test modules; the focus of each is on two or more specific, thematically linked Assessment Anchors and Eligible Content. The content in each module remains separate, and items measuring the Eligible Content in a module appear only in that module. The module design is identical in the print (paper-and-pencil) and online modes of delivery.

Each exam section is administered in an exam session. Local districts must schedule the two modules as two separate exam sessions (morning and afternoon or two separate days), and an individual module must be completed in one exam session.

Each test session is to be completed within a prescribed testing window. The testing windows below reflect both online and paper-based administrations in the 2020–2021 school year. The testing windows also include all make-up testing. Schools were able to choose one of the two testing windows (“waves”) for the winter administration. Two windows were provided to accommodate different semester end dates for schools with block scheduling.

Table 7–2. Winter 2020/2021 Operational Keystone Exam Testing Windows

Exams	Wave 1 Dates	Wave 2 Dates
Algebra I, Biology, Literature	December 1–15, 2020	January 4–March 31, 2021

Table 7–3. Spring 2021 Operational Keystone Exam Testing Window

Exams	Dates
Algebra I, Biology, Literature	May 17–September 30, 2021

Table 7–4. Summer 2021 Operational Keystone Exam Testing Window

Exams	Dates
Algebra I, Biology, Literature	

TIMING

In general, the estimated testing times allow 1–2 minutes per MC item on the Keystone Exams, depending on the exam. The CR items are estimated to take approximately 5–10 minutes per item, also depending on the exam. Each stimulus passage on the Literature exam is estimated to take about 10 minutes to read. There was no difference in the timing for online and print forms of delivery.

Test administrators were instructed that each section (module) in a form should be scheduled as a separate exam session. Exam modules were not to have been scheduled back-to-back in the morning (or in the afternoon). Instead, the exam modules were to be divided across two days or divided across the morning and afternoon of the same day.

Since not all students are expected to finish the exam sections at the same time, test administrators are advised to use the flexibility of the time limits to the students' advantage. For example, test administrators manage the testing time so that students do not feel rushed while they are taking any assessment section, and no student is penalized because he or she works slowly. It is also stressed to test administrators that a student should not be given an opportunity to waste time. Students are told to close their exam materials when they have finished the section of the exam in which they have been working. Students who finish early are allowed to sit quietly or read for pleasure until all students have finished. Students with special requirements and/or abilities (i.e., physical, visual, auditory, or learning disabilities as defined by their IEP or service contracts) and students who just work slowly may require extended time. Special assessment situations are arranged for these students. When all students in a testing session indicate that they have finished an exam section, test administrators end the section.

Scheduled extended time is provided by a test administrator, and students are allowed to request extended time if they indicate that they have not completed the task. Such requests are granted if the test administrator finds the request to be educationally valid. Test administrators are advised that not permitting ample time for students to complete the assessment might impact the students' and school's performances.

As a general guideline, however, when all students indicate that they have finished a section, that section is closed. Students requiring time beyond the majority of the student population are allowed to continue immediately following the regularly scheduled session in another setting. When such accommodations are made, school personnel ensure that students are monitored at all times to prevent sharing of information. Students are not permitted to continue a section of the assessment after a significant lapse of time from the original session.

Table 7–5. Keystone Testing Load and Duration by Exam

Exam	Total No. of MC Items per Form per Administration	Total No. of CR Items per Form per Administration	Total Estimated Testing Time per Form (in minutes)	Total Estimated Administration Time per Form (in minutes)
Algebra I	46	8	150	170–180
Biology	64	8	144	164–174
Literature	46	8	146	166–176

Table 7–6. Keystone Testing Load and Duration by Type per Unit (in minutes)

Exam	Administration Tasks	Stimulus Passages	MC Points per Minute [PPM]*	CR [PPM]*	Estimated Overall PPM**
Algebra I	24	–	1.5 [0.670]	10 [0.400]	0.400
Biology	24	–	1.25 [0.800]	8 [0.375]	0.458
Literature	24	–	1 [1.000]	5 [0.600]	0.356

*Based on rates per item type

**Based on total testing time

Prior to beginning the exam, students were asked to verify that they understood the *Code of Conduct for Test Takers* by marking the circle in the exam. Additionally, an Attention statement was added to the beginning of the exams to notify students of the penalties incurred if exam materials are copied.

LAYOUT

The layout of the operational Keystone Exams follows a general sequence regardless of the exam. Each exam is divided into thematically linked sets of content called modules. Within each module, there are core (common) items and field test items. Both core and field test items are represented though MC and CR items.

Stimulus material (like passages), text for MC items, answer options, and any stimulus materials associated with MC items or answer options appear in the test book. Answer bubbles, text for CR items, and associated response spaces appear in the answer book.

Within a non-passage-based module (like Algebra I and Biology), the sequencing of items follows this pattern:

- 1st: Approximately half of the MC items
- 2nd: Half of the CR items
- 3rd: Remaining half of the MC items
- 4th: Remaining CR items

Within a passage-based module (like Literature), the sequencing of items follows this pattern:

- 1st: Stimulus Passage X
- 2nd: MC items associated with Passage X
- 3rd: CR items associated with Passage X
- 4th: Stimulus Passage Y
- 5th: MC items associated with Passage Y
- 6th: CR items associated with Passage Y
- 7th: Stimulus Passage Z
- 8th: MC items associated with Passage Z
- 9th: CR items associated with Passage Z

Regardless of sequencing pattern, the field test items appear in the relative middle of each module, and item sequencing is self-contained within a module.

For more information about the test layout of the operational Keystone Exams, see Appendix G.

SHIPPING, PACKAGING, AND DELIVERY OF MATERIALS

Participating sites receive the *Handbook for Assessment Coordinators*, the *Directions for Administration Manuals*, the administrative materials (e.g., Return Shipping labels, District/School labels, Do Not Score labels, Student Precode labels) and secure materials (e.g., consumable test/answer books) for each tested subject in Algebra I, Biology, and Literature Keystone Exams. All materials arrive at least two weeks prior to the start of the testing window.

DRC ensured that all exam materials were assembled correctly prior to shipping. DRC operations staff used the automated Operations Materials Management System (Ops MMS) to assign secure materials to a school at the time of ship out. This system used barcode technology to provide an automated quality check between items requested for a site and items shipped to a site. A shipment box manifest was produced for and placed in each box shipped. DRC operations staff double-checked all box contents with the box manifest prior to sealing the box for shipping to ensure accurate delivery of materials. DRC operations staff performed lot acceptance sampling on both shipments. Districts and schools were selected at random and examined for correct and complete packaging and labeling. This sampling represented a minimum of 10 percent of all shipping sites.

DRC's materials management system, along with the systems of shippers, allowed DRC to track materials from DRC's warehouse facility to receipt at the district, school, or testing site. All DRC shipping facilities, materials processing facilities, and storage facilities are secure. Access is restricted by security code. Non-DRC personnel are escorted by a DRC employee at all times. Only DRC inventory control personnel have access to stored secure materials. DRC employees are trained in and made aware of the high level of security that is required.

DRC used United Parcel Service (UPS) to deliver the secure materials to the testing sites.

ONLINE TESTING

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

Student information from the Pennsylvania Information Management System (PIMS) is imported into the DRC INSIGHT Portal Test Setup application via file transfer. If a record was not transferred via the PIMS file, LEAs also have the opportunity to upload a student(s) directly into the DRC INSIGHT Portal so the student can be included in a test session.

Once the student data is loaded into Test Setup, users organize students into test sessions. Test sessions can be created by class, grade, or school. Through Test Setup, users can also update student accommodation information, print test tickets, and monitor student testing status.

The student login ticket contains unique login credentials used by the student to access the testing software. For a selected test session, users can download and print a PDF document containing instructions, a roster of student tickets being printed, and the actual test tickets. Student test tickets are considered secure materials and LEAs are required to keep printed tickets in a predetermined, locked, secure storage area.

The web-based test engine, DRC INSIGHT Online Learning System, is downloaded onto computers that students will access during the assessment. Test items and forms can only be accessed using a valid test ticket. During testing, responses are sent to a DRC server each time the student navigates away from an item or clicks the *Next* button to submit an answer. The system is configured to allow students to review answers before submitting their test.

TEST SECURITY MEASURES

Test security is essential to obtaining reliable and valid scores for accountability purposes. Test Security Certifications were required to be signed by each building Principal, School Assessment Coordinator, District Assessment Coordinator, Test Administrator, and Proctor after the assessment is administered. All signed Certifications were returned to the Chief School Administrator who must retain the Certifications for three years. The purpose of the Certifications was to serve as a tool to document that the individuals responsible for administering the assessments both understood and acknowledged the importance of test security and accountability. Additional details can be found in the *Handbook for Assessment Coordinators*. A screen shot of the Test Administrator Certificate is provided in Figure 7–1.

Figure 7–1. Test Administrator and Proctor Test Security Certification

Keystone Exam Test Security Certification Form

(Test Administrator and Proctor)

District: _____

School: _____

AUN: _____

Maintaining the security and integrity of all assessment materials, preventing any dishonest or fraudulent behavior in the administration and handling of the assessment, and promoting a fair and equitable testing environment are essential in order to obtain reliable and valid student scores. In that regard, I certify the following:

Prior to the administration of the assessment, I completed the Pennsylvania State Test Administration Training, and I understand that the assessment materials are secure, confidential, and proprietary documents owned by the Pennsylvania Department of Education.

I have not reviewed, discussed, disseminated, described, or otherwise revealed the contents of the assessment to anyone. I have not removed any assessment materials from the school building unless I was specifically authorized to administer the assessment to a student on homebound instruction. I have not kept, copied, reproduced, released, or used any assessment, assessment question, specific assessment content, or examinee response to any item or any section of the secure assessment in any manner that is inconsistent with the instructions provided by or through the Pennsylvania Department of Education. I have not provided any examinee with an answer to an assessment question or in any way influenced an examinee's response to any assessment question. I have not in any manner altered or caused the alteration of any examinee response, assessment booklet, or papers used by examinees.

I understand that any breach in assessment security could result in the invalidation of assessment results, professional discipline, and/or criminal prosecution.

I understand that false statements herein are made subject to the penalties of 18 Pa.C.S. § 4904.

Administrator/Proctor Name

Administrator/Proctor Signature

Date of Signature

SAMPLE MANUALS

Copies of the *Handbook for Assessment Coordinators* and the *Directions for Administration Manuals* can be found on the PDE website at www.education.pa.gov.

TESTING WINDOW ASSESSMENT ACCOMMODATIONS

PDE develops an *Accommodations Guidelines* handbook for use with the Keystone Exams. This manual can be found on the PDE website at www.education.pa.gov. Additional information regarding assessment accommodations can be found in Chapter Four and Six of this report.

CHAPTER EIGHT: PROCESSING AND SCORING

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

RECEIPT OF MATERIALS

Receipt of Pennsylvania Keystone Exams' test materials began five days after the start of the test window. DRC's Operations Materials Management System (Ops MMS) was utilized to receive assessment materials securely, accurately, and efficiently. This system features innovative automation and advanced barcode scanners. Captured data was organized into reports, which provided timely information with respect to suspected missing material.

The first step in Ops MMS was Box Receipt. When a shipment arrived at DRC, the boxes were removed from the carrier's truck and passed under a barcode reader, which read the barcode printed on the return label and identified the district and school. The number of boxes was immediately compared to what was picked up at the district. The data collected in this process was stored in the Ops MMS database. After the barcode data was captured, the boxes were placed on a pallet and assigned a corresponding pallet number.

Once the Box Receipt process was completed, the Materials Separation phase began. Warehouse personnel opened the boxes and sorted materials by grade, subject, and status (used and/or unused booklets) into scanning boxes. Every booklets' security barcode and precode barcode were hand scanned to link each document to the original box. As the booklets were sorted, Ops MMS guided the floor operator to the box in which to place the document. Ops MMS kept count and record of the materials placed in each box. This count remained correlated to the box as an essential quality-control step throughout the secure booklet processing and provided a target number for all steps of the check-in process. Once a box was closed, an MMS Processing Label was placed on that box.

Once labeled, the sorted and counted boxes proceeded to Quality Assurance, where a secure booklet check-in operator used a hand scanner to scan the MMS Processing Label. This procedure identified the material type and quantity parameters for what Ops MMS should expect within a box. The box contents were then loaded into the stream feeder.

The documents were fed past oscillating scanners that captured both the security code and precode from the booklets. A human operator monitored an Ops MMS screen that displayed scan errors, an ordered accounting of what was successfully scanned, and the document count for each box. The system ensured that all material within the box matched the information obtained from the original hand-scanning process.

When all materials were scanned and the correct document count was confirmed, the box was sealed and placed on a pallet. If the correct document count was not confirmed, or if the operator encountered difficulties with material scanning, the box and its contents were delivered to an exception-handling station for resolution.

This check-in process occurred immediately upon receipt of materials; therefore, DRC provided feedback to districts and schools regarding any missing materials based on actual receipt versus expected receipt. Sites that had 100 percent of their materials missing after the date they were due to DRC were contacted, and any issues were resolved.

Throughout the process of secure booklet check-in, DRC project management ran a daily Missing Materials Report. Every site that was missing any number of booklets was contacted by DRC. Results of these correspondences were recorded for inclusion in the final Missing Materials Report if the missing booklets were not returned by the testing site. DRC produced the Missing Materials Report for PDE upon completion of secure booklet check-in. The report listed all schools in each participating district, along with security barcodes for any booklets not returned to DRC.

After scannable materials (used answer booklets) were processed through booklet check-in, the materials became available to the DRC Document Processing log-in staff for document log-in. The booklets were logged in using the following process:

- A DRC scannable barcode batch header was scanned, and a batch number was assigned to each box of booklets.
- The DRC box label barcode was scanned into the system to link the box and booklets to the newly created batch and to create a Batch Control Sheet.
- The DRC box label barcode number and the number of booklets in the box were printed on the Batch Control Sheet for document-tracking purposes. All booklets linked to the box barcode were assigned to the batch number and tracked through all processing steps. As booklets were processed, DRC staff dated and initialed the Batch Control Sheet to indicate that proper processing and controls were observed.

Before the booklets were scanned, all batches went through a quality inspection to ensure batch integrity and correct document placement.

After a quality check-in at the DRC Document Processing log-in area, the spines were cut off the scannable documents, and the pages were sent to DRC's Imaging and Scoring System.

SCANNING OF MATERIALS

Customized scanning programs for all scannable documents were prepared to read the books and to format the scanned information electronically. Before materials arrived, all image-scanning programs went through a quality review process that included scanning of mock data from production books to ensure proper data collection.

DRC's image scanners were calibrated using a standard deck of scannable pages with 16 known levels of gray. On a predefined page location, the average pixel darkness was compared to the standard calibration to determine the level of gray. Marks with an average darkness level of 4 or above on a scale of 16 (0 through F) were determined to be valid responses, per industry standards. If multiple marks were read for a single item and the difference between the grayscale reads was greater than four levels, the lighter mark was discarded. If the multiple marks had fewer than four levels of grayscale difference, the response was flagged and forwarded to an editor for resolution.

DRC's image scanners read selected-response, demographic, and identification information. The image scanners also used barcode readers to read preprinted barcodes from a label on the book.

The scannable documents were automatically fed into the image scanners where predefined processing criteria determined which fields were to be captured electronically. Open-ended (OE) response images were separated out for image-based scoring.

During scanning, a unique serial number was printed on each sheet of paper. This serial number was used to ensure document integrity and to maintain sequencing within a batch of books.

A monitor randomly displayed images, and the human operator adjusted or cleaned the scanner when the scanned image did not meet DRC's strict quality standards for image clarity.

All images passed through a process and a software clean-up program that despeckled, deskewed, and desmeared the images. A random sample of images was reviewed for image quality approval. If any document failed to meet image quality standards, the document was returned for rescanning.

Page-scan verification was performed to ensure that all predefined portions of the booklets were represented in their entirety in the image files. If a page was missing, the entire book was flagged for resolution.

After each batch was scanned, books were processed through a computer-based editing program to detect potential errors as a result of smudges, multiple marks, and omissions in predetermined fields. Marks that did not meet the predefined editing standards were routed to editors for resolution.

Experienced DRC Document Processing editing staff reviewed all potential errors detected during scanning and made necessary corrections to the data file. The imaging system displayed each suspected error. The editing staff then inspected the image and made any necessary corrections using the unique serial number printed on the document during scanning.

Upon completion of editing, quality control reports were run to ensure that all detected potential errors were reviewed again and a final disposition was determined.

Before batches of books were extracted for scoring, a final edit was performed to ensure that all requirements for final processing were met. If a batch contained errors, it was flagged for further review before being extracted for scoring and reporting.

During this processing step, the actual number of documents scanned was compared to the number of books assigned to the box during book receipt. Count discrepancies between book receipt and books scanned were resolved at this time.

Once all requirements for final processing were met, the batch was released for scoring and student level processing.

Table 8–1 shows the number of answer books received through book check-in, the number of books that contained student responses that were scanned and scored, the number of test books received, and the total number of books received for the Algebra I, Biology, and Literature Keystone Exams.

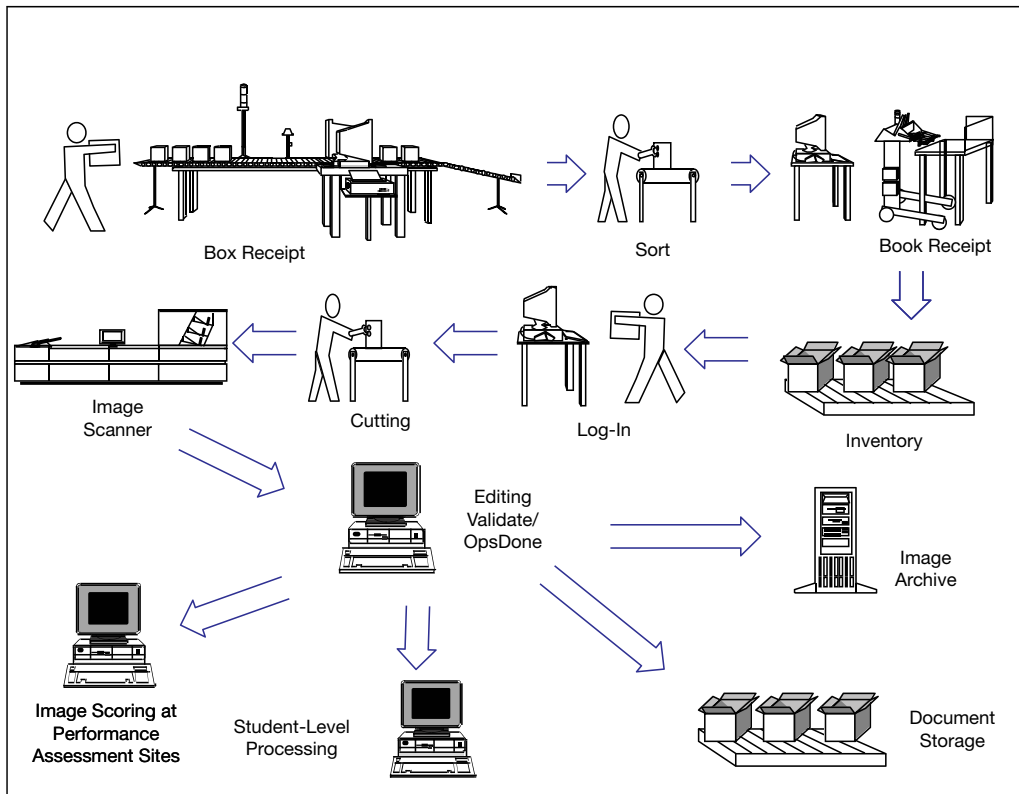
Table 8–1. Counts of 2021 Keystone Exams Materials Received: Algebra I, Biology, and Literature

Exam	Answer Books Received	Used Answer Books Received	Test Books Received	Total Books Received	Total Books Shipped
Algebra I (Winter)	88,392	5,147	88,393	176,785	176,788
Biology (Winter)	76,950	3,838	76,950	153,900	153,914
Literature (Winter)	78,928	4,436	78,928	157,856	157,856
Algebra I (Spring)	189,837	94,738	189,799	379,636	380,088
Biology (Spring)	157,385	84,114	157,410	314,795	315,438
Literature (Spring)	153,842	84,072	153,799	307,641	308,274
Algebra I (Summer)					
Biology (Summer)					
Literature (Summer)					

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Figure 8–1 illustrates the production workflow for DRC’s Ops MMS and Image Scanning and Scoring System from receipt of materials through all processing of materials and the presentation of scanned images for scoring.

Figure 8–1. Workflow System



MATERIALS STORAGE

Upon completion of processing, student response documents were boxed for security purposes and final storage as follows:

- Project-specific box labels were created containing unique customer and project information, material type, batch number, pallet/box number, and the number of boxes for a given batch.
- Boxes were stacked on pallets that were labeled with the project information and a list of the pallet's contents before delivery to the Materials Distribution Center for final secure storage.

Materials will be destroyed one year after the contract year ends with PDE written approval.

ONLINE TESTING

The DRC INSIGHT test engine runs on a custom web browser that is designed to ensure a fully secure environment during testing. The secure browser “locks down” the student’s testing device, preventing the student from accessing the desktop, the Internet, and other external programs. For non-secure testing such as practice and training sessions, students can use the Online Tools Training (OTT) environment, which runs on a standard web browser.

The custom browser software is downloaded from the DRC INSIGHTS Portal and installed onto student testing devices. The secure browser can be installed on computers individually, or it can be downloaded to a central location, copied, and distributed to multiple computers simultaneously using common network distribution tools. Everything needed for testing is found within the secure browser, eliminating the need for districts to coordinate updates to third-party software.

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

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

Responses are saved automatically every 45 seconds during testing, or when the student navigates away from an item or answers a selected-response item (whichever comes first). If a particular question takes the student longer than 45 seconds to answer, then the partial, incomplete responses are submitted at 45-second intervals until the student completes the item. This auto-save helps safeguard against students losing their work on longer items, such as constructed-response items. When the student returns to the test after a break or interruption, the student is returned to the point that they left off without having to navigate through all previously answered questions.

Table 8–2. Counts of 2021 Keystone Exams Online Assessments

Grade/Subject	Total Online Assessments Completed
Algebra I (Winter)	5,466
Biology (Winter)	5,490
Literature (Winter)	4,924
Algebra I (Spring)	32,459
Biology (Spring)	32,115
Literature (Spring)	29,691
Algebra I (Summer)	
Biology (Summer)	
Literature (Summer)	

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Figure 8–2 illustrates the secure transfer of online test responses between the student and DRC.

Figure 8–2. Architecture of the Student Testing Experience



SCORING MULTIPLE-CHOICE ITEMS

For both online and paper-and-pencil modes, the scoring process included the scoring of multiple-choice (MC) items against the answer key and the aggregation of raw scores from the OE responses. A student's raw score is the actual number of points achieved for tested elements of an assessment. From the raw scores, the scale scores were calculated.

The student file was scored against the final and approved MC answer key. Items were scored as right, wrong, omitted, or double-gridded (more than one answer was bubbled for an item). Sections of the exam were evaluated as a whole, and an attempt status was determined for each student for each subject. The score program defined all data elements for reporting at the student level.

HANDSCORING OPEN-ENDED ITEMS

During 2021, Coronavirus (COVID-19) mitigation efforts were in place for some of the Keystone administrations:

- Winter Keystone – Handscoring dates shifted to allow for an extended testing window.
- Spring Keystone – Handscoring dates shifted significantly to allow for an extended testing window.
- Summer Keystone – The Summer administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021.

RANGEFINDING

Due to Coronavirus (COVID-19), we did not have rangefinding this year. However, the training materials for all operational items are based on the rangefinding that was initially done for that item when it was field tested.

After student answer documents were received and processed, DRC's Performance Assessment Services (PAS) staff assembled groups of responses that exemplified the different score points for each subject. The score point ranges were represented by the following scoring guidelines:

- 0–3 item-specific scoring guidelines for Literature
- 0–4 item-specific scoring guidelines for Algebra I (some items were divided into separate parts that were scored on a 0–1, 0–2, or 0–3 point scale, but the sum of the parts always resulted in an overall score of 0–4 for each item)
- 0–3 item-specific scoring guidelines for Biology

Responses are pulled from the embedded field test portion of the Keystone Exams for each subject. Once examples of all score points are selected for each item, sets are assembled for rangefinding and copies are made for each rangefinding participant. Rangefinding committees consist of Pennsylvania educators, PDE staff members, DRC Test Development staff, and DRC Performance Assessment Services staff.

Each rangefinding meeting begins in a joint session with a review of the history of the assessment as well as a discussion of the purpose of the rangefinding meeting and the role rangefinding plays within the item development process. The session then breaks into subject/grade-specific committees. Sets of student responses are presented to the committees, one item at a time. Each committee initially reviews and scores student responses as a group to ensure consistency in the interpretation of the scoring guidelines. Committee members then go on to score responses independently. For each student response, committee members' scores are discussed until a consensus is reached. Only those responses for which there is strong agreement among committee members are chosen for inclusion in training materials for DRC raters.

Discussions of student responses include the mandatory use of scoring guideline language. This ensures that committee members remain focused on the specific requirements of each score level. DRC PAS staff takes notes addressing how and why the committees arrived at score point decisions, and this information is used by the scoring directors in rater training.

DRC and PDE discuss scoring guideline edits suggested by the rangefinding committees. Changes approved by PDE are incorporated into the scoring guidelines by DRC Test Development staff. The edited scoring guidelines are used in the preparation of materials and the training of raters.

RATER RECRUITMENT/QUALIFICATIONS

DRC retains a number of raters from year to year; the overall return rate in 2021 was 60%. This pool of experienced raters was drawn from to staff the scoring of the 2020 Keystone Exams. To complete the rater staffing for this project, recruiting events were held and applications for rater positions were screened by DRC's recruiting staff. Candidates were personally interviewed by DRC staff. In addition, each candidate was required to provide an on-demand writing sample, an on-demand math sample, references, and proof of a four-year college degree. In this screening process, preference was given to candidates with previous experience scoring large-scale assessments and degrees emphasizing expertise in the subjects being scored. In some locations, staffing partners were used to augment hiring using the same practices as those employed by DRC. The rater pool consisted of educators and other professionals with content-specific backgrounds. These individuals were valued for their content-specific knowledge, but they were required to set aside their own biases about student performance and accept the scoring standards of the Keystone Exams.

LEADERSHIP RECRUITMENT/QUALIFICATIONS

Scoring directors and team leaders were selected from a pool of employees who displayed expertise as raters and leaders on previous DRC projects. These individuals had strong backgrounds in mathematics, English language arts, or science, and demonstrated organizational, leadership, communication, and management skills. All scoring directors had previous leadership experience working on large scale assessments. All scoring directors, team leaders, and raters were required to sign confidentiality agreements before handling secure materials.

Each group of raters was assigned a scoring director. All handscoring activities were led by a scoring director for the duration of the project. Scoring directors assisted in rangefinding, worked with supervisors to create training materials, conducted team leader training, and were responsible for training the raters. The scoring director made sure that handscoring reports were available and interpreted those reports for the raters. The scoring director also supervised the team leaders. Scoring directors were monitored by the project managers throughout the project.

Team leaders assisted the scoring director with rater training by answering individual questions that raters may not have felt comfortable asking in a large group. Once raters were qualified, team leaders were responsible for monitoring and maintaining the accuracy and workload of each team member. Ongoing monitoring identified those individuals having difficulty maintaining accuracy. These raters received one-on-one retraining from the team leader or scoring director. Any rater who could not be successfully retrained had his/her scores purged and was released from the project.

TRAINING

As part of preparation for the 2021 Keystone Exams, DRC's PAS staff assembled the PDE-approved scoring guidelines and scored student responses approved by rangefinding committees into sets used for training raters. These item-specific scoring guidelines served as the raters' constant reference. Responses that were relevant in terms of the scoring concepts they illustrated were annotated and included in an anchor set. The full range of each score point was clearly represented and annotated in the anchor set, which was used for reference by raters throughout the project.

Training sets and qualifying sets contained student responses consensus-scored by rangefinding committee members. Raters were instructed on how to apply the scoring guidelines and were required to demonstrate a clear comprehension of each anchor set by performing well on the associated training materials. Responses were selected for training to show raters the range of each score point (e.g., high, mid, and low 2s). Examples of 0s were also included for all items. This process helped raters recognize the various ways that a student could respond in order to earn each score point outlined and defined in the item-specific scoring guidelines.

The scoring director conducted a team leader training session before training the raters. This session followed the same procedures as rater training, but was more rigorous and in-depth due to the extra responsibilities required of team leaders. During team leader training, all pertinent training materials were reviewed and discussed. Team leaders were given access to fully annotated training materials with committee justifications from the rangefinding

meetings. To facilitate scoring consistency, it was imperative that all team leaders imparted the same rationale for each response. Once the team leaders were qualified, leadership responsibilities were reviewed and team assignments were given. A ratio of one team leader per 7–10 raters ensured sufficient monitoring rates for team members.

Rater training began with the scoring director providing an intensive review of the scoring guidelines and anchor papers. Next, raters practiced by independently scoring the responses in the training sets. After each training set, the scoring director led a thorough discussion of the responses.

Once the scoring guidelines, anchor sets, and training sets were thoroughly discussed, each rater was required to demonstrate understanding of the scoring criteria by qualifying (i.e., scoring with acceptable agreement to the true scores) on at least one of the qualifying sets. Raters who failed to achieve at least 70 percent exact agreement on the first qualifying set were given additional training, either individually or in a small group setting. Raters who did not perform at the required level of agreement by the end of the qualifying process were not allowed to score any student responses. These individuals were removed from the pool of potential raters in DRC’s imaging system and released from the project.

DRC’s remote scoring is designed to very closely emulate the work that is done in our physical scoring locations. The platform, content, and expectations for quality remain the same, and interactive technology and content training and discussions are conducted live (virtually). The differences come with the method through which training is delivered (online), and in the modes of communication that are used (web screen sharing, webcast, video chat, and chat). Our scoring leaders are equipped with a variety of tools to ensure every scorer is successful in understanding and applying scoring criteria to student responses.

The 2021 assessment included the opportunity for students to respond in Spanish to Algebra I and Biology items. Rater training for the Spanish language response scoring was conducted by Tri-Lin Integrated Services in San Antonio, Texas, and was overseen by a DRC project manager, who is a Spanish language speaker with a strong handscoring background. All Spanish raters were bilingual and hired specifically to score the Spanish portion of the assessment and were required to meet the same standards set for raters of the English language version of the assessment.

Table 8–3. Qualification Rates for 2021 Keystone Open-Ended Response Items – Winter

Subject	% Qualifying	% That Did Not Qualify
Algebra I	100	0
Biology	95	5
Literature	98	2

Table 8–4. Qualification Rates for 2021 Keystone Open-Ended Response Items – Spring

Subject	% Qualifying	% That Did Not Qualify
Algebra I	100	0
Biology	99	1
Literature	100	0

Table 8–5. Qualification Rates for 2021 Keystone Open-Ended Response Items – Summer

Subject	% Qualifying	% That Did Not Qualify
Algebra I		
Biology		
Literature		

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

HANDSCORING PROCESS

Student responses were scored independently. All responses were scored once, and ten percent of the responses were scored a second time. The data collected from the ten-percent double-read portion was used to calculate the exact and adjacent agreement rates in the Scoring Summary Reports. The responses that were used for the ten percent read behind were randomly chosen by the imaging system at the item level. Additional read behinds by the team leaders and scoring directors were done to further ensure reliability.

Raters scored the imaged student responses at scoring locations in Cincinnati, Ohio; Columbus, Ohio; Plymouth, Minnesota; Woodbury, Minnesota; Philadelphia, Pennsylvania; and San Antonio, Texas. Raters scored either on PC monitors at the above locations or remotely, due to some of our scoring centers being closed because of Covid-19.

In all locations for on-site scoring, raters were seated at tables with individual imaging stations. In the case of remote scoring, raters worked in a secure location in their homes. Image distribution was controlled, ensuring that student images were sent only to designated groups of raters qualified to score those items. Imaged student responses were electronically separated for routing to individual raters by item. Raters were only provided with student responses for items that they were qualified to score. Scores were keyed into DRC’s imaging system.

To handle possible alerts (i.e., student responses indicating potential issues related to students’ safety and well-being that sometimes require attention at the state or local level), DRC’s imaging system allows raters to forward responses needing attention to the scoring director. These alerts are reviewed by project management, who then notifies the students’ schools and PDE of the occurrences. PDE does not receive any identifying information about the students. At no point in the alerts process do raters, or other DRC handscoring staff, acquire any knowledge concerning a student’s personal identity.

HANDSCORING VALIDITY PROCESS

One of the training tools PAS utilized to ensure rater accuracy was the validity process. The goal of the validity process is to ensure that scoring standards are maintained. Specifically, the objective is to make sure that raters score student responses in a manner consistent with statewide standards both within a single administration of the Keystones and across consecutive administrations. In scoring the 2021 Keystone Exams, scoring consistency was maintained, in part, through the validity process.

The validity process began with the selection of scored responses. Forty validity papers were selected for each core open-ended (OE) item. These 40 papers were drawn from a pool of exemplars (responses that are representative of a particular score point and have been verified by the scoring director). The scores on validity responses are considered true scores.

The validity papers were then implemented to test rater accuracy. The responses were selected within the imaging system and dispersed intermittently to the raters. By the end of the project, raters had scored all 40 validity papers for any items they were qualified to score. Raters were unaware of when they were being dealt pre-scored validity responses and assumed that they were scoring live student responses. This helped bolster the internal validity of the process. All raters who received validity papers had already successfully completed the training/qualifying process.

The scores that the raters assigned to the validity papers were compared to the true scores in order to determine the validity of the raters' scores. For each item, the percentage of exact agreement as well as the percentage of high and low scores was computed. This data was accessed through the Validity Item Detail Report. The same sort of data was also computed for each specific rater. This data was accessed through the Validity Reader Detail Report. Both of these may be run as daily or cumulative reports.

The Validity Reader Detail Report was used to identify particular raters for retraining. If a rater on a certain day generated a lower rate of agreement on a group of validity papers, it was immediately apparent in the Validity Reader Detail Report. A lower rate of agreement was defined as anything below 70 percent exact agreement with the true scores. Any time a rater's validity agreement rate fell below 70 percent, the scoring director was cued to examine that rater's scoring. First, the scoring director attempted to ascertain what kind of validity papers the rater was scoring incorrectly. This was done to determine whether there was any sort of a trend (e.g., trending low on the 1–2 line). Once the source of the low agreement rate was determined, the rater was retrained. If it was determined that the rater had been scoring live responses inaccurately, then his/her scores were purged for that day, and the responses were re-circulated and scored by other raters.

The cumulative Validity Item Detail Report was utilized to identify potential room-wide trends in need of correction. For instance, if a particular validity response with a true score of 3 was given a score of 2 by a significant number of raters within the room, that trend would be revealed in the Validity Item Detail Report. To correct a trend of this sort, the scoring director would look for student responses similar to the validity response being scored incorrectly. Once located, these responses would be used in room-wide re-training, usually in the form of an annotated handout or a short set of responses without printed scores given to raters as a recalibration test.

Validity was employed on all core Algebra I, Biology, and Literature OE items. Each 40-paper validity set was formulated to mirror the score point distribution that the item generated during its previous administration. Each validity set included at least five examples of each score point. Examples of different types of responses were included to ensure that raters were tested on the full spectrum of response types.

The exact rater agreement rate generated during the validity process was often higher than the inter-rater agreement rate for the same item. The reason for this discrepancy has to do with how validity sets are formulated. The 40 validity responses for each item are intended to cover the full breadth of each score point. For example, each validity set contains examples of high, mid, and low 2s. This scope ensures that the validity process is truly valid in terms of addressing the complete spectrum of response types. However, certain types of responses are generally not included in validity sets. These include line responses (i.e., examples of score points that are so close to the adjacent score point that raters are instructed to consult with a supervisor before assigning a score) and responses that, because of poor word choice/writing, are difficult to understand. The reason for these exclusions is that confusing/line/illegible papers often do not impart a teachable lesson. Since these types of responses are usually unique, any potential lesson the response might teach would apply only to that particular response. Conversely, the responses in validity sets are chosen because they represent common response types and teach lessons that can be applied to other similar papers. Due to this distinction, validity sets often generate a slightly higher agreement rate than is typically generated during operational scoring.

QUALITY CONTROL

Rater accuracy was monitored throughout the scoring session by means of daily and on-demand reports. These reports ensured that an acceptable level of scoring accuracy was maintained throughout the project. Inter-rater reliability was tracked and monitored with multiple quality control reports that were reviewed by quality assurance analysts. These reports and other quality control documents were generated at the scoring centers, where they were reviewed by the scoring directors, team leaders, and project managers. The following reports and documents were used during the scoring of the open-ended items:

The Scoring Summary Report (includes two related reports).

- The Reader Monitor Report monitored how often raters were in exact agreement with one another and ensured that an acceptable agreement rate was maintained throughout the project. This report provided daily and cumulative exact and adjacent inter-rater agreement on the ten percent that was double read.

- The Score Point Distribution Report monitored the percentage of responses given each of the score points. For example, the Algebra I daily and cumulative reports showed what percentage of 0s, 1s, 2s, 3s, and 4s a rater—or room of raters—had given to all the responses scored at the time the report was produced. It also indicated the number of responses read by each rater so that production rates could be monitored.
- The Item Status Report monitored the progress of handscoring. This report tracked each response and indicated the status (e.g., not read, complete, awaiting supervisor review, etc.). This report ensured that all responses were scored by the end of the project.

The Reader Score Report identified all responses scored by an individual rater. This report was useful if any responses needed rescoring due to possible rater drift.

The Validity Reports (addressed in detail on previous page) tracked how raters performed by comparing pre-scored responses to raters' scores for the same responses. If a rater's scoring fell below the 70 percent determined agreement rate, remediation occurred. Raters who did not retrain to the required level of agreement were released from the project. The Read-Behind Log was used by the team leader/scoring director to monitor individual rater reliability. Team leaders read randomly-selected, scored responses from each team member on a daily basis. If the team leader disagreed with a rater's score, remediation occurred. This was a particularly effective form of feedback because it was performed in real time with live student responses scored by each rater.

Recalibration Sets were used throughout the scoring sessions to ensure accuracy by comparing each rater's scores with the true scores on a pre-selected set of responses. Recalibration sets helped to refocus raters on Pennsylvania scoring standards. These checks made sure there was no change in the scoring pattern as the project progressed. Raters failing to achieve 70 percent agreement with the recalibration true scores were given additional training to achieve the highest degree of accuracy possible. Raters who were unable to recalibrate were released from the project. The process used for creating and administering recalibration sets was similar to the process employed for creating and administering training sets.

Table 8–6. Inter-Rater Agreement and Percentage Awarded for Each Score Point for CR Items Winter 2021

Exam	Module	Item ID	Item Part	Score Point Range	Inter-Rater Agreement % Exact	Inter-Rater Agreement % Adjacent	% Validity Agreement	% 0s	% 1s	% 2s	% 3s	% 4s	% B/NS
Alg. 1	1	818616	A	0–1	99	1	99	44	36	NA	NA	NA	20
Alg. 1	1	818616	B	0–2	99	1	100	52	12	17	NA	NA	20
Alg. 1	1	818616	C	0–1	100	0	100	78	3	NA	NA	NA	20
Alg. 1	1	877382		0–4	97	3	97	40	28	7	3	1	20
Alg. 1	1	673351	A	0–1	100	0	100	30	50	NA	NA	NA	21
Alg. 1	1	673351	B	0–1	99	1	99	62	17	NA	NA	NA	21
Alg. 1	1	673351	C	0–1	99	1	100	46	33	NA	NA	NA	21
Alg. 1	1	673351	D	0–1	99	1	100	55	25	NA	NA	NA	21
Alg. 1	2	818665		0–4	95	5	96	44	23	7	5	1	20
Alg. 1	2	818209	A	0–1	100	0	100	37	42	NA	NA	NA	20
Alg. 1	2	818209	B	0–1	100	0	100	60	19	NA	NA	NA	20
Alg. 1	2	818209	C	0–1	100	0	100	74	5	NA	NA	NA	20
Alg. 1	2	818209	D	0–1	100	0	98	74	5	NA	NA	NA	20
Alg. 1	2	734697		0–4	89	11	98	10	28	23	16	1	22
Bio.	1	741576		0–3	91	9	92	43	18	13	10	NA	15
Bio.	1	877366		0–3	93	7	97	10	37	37	2	NA	14
Bio.	1	978211		0–3	92	7	97	16	30	28	12	NA	14
Bio.	2	702742		0–3	82	17	77	30	33	18	4	NA	14
Bio.	2	966775		0–3	89	11	89	26	20	18	21	NA	14
Bio.	2	819535		0–3	79	20	78	16	22	26	22	NA	14
Lit.	1	742085		0–3	82	18	83	4	35	37	10	NA	14
Lit.	1	994603		0–3	80	20	75	5	32	33	15	NA	14
Lit.	1	994606		0–3	85	15	89	6	32	34	11	NA	17
Lit.	2	986358		0–3	90	10	88	3	30	41	11	NA	14
Lit.	2	704766		0–3	82	18	77	10	30	33	11	NA	16
Lit.	2	704767		0–3	92	8	87	7	25	39	11	NA	17

Notes. B = blank; N = non-scorable. NA= non-applicable. Algebra I responses received a possible total of 0–4 points. For some Algebra I items, readers applied a single score of 0, 1, 2, 3, or 4; however, many Algebra I items were divided into separate parts that were scored on 0–1, 0–2, or 0–3-point scales, the sum of which always resulted in an overall score of 0–4 points. For example, an Algebra I item might have a part A, a part B, a part C, and a part D, each of which was scored on a 0–1-point scale, resulting in a summed 0–4-point total score scale. Additionally, some Algebra I items with multiple parts could receive up to one point for “minimal understanding” (MU) even if the student did not receive a point, or points, for any of the item’s individual parts.

Table 8–7. Inter-Rater Agreement and Percentage Awarded for Each Score Point for CR Items Spring 2021

Exam	Module	Item ID	Item Part	Score Point Range	Inter-Rater Agreement % Exact	Inter-Rater Agreement % Adjacent	% Validity Agreement	% 0s	% 1s	% 2s	% 3s	% 4s	% B/NS
Alg. 1	1	821569		0–4	90	10	94	25	37	16	7	2	14
Alg. 1	1	905404		0–4	95	5	94	42	35	5	2	0	16
Alg. 1	1	821550	A	0–1	100	0	100	15	70	NA	NA	NA	15
Alg. 1	1	821550	B	0–1	98	2	99	34	51	NA	NA	NA	15
Alg. 1	1	821550	C	0–1	99	1	98	78	7	NA	NA	NA	15
Alg. 1	1	821550	D	0–1	100	0	100	69	16	NA	NA	NA	15
Alg. 1	2	817339	A	0–1	100	0	100	50	37	NA	NA	NA	14
Alg. 1	2	817339	B	0–1	100	0	100	39	47	NA	NA	NA	14
Alg. 1	2	817339	C	0–1	100	0	100	62	24	NA	NA	NA	14
Alg. 1	2	817339	D	0–1	100	0	99	86	1	NA	NA	NA	14
Alg. 1	2	820072	A	0–1	99	1	99	55	30	NA	NA	NA	14
Alg. 1	2	820072	B	0–1	100	0	100	77	9	NA	NA	NA	14
Alg. 1	2	820072	C	0–1	99	1	98	74	12	NA	NA	NA	14
Alg. 1	2	820072	D	0–1	99	1	99	74	12	NA	NA	NA	14
Alg. 1	2	969452		0–4	90	10	95	33	24	14	9	4	16
Bio.	1	978647		0–3	85	14	94	23	24	22	17	NA	13
Bio.	1	966416		0–3	90	10	96	26	22	21	17	NA	14
Bio.	1	741444		0–3	96	4	99	29	27	19	12	NA	14
Bio.	2	741581		0–3	92	8	96	25	22	20	13	NA	19
Bio.	2	983824		0–3	92	8	97	26	28	20	12	NA	13
Bio.	2	978209		0–3	85	15	92	19	30	23	11	NA	16
Lit.	1	994432		0–3	82	18	89	3	22	44	16	NA	14
Lit.	1	824935		0–3	83	17	86	9	29	38	10	NA	14
Lit.	1	824936		0–3	83	17	85	7	27	38	11	NA	17
Lit.	2	984231		0–3	87	13	91	5	36	38	5	NA	15
Lit.	2	826264		0–3	87	13	86	9	25	43	8	NA	15
Lit.	2	826283		0–3	82	18	82	8	24	38	13	NA	18

Notes. B = blank; NS = non-scorable. NA= non-applicable. Algebra I responses received a possible total of 0–4 points. For some Algebra I items, readers applied a single score of 0, 1, 2, 3, or 4; however, many Algebra I items were divided into separate parts that were scored on 0–1, 0–2, or 0–3-point scales, the sum of which always resulted in an overall score of 0–4 points. For example, an Algebra I item might have a part A, a part B, a part C, and a part D, each of which was scored on a 0–1-point scale, resulting in a summed 0–4-point total score scale. Additionally, some Algebra I items with multiple parts could receive up to one point for “minimal understanding” (MU) even if the student did not receive a point, or points, for any of the item’s individual parts.

Table 8–8. Inter-Rater Agreement and Percentage Awarded for Each Score Point for CR Items Summer 2021

Exam	Module	Item ID	Item Part	Score Point Range	Inter-Rater Agreement % Exact	Inter-Rater Agreement % Adjacent	% Validity Agreement	% 0s	% 1s	% 2s	% 3s	% 4s	% B/ NS
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Alg. 1													
Bio.													
Bio.													
Bio.													
Bio.													
Bio.													
Bio.													
Bio.													
Lit.													
Lit.													
Lit.													
Lit.													
Lit.													
Lit.													

Notes. B = blank; NS = non-scorable. NA= non-applicable. Algebra I responses received a possible total of 0–4 points. For some Algebra I items, readers applied a single score of 0, 1, 2, 3, or 4; however, many Algebra I items were divided into separate parts that were scored on 0–1, 0–2, or 0–3-point scales, the sum of which always resulted in an overall score of 0–4 points. For example, an Algebra I item might have a part A, a part B, a part C, and a part D, each of which was scored on a 0–1-point scale, resulting in a summed 0–4-point total score scale. Additionally, some Algebra I items with multiple parts could receive up to one point for “minimal understanding” (MU) even if the student did not receive a point, or points, for any of the item’s individual parts.

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

CHAPTER NINE: DESCRIPTION OF DATA SOURCES

This section describes the filtering process and data sources used for the various analysis procedures discussed in the remaining sections of this report. Psychometric analyses were conducted at several points for the winter, spring, and summer. Pennsylvania Keystone Exams in Algebra I, Biology, and Literature: 1) key verification analyses for quality-control purposes; 2) pre-equating verification; 3) item analysis and calibration of field-test items embedded in the spring forms; and 4) analyses for this technical report.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

STUDENT FILTERING CRITERIA

Students' records included used for psychometric analyses needed to meet at least the following criteria:

- Module 1 Attempted Status = 1 (1 represent if the student attempted a minimum of five items in Module 1)
- Module 2 Attempted Status = 1 (1 represents if the student attempted a minimum of five items in Module 2)
- Module 1 Invalidated = N (N represents if the student's score was not invalidated)
- Module 2 Invalidated = N (N represents if the student's score was not invalidated)
- Student Duplication Status = N (N represents if no record duplication)
- Module 1 Form Number = Module 2 Form Number
- Module_1_Form_Name \neq 01V (exclude students administered the VSL form)
- Module_2_Form_Name \neq 01V (exclude students administered the VSL form)

For each specific analysis conducted at different times, additional criteria might be needed to filter students. For example, the following criteria were used in addition to the ones listed above for the pre-equating verification, since the analyses were conducted during the scoring window:

- Module 1 Complete Status = 01
- Module 2 Complete Status = 01

The value 01 represents the response string which includes scores on the multiple-choice (MC) and constructed-response (CR) operational items. When the analyses were conducted by using the final data files, these criteria were no longer necessary since all operational CR items had been scored.

Item analysis and calibration of embedded field-test items were conducted using the first-time testers only (i.e., retester = N). The classical item statistics for the field-test items analyzed by using the first-time testers were more comparable to the results of the spring 2011 Keystone Exams, which were given to the first-time test takers.

Because a large number of students took the Keystone Exams, only a representative sample of students' responses on field-test CR items was scored within each content area¹. For the item analysis of field-test CR items, the following additional criteria were used to select only those who were sampled for hand-scoring:

- Module 1 Complete Status = 02
- Module 2 Complete Status = 02

¹ In Spring 2021, field-test CR items were not scored. Therefore, no item-level information is available for field-test CR items in this report.

KEY VERIFICATION DATA

The key verification data are mentioned only for completeness, as no formal results are provided in this technical document. Key verification is often conducted early in the scoring process to ensure the keys for the MC items are applied correctly. The data files used for the key verification analysis are usually (but not always) based on the student data from early-return schools. The sample representativeness is not required for this internal quality check. Available student data typically suffices as long as there is reasonable variability in the total-test scores of students. The details about the sample sizes for the winter, spring, and summer administrations can be found in Table 9–1.

PRE-EQUATING VERIFICATION

The pre-equating verification data included all students who met the inclusion criteria and were scored by 04/13/2021 and 08/23/2021 for the winter and spring administrations, respectively². Pre-equating verification data included students who had testing accommodations, except those who utilized alternate video sign language forms.

FINAL DATA

The final data files were used to conduct item analyses for the operational items and analyses conducted for Chapters Sixteen through Nineteen in this technical report. The final data contained students' responses to both the MC and CR items. All students' responses included in the analyses met the filtering criteria. The final sample sizes (or *n*-counts) can be found in the column labeled "Final" in Table 9–1.

Table 9–1. Data Source *N*-Counts

Administration	Content Area	Key Verification	Pre-Equating Verification	Final
Winter	Algebra I	5791	13201	13200
Winter	Biology	4361	12143	12144
Winter	Literature	7027	11714	11722
Spring	Algebra I	94889	94861	109710
Spring	Biology	86628	86599	100976
Spring	Literature	84038	84036	97928
Summer	Algebra I			
Summer	Biology			
Summer	Literature			

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

SPIRALING OF FORMS

During the administration of Keystone Exams, test forms were spiraled at the student level. The goal of spiraling is to achieve equivalent samples of students across forms so the classical statistics (e.g., *p*-value and point-biserial correlation) for all the field-test items can be compared. Given that the field-test items were embedded in the spring administration only, the equivalence of samples was checked for the spring administration instead of all administrations. When spiraling achieves randomly equivalent samples, the forms will have equal means (within sampling error) over the operational items.

² The winter 2020–2021 and spring 2021 testing windows were both extended. The pre-equating verification for the spring 2021 administration was based only on the first wave of test-takers who tested through mid-July 2021.

Appendix H provides summary statistics for all the spring forms for each content area exam. The tables provide the form number (Form), number of students (N), test length in items (L), total points (Pts.), minimum (Min) score, maximum (Max) score, mean (Mean) score, median (Med) score, and standard deviation (SD). The extent to which the mean raw scores across forms are similar indicates the extent to which the student populations taking each form are of approximately equal ability. This equivalence of ability distributions across forms is the desired outcome of spiraling and allows for optimum analysis of the embedded field-test items.

Figure 9–1 presents the mean raw scores for all forms by content area and is based on the final data; in addition, accommodated forms (i.e., Spanish, Braille, large print, VSL) were removed to best represent comparisons among similar groups of test-takers. The spring form mean raw scores are plotted (circle-shaped marker) with standard error of mean lines. For each form, the standard error of mean was computed by taking the standard deviation of all student scores (assumed as the population standard deviation divided by the square root of the form *n*-count). The mean score across all forms is indicated by the red horizontal broken line. If the three-standard error band captures the horizontal line, then that suggests only random differences exist between the form mean and the population mean. This is true for all forms in all content areas.

Figure 9–1. Spring Form Mean Scores with +/- Standard Error (SE) Bands

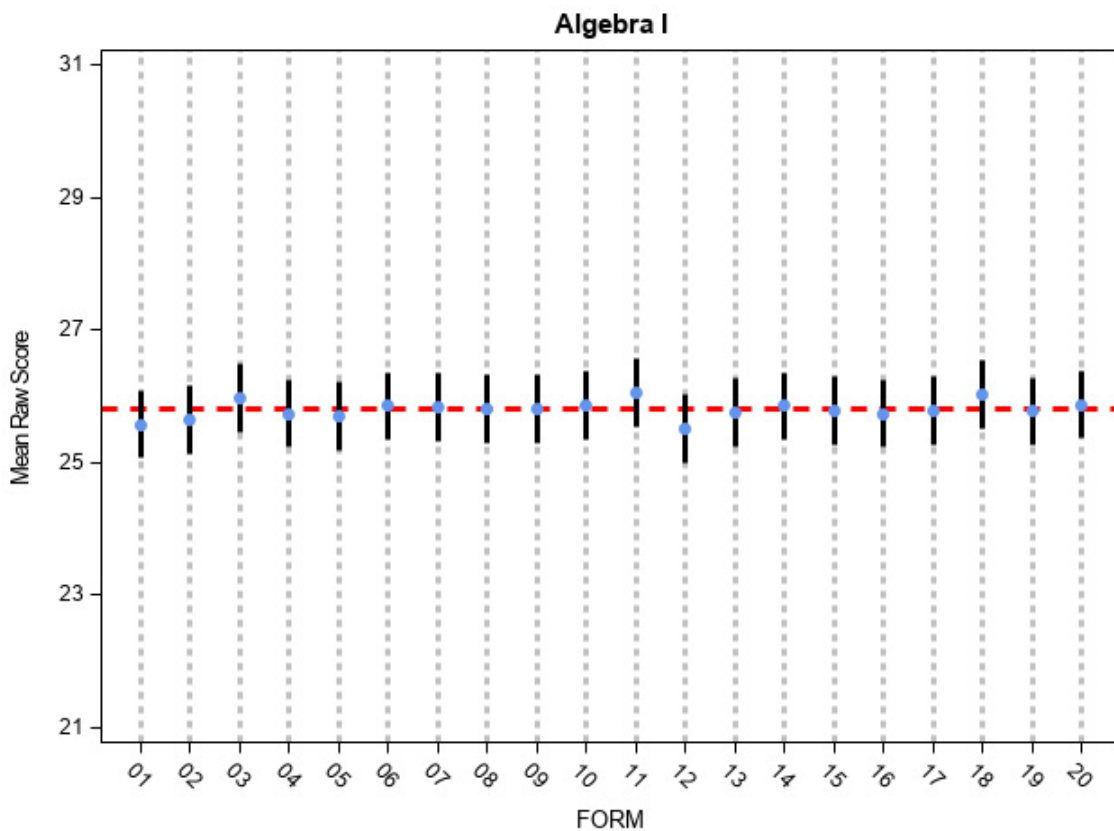
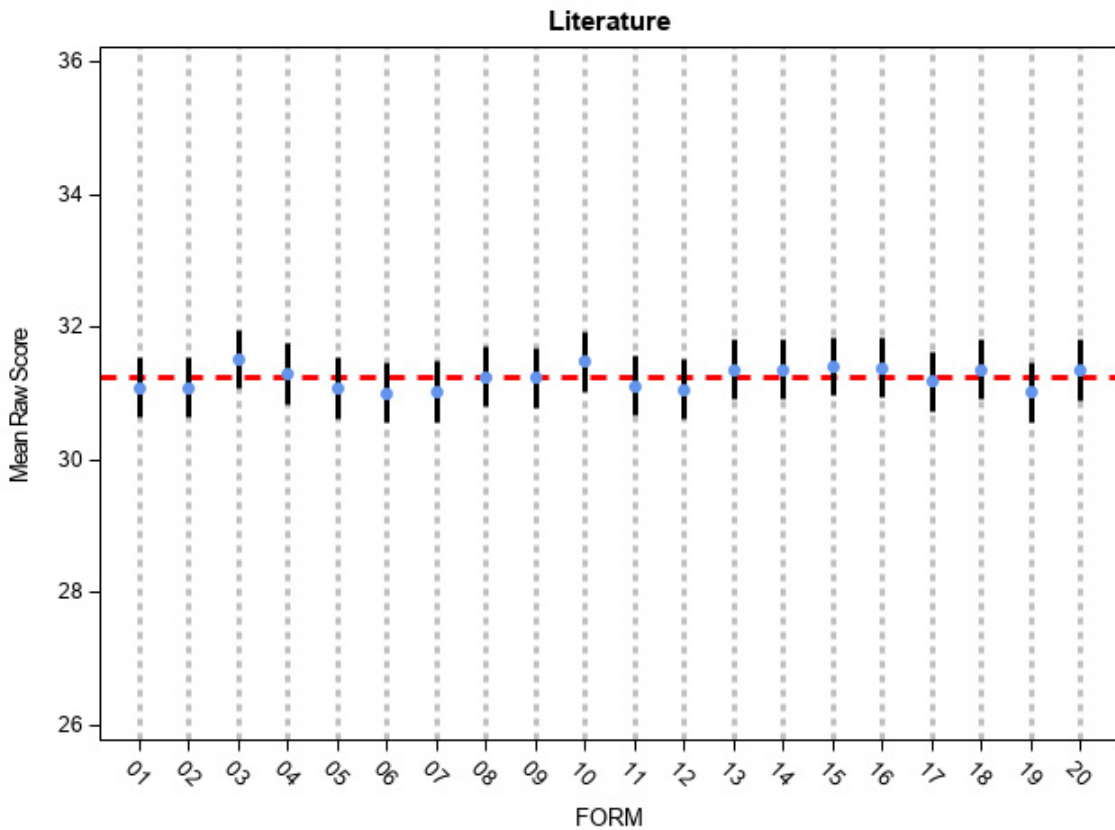
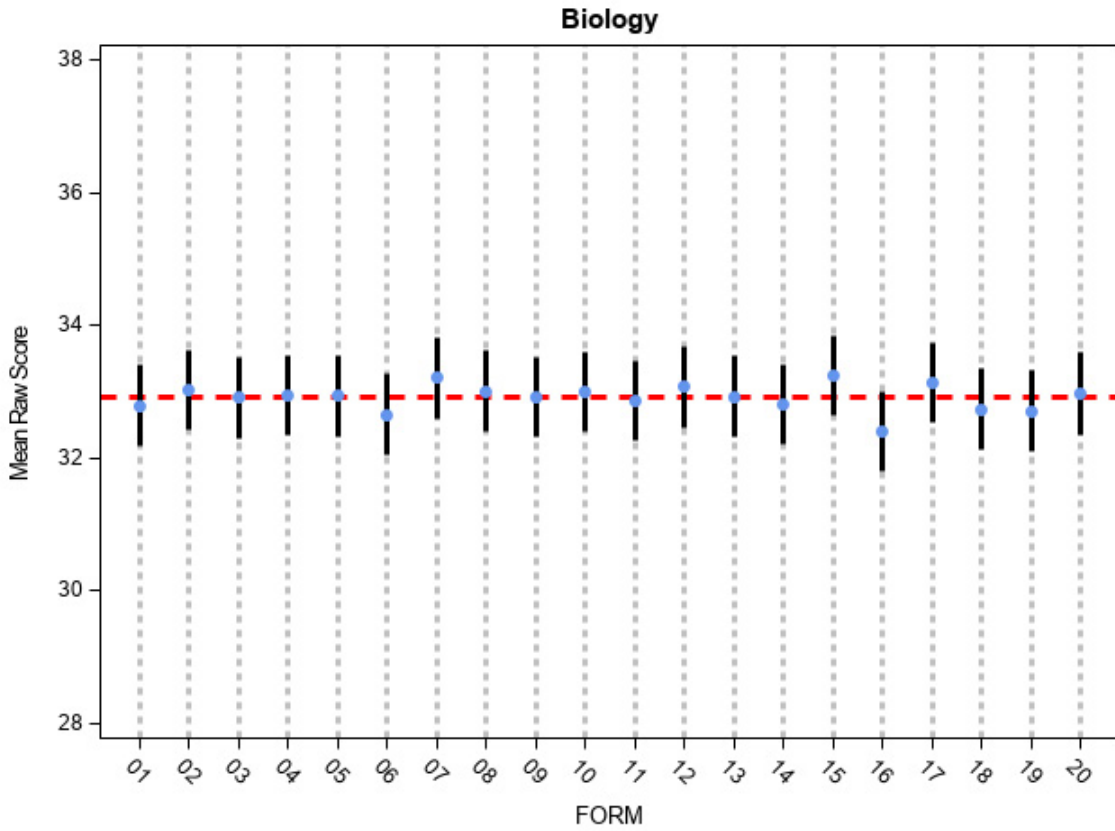


Figure 9–1 (continued). Spring Form Mean Scores with +/- Standard Error (SE) Bands



CHAPTER TEN: SUMMARY DEMOGRAPHIC AND ACCOMMODATION DATA FOR SPRING 2021 KEYSTONE EXAMS

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information. The corresponding Appendix I presents the results for the winter administration only due to the cancellation of the 2021 summer administration.

ASSESSED STUDENTS

Students assessed on the Keystone Exams include students from public schools who are required to participate by virtue of being in the graduating class of 2023, students in a school district planning to use the Keystone Exams to meet graduation requirements, and students enrolled in an Algebra I, Biology, or Literature course during the 2020–2021 school year. The operational Keystone Exams were administered in both paper-and-pencil test (PPT) and computer-based test (CBT) formats.

Statistical tables and associated commentary embedded in this chapter are based on data from the Spring test administration of the Keystone Exams. There were two other administrations during the school year, each of which involved fewer students than the spring. One occurred during winter and the other in the summer. Tables summarizing results from these two administrations can be found in Appendix I.

Results for this chapter are presented in sets of tables for the three Keystone Exams administered in Spring (Algebra I, Biology, and Literature). Accompanying each numbered table is a letter (A, B, or L) to designate the content area. Tables 10–1A through 10–1L provides a summary of tests processed and scored, which are displayed separately by student grade level. The first two rows present the number processed for each administration mode (PPT and CBT). The total number of tests processed is presented on the third row. The fourth and fifth rows show the number and percentage of students with a Keystone Exam score, while the sixth and seventh row presents the number and percentage not receiving a score. Please note that the percent of students assessed (received a total score) is typically in the high 90s across grade levels.

Table 10–1A. Students Assessed on the Spring 2021 Keystone Exam: Algebra I

Description	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)	48	229	4,128	19,768	43,755	16,218	6,406	167	90,719
Total number of CBT processed (Number)	2	71	2,471	9,115	13,754	5,461	1,551	34	32,459
Total number of tests processed (Number)	50	300	6,599	28,883	57,509	21,679	7,957	201	123,178
Total number of tests processed with a score (Number)	28	289	6,278	27,012	51,099	17,566	5,766	132	108,170
Total number of tests processed with a score (Percent)	56	96.3	95.1	93.5	88.9	81	72.5	65.7	87.8
Total number of tests processed without a score (Number)	22	11	321	1,871	6,410	4,113	2,191	69	15,008
Total number of tests processed without a score (Percent)	44	3.7	4.9	6.5	11.1	19	27.5	34.3	12.2

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table 10–1B. Students Assessed on the Spring 2021 Keystone Exam: Biology

Description	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)	42	273	34,665	39,453	5,955	193	80,581
Total number of CBT processed (Number)	0	27	13,199	17,190	1,662	37	32,115
Total number of tests processed (Number)	42	300	47,864	56,643	7,617	230	112,696
Total number of tests processed with a score (Number)	20	282	43,880	49,755	5,653	172	99,762
Total number of tests processed with a score (Percent)	47.6	94	91.7	87.8	74.2	74.8	88.5
Total number of tests processed without a score (Number)	22	18	3,984	6,888	1,964	58	12,934
Total number of tests processed without a score (Percent)	52.4	6	8.3	12.2	25.8	25.2	11.5

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table 10–1L. Students Assessed on the Spring 2021 Keystone Exam: Literature

Description	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)	41	23	4,935	69,501	5,285	181	79,966
Total number of CBT processed (Number)	0	1	604	26,517	2,519	50	29,691
Total number of tests processed (Number)	41	24	5,539	96,018	7,804	231	109,657
Total number of tests processed with a score (Number)	24	22	4,861	86,129	5,417	168	96,621
Total number of tests processed with a score (Percent)	58.5	91.7	87.8	89.7	69.4	72.7	88.1
Total number of tests processed without a score (Number)	17	2	678	9,889	2,387	63	13,036
Total number of tests processed without a score (Percent)	41.5	8.3	12.2	10.3	30.6	27.3	11.9

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

REASONS FOR STUDENT NON-ASSESSMENT

Typically, a small percent of students were not assessed. Although there are a variety of reasons for this, the major ones pertain to:

- Extended absence from school that continued beyond the assessment window.
- Failure to meet the attempt criteria on one or more test modules and no exclusion code marked by school personnel. The attempt criteria required a minimum of five items to be completed in each module.
- Medical emergency.
- Parental request in which the student’s parent/guardian reviewed the assessment, found it to be in conflict with his/her religious belief, and requested in writing that the student be excluded from participation.

- Parental request in which the student’s parent/guardian chose to have his/her child excluded from participation based on reasons other than conflict with religious belief, even though there is no provision for this exclusion in Pennsylvania regulation.
- Other reasons.

The number of students without a total test score for each of these reasons is provided in Tables 10–2A through 10–2L. Associated with this number is the percent of the total of non-assessed students in each column (grade level) attributed to a particular reason.

Table 10–2A. Counts/Percentages of Students without Scores on the Spring 2021 Keystone Exam: Algebra I

Reason for Non-Assessment	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)	0	2	41	384	1,951	1,389	762	38	4,567
Extended absence from school (Percent)	0	18.2	12.8	20.5	30.4	33.8	34.8	55.1	30.4
Non-attempt (Number)	5	0	4	69	1,092	614	273	5	2,062
Non-attempt (Percent)	22.7	0	1.2	3.7	17	14.9	12.5	7.2	13.7
Medical emergency (Number)	1	0	8	54	106	64	19	1	253
Medical emergency (Percent)	4.5	0	2.5	2.9	1.7	1.6	.9	1.4	1.7
Parental request - Chapter 4 (Number)	0	2	36	125	376	176	71	5	791
Parental request - Chapter 4 (Percent)	0	18.2	11.2	6.7	5.9	4.3	3.2	7.2	5.3
Parental request - Other reasons (Number)	0	5	159	960	1,568	904	428	2	4,026
Parental request - Other reasons (Percent)	0	45.5	49.5	51.3	24.5	22	19.5	2.9	26.8
Other reasons (Number)	16	2	73	279	1,317	966	638	18	3,309
Other reasons (Percent)	72.7	18.2	22.7	14.9	20.5	23.5	29.1	26.1	22
Total not assessed (Number)	22	11	321	1,871	6,410	4,113	2,191	69	15,008

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table 10–2B. Counts/Percentages of Students without Scores on the Spring 2021 Keystone Exam: Biology

Reason for Non-Assessment	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)	0	0	970	2,310	713	28	4,021
Extended absence from school (Percent)	0	0	24.3	33.5	36.3	48.3	31.1
Non-attempt (Number)	3	0	404	953	227	5	1,592
Non-attempt (Percent)	13.6	0	10.1	13.8	11.6	8.6	12.3
Medical emergency (Number)	0	0	71	94	28	0	193
Medical emergency (Percent)	0	0	1.8	1.4	1.4	0	1.5
Parental request - Chapter 4 (Number)	0	7	189	451	70	5	722
Parental request - Chapter 4 (Percent)	0	38.9	4.7	6.5	3.6	8.6	5.6
Parental request - Other reasons (Number)	2	4	1,190	1,501	306	3	3,006
Parental request - Other reasons (Percent)	9.1	22.2	29.9	21.8	15.6	5.2	23.2
Other reasons (Number)	17	7	1,160	1,578	620	17	3,399
Other reasons (Percent)	77.3	38.9	29.1	22.9	31.6	29.3	26.3
Total not assessed (Number)	22	18	3,984	6,888	1,964	58	12,934

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table 10–2L. Counts/Percentages of Students without Scores on the Spring 2021 Keystone Exam: Literature

Reason for Non-Assessment	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)	0	2	204	3,482	608	29	4,325
Extended absence from school (Percent)	0	100	30.1	35.2	25.5	46	33.2
Non-attempt (Number)	2	0	84	1,354	271	10	1,721
Non-attempt (Percent)	11.8	0	12.4	13.7	11.4	15.9	13.2
EL in first year in U.S. schools (Number)	0	0	0	33	5	0	38
EL in first year in U.S. schools (Percent)	0	0	0	.3	.2	0	.3
Medical emergency (Number)	0	0	13	177	19	1	210
Medical emergency (Percent)	0	0	1.9	1.8	.8	1.6	1.6
Parental request - Chapter 4 (Number)	0	0	23	594	41	3	661
Parental request - Chapter 4 (Percent)	0	0	3.4	6	1.7	4.8	5.1
Parental request - Other reasons (Number)	0	0	123	2,087	1,080	4	3,294
Parental request - Other reasons (Percent)	0	0	18.1	21.1	45.2	6.3	25.3
Other reasons (Number)	15	0	231	2,161	363	16	2,786
Other reasons (Percent)	88.2	0	34.1	21.9	15.2	25.4	21.4
Total not assessed (Number)	17	2	678	9,889	2,387	63	13,036

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

DEMOGRAPHIC CHARACTERISTICS OF STUDENTS RECEIVING TEST SCORES

COMPOSITION OF SAMPLE USED IN SUBSEQUENT TABLES

The following state summary statistic data analyses were completed using the final individual student data file containing records from the Spring administration, which are typically provided to the Pennsylvania Department of Education in July¹. State summary statistics were based on students who received a total test score on the Spring administration with the exception of students who attended non-public schools or those who were home schooled. Also excluded were students who were non-Keystone proficient.

Demographic data for students taking the Keystone Exams is presented separately for each course (Tables 10–3A, 10–3B, 10–3L). Results for accommodations received were collected separately by course and are presented in separate tables as well. For example, tables involving accommodations for Biology are found in Tables 10–4B, 10–5B, 10–6B, and 10–7B.

COLLECTION OF STUDENT DEMOGRAPHIC INFORMATION

Data for analyses involving demographic characteristics were obtained primarily from information supplied by school district personnel through the Pennsylvania Information Management System (PIMS) and subsequently transmitted to DRC. Some data such as accommodation information are recorded by school personnel directly on the student answer document (PPT) or in the DRC INSIGHT Portal Test Setup (CBT) at the time a Keystone Exam is administered.

DEMOGRAPHIC CHARACTERISTICS

Frequency data for each demographic category is presented in Tables 10–3A through 10–3L. Data is presented by grade level with PPT and CBT formats combined into a single composite. Shown at the bottom of the appropriate table is the number of assessed students contributing to summary statistics on which the column percentages are based.

¹ Due to the extension to the Spring 2021 testing window, 2021 files were delivered in January 2022.

Table 10–3A. Demographic Characteristics of Students taking the Spring 2021 Keystone Exam: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)	1	88	2,888	14,081	25,423	7,808	2,670	48	53,007
Female (Percent)	3.6	30.4	46	52.1	49.8	44.4	46.3	36.4	49
Male (Number)	12	201	3,390	12,923	25,655	9,742	3,094	81	55,098
Male (Percent)	42.9	69.6	54	47.8	50.2	55.5	53.7	61.4	50.9
American Indian/Alaskan Native (not Hispanic) (Number)	0	1	4	37	77	58	8	1	186
American Indian/Alaskan Native (not Hispanic) (Percent)	0	.3	.1	.1	.2	.3	.1	.8	.2
Asian (not Hispanic) (Number)	6	122	820	1,581	1,858	445	133	2	4,967
Asian (not Hispanic) (Percent)	21.4	42.2	13.1	5.9	3.6	2.5	2.3	1.5	4.6
Black or African American (not Hispanic) (Number)	0	3	113	853	7,823	2,988	1,111	61	12,952
Black or African American (not Hispanic) (Percent)	0	1	1.8	3.2	15.3	17	19.3	46.2	12
Hispanic (any race) (Number)	0	4	210	1,515	6,645	2,697	915	17	12,003
Hispanic (any race) (Percent)	0	1.4	3.3	5.6	13	15.4	15.9	12.9	11.1
Multi-Racial (not Hispanic) (Number)	0	12	206	802	1,996	662	210	4	3,892
Multi-Racial (not Hispanic) (Percent)	0	4.2	3.3	3	3.9	3.8	3.6	3	3.6
White (not Hispanic) (Number)	6	147	4,922	22,196	32,628	10,684	3,384	44	74,011
White (not Hispanic) (Percent)	21.4	50.9	78.4	82.2	63.9	60.8	58.7	33.3	68.4
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)	0	0	3	20	51	16	3	0	93
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)	0	0	0	.1	.1	.1	.1	0	.1
IEP (not gifted) (Number)	3	4	136	931	7,689	5,359	1,985	60	16,167
IEP (not gifted) (Percent)	10.7	1.4	2.2	3.4	15	30.5	34.4	45.5	14.9
Student exited IEP in last 2 years (Number)	0	12	141	680	977	351	68	0	2,229
Student exited IEP in last 2 years (Percent)	0	4.2	2.2	2.5	1.9	2	1.2	0	2.1
Title I (Number)	4	12	526	4,070	13,409	4,915	1,617	50	24,603
Title I (Percent)	14.3	4.2	8.4	15.1	26.2	28	28	37.9	22.7
Title III served (Number)	0	0	8	82	1,887	1,205	421	9	3,612
Title III served (Percent)	0	0	.1	.3	3.7	6.9	7.3	6.8	3.3
Title III not served (Number)	0	0	0	0	0	0	0	0	0
Title III not served (Percent)	0	0	0	0	0	0	0	0	0
Migrant student (Number)	0	0	0	10	49	47	13	1	120
Migrant student (Percent)	0	0	0	0	.1	.3	.2	.8	.1
EL enrolled first year (Number)	0	0	0	11	199	106	56	2	374
EL enrolled first year (Percent)	0	0	0	0	.4	.6	1	1.5	.3
EL enrolled not first year (Number)	0	0	8	94	1,802	1,126	384	9	3,423
EL enrolled not first year (Percent)	0	0	.1	.3	3.5	6.4	6.7	6.8	3.2

Table 10–3A (continued). Demographic Characteristics of Students taking the Spring 2021 Keystone Exam: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in first year of monitoring (Number)	0	0	0	37	179	36	11	0	263
Exited ESL/bilingual program and in first year of monitoring (Percent)	0	0	0	.1	.4	.2	.2	0	.2
Exited ESL/bilingual program and in 2nd year of monitoring (Number)	0	0	14	38	157	61	15	0	285
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)	0	0	.2	.1	.3	.3	.3	0	.3
Former EL no longer monitored (Number)	0	3	85	462	1,108	282	75	1	2,016
Former EL no longer monitored (Percent)	0	1	1.4	1.7	2.2	1.6	1.3	.8	1.9
LIFE first year (Number)	0	0	0	0	3	2	0	0	5
LIFE first year (Percent)	0	0	0	0	0	0	0	0	0
LIFE not first year (Number)	0	0	0	3	16	42	13	0	74
LIFE not first year (Percent)	0	0	0	0	0	.2	.2	0	.1
Former EL exited and in 3rd year of monitoring (Number)	0	0	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Percent)	0	0	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Number)	0	0	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Percent)	0	0	0	0	0	0	0	0	0
Foreign exchange student (Number)	0	0	0	0	0	0	0	0	0
Foreign exchange student (Percent)	0	0	0	0	0	0	0	0	0
Economically disadvantaged (Number)	0	10	677	5,575	22,503	9,498	3,152	90	41,505
Economically disadvantaged (Percent)	0	3.5	10.8	20.6	44	54.1	54.7	68.2	38.4
Historically Underperforming Subgroup (Number)	3	14	801	6,306	26,791	12,127	4,097	110	50,249
Historically Underperforming Subgroup (Percent)	10.7	4.8	12.8	23.3	52.4	69	71.1	83.3	46.5
Enrollment in school of residence after Oct 1 (Number)	0	4	38	279	1,422	676	282	13	2,714
Enrollment in school of residence after Oct 1 (Percent)	0	1.4	.6	1	2.8	3.8	4.9	9.8	2.5
Enrollment in district of residence after Oct 1 (Number)	0	1	29	240	1,147	561	236	11	2,225
Enrollment in district of residence after Oct 1 (Percent)	0	.3	.5	.9	2.2	3.2	4.1	8.3	2.1
Enrollment as PA resident after Oct 1 (Number)	0	0	11	114	589	272	107	3	1,096
Enrollment as PA resident after Oct 1 (Percent)	0	0	.2	.4	1.2	1.5	1.9	2.3	1
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)	4	116	2,352	4,863	28,617	4,415	1,408	33	41,808

Table 10–3A (continued). Demographic Characteristics of Students taking the Spring 2021 Keystone Exam: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)	14.3	40.1	37.5	18	56	25.1	24.4	25	38.7
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)	1	67	800	2,843	8,313	2,391	845	28	15,288
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)	3.6	23.2	12.7	10.5	16.3	13.6	14.7	21.2	14.1
Military family (Number)	0	0	38	122	249	85	23	1	518
Military family (Percent)	0	0	.6	.5	.5	.5	.4	.8	.5
Homeless (Number)	0	0	0	0	0	0	0	0	0
Homeless (Percent)	0	0	0	0	0	0	0	0	0
Foster (Number)	0	0	1	22	204	123	49	7	406
Foster (Percent)	0	0	0	.1	.4	.7	.8	5.3	.4
Home schooled (Number)	0	0	0	0	0	0	0	0	0
Home schooled (Percent)	0	0	0	0	0	0	0	0	0
Court/agency placed (Number)	0	0	0	0	36	40	24	19	119
Court/agency placed (Percent)	0	0	0	0	.1	.2	.4	14.4	.1
Number of assessed students (Number)	28	289	6,278	27,012	51,099	17,566	5,766	132	108,170

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table 10–3B. Demographic Characteristics of Students Taking the Spring 2021 Keystone Exam: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)	4	138	22,564	23,444	2,572	72	48,794
Female (Percent)	20	48.9	51.4	47.1	45.5	41.9	48.9
Male (Number)	2	144	21,311	26,307	3,070	98	50,932
Male (Percent)	10	51.1	48.6	52.9	54.3	57	51.1
American Indian/Alaskan Native (not Hispanic) (Number)	0	0	51	99	12	2	164
American Indian/Alaskan Native (not Hispanic) (Percent)	0	0	.1	.2	.2	1.2	.2
Asian (not Hispanic) (Number)	0	40	2,684	1,658	209	3	4,594
Asian (not Hispanic) (Percent)	0	14.2	6.1	3.3	3.7	1.7	4.6
Black or African American (not Hispanic) (Number)	1	3	3,953	6,437	1,195	89	11,678
Black or African American (not Hispanic) (Percent)	5	1.1	9	12.9	21.1	51.7	11.7
Hispanic (any race) (Number)	0	3	3,372	6,339	918	17	10,649
Hispanic (any race) (Percent)	0	1.1	7.7	12.7	16.2	9.9	10.7
Multi-Racial (not Hispanic) (Number)	0	14	1,546	1,610	195	3	3,368
Multi-Racial (not Hispanic) (Percent)	0	5	3.5	3.2	3.4	1.7	3.4
White (not Hispanic) (Number)	5	222	32,223	33,571	3,109	56	69,186
White (not Hispanic) (Percent)	25	78.7	73.4	67.5	55	32.6	69.4
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)	0	0	46	36	3	0	85
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)	0	0	.1	.1	.1	0	.1
IEP (not gifted) (Number)	0	9	4,392	9,534	1,452	54	15,441
IEP (not gifted) (Percent)	0	3.2	10	19.2	25.7	31.4	15.5
Student exited IEP in last 2 years (Number)	0	4	833	859	71	0	1,767
Student exited IEP in last 2 years (Percent)	0	1.4	1.9	1.7	1.3	0	1.8
Title I (Number)	3	42	8,439	10,530	1,841	57	20,912
Title I (Percent)	15	14.9	19.2	21.2	32.6	33.1	21
Title III served (Number)	0	1	583	1,982	467	10	3,043
Title III served (Percent)	0	.4	1.3	4	8.3	5.8	3.1
Title III not served (Number)	0	0	0	0	0	0	0
Title III not served (Percent)	0	0	0	0	0	0	0
Migrant student (Number)	0	0	9	80	9	0	98
Migrant student (Percent)	0	0	0	.2	.2	0	.1
EL enrolled first year (Number)	0	0	56	151	52	1	260
EL enrolled first year (Percent)	0	0	.1	.3	.9	.6	.3
EL enrolled not first year (Number)	0	1	599	1,887	429	11	2,927
EL enrolled not first year (Percent)	0	.4	1.4	3.8	7.6	6.4	2.9

Table 10–3B (continued). Demographic Characteristics of Students Taking the Spring 2021 Keystone Exam: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in first year of monitoring (Number)	0	0	94	124	15	0	233
Exited ESL/bilingual program and in first year of monitoring (Percent)	0	0	.2	.2	.3	0	.2
Exited ESL/bilingual program and in 2nd year of monitoring (Number)	0	1	102	154	13	0	270
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)	0	.4	.2	.3	.2	0	.3
Former EL no longer monitored (Number)	0	10	941	969	92	0	2,012
Former EL no longer monitored (Percent)	0	3.5	2.1	1.9	1.6	0	2
LIFE first year (Number)	0	0	1	2	0	0	3
LIFE first year (Percent)	0	0	0	0	0	0	0
LIFE not first year (Number)	0	0	6	40	10	1	57
LIFE not first year (Percent)	0	0	0	.1	.2	.6	.1
Former EL exited and in 3rd year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Percent)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Percent)	0	0	0	0	0	0	0
Foreign exchange student (Number)	0	0	0	1	1	1	3
Foreign exchange student (Percent)	0	0	0	0	0	.6	0
Economically disadvantaged (Number)	0	28	13,340	21,223	2,918	104	37,613
Economically disadvantaged (Percent)	0	9.9	30.4	42.7	51.6	60.5	37.7
Historically Underperforming Subgroup (Number)	0	37	15,864	26,051	3,645	128	45,725
Historically Underperforming Subgroup (Percent)	0	13.1	36.2	52.4	64.5	74.4	45.8
Enrollment in school of residence after Oct 1 (Number)	1	1	752	1,286	305	17	2,362
Enrollment in school of residence after Oct 1 (Percent)	5	.4	1.7	2.6	5.4	9.9	2.4
Enrollment in district of residence after Oct 1 (Number)	1	1	658	1,048	247	13	1,968
Enrollment in district of residence after Oct 1 (Percent)	5	.4	1.5	2.1	4.4	7.6	2
Enrollment as PA resident after Oct 1 (Number)	0	0	294	463	129	3	889
Enrollment as PA resident after Oct 1 (Percent)	0	0	.7	.9	2.3	1.7	.9
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)	1	46	21,173	12,892	1,546	34	35,692

Table 10–3B (continued). Demographic Characteristics of Students Taking the Spring 2021 Keystone Exam: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)	5	16.3	48.3	25.9	27.3	19.8	35.8
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)	1	17	5,952	6,609	746	26	13,351
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)	5	6	13.6	13.3	13.2	15.1	13.4
Military family (Number)	0	1	207	242	44	1	495
Military family (Percent)	0	.4	.5	.5	.8	.6	.5
Homeless (Number)	0	0	0	0	0	0	0
Homeless (Percent)	0	0	0	0	0	0	0
Foster (Number)	0	0	97	247	53	6	403
Foster (Percent)	0	0	.2	.5	.9	3.5	.4
Home schooled (Number)	0	0	0	0	0	0	0
Home schooled (Percent)	0	0	0	0	0	0	0
Court/agency placed (Number)	0	0	35	46	28	16	125
Court/agency placed (Percent)	0	0	.1	.1	.5	9.3	.1
Number of assessed students (Number)	20	282	43,880	49,755	5,653	172	99,762

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade.

Table 10–3L. Demographic Characteristics of Students taking the Spring 2021 Keystone Exam: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)	3	9	2,362	42,645	2,492	65	47,576
Female (Percent)	12.5	40.9	48.6	49.5	46	38.7	49.2
Male (Number)	3	13	2,498	43,471	2,908	98	48,991
Male (Percent)	12.5	59.1	51.4	50.5	53.7	58.3	50.7
American Indian/Alaskan Native (not Hispanic) (Number)	0	0	4	146	19	1	170
American Indian/Alaskan Native (not Hispanic) (Percent)	0	0	.1	.2	.4	.6	.2
Asian (not Hispanic) (Number)	0	0	159	4,112	209	0	4,480
Asian (not Hispanic) (Percent)	0	0	3.3	4.8	3.9	0	4.6
Black or African American (not Hispanic) (Number)	1	1	470	9,818	1,053	81	11,424
Black or African American (not Hispanic) (Percent)	4.2	4.5	9.7	11.4	19.4	48.2	11.8
Hispanic (any race) (Number)	1	0	435	8,794	1,001	23	10,254
Hispanic (any race) (Percent)	4.2	0	8.9	10.2	18.5	13.7	10.6
Multi-Racial (not Hispanic) (Number)	0	2	161	2,862	179	2	3,206
Multi-Racial (not Hispanic) (Percent)	0	9.1	3.3	3.3	3.3	1.2	3.3
White (not Hispanic) (Number)	4	19	3,626	60,314	2,933	56	66,952
White (not Hispanic) (Percent)	16.7	86.4	74.6	70	54.1	33.3	69.3
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)	0	0	5	71	5	0	81
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)	0	0	.1	.1	.1	0	.1
IEP (not gifted) (Number)	3	1	727	12,521	1,477	50	14,779
IEP (not gifted) (Percent)	12.5	4.5	15	14.5	27.3	29.8	15.3
Student exited IEP in last 2 years (Number)	0	0	54	1,555	59	0	1,668
Student exited IEP in last 2 years (Percent)	0	0	1.1	1.8	1.1	0	1.7
Title I (Number)	0	20	1,068	17,005	1,943	59	20,095
Title I (Percent)	0	90.9	22	19.7	35.9	35.1	20.8
Title III served (Number)	0	0	90	2,072	473	15	2,650
Title III served (Percent)	0	0	1.9	2.4	8.7	8.9	2.7
Title III not served (Number)	0	0	0	0	0	0	0
Title III not served (Percent)	0	0	0	0	0	0	0
Migrant student (Number)	0	0	2	76	14	1	93
Migrant student (Percent)	0	0	0	.1	.3	.6	.1
EL enrolled first year (Number)	0	0	13	118	35	0	166
EL enrolled first year (Percent)	0	0	.3	.1	.6	0	.2
EL enrolled not first year (Number)	1	0	90	2,056	468	17	2,632
EL enrolled not first year (Percent)	4.2	0	1.9	2.4	8.6	10.1	2.7

Table 10–3L (continued). Demographic Characteristics of Students taking the Spring 2021 Keystone Exam: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in first year of monitoring (Number)	0	0	5	184	22	0	211
Exited ESL/bilingual program and in first year of monitoring (Percent)	0	0	.1	.2	.4	0	.2
Exited ESL/bilingual program and in 2nd year of monitoring (Number)	0	0	5	254	18	0	277
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)	0	0	.1	.3	.3	0	.3
Former EL no longer monitored (Number)	0	0	59	1,898	97	1	2,055
Former EL no longer monitored (Percent)	0	0	1.2	2.2	1.8	.6	2.1
LIFE first year (Number)	0	0	0	2	0	0	2
LIFE first year (Percent)	0	0	0	0	0	0	0
LIFE not first year (Number)	0	0	0	43	3	0	46
LIFE not first year (Percent)	0	0	0	0	.1	0	0
Former EL exited and in 3rd year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Percent)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Percent)	0	0	0	0	0	0	0
Foreign exchange student (Number)	0	0	0	2	0	0	2
Foreign exchange student (Percent)	0	0	0	0	0	0	0
Economically disadvantaged (Number)	0	19	1,813	31,343	2,942	103	36,220
Economically disadvantaged (Percent)	0	86.4	37.3	36.4	54.3	61.3	37.5
Historically Underperforming Subgroup (Number)	4	19	2,141	37,994	3,656	121	43,935
Historically Underperforming Subgroup (Percent)	16.7	86.4	44	44.1	67.5	72	45.5
Enrollment in school of residence after Oct 1 (Number)	0	0	162	1,740	301	19	2,222
Enrollment in school of residence after Oct 1 (Percent)	0	0	3.3	2	5.6	11.3	2.3
Enrollment in district of residence after Oct 1 (Number)	0	0	126	1,439	255	15	1,835
Enrollment in district of residence after Oct 1 (Percent)	0	0	2.6	1.7	4.7	8.9	1.9
Enrollment as PA resident after Oct 1 (Number)	0	0	48	639	122	4	813
Enrollment as PA resident after Oct 1 (Percent)	0	0	1	.7	2.3	2.4	.8
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	2,313	21,247	1,466	34	25,060

Table 10–3L (continued). Demographic Characteristics of Students taking the Spring 2021 Keystone Exam: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	47.6	24.7	27.1	20.2	25.9
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	473	10,353	624	22	11,472
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	9.7	12	11.5	13.1	11.9
Military family (Number)	0	0	26	421	23	1	471
Military family (Percent)	0	0	.5	.5	.4	.6	.5
Homeless (Number)	0	0	0	0	0	0	0
Homeless (Percent)	0	0	0	0	0	0	0
Foster (Number)	0	0	22	301	51	8	382
Foster (Percent)	0	0	.5	.3	.9	4.8	.4
Home schooled (Number)	0	0	0	0	0	0	0
Home schooled (Percent)	0	0	0	0	0	0	0
Court/agency placed (Number)	0	0	37	44	28	21	130
Court/agency placed (Percent)	0	0	.8	.1	.5	12.5	.1
Number of assessed students (Number)	24	22	4,861	86,129	5,417	168	96,621

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

PARTICIPATION BY ADMINISTRATION MODE

Although there are two administration modes, paper/pencil test (PPT) and computer-based test (CBT), most students are administered PPT.

TEST ACCOMMODATIONS PROVIDED

School personnel supplied information regarding accommodations that a student may have received while taking the Keystone Exams. Accommodations are classified in terms of presentation, response, setting, and timing to enable students to better manage disabilities that hinder their ability to learn and respond to assessments. An accommodations manual entitled, *Accommodations Guidelines* was updated for use with the PSSA and Keystone Exams. This manual may be found on the PDE website at www.education.pa.gov. A glossary of accommodation terms as applied to the Keystone Exams is provided in Table 10–10 at the end of this chapter.

The frequency with which accommodations were utilized for PPT and CBT formats is summarized separately for each course exam in Tables 10–4A through 10–7L. Tabled values are based on all students whose scores contributed to state summary statistics in a given Keystone Exam. Because of the very small incidence of usage of many accommodations, combined with the fact that a number of accommodations are primarily accessed by only one of the two administration modes, meaningful comparisons between modes are rather limited. In the following tables, an NA denotes those instances in which a particular accommodation does not apply to one of the testing modes.

PRESENTATION ACCOMMODATIONS RECEIVED

Presentation accommodations are those that provide alternate ways for students to access and process printed instructional material and assessments. These include auditory, tactile, visual, and combined auditory/visual modes of presentation. The number of presentation accommodations provided in the Keystone Exams varied by content area and test administration mode.

As depicted in Tables 10–4A through 10–4L, the actual frequencies were quite low, generally representing less than one-tenth of one percent of assessed students in each course. The most notable exceptions, applicable to Algebra I and Biology only, were “All items/questions read aloud” and “Some items/questions read aloud.” Among accommodations specific to CBT, the use of audio was the most frequent.

Table 10–4A. Incidence of Presentation Accommodations Received on the Spring 2021 Keystone Exam: Algebra I

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)	9	N/A	9
Braille format (Percent)	0	N/A	0
Large print format (Number)	40	N/A	40
Large print format (Percent)	.1	N/A	0
Computer Assistive Technology (Number)	4	N/A	4
Computer Assistive Technology (Percent)	0	N/A	0
Some test items/questions read aloud (Number)	461	327	788
Some test items/questions read aloud (Percent)	.6	1.1	.7
All test items/questions read aloud (Number)	334	330	664
All test items/questions read aloud (Percent)	.4	1.1	.6
Test items/questions signed (Number)	11	2	13
Test items/questions signed (Percent)	0	0	0
Test items/questions interpreted for EL student (Number)	6	9	15
Test items/questions interpreted for EL student (Percent)	0	0	0
Amplification device (Number)	10	5	15
Amplification device (Percent)	0	0	0
Magnification device (Number)	4	7	11
Magnification device (Percent)	0	0	0
Color overlay (Number)	3	N/A	3
Color overlay (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	55	21	76
Other (per Accommodations Guidelines) (Percent)	.1	.1	.1
Spanish version (Number)	515	N/A	515
Spanish version (Percent)	.7	N/A	.5
Audio (Number)	N/A	1,090	1,090
Audio (Percent)	N/A	3.7	1
Color Chooser (Number)	N/A	51	51
Color Chooser (Percent)	N/A	.2	0
Contrasting Text Chooser (Number)	N/A	51	51
Contrasting Text Chooser (Percent)	N/A	.2	0
Reverse Contrast (Number)	N/A	2	2
Reverse Contrast (Percent)	N/A	0	0
Refreshable Braille (Number)	N/A	0	0
Refreshable Braille (Percent)	N/A	0	0
Video Sign Language (Number)	N/A	6	6
Video Sign Language (Percent)	N/A	0	0
Number of assessed students (Number)	78,545	29,625	108,170

Table 10–4B. Incidence of Presentation Accommodations Received on the Spring 2021 Keystone Exam: Biology

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)	8	N/A	8
Braille format (Percent)	0	N/A	0
Large print format (Number)	40	N/A	40
Large print format (Percent)	.1	N/A	0
Computer Assistive Technology (Number)	8	N/A	8
Computer Assistive Technology (Percent)	0	N/A	0
Some test items/questions read aloud (Number)	354	404	758
Some test items/questions read aloud (Percent)	.5	1.4	.8
All test items/questions read aloud (Number)	380	363	743
All test items/questions read aloud (Percent)	.5	1.2	.7
Test items/questions signed (Number)	7	3	10
Test items/questions signed (Percent)	0	0	0
Test items/questions interpreted for EL student (Number)	7	9	16
Test items/questions interpreted for EL student (Percent)	0	0	0
Amplification device (Number)	6	6	12
Amplification device (Percent)	0	0	0
Magnification device (Number)	7	7	14
Magnification device (Percent)	0	0	0
Color overlay (Number)	4	N/A	4
Color overlay (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	43	39	82
Other (per Accommodations Guidelines) (Percent)	.1	.1	.1
Spanish version (Number)	421	N/A	421
Spanish version (Percent)	.6	N/A	.4
Audio (Number)	N/A	1,165	1,165
Audio (Percent)	N/A	3.9	1.2
Color Chooser (Number)	N/A	24	24
Color Chooser (Percent)	N/A	.1	0
Contrasting Text Chooser (Number)	N/A	23	23
Contrasting Text Chooser (Percent)	N/A	.1	0
Reverse Contrast (Number)	N/A	1	1
Reverse Contrast (Percent)	N/A	0	0
Refreshable Braille (Number)	N/A	0	0
Refreshable Braille (Percent)	N/A	0	0
Video Sign Language (Number)	N/A	6	6
Video Sign Language (Percent)	N/A	0	0
Number of assessed students (Number)	70,193	29,569	99,762

Table 10–4L. Incidence of Presentation Accommodations Received on the Spring 2021 Keystone Exam: Literature

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)	10	N/A	10
Braille format (Percent)	0	N/A	0
Large print format (Number)	53	N/A	53
Large print format (Percent)	.1	N/A	.1
Computer Assistive Technology (Number)	11	N/A	11
Computer Assistive Technology (Percent)	0	N/A	0
Amplification device (Number)	11	5	16
Amplification device (Percent)	0	0	0
Magnification device (Number)	5	5	10
Magnification device (Percent)	0	0	0
Color overlay (Number)	3	N/A	3
Color overlay (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	52	22	74
Other (per Accommodations Guidelines) (Percent)	.1	.1	.1
Color Chooser (Number)	N/A	31	31
Color Chooser (Percent)	N/A	.1	0
Contrasting Text Chooser (Number)	N/A	30	30
Contrasting Text Chooser (Percent)	N/A	.1	0
Reverse Contrast (Number)	N/A	4	4
Reverse Contrast (Percent)	N/A	0	0
Refreshable Braille (Number)	N/A	0	0
Refreshable Braille (Percent)	N/A	0	0
Number of assessed students (Number)	69,696	26,925	96,621

RESPONSE ACCOMMODATIONS RECEIVED

Response accommodations permit students to complete assignments, tests, and activities in different ways and to solve or organize problems using some type of assistive device or organizer. The number of response accommodations provided on the Spring Keystone Exams varied by subject.

The frequency with which these accommodations were utilized is summarized in Tables 10–5A through 10–5L. The actual frequencies are quite low, representing less than one-tenth of one percent of assessed students in nearly all instances, regardless of administration mode.

Table 10–5A. Incidence of Response Accommodations Received on the Spring 2021 Keystone Exam: Algebra I

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student’s direction (Number)	50	2	52
Test administrator marked multiple-choice responses at student’s direction (Percent)	.1	0	0
Test administrator scribed open-ended responses at student’s direction (Number)	26	4	30
Test administrator scribed open-ended responses at student’s direction (Percent)	0	0	0
Test administrator transcribed student responses (Number)	57	7	64
Test administrator transcribed student responses (Percent)	.1	0	.1
Qualified interpreter translated, transcribed, and/or scribed student’s signed responses (Number)	10	1	11
Qualified interpreter translated, transcribed, and/or scribed student’s signed responses (Percent)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Number)	4	7	11
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Percent)	0	0	0
Keyboard, word processor, or computer (Number)	15	N/A	15
Keyboard, word processor, or computer (Percent)	0	N/A	0
Braille/Notetaker (Number)	5	N/A	5
Braille/Notetaker (Percent)	0	N/A	0
Augmentative communication device (Number)	3	0	3
Augmentative communication device (Percent)	0	0	0
Computer Assistive Technology (Number)	2	N/A	2
Computer Assistive Technology (Percent)	0	N/A	0
Translation dictionary for EL student (Number)	94	21	115
Translation dictionary for EL student (Percent)	.1	.1	.1
Other (per Accommodations Guidelines) (Number)	49	19	68
Other (per Accommodations Guidelines) (Percent)	.1	.1	.1
Number of assessed students (Number)	78,545	29,625	108,170

Table 10–5B. Incidence of Response Accommodations Received on the Spring 2021 Keystone Exam: Biology

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student's direction (Number)	38	3	41
Test administrator marked multiple-choice responses at student's direction (Percent)	.1	0	0
Test administrator scribed open-ended responses at student's direction (Number)	46	8	54
Test administrator scribed open-ended responses at student's direction (Percent)	.1	0	.1
Test administrator transcribed student responses (Number)	62	5	67
Test administrator transcribed student responses (Percent)	.1	0	.1
Qualified interpreter translated, transcribed, and/or scribed student's signed responses (Number)	5	2	7
Qualified interpreter translated, transcribed, and/or scribed student's signed responses (Percent)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Number)	3	8	11
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Percent)	0	0	0
Keyboard, word processor, or computer (Number)	30	N/A	30
Keyboard, word processor, or computer (Percent)	0	N/A	0
Braille/Notetaker (Number)	2	N/A	2
Braille/Notetaker (Percent)	0	N/A	0
Augmentative communication device (Number)	2	0	2
Augmentative communication device (Percent)	0	0	0
Computer Assistive Technology (Number)	6	N/A	6
Computer Assistive Technology (Percent)	0	N/A	0
Translation dictionary for EL student (Number)	96	35	131
Translation dictionary for EL student (Percent)	.1	.1	.1
Other (per Accommodations Guidelines) (Number)	33	20	53
Other (per Accommodations Guidelines) (Percent)	0	.1	.1
Number of assessed students (Number)	70,193	29,569	99,762

Table 10–5L. Incidence of Response Accommodations Received on the Spring 2021 Keystone Exam: Literature

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student’s direction (Number)	29	2	31
Test administrator marked multiple-choice responses at student’s direction (Percent)	0	0	0
Test administrator scribed open-ended responses at student’s direction (Number)	35	4	39
Test administrator scribed open-ended responses at student’s direction (Percent)	.1	0	0
Test administrator transcribed student responses (Number)	99	3	102
Test administrator transcribed student responses (Percent)	.1	0	.1
Keyboard, word processor, or computer (Number)	61	N/A	61
Keyboard, word processor, or computer (Percent)	.1	N/A	.1
Braille/Notetaker (Number)	8	N/A	8
Braille/Notetaker (Percent)	0	N/A	0
Augmentative communication device (Number)	4	0	4
Augmentative communication device (Percent)	0	0	0
Computer Assistive Technology (Number)	2	N/A	2
Computer Assistive Technology (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	15	6	21
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	69,696	26,925	96,621

SETTING ACCOMMODATIONS RECEIVED

Setting accommodations permit a change in the location in which a student receives instruction or participates in an assessment. In the Keystone Exam administration, there were four categories of setting accommodations, which applied to both administration modes and to each course exam. As depicted in Tables 10–6A through 10–6L, the most common accommodation was small group setting for both PPT and CBT modes of administration.

Table 10–6A. Incidence of Setting Accommodations Received on the Spring 2021 Keystone Exam: Algebra I

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)	12	0	12
Hospital/home setting (Percent)	0	0	0
One-on-one setting (Number)	140	16	156
One-on-one setting (Percent)	.2	.1	.1
Small group setting (Number)	5,532	2,138	7,670
Small group setting (Percent)	7	7.2	7.1
Other (per Accommodations Guidelines) (Number)	45	8	53
Other (per Accommodations Guidelines) (Percent)	.1	0	0
Number of assessed students (Number)	78,545	29,625	108,170

Table 10–6B. Incidence of Setting Accommodations Received on the Spring 2021 Keystone Exam: Biology

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)	14	0	14
Hospital/home setting (Percent)	0	0	0
One-on-one setting (Number)	150	12	162
One-on-one setting (Percent)	.2	0	.2
Small group setting (Number)	5,301	2,470	7,771
Small group setting (Percent)	7.6	8.4	7.8
Other (per Accommodations Guidelines) (Number)	31	12	43
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	70,193	29,569	99,762

Table 10–6L. Incidence of Setting Accommodations Received on the Spring 2021 Keystone Exam: Literature

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)	11	0	11
Hospital/home setting (Percent)	0	0	0
One-on-one setting (Number)	123	12	135
One-on-one setting (Percent)	.2	0	.1
Small group setting (Number)	5,190	1,948	7,138
Small group setting (Percent)	7.4	7.2	7.4
Other (per Accommodations Guidelines) (Number)	28	13	41
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	69,696	26,925	96,621

TIMING ACCOMMODATIONS RECEIVED

Timing accommodations involve a change in the allowable length of time to complete assignments or assessments, including the way in which time is organized. There were four categories of timing accommodations, which applied to both administration modes and to each course exam. As depicted in Tables 10–7A through 10–7L, the most common accommodation was extended time for both PPT and CBT administration modes with slightly higher percentages for PPT than CBT in Algebra I and Literature.

Table 10–7A. Incidence of Timing Accommodations Received on the Spring 2021 Keystone Exam: Algebra I

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)	5,969	2,088	8,057
Extended time (Percent)	7.6	7	7.4
Frequent breaks (Number)	387	497	884
Frequent breaks (Percent)	.5	1.7	.8
Changed test schedule (Number)	126	22	148
Changed test schedule (Percent)	.2	.1	.1
Other (per Accommodations Guidelines) (Number)	48	5	53
Other (per Accommodations Guidelines) (Percent)	.1	0	0
Number of assessed students (Number)	78,545	29,625	108,170

Table 10–7B. Incidence of Timing Accommodations Received on the Spring 2021 Keystone Exam: Biology

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)	1,712	1,533	3,245
Extended time (Percent)	2.4	5.2	3.3
Frequent breaks (Number)	314	601	915
Frequent breaks (Percent)	.4	2	.9
Changed test schedule (Number)	156	14	170
Changed test schedule (Percent)	.2	0	.2
Other (per Accommodations Guidelines) (Number)	58	6	64
Other (per Accommodations Guidelines) (Percent)	.1	0	.1
Number of assessed students (Number)	70,193	29,569	99,762

Table 10–7L. Incidence of Timing Accommodations Received on the Spring 2021 Keystone Exam: Literature

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)	4,575	1,319	5,894
Extended time (Percent)	6.6	4.9	6.1
Frequent breaks (Number)	318	439	757
Frequent breaks (Percent)	.5	1.6	.8
Changed test schedule (Number)	147	19	166
Changed test schedule (Percent)	.2	.1	.2
Other (per Accommodations Guidelines) (Number)	36	5	41
Other (per Accommodations Guidelines) (Percent)	.1	0	0
Number of assessed students (Number)	69,696	26,925	96,621

ACCOMMODATION RATE FOR NON-IEP AND IEP STUDENTS

A comparison between students without an IEP (non-IEP Students) and those with an IEP (IEP Students) with regard to having received an accommodation is provided in Table 10–8. In this data, accommodated means that a student received one or more of the total number of accommodations available for a given course; however, this varies somewhat with administration mode. The category of non-accommodated indicates that a student did not receive any accommodations during testing.

Typically, the general pattern of findings reveals a consistent and substantially higher percentage of IEP Students receiving an accommodation, in contrast to non-IEP Students. This same pattern typically holds true regardless of test administration mode for the Keystone Exams. In prior spring administrations, the comparisons between administration modes revealed that the accommodation rates for IEP students taking a PPT are somewhat lower than accommodation rates for IEP students taking a CBT for Algebra I and Biology, but not Literature.

Table 10–8A. Accommodation Rate for Non-IEP and IEP Students on the Spring 2021 Keystone Exams: Algebra I

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)	67,012	24,991	92,003
Non-Accommodated (Number)	61,263	23,928	85,191
Non-Accommodated (Percent)	91.4	95.7	92.6
Accommodated (Number)	5,749	1,063	6,812
Accommodated (Percent)	8.6	4.3	7.4
IEP Students (Number)	11,533	4,634	16,167
Non-Accommodated (Number)	6,362	2,098	8,460
Non-Accommodated (Percent)	55.2	45.3	52.3
Accommodated (Number)	5,171	2,536	7,707
Accommodated (Percent)	44.8	54.7	47.7

Table 10–8B. Accommodation Rate for Non-IEP and IEP Students on the Spring 2021 Keystone Exams: Biology

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)	59,730	24,591	84,321
Non-Accommodated (Number)	57,769	24,221	81,990
Non-Accommodated (Percent)	96.7	98.5	97.2
Accommodated (Number)	1,961	370	2,331
Accommodated (Percent)	3.3	1.5	2.8
IEP Students (Number)	10,463	4,978	15,441
Non-Accommodated (Number)	5,638	2,133	7,771
Non-Accommodated (Percent)	53.9	42.8	50.3
Accommodated (Number)	4,825	2,845	7,670
Accommodated (Percent)	46.1	57.2	49.7

Table 10–8L. Accommodation Rate for Non-IEP and IEP Students on the Spring 2021 Keystone Exams: Literature

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)	59,299	22,543	81,842
Non-Accommodated (Number)	54,966	22,136	77,102
Non-Accommodated (Percent)	92.7	98.2	94.2
Accommodated (Number)	4,333	407	4,740
Accommodated (Percent)	7.3	1.8	5.8
IEP Students (Number)	10,397	4,382	14,779
Non-Accommodated (Number)	5,538	2,429	7,967
Non-Accommodated (Percent)	53.3	55.4	53.9
Accommodated (Number)	4,859	1,953	6,812
Accommodated (Percent)	46.7	44.6	46.1

THE INCIDENCE OF ACCOMMODATIONS AND IEP AND EL STATUS

Students with an IEP typically receive an accommodation of some type far more often than non-IEP students. Certain accommodations with very low frequencies are specific to particular disabilities while others are far more common and may also apply to students classified as English Learners (EL). Because the accommodations with the largest frequencies can potentially supply the most stable data when separated out for subgroup analysis, those in most common use are typically selected for display in Tables 10–9A through 10–9L. The most frequently occurring accommodations for assessed students were:

- Some test items/questions read aloud (Algebra I and Biology only)
- All test items/questions read aloud (Algebra I and Biology only)
- Small group setting
- Extended time
- Frequent breaks

Coding for IEP is dichotomous, as students are classified IEP and non-IEP. For purposes of this analysis, an English Learner (EL) is an assessed student classified EL and enrolled in a U.S. school fewer than 12 cumulative months. All other assessed students, including those who have exited an ESL/bilingual program and are in the first or second year of monitoring, are regarded as non-EL.

Customarily, a considerably larger percentage of IEP students receive a given accommodation than non-IEP students. Although less frequent, certain accommodations also have a high frequency rate for EL students. To separate out the effect of being classified IEP or EL, four possible combinations are presented in Tables 10–9A through 10–9L. These include general education students (who are neither IEP nor EL), students who are IEP but non-EL, students who are EL but non-IEP, and students who are both IEP and EL. The bottom row for each administration mode provides the total number of assessed students in each of the four classifications.

For purposes of descriptively comparing the four IEP/EL subgroups with respect to whether a subgroup displayed a larger percentage rate than another subgroup, a choice was made to use a difference of five or more percentage points as a criterion for judging importance. In many instances, the percentage difference between subgroups was of little practical significance (from zero to only several percentage points).

Although the separate presentation of data for PPT and CBT modes provides an impression of overall findings, the much smaller n-counts and accommodation rate by students taking a CBT renders an administration mode comparison less meaningful. Nevertheless, it is possible to make some cautious observations when sufficient n-counts and consistency are present as noted in the summary of findings below. Please refer to Appendix I for winter and summer Keystone Exam results.

SUBGROUP COMPARISONS FOR PPT ADMINISTRATION MODE

Subgroup comparisons were regarded as viable for the PPT administration. There was little differentiation across subgroups for the two accommodations involving items/questions read aloud (Algebra I and Biology) and for frequent breaks (Algebra I, Biology, and Literature). Small group setting are typically the most prevalent accommodation for the IEP/non-EL subgroup followed by the IEP/EL and EL/non-IEP subgroups. This pattern was consistent across all three course exams. Another consistent pattern was observed for extended time, which was more prevalent for the IEP/non-EL, EL/non-IEP, and IEP/EL subgroups than for the General Education subgroup.

SUBGROUP COMPARISONS FOR CBT ADMINISTRATION MODE

For the CBT administration the EL/Non-IEP and IEP/EL subgroup n-counts were low in prior administrations. Consequently, only the General Education and IEP/non-EL subgroups had a sufficient sample size to support reasonable comparisons. A consistent pattern noted for all three course exams was the greater prevalence of small group setting, extended time, and frequent breaks by the IEP/non-EL subgroup than for the General Education subgroup.

COMPARISONS BETWEEN PPT AND CBT

The only subgroups for which comparisons between PPT and CBT administration modes were deemed reasonable based on sample size were within the General Education and IEP/non-EL subgroups. The findings are summarized below.

- The General Education subgroup displayed a very low incidence of accommodations, typically less than one percent, in nearly all instances for both PPT and CBT administrations. The accommodation students mostly received is extended time.
- For the IEP/non-EL subgroup, small group setting was the only accommodation for which PPT administration consistently exceeded CBT by more than five percentage points in all three course exams. The instances in which students tested by CBT exceeded those responding by PPT were extended time and frequent breaks.

Table 10–9A. Incidence of IEP and EL Students Receiving Selected Accommodations on the Spring 2021 Keystone Exam: Algebra I

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Some test items/questions read aloud (Number)	3	32	32	394
PPT - Some test items/questions read aloud (Percent)	1	1.2	0	3.5
PPT - All test items/questions read aloud (Number)	5	2	19	308
PPT - All test items/questions read aloud (Percent)	1.7	.1	0	2.7
PPT - Small group setting (Number)	80	271	602	4,579
PPT - Small group setting (Percent)	26.8	10.2	.9	40.8
PPT - Extended time (Number)	22	212	4,649	1,086
PPT - Extended time (Percent)	7.4	8	7.2	9.7
PPT - Frequent breaks (Number)	7	6	47	327
PPT - Frequent breaks (Percent)	2.3	.2	.1	2.9
PPT - Number assessed (Number)	299	2,646	64,366	11,234
CBT - Some test items/questions read aloud (Number)	9	2	6	310
CBT - Some test items/questions read aloud (Percent)	7.8	.3	0	6.9
CBT - All test items/questions read aloud (Number)	7	2	8	313
CBT - All test items/questions read aloud (Percent)	6.1	.3	0	6.9
CBT - Small group setting (Number)	36	29	109	1,964
CBT - Small group setting (Percent)	31.3	3.9	.4	43.5
CBT - Extended time (Number)	24	41	900	1,123
CBT - Extended time (Percent)	20.9	5.6	3.7	24.9
CBT - Frequent breaks (Number)	10	0	30	457
CBT - Frequent breaks (Percent)	8.7	0	.1	10.1
CBT - Number assessed (Number)	115	737	24,254	4,519
Total - Some test items/questions read aloud (Number)	12	34	38	704
Total - Some test items/questions read aloud (Percent)	2.9	1	0	4.5
Total - All test items/questions read aloud (Number)	12	4	27	621
Total - All test items/questions read aloud (Percent)	2.9	.1	0	3.9
Total - Small group setting (Number)	116	300	711	6,543
Total - Small group setting (Percent)	28	8.9	.8	41.5
Total - Extended time (Number)	46	253	5,549	2,209
Total - Extended time (Percent)	11.1	7.5	6.3	14
Total - Frequent breaks (Number)	17	6	77	784

Table 10–9A (continued). Incidence of IEP and EL Students Receiving Selected Accommodations on the Spring 2021 Keystone Exam: Algebra I

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
Total - Frequent breaks (Percent)	4.1	.2	.1	5
Total - Number assessed (Number)	414	3,383	88,620	15,753

Table 10–9B. Incidence of IEP and EL Students Receiving Selected Accommodations on the Spring 2021 Keystone Exam: Biology

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Some test items/questions read aloud (Number)	4	13	20	317
PPT - Some test items/questions read aloud (Percent)	1.6	.6	0	3.1
PPT - All test items/questions read aloud (Number)	4	1	6	369
PPT - All test items/questions read aloud (Percent)	1.6	0	0	3.6
PPT - Small group setting (Number)	71	214	564	4,452
PPT - Small group setting (Percent)	28.1	9.9	1	43.6
PPT - Extended time (Number)	16	81	967	648
PPT - Extended time (Percent)	6.3	3.7	1.7	6.3
PPT - Frequent breaks (Number)	4	1	41	268
PPT - Frequent breaks (Percent)	1.6	0	.1	2.6
PPT - Number assessed (Number)	253	2,163	57,567	10,210
CBT - Some test items/questions read aloud (Number)	8	0	10	386
CBT - Some test items/questions read aloud (Percent)	6.4	0	0	8
CBT - All test items/questions read aloud (Number)	13	2	8	340
CBT - All test items/questions read aloud (Percent)	10.4	.3	0	7
CBT - Small group setting (Number)	48	18	130	2,274
CBT - Small group setting (Percent)	38.4	2.8	.5	46.9
CBT - Extended time (Number)	28	22	214	1,269
CBT - Extended time (Percent)	22.4	3.4	.9	26.1
CBT - Frequent breaks (Number)	14	0	37	550
CBT - Frequent breaks (Percent)	11.2	0	.2	11.3
CBT - Number assessed (Number)	125	646	23,945	4,853
Total - Some test items/questions read aloud (Number)	12	13	30	703
Total - Some test items/questions read aloud (Percent)	3.2	.5	0	4.7
Total - All test items/questions read aloud (Number)	17	3	14	709
Total - All test items/questions read aloud (Percent)	4.5	.1	0	4.7
Total - Small group setting (Number)	119	232	694	6,726
Total - Small group setting (Percent)	31.5	8.3	.9	44.7
Total - Extended time (Number)	44	103	1,181	1,917
Total - Extended time (Percent)	11.6	3.7	1.4	12.7
Total - Frequent breaks (Number)	18	1	78	818
Total - Frequent breaks (Percent)	4.8	0	.1	5.4
Total - Number assessed (Number)	378	2,809	81,512	15,063

Table 10–9L. Incidence of IEP and EL Students Receiving Selected Accommodations on the Spring 2021 Keystone Exam: Literature

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Small group setting (Number)	59	127	594	4,410
PPT - Small group setting (Percent)	23.9	7	1	43.4
PPT - Extended time (Number)	17	118	3,433	1,007
PPT - Extended time (Percent)	6.9	6.5	6	9.9
PPT - Frequent breaks (Number)	3	6	38	271
PPT - Frequent breaks (Percent)	1.2	.3	.1	2.7
PPT - Number assessed (Number)	247	1,826	57,473	10,150
CBT - Small group setting (Number)	39	12	113	1,784
CBT - Small group setting (Percent)	34.8	2	.5	41.8
CBT - Extended time (Number)	20	16	289	994
CBT - Extended time (Percent)	17.9	2.6	1.3	23.3
CBT - Frequent breaks (Number)	5	0	23	411
CBT - Frequent breaks (Percent)	4.5	0	.1	9.6
CBT - Number assessed (Number)	112	613	21,930	4,270
Total - Small group setting (Number)	98	139	707	6,194
Total - Small group setting (Percent)	27.3	5.7	.9	43
Total - Extended time (Number)	37	134	3,722	2,001
Total - Extended time (Percent)	10.3	5.5	4.7	13.9
Total - Frequent breaks (Number)	8	6	61	682
Total - Frequent breaks (Percent)	2.2	.2	.1	4.7
Total - Number assessed (Number)	359	2,439	79,403	14,420

GLOSSARY OF ACCOMMODATION TERMS

Table 10–10 provides a brief description of accommodation terms as used in the PSSA and Keystone Exams. Accommodation data was supplied by school personnel as noted in the left column of the table. The right column contains an explanation derived from the PDE publication, *Accommodations Guidelines*. This manual may be found on the [PDE website](http://www.education.pa.gov) at www.education.pa.gov.

Table 10–10. Glossary of Accommodation Terms as Applied in the PSSA and 2021 Keystone Exams

Type of Testing Accommodation	Explanation
Student used the following Online Presentation Accommodations	
Braille format	Students may use a Braille format of the test. Answers must then be transcribed into the answer booklet without alteration.
Large print format	Students with visual impairments may use a large print format. Answers must then be transcribed into the answer booklet without alteration.
Magnification device	Devices to magnify print may be used for students with visual impairments and/or print disabilities.
Color overlay	Students with visual impairments may place a color overlay on a printed page of the test document to make text more readable.
Computer assistive technology (e.g., electronic screen reader) (PDE approval required)	Students with severe visual disabilities that prevent them from accessing instructional material or performing the skill may use computer assistive technology; however, PDE must approve the program and functions prior to the test window.
Test items/questions/text-dependent analysis signed	Deaf/hearing impaired students may receive test directions from a qualified interpreter. Signing is also permitted for PSSA ELA writing section multiple-choice items, and text-dependent analysis questions and all items in PSSA mathematics and science and for Keystone Algebra and Biology.
Test items/questions/text-dependent analysis interpreted for EL	A qualified interpreter may translate directions or clarify instructions for the assessments. The interpreter may translate but not define specific words or test questions on the PSSA mathematics, science, ELA writing section multiple-choice items, and text-dependent analysis questions and Keystone Algebra and Biology exams.
Some or all test items/questions/text-dependent analysis read aloud	Students unable to decode text visually may have items/questions read aloud for PSSA ELA writing section multiple-choice items, and text-dependent analysis questions and all items in PSSA mathematics and science and for Keystone Algebra and Biology; however, words may not be defined.
Amplification device	In addition to using hearing aids, an amplification device to enhance clarity may be required.
Other (PDE approval required)	Other presentation accommodations indicated in the <i>Accommodations Guidelines</i> may be provided; however, PDE approval is required prior to the test window.
Spanish version for PSSA (Math and Science) and Keystone (Algebra and Biology)	Students whose first language is Spanish and who have been enrolled in U.S. schools for fewer than three years may take this version.
Student used the following Online Presentation Accommodations	
Audio	The online test form reads permissible test directions and items for a student unable to decode text. The accommodation must be marked within the test engine system. The accommodation is available on PSSA mathematics, science, ELA writing section multiple-choice items, and text-dependent analysis questions and Keystone Algebra and Biology exams.
Video sign language (per <i>Accommodations Guidelines</i>)	Eligible students who use a sign language accommodation during instructional periods may use a VSL on the PSSA mathematics and science assessments, or Keystone Algebra I and Biology.
Color chooser or contrasting text chooser	The use of this accommodation enables a visually impaired student to change the background color or text color to make text more readable.
Refreshable Braille	This accommodation allows students to use a screen reader to produce a Braille translation output.

Table 10–10 (continued). Glossary of Accommodation Terms as Applied in the PSSA and 2021 Keystone Exams

Type of Testing Accommodation	Explanation
Student used the following Response Accommodations	
Braille/Note taker (per <i>Accommodations Guidelines</i>)	Students using this device as part of their regular instructional program may use it on the assessments; however, without thesaurus, spelling, or grammar checker.
Test administrator scribed open-ended responses at student’s direction	A test administrator may record word-for-word exactly what a student dictated directly into the test booklet. This includes MC and OE responses Keystone Algebra, Biology, and Literature tests and PSSA mathematics and science.
Test administrator marked multiple-choice responses at student’s direction	A test administrator may mark an answer booklet at the direction of a student (e.g., a student may point to an MC answer with the test administrator marking the response in the answer booklet).
Test administrator transcribed student responses (per <i>Accommodations Guidelines</i>)	A test administrator may transcribe (copy) a student’s written, typed, or keyed response into a standard answer booklet.
Qualified Interpreter translated, transcribed, and/or scribed student’s signed responses	A qualified interpreter may interpret a student’s signed responses into written English for Keystone Algebra and Biology exams, and PSSA mathematics and science assessments. Interpreters are not permitted to make corrections or change the meaning of the response.
Qualified Interpreter translated, transcribed, and/or scribed EL student responses	A qualified interpreter may interpret a student’s non-English oral responses into written English for Keystone Algebra and Biology exams, and PSSA mathematics and science assessments. Interpreters are not permitted to make corrections or change the meaning of the response.
Augmentative communication device	Students with severe communication difficulties may use a special device to convey responses, which must be transcribed into the answer booklet by the test administrator.
Keyboard, word processor, or computer (per <i>Accommodations Guidelines</i>)	This is an allowable accommodation as a typing function only for students with the identified need. Online test should be considered for students who prefer/need to type open-ended responses. Supports such as dictionaries, thesauri, spell checkers, and grammar checkers must be turned off. Answers must then be transcribed into the answer booklet without alteration.
Translation dictionary for EL student	A word-to-word dictionary that translates native language to English (or vice versa) without word definitions or pictures is allowed on any portion of the Keystone Algebra and Biology exams, and PSSA mathematics and science tests.
Computer assistive technology (e.g., electronic screen reader) (PDE approval required)	Students with blindness or extremely low vision may use dictate text into a computer. Responses must be transcribed verbatim into student’s regular answer booklet.
Other (per <i>Accommodations Guidelines</i> or PDE approval)	Other accommodations may be appropriate and available if they do not compromise the integrity of the assessment. Documentation must be provided to PDE.
Student used the following Setting Accommodations	
Hospital/home testing	A student who is confined to a hospital or to home during the testing window may be tested in that environment.
One-on-one setting	One-on-one settings are necessitated in certain instances, such as to reduce distraction or in the use of certain devices. A separate room may be used to reduce distraction.
Small group setting	Some students may require a test setting with fewer students or a setting apart from all other students to minimize distraction.
Other (per <i>Accommodations Guidelines</i> or PDE approval)	Other accommodations may be appropriate and available if they do not compromise the integrity of the assessment. Documentation must be provided to PDE.

Table 10–10 (continued). Glossary of Accommodation Terms as Applied in the PSSA and 2021 Keystone Exams

Type of Testing Accommodation	Explanation
Student used the following Timing Accommodations	
Extended time	Extended time may be allotted for each section of the test as a planned accommodation to enable students to finish.
Frequent breaks	Frequent breaks (breaks within a test section) may be scheduled for the completion of each test section; however, a test section must be completed within one school day.
Changed test schedule	Students whose disabilities prevent them from following a regular, planned test schedule may follow an individual schedule that enables test completion.
Other (per <i>Accommodations Guidelines</i> or PDE approval)	Other accommodations may be appropriate and available if they do not compromise the integrity of the assessment. Documentation must be provided to PDE.

CHAPTER ELEVEN: CLASSICAL ITEM STATISTICS

This chapter provides an overview of the two most familiar item-level statistics obtained from classical (traditional) item analysis: item difficulty and item discrimination. The following results pertain not only to the operational Keystone Exams items but also to the embedded field test items. Other statistics such as Rasch item statistics and test-level statistics are discussed in Chapter Twelve and Chapter Seventeen, respectively.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

ITEM-LEVEL STATISTICS

Appendix J provides classical item statistics for all items (i.e., operational and embedded field test items) in the Algebra I, Biology, and Literature Keystone Exams. Results are organized by administration and content area. These statistics represent the item characteristics most often used to determine whether an item functioned properly and/or how a group of students performed on a particular item. The item statistics in Appendix J include N , the number of students taking the test form for which there are valid test scores; p -values (denoted as P_{Val}) for multiple-choice (MC) items and item means (denoted as Mean) for constructed-response (CR) items (indicators of item difficulty); proportions of students who chose each response option for MC items (denoted as $P(A)$, $P(B)$, $P(C)$, and $P(D)$) or gained each score point for CR items (denoted as $P(0)$, $P(1)$, $P(2)$, $P(3)$, and $P(4)$); proportions of students who omitted an item (denoted as $P(-)$ for MC items and $P(B)$ for CR items); item-total correlations (denoted as Total, indicators of item discrimination); item-total correlations for each response option for MC items (denoted as $PT(A)$, $PT(B)$, $PT(C)$, and $PT(D)$) and gained score point for CR items (denoted as $PT(1)$, $PT(2)$, $PT(3)$, and $PT(4)$).

Appendix J also provides the Rasch measurement-based statistics in columns Rasch, Infit, and Outfit. Detailed explanations of these statistics can be found in Chapter Twelve. The differential item functioning (DIF) analysis on the embedded field test items is provided as well. The detailed explanation of DIF codes can be found in Chapter Five.

ITEM DIFFICULTY

At the most general level, an item's difficulty is indicated by its mean score in some specified group (e.g., grade level).

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students (n). For MC items, student scores are represented by 0's and 1's (0 = wrong, 1 = right). With 0/1 scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. So, this is also the *proportion correct* for the item, or as it is better known, the p -value. In theory, p -values can range from 0.00¹ to 1.00 on the proportion-correct scale. For example, if an item has a p -value of 0.89, it means 89 percent of the students answered the item correctly. Additionally, this value might also suggest that the item is relatively easy and/or the students who attempted the item are relatively high achievers. In other words, item difficulty and student ability are somewhat confounded.

For CR items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score (e.g., four points in the case of Algebra I CR items and three points in the case of Biology and Literature CR items). Sometimes a *pseudo p*-value is provided for a CR item by dividing the mean item score by the maximum possible item score.

¹ For MC items with four response options, pure random guessing would lead to an expected p -value of 0.25.

The minimum and maximum extremes of the difficulty scale are virtually never seen in applied practice. However, understanding what those values are helps illustrate that relatively lower values correspond to more difficult items and that relatively higher values correspond to easier items. (Because of this, some assert that this index would be better referred to as the item's *easiness*.)

Item difficulty is an important consideration for the Keystone Exams because of the various student achievement levels in Pennsylvania (Below Basic, Basic, Proficient, and Advanced). Items that are either very hard or very easy provide little information about student differences in achievement. However, an item answered correctly by a high percentage of students would suggest that the knowledge or skill the item taps has been mastered by most students. Conversely, an item answered correctly by a low percentage of students would suggest that few students have mastered the knowledge or skill the item taps. So, on a criteria-referenced test like the Keystone Exams, a test development goal is to include a wide range of item difficulties.

Utilizing the proportion of students who chose each option can be helpful for verifying keys. For example, if a large proportion of students chose a distractor instead of the correct answer of an MC item, it may indicate that the key is not correct. Proportion of students omitting or not reaching an item is useful for identifying issues related to testing time and item/test layout. Keystone Exams are not speed tests. Therefore, students should have enough time to take the exams. An omit rate greater than 5% for a single item could be an indication that students were not given enough time to take the test or an indication of an item/test layout problem. For example, some students might accidentally skip an item that follows a lengthy stem.

ITEM DISCRIMINATION

At the most general level, item discrimination² indicates an item's ability to differentiate between high and low achievers. It is expected that students with high ability (i.e., those who perform well on the Keystone Exams overall) would be more likely to answer any given item correctly, while students with low ability (i.e., those who perform poorly on the Keystone Exams overall) would be more likely to answer the same item incorrectly. For the Keystone Exams, Pearson's product-moment correlation coefficient between item scores and test scores is used to indicate discrimination. As commonly practiced, Data Recognition Corporation (DRC) removes the item score from the total score so that the resulting correlations will not be spuriously high. The correlation coefficient can range from negative 1.0 to positive 1.0. If the aforementioned expectation is met (high-scoring students tend to get the item right while low-scoring students do not), the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., well above zero) indicating that the item is a good discriminator between high- and low-ability students.

Item-total correlation for each option is another indicator of an item's ability to differentiate between high and low achievers. It is expected that students with high ability would be less likely to choose any distractors, while students with low ability would be more likely to choose a distractor. In other words, the item-total correlations for the distractors are expected to be negative.

In summary, the correlation will be positive in value when the mean test score of the students answering the item correctly is higher than the mean test score of the students answering the item incorrectly.³ In other words, students who did well on the total test tended to do well on the item as well. However, an interaction can exist between item discrimination and item difficulty. Items answered correctly (or incorrectly) by a large proportion of examinees (i.e., items with extreme *p*-values) can have reduced power to discriminate and thus can have lower correlations.

Discrimination is an important consideration for the operational Keystone Exams because the use of more discriminating items on a test is associated with more reliable test scores. This in turn means that score estimates will be more precise (i.e., there will be smaller confidence intervals around the scores) and, perhaps more importantly, that more accurate performance level placements will be made. The issues of reliability, confidence intervals, and performance level classifications are further discussed in Chapter Eighteen.

² As noted earlier, the discrimination index for dichotomous MC items is typically referred to as the *point-biserial correlation coefficient*. For CR items, the term *item-test correlation* is sometimes used.

³ It is legitimate to view the point-biserial correlation as a standardized mean. A positive value indicates that students who chose that response had a higher mean score than the average score; a negative value indicates that students who chose that response had a lower mean score than the average score.

SCATTER PLOTS OF ITEM DISCRIMINATION AND DIFFICULTY

Figure 11–1 contains a series of scatter plots showing item discrimination (i.e., item-total correlation on y -axis) on the item difficulty (i.e., p -value on x -axis) for the operational items on each exam by test administration. The summer plots are not included due to the elongated spring testing window. These plots provide information about item discrimination and difficulty in a single visual image for each Keystone Exam. This is because the x - and y -axes visually represent many important distributional indices:

- The minimum and maximum values are listed.
- Mean and median scores are indicated by the red dash lines.
- The first and third quartile (Q1 and Q3) are indicated by the red lines.
- Marginal histogram indicates the density of the individual data points.

It should be noted that pseudo p -values are used for CR items in these plots. Of course, the bivariate relationship between discrimination and difficulty is also presented. One does not usually expect any type of trend here. However, as noted earlier, it is often the case that items with extreme difficulties can have lower discrimination values, so this can be revealed in such a plot.

Figure 11–1. Scatter Plots of Item Discrimination and Difficulty

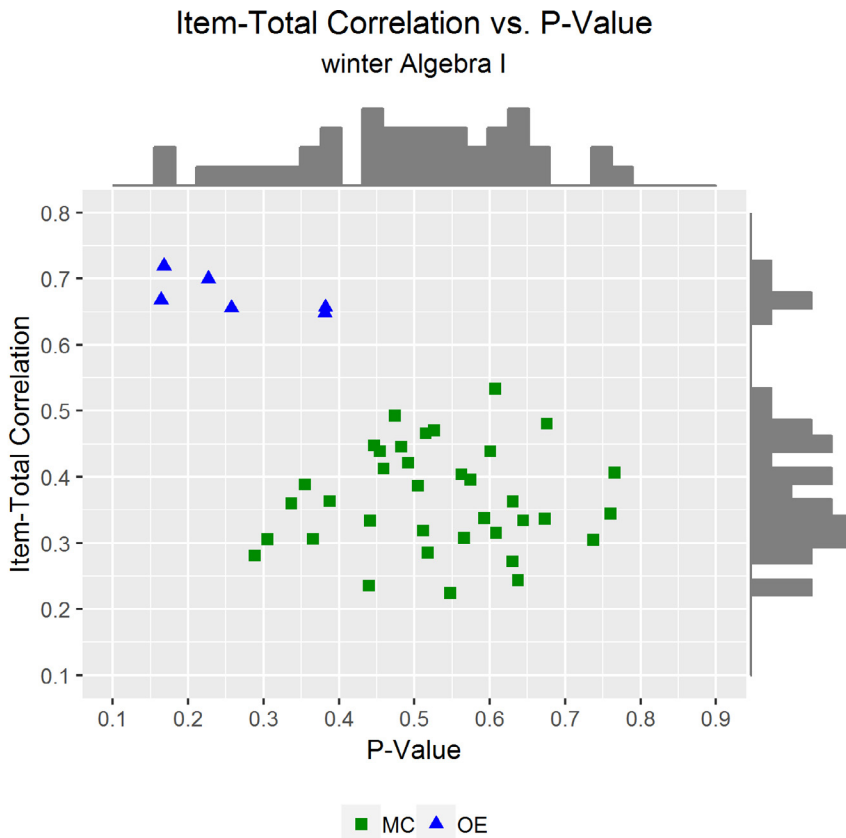
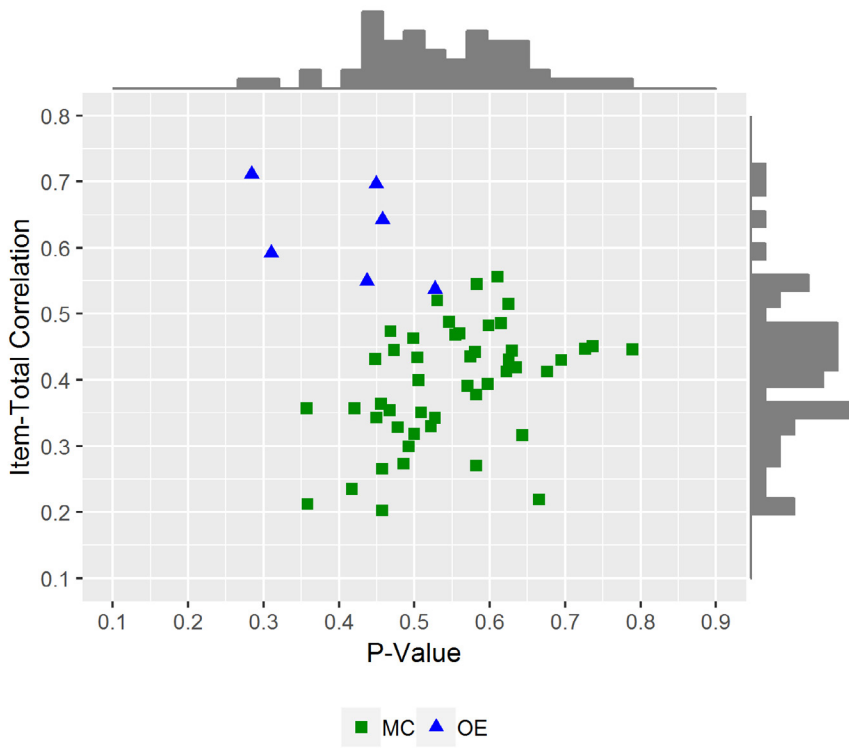


Figure 11–1 (continued). Scatter Plots of Item Discrimination and Difficulty

Item-Total Correlation vs. P-Value

winter Biology



Item-Total Correlation vs. P-Value

winter Literature

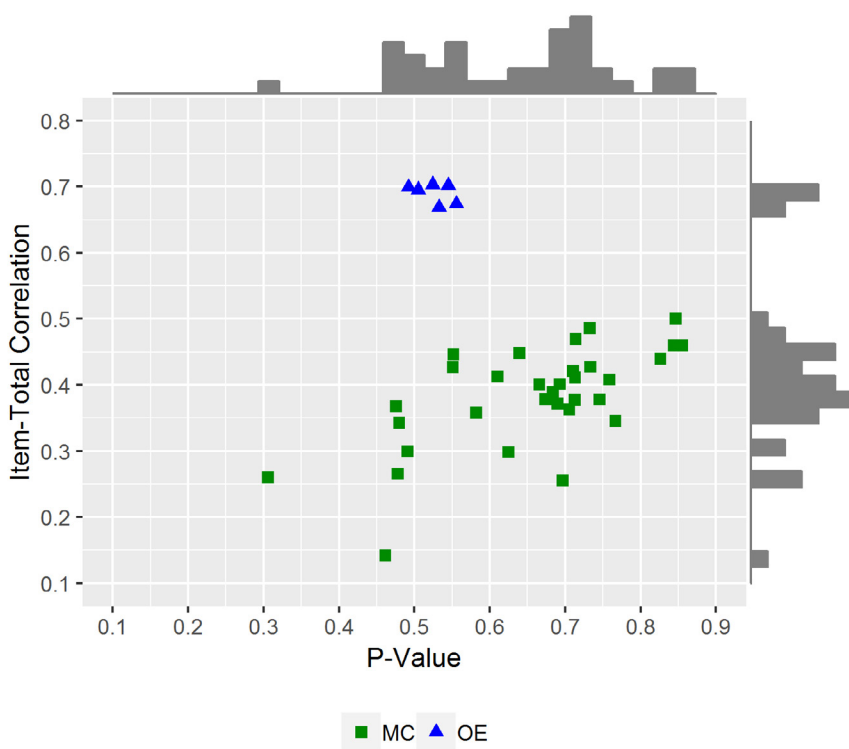
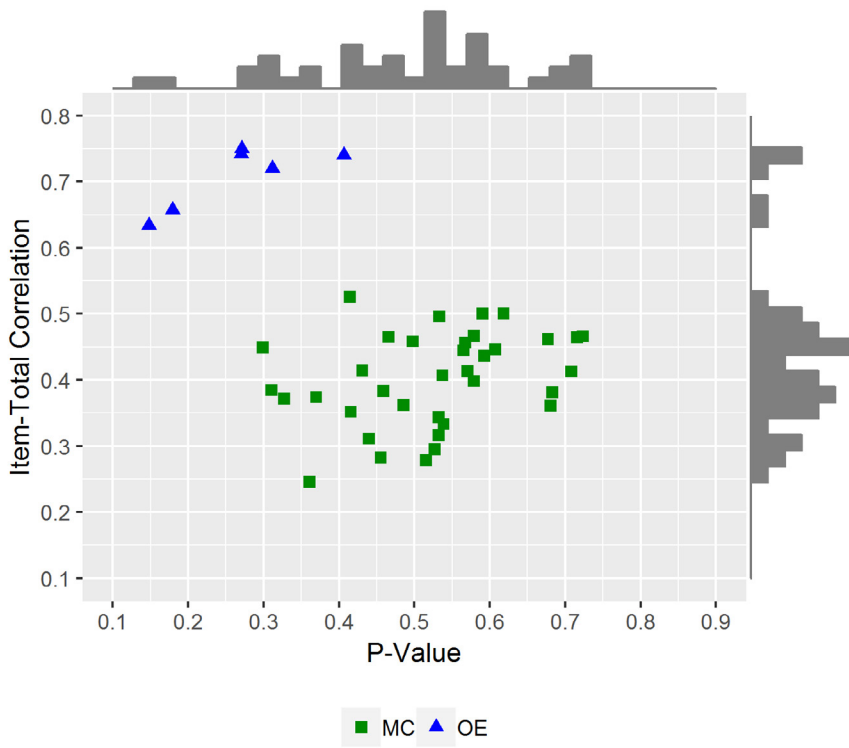


Figure 11–1 (continued). Scatter Plots of Item Discrimination and Difficulty

Item-Total Correlation vs. P-Value

Spring Algebra I



Item-Total Correlation vs. P-Value

Spring Biology

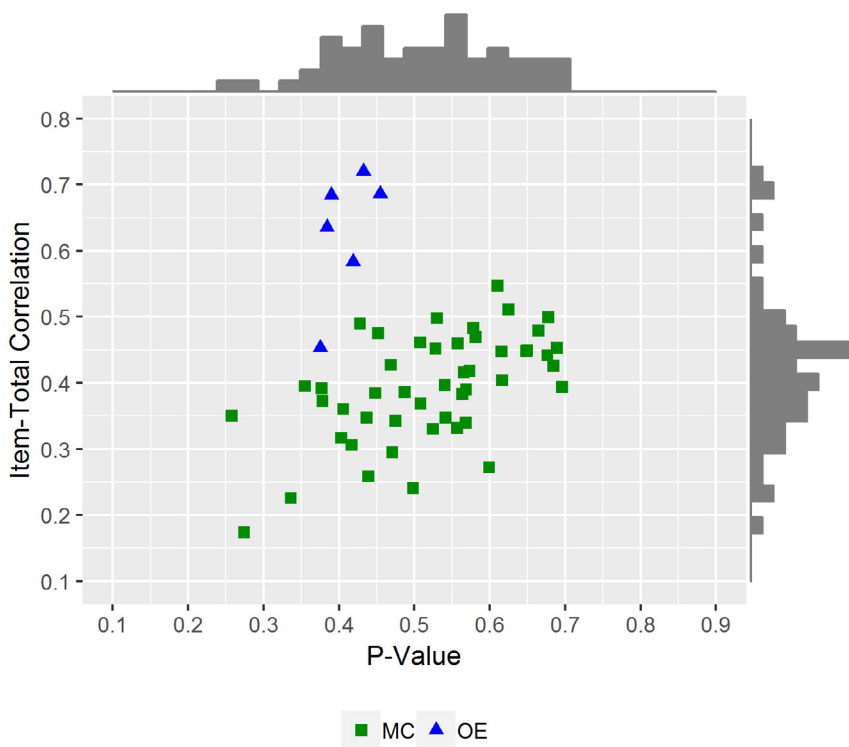
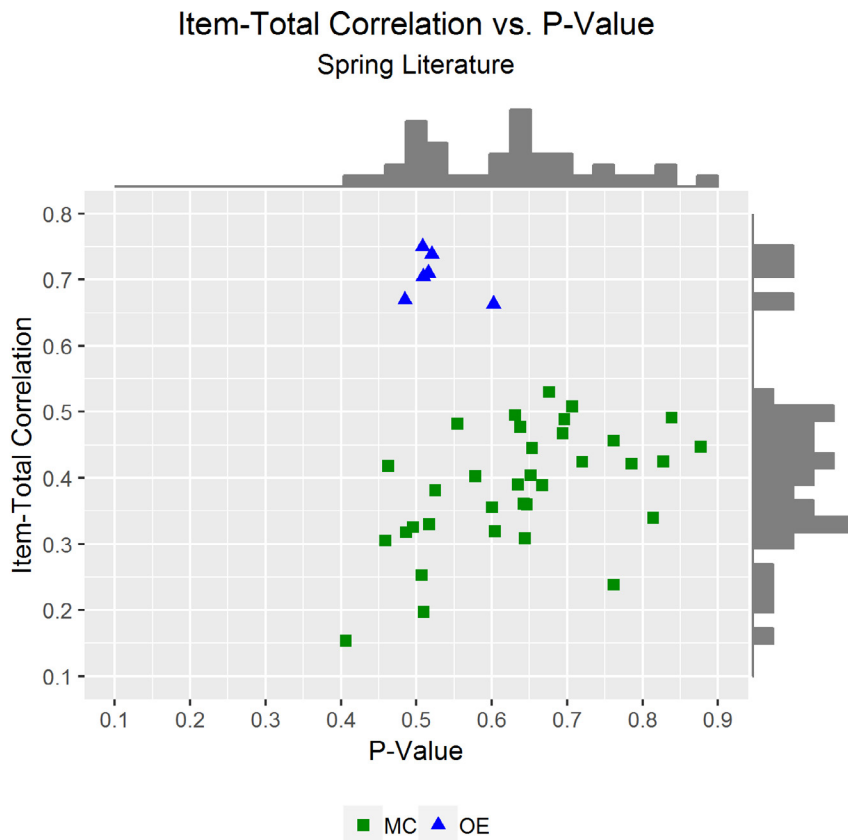


Figure 11–1 (continued). Scatter Plots of Item Discrimination and Difficulty



OBSERVATIONS AND INTERPRETATIONS

Table 11–1 provides the mean and median p -values and median⁴ item-total correlations for the operational MC and CR items in each content area. The mean p -value for the operational MC items ranged from 0.53 to 0.67 with standard deviation (SD) ranging from 0.10 to 0.14, whereas the mean p -values for the CR items ranged from about 0.26 to 0.54 with standard deviation ranging from 0.02 to 0.10. The median item-test correlations ranged from 0.35 to 0.41 and 0.65 to 0.72 for the MC and CR items, respectively. The CR correlations tended to be higher than the MC correlations, which is not surprising because the CR items include more score points and tend to be better at discriminating between low and high achieving students than MC items.

It is impossible to make global conclusions about the overall test quality from these item statistics alone. With that caveat in mind, the results presented in this chapter indicate that the item difficulties and discriminations were within expected and acceptable ranges.

⁴ Given that the value of the item-total correlation coefficient is not a linear function of the magnitude of the relation between the item and total test scores, the median instead of the mean of the item-total correlation was calculated for this statistic.

Table 11–1. Mean and Median Statistics for Operational MC and CR Items

Administration	Content Area	MC Items Mean <i>p</i> -Value	MC Items SD <i>p</i> -Value	MC Items Median <i>p</i> -Value	MC Items Median I-T Corr.	CR Items Mean <i>p</i> -Value	CR Items SD <i>p</i> -Value	CR Items Median <i>p</i> -Value	CR Items Median I-T Corr.
Winter	Algebra I	0.53	0.12	0.52	0.35	0.26	0.10	0.24	0.66
Winter	Biology	0.55	0.10	0.55	0.41	0.41	0.09	0.44	0.62
Winter	Literature	0.67	=0.14	0.69	0.39	0.53	0.02	0.53	0.70
Spring	Algebra I	0.55	0.12	0.55	0.40	0.28	0.09	0.29	0.72
Spring	Biology	0.54	0.11	0.56	0.40	0.44	0.04	0.43	0.65
Spring	Literature	0.65	0.12	0.66	0.40	0.54	0.04	0.53	0.69
Summer	Algebra I								
Summer	Biology								
Summer	Literature								

Note. I-T Corr. is the item-total test score correlation; SD represents the standard deviation.

Note. Summer 2021 data is not available due to the elongated spring testing window.

CHAPTER TWELVE: RASCH ITEM CALIBRATION

The particular item response theory (IRT) model used for the Keystone Exams is based on the work of Georg Rasch. Rasch models have had a long-standing presence in applied testing programs and have been the methodology continually used to calibrate the Pennsylvania System of School Assessment (PSSA) items in recent history. Consequently, this model was also chosen for the Keystone Exams. IRT has several advantages over classical test theory, so it has become the standard procedure for analyzing item response data in large-scale assessments. However, IRT models make several strong assumptions related to dimensionality, local item independence, and model-data fit. Resulting inferences derived from any application of IRT rest strongly on the degree to which the underlying assumptions are met.

This chapter outlines the procedures used for calibrating the operational Keystone Exams items. Generally, item calibration is the process of assigning a difficulty-parameter estimate to each item on an assessment so that they are placed on a common scale. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes the Rasch item statistics for the Keystone Exams in Algebra I, Biology, and Literature.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

DESCRIPTION OF THE RASCH MODEL

The Rasch partial credit model (RPCM; Wright & Masters, 1982) was used to calibrate Keystone Exams item response data because both multiple-choice (MC) and constructed-response (CR) items were part of the assessment. The RPCM extends the Rasch model (Rasch, 1960) for dichotomous (0, 1) items to accommodate polytomous CR items. Under the RPCM, for a given item i with m_i score categories, the probability of person n scoring x ($x = 0, 1, 2, \dots, m_i$) is given by:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^x (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})}, \quad x = 0, 1, \dots, m_i$$

where β_n represents a student's proficiency (ability) level, and δ_{ij} is the step difficulty of the j th step on item i . For dichotomous MC items, the RPCM reduces to the standard Rasch model and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person n getting item i correct as follows:

$$\Phi_{ni}(X = 1 | \beta_n) = \frac{\exp(\beta_n - \delta_{ij})}{1 + \exp(\beta_n - \delta_{ij})}.$$

The Rasch model places both student ability and item difficulty (estimated in terms of log-odds or logits) on the same continuum. When the model assumptions are met, it also provides person ability estimates that are independent of the items employed in the assessment, and, conversely, estimates item difficulty independently of the sample of examinees. (As noted in Chapter Eleven, interpretation of item p -values confounds item difficulty and student ability.)

SOFTWARE AND ESTIMATION ALGORITHM

Item calibration was implemented via WINSTEPS computer program (Linacre & Wright, 2013), which employs unconditional (UCON), joint-maximum-likelihood estimation (JMLE).

SAMPLE CHARACTERISTICS

The characteristics of calibration samples are reported in Chapter Nine. These samples only include the students who attempted the tests. All omitted and multiple responses (more than one response selected) for MC items were scored as incorrect answers (coded as 0s) for calibration purposes.

CHECKING RASCH ASSUMPTIONS

Because the Rasch model was the basis of all calibration and equating analyses associated with the Keystone Exams, the validity of the inferences from these results depends on the degree to which the assumptions of the model are met and how well the test data fits the model. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and model-data fit at the item level. Though a variety of methods are available for assessing these issues, the Rasch analyses and criteria available from WINSTEPS were used here. It should be noted that only operational items were analyzed since they are the basis of student scores.

Given Keystone Exams use a pre-equating design (see details in Chapter Fifteen), calibrations with and without anchoring all the item parameter estimates were conducted to check the stability of item parameters. After reviewing the analyses results for the winter, spring, and summer administrations, a decision was made to use the item difficulty estimated from the field-test data to generate the raw-to-scaled score conversion tables. In this chapter, the adequacy of the Rasch calibration assumptions was checked with all the item difficulties anchored to the pre-equated values.

UNIDIMENSIONALITY

Rasch models assume that one dominant dimension determines the difference in students' performances. WINSTEPS provides results from a principal components analysis (PCA) that can be used to assess the unidimensionality assumption. Different from standard applications of PCA, WINSTEPS conducts its PCA on the response residuals, not the original observations. That is, the primary dimension from the Rasch model is removed first and then the residual variance is analyzed. The purpose of the analysis is to verify whether any other dominant components exist among the residuals (i.e., they account for a practically significant amount of residual variance). If any other dimensions are found, the unidimensionality assumption would be violated.

For Keystone Exams, the standardized residuals were used to conduct the PCA because simulation studies indicate that it gives the most accurate reflection of secondary dimensions in the items (Linacre, 2013). Table 12–1 presents the PCA results by administration for each content area. The results include the eigenvalues and variance explained by each component. As can be seen from the table, the eigenvalues for the first component are much larger than those for the rest of the components. The first component explained about 31.0 to 48.6 percent of the total variance. The rest of the components explained only a small percentage of variance (less than 4.6%). These results suggest that each of the Keystone Exams essentially measure a single dominant dimension.

Table 12–1. Results from PCA of Residuals in WINSTEPS

Administration/ Content Area	Component	Eigenvalue	Variance Explained
Winter Algebra I	1	36.1	46.2%
Winter Algebra I	2	1.9	2.4%
Winter Algebra I	3	1.4	1.7%
Winter Algebra I	4	1.3	1.6%
Winter Algebra I	5	1.2	1.5%
Winter Biology	1	25.4	32.0%
Winter Biology	2	1.7	2.2%
Winter Biology	3	1.7	2.1%
Winter Biology	4	1.4	1.8%
Winter Biology	5	1.3	1.6%
Winter Literature	1	24.5	38.0%
Winter Literature	2	2.6	4.0%
Winter Literature	3	1.3	2.0%
Winter Literature	4	1.3	2.0%
Winter Literature	5	1.3	2.0%
Spring Algebra I	1	39.7	48.6%
Spring Algebra I	2	1.8	2.2%
Spring Algebra I	3	1.4	1.7%
Spring Algebra I	4	1.2	1.5%
Spring Algebra I	5	1.2	1.5%
Spring Biology	1	24.2	31.0%
Spring Biology	2	1.7	2.1%
Spring Biology	3	1.5	1.9%
Spring Biology	4	1.3	1.6%
Spring Biology	5	1.2	1.5%
Spring Literature	1	21.4	34.9%
Spring Literature	2	2.8	4.6%
Spring Literature	3	1.5	2.5%
Spring Literature	4	1.2	2.0%
Spring Literature	5	1.2	1.9%
Summer Algebra I	1		
Summer Algebra I	2		
Summer Algebra I	3		
Summer Algebra I	4		
Summer Algebra I	5		
Summer Biology	1		
Summer Biology	2		
Summer Biology	3		
Summer Biology	4		
Summer Biology	5		

Table 12–1 (continued). Results from PCA of Residuals in WINSTEPS

Administration/ Content Area	Component	Eigenvalue	Variance Explained
Summer Literature	1		
Summer Literature	2		
Summer Literature	3		
Summer Literature	4		
Summer Literature	5		

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

LOCAL INDEPENDENCE

Local independence (LI) is a fundamental assumption of IRT. No relationship should exist between examinees' responses to different items after accounting for the abilities measured by a test. In formal statistical terms, a test X that is composed of items X_1, X_2, \dots, X_I is locally independent with respect to the latent variable δ if, for all $x = (x_1, x_2, \dots, x_I)$ and δ ,

$$P_n(\mathbf{X} = \mathbf{x} | \delta_n) = \prod_{i=1}^I P(X_i = x_i | \delta_n).$$

This formula essentially states that the probability of any pattern of responses across all items (x), after conditioning on the abilities (δ) measured by the test, should be equal to the product of the conditional probabilities across each item (cf. the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of local independence. A *weak form* of local independence (WLI) was proposed by McDonald (1979). The distinction is important as many indicators of local dependency are actually framed by WLI. The requirement here would be for the conditional covariances of all pairs of item responses, conditioned on the abilities, to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on abilities, is the product of the probabilities of responses to these two items, as shown below. (This is a *weaker* form because higher-order dependencies among items are allowed.) Based on the WLI, the following expression can be derived:

$$P(X_i = x_i, X_j = x_j | \delta_n) = P(X_i = x_i | \delta_n)P(X_j = x_j | \delta_n).$$

Marais and Andrich (2008) pointed out that local item dependence in the Rasch model can occur in two ways that some may not distinguish. The first way occurs when the assumption of unidimensionality is violated. Here, other nuisance dimensions besides a dominant dimension also determine students' performance (this can be called *trait dependence*). The second violation occurs when responses to an item depend on responses to another. This is a violation of statistical independence and can be called *response dependence*. Many people treat the assumptions of *unidimensionality* and *local independence* as one phenomenon and believe that once unidimensionality holds, that local independence also holds. By distinguishing the two sources of local dependence, one can see that while local independence can be related to unidimensionality, the two are different assumptions, and therefore, require different tests.

Residual item correlations estimated in WINSTEPS for each item pair were used to assess the local dependence among the Keystone Exams items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using ability and item parameter estimates. Next, deviations (residuals) between the examinees' expected and observed performance is determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Two types of residual correlations are available in WINSTEPS: raw and standardized residuals. It should be noted that the raw score residual correlation essentially corresponds to Yen's Q_3 index (Yen, 1993), a popular LI statistic. The expected value for the Q_3 statistic is approximately $-1/(k-1)$ when no local dependence exists, where k is test length. Thus, the expected Q_3 values should be approximately -0.026 or larger for the Keystone Exams (since Literature is the shortest test with 40 items). Index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997). Since the two residual correlations are very similar, the default *standardized residual correlation* in WINSTEPS was used for these analyses. Table 12–2 shows the summary statistics—mean, SD, minimum (Min), maximum (Max), and several percentiles (P_{10} , P_{25} , P_{50} , P_{75} , P_{90})—for all the residual correlations for each test. The total number of item pairs (N) and the number of pairs with residual correlations greater than 0.20 are also reported in this table. The mean residual correlations were slightly negative and the values were -0.02 after rounding. The vast majority of the correlations were very small, suggesting local item independence generally holds for the Keystone Exams in Algebra I, Biology, and Literature.

Table 12–2. Summary of Item Residual Correlations

Administration	Content Area	N	Mean	SD	Min	P10	P23	P50	P75	P90	Max	Stats >0.20
Winter	Algebra I	861	-0.02	0.03	-0.11	-0.06	-0.04	-0.02	-0.01	0.01	0.12	0
Winter	Biology	1431	-0.02	0.02	-0.10	-0.05	-0.03	-0.02	0.00	0.01	0.13	0
Winter	Literature	780	-0.02	0.05	-0.14	-0.09	-0.04	-0.02	0.00	0.02	0.26	6
Spring	Algebra I	861	-0.02	0.03	-0.11	-0.05	-0.04	-0.02	-0.01	0.01	0.21	1
Spring	Biology	1431	-0.02	0.02	-0.09	-0.04	-0.03	-0.02	-0.01	0.01	0.11	0
Spring	Literature	780	-0.02	0.05	-0.15	-0.09	-0.05	-0.02	0.00	0.01	0.29	8
Summer	Algebra I											
Summer	Biology											
Summer	Literature											

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

ITEM FIT

WINSTEPS estimates two item-fit statistics (infit and outfit) for evaluating the degree to which the Rasch model predicts the observed item responses. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (Z-standardized with mean = 0 and variance = 1). Mean-square values are more oriented toward practical significance, while Z-standardized values are more oriented toward statistical significance. Though both are informative, the Z-standardized values are very likely too sensitive to the large sample sizes observed on the Keystone Exams. In this situation it is recommended that the Z-standardized values be ignored if the mean-square values are acceptable (Linacre, 2009).

Both infit and outfit mean-square represent the average of standardized residual variance (the difference between the observed score and the Rasch estimated score divided by the square root of the Rasch model variance). The difference is that the outfit statistic gives all examinees equal weight in computing the fit and tends to be affected more by unexpected responses far from the person, item, or rating scale category measure (i.e., it is more sensitive to outlying, off-target, low-information responses). The infit statistic is weighted by the examinee locations relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., informative, on-target responses). Some feel that extreme infit values are a greater threat to the measurement process than extreme outfit values since most tests intend to measure the on-target population rather than extreme outliers.

The expected mean-square value is 1.0, and it can range from 0 to infinity. Deviation in excess of the expected value can be interpreted as noise or lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (too unpredictable, too much noise). Rules of thumb regarding practically significant mean-square values vary. More conservative users might prefer items with mean-square values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values

from 0.5 to 1.5. In the results below, values outside of 0.7 to 1.3 are given practical importance.

Table 12–3 presents the summary statistics of infit and outfit mean-square statistics for the Keystone Exams in Algebra I, Biology, and Literature, including the mean, SD, and minimum and maximum values. The number of items within the range of [0.7, 1.3] is also reported in Table 12–3. As can be seen, the mean values for both fit statistics were close to 1.00 for all the exams. Most of the items had mean-squared fit statistics ranging from 0.7 to 1.3.

Table 12–3. Summary of Infit and Outfit Mean-Square (MnSq) Statistics

Admin	Content Area	N	Infit MnSq Mean	Infit MnSq SD	Infit MnSq Min	Infit MnSq Max	Infit MnSq [0.7, 1.3]	Outfit MnSq Mean	Outfit MnSq SD	Outfit MnSq Min	Outfit MnSq Max	Outfit MnSq [0.7, 1.3]
Winter	Algebra I	42	0.99	0.10	0.74	1.22	42	1.00	0.14	0.67	1.30	40
Winter	Biology	54	1.00	0.11	0.79	1.23	54	1.00	0.17	0.68	1.38	48
Winter	Literature	40	0.95	0.16	0.60	1.30	37	0.97	0.24	0.49	1.49	30
Spring	Algebra I	42	0.99	0.12	0.69	1.25	41	1.01	0.17	0.70	1.37	39
Spring	Biology	54	1.01	0.11	0.79	1.37	53	1.02	0.16	0.76	1.41	51
Spring	Literature	40	1.01	0.16	0.60	1.31	38	1.03	0.23	0.59	1.56	34
Summer	Algebra I											
Summer	Biology											
Summer	Literature											

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

RASCH ITEM STATISTICS

As noted earlier, the Rasch model expresses item difficulty (and student ability) in units referred to as logits, rather than on the percent-correct metric. In the simplest case, a logit is a transformed p -value with the average p -value becoming a logit of zero. In this form, logits resemble z -scores or standard normal deviates; a very difficult item might have a logit of +4.0 and a very easy item might have a logit of –4.0. However, they have no formal relationship to the normal distribution.

The logit metric has several mathematical advantages over p -values. Logits have an interval scale, meaning that two items with logits of 0.0 and +1.0, respectively, are the same distance apart as two items with logits of +3.0 and +4.0. Logits are not dependent on the ability level of the students. For example, a test form can have a mean logit of zero, regardless of if the average item p -value for the student sample is 0.8 or 0.3.

The standard Rasch calibration procedure arbitrarily sets the mean difficulty of the items on any form at zero. Under normal circumstances where all students are administered the same set of items, any item with a p -value lower than the average item on the form receives a positive logit and any item with a p -value higher than the average receives a negative logit. Consequently, the logits for any calibration relate to an arbitrary origin defined by the center of items on that form. Logits for both item difficulties and student abilities are placed on the same scale and relate to the same mean item difficulty.

There are a number of other choices that could be made for centering the item difficulties. Rather than using all the items, the origin could be defined by content. For the Keystone Exams, all test forms in a particular content area share the same operational item set. All items on each form can then be easily adjusted to a single origin by defining the origin as the mean of the operational items. With this done, the origins for all the forms will be statistically equal. For example, items on any two forms that are equally difficult will now have statistically equal logit difficulties.

Appendix J reports the item statistics including classical and Rasch logit difficulties for all the operational items and the field-test items embedded in the spring forms¹. Table 12–4 summarizes the Rasch logit difficulties of the operational items on each test for each administration. The mean of MC item difficulty was no longer equal to zero as it was for the 2011 administration. This is because all the item parameter estimates were anchored to the pre-equated values. The mean item difficulties for MC items were less than that of CR items. Table 12–4 also shows the mean standard errors (SE) of the item difficulties, which were relatively small, suggesting that items were calibrated with very small errors. The minimum (Min) and maximum (Max) values and standard deviations (SD) suggest the Keystone Exams items covered a relatively wide range of difficulties.

Table 12–4. Summary of Rasch Item Difficulties

Administration/Content Area	Item Types	N	Mean Item Difficulty	Mean SE	SD Item Difficulty	Min Item Difficulty	Max Item Difficulty
Winter Algebra I	All	42	0.41	0.02	0.80	-1.15	2.62
Winter Algebra I	MC	36	0.17	0.02	0.53	-1.15	1.41
Winter Algebra I	CR	6	1.82	0.01	0.71	0.83	2.62
Winter Biology	All	54	0.27	0.02	0.54	-0.98	1.56
Winter Biology	MC	48	0.19	0.02	0.50	-0.98	1.17
Winter Biology	CR	6	0.86	0.01	0.51	0.21	1.56
Winter Literature	All	40	0.19	0.02	0.79	-1.88	1.84
Winter Literature	MC	34	0.07	0.02	0.80	-1.88	1.84
Winter Literature	CR	6	0.86	0.01	0.16	0.72	1.16
Spring Algebra I	All	42	0.39	0.01	0.85	-1.07	2.51
Spring Algebra I	MC	36	0.14	0.01	0.60	-1.07	1.46
Spring Algebra I	CR	6	1.84	0.00	0.62	0.87	2.51
Spring Biology	All	54	0.30	0.01	0.61	-0.97	1.59
Spring Biology	MC	48	0.24	0.01	0.61	-0.97	1.59
Spring Biology	CR	6	0.77	0.00	0.22	0.49	1.01
Spring Literature	All	40	0.30	0.01	0.75	-1.40	1.54
Spring Literature	MC	34	0.21	0.01	0.77	-1.40	1.54
Spring Literature	CR	6	0.83	0.01	0.23	0.44	1.12
Summer Algebra I	All	42					
Summer Algebra I	MC	36					
Summer Algebra I	CR	6					
Summer Biology	All	54					
Summer Biology	MC	48					
Summer Biology	CR	6					
Summer Literature	All	40					
Summer Literature	MC	34					
Summer Literature	CR	6					

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

¹ In Spring 2021, field-test constructed-response items were not scored. Therefore, these item statistics are not reported in Appendix J.

ITEM DIFFICULTY-STUDENT ABILITY MAP

The distributions of the Rasch item logits (item difficulty estimates) are shown on the item difficulty-student ability maps presented in Figure 12–1. In each item-student map, the top bar graph displays the student distribution on the logit scale, and the bottom displays markers of item difficulty parameter estimates. MC items are represented by a circle and CR items are represented by a square. As noted earlier, the Rasch model enables placement of both items and students on the same scale. Consequently, one can easily visualize information regarding the relationship between the distributions of item difficulty and student ability. The vertical red lines (from left to right) represent the performance cut-points: below basic/basic, basic/proficient, and proficient/advanced. On the top plot, the logit represents lower abilities (negative values) to higher abilities (positive values), whereas on the bottom plot the logit represents easier items (negative values) to harder items (positive values). To achieve precise measures of student ability, the student distribution should mirror the item distribution.

The common pattern seen across all maps was that the item difficulties were comparable to the student ability levels. It is also important to understand where the items are providing more-accurate measurement. This issue is addressed more fully in Chapter Eighteen (see Figure 18–2).

Note that summer figures are not presented due to the cancellation of summer Keystone Exams in 2021.

Figure 12–1. Item Difficulty-Student Ability Maps

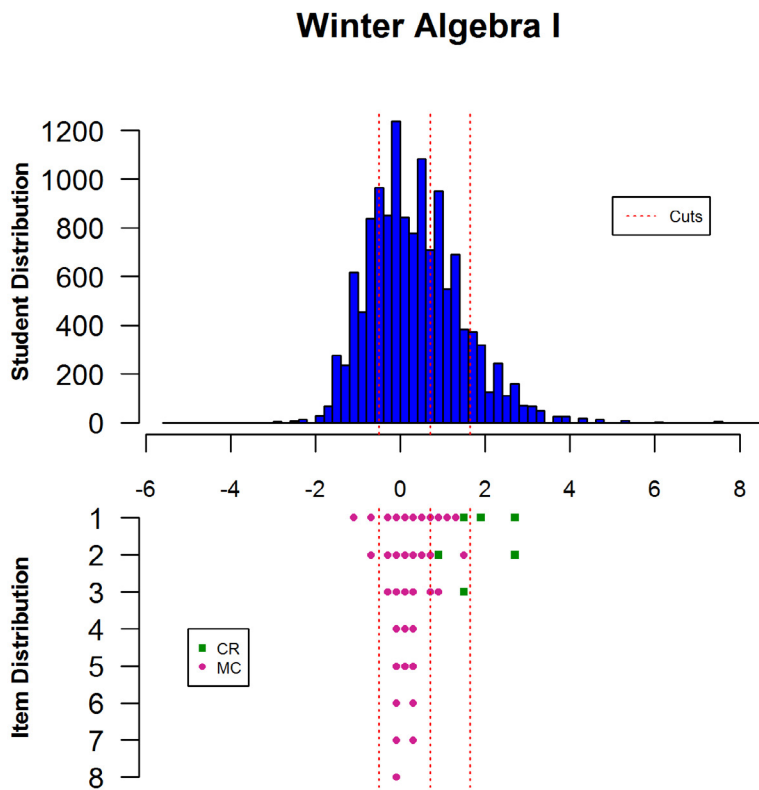
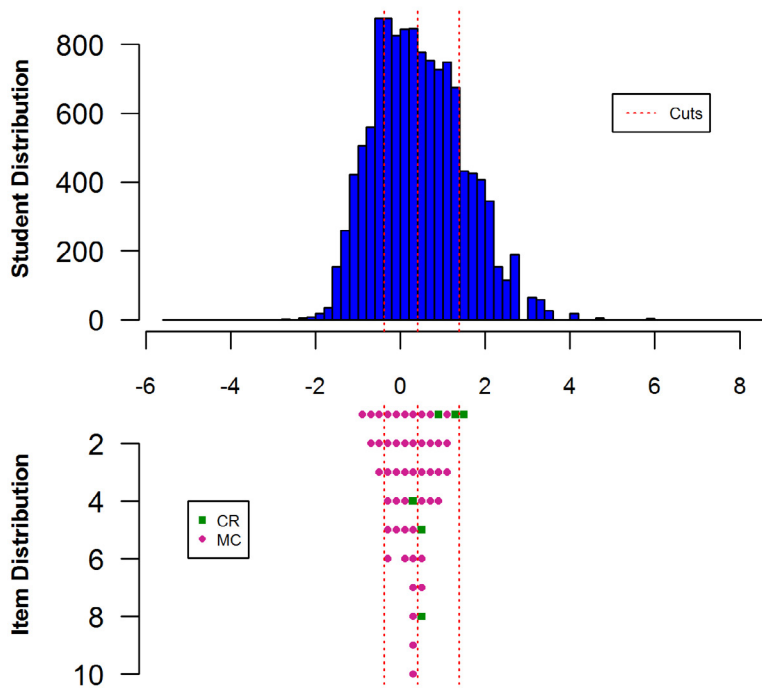


Figure 12–1 (continued). Item Difficulty-Student Ability Maps

Winter Biology



Winter Literature

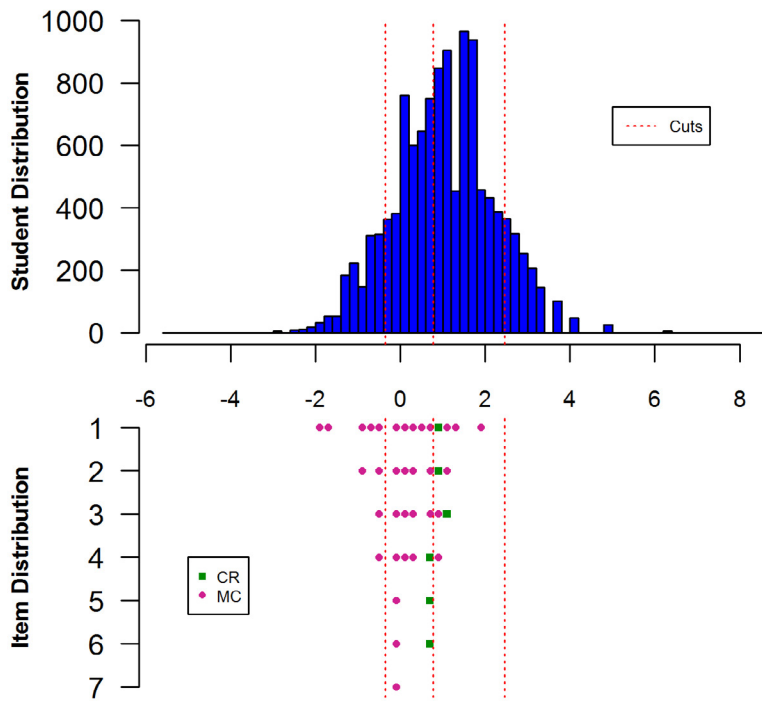
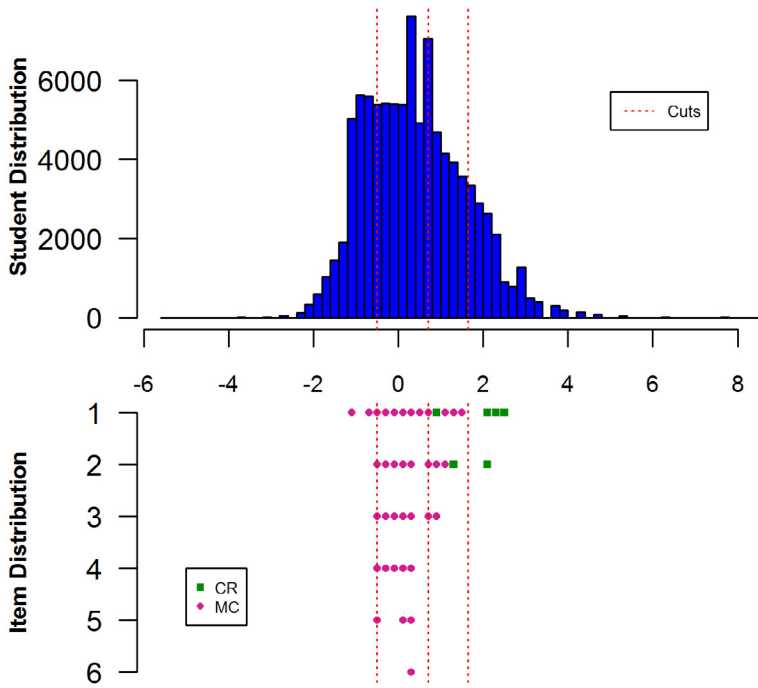


Figure 12–1 (continued). Item Difficulty-Student Ability Maps

Spring Algebra I



Spring Biology

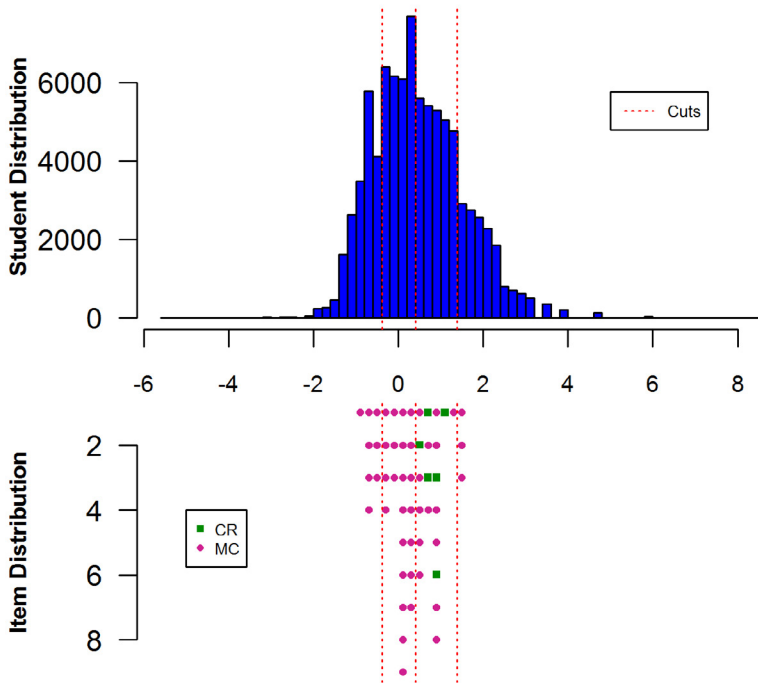
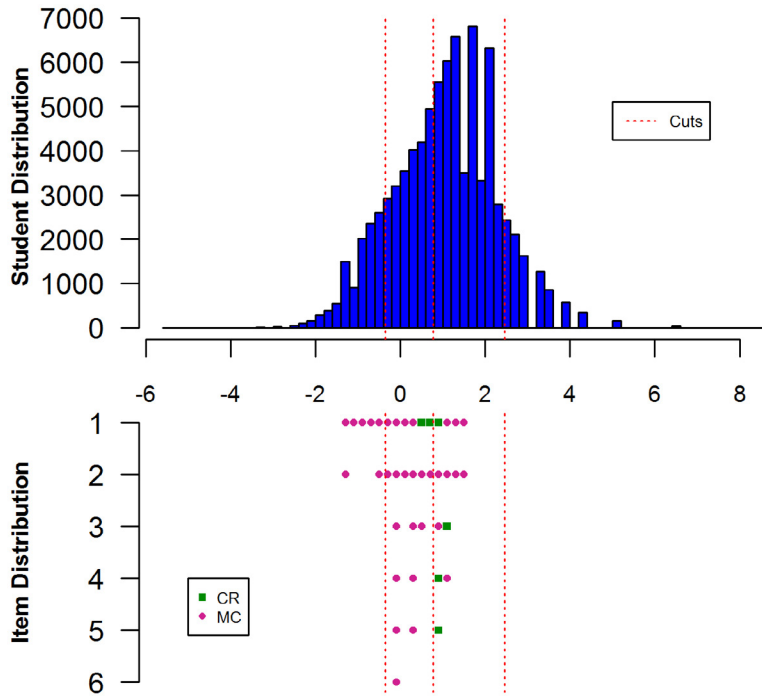


Figure 12–1 (continued). Item Difficulty-Student Ability Maps
Spring Literature



CHAPTER THIRTEEN: STANDARD SETTING

STANDARD SETTING AND PERFORMANCE LEVEL DESCRIPTORS

The Keystone Performance Level Descriptors (PLDs) are paragraphs that describe the knowledge and skills expected at different performance levels with respect to the content standards (Pennsylvania Keystone Exams Assessment Anchor Content Standards and Eligible Content) for each of the Keystone Exams. Descriptors must be clearly written to ensure that all stakeholders have a common understanding of what describes expected performance at the various levels (i.e., Below Basic, Basic, Proficient, and Advanced). PLDs were developed, reviewed, and finalized by the PDE/QRT¹ and committees of Pennsylvania educators as required by the Chapter 4 Regulations. After the development and final review by PDE/QRT and Pennsylvania educators, the descriptors were prepared for use during the standard setting workshop. During this meeting, the descriptors were used to guide the standard setting process for each of the Keystone Exams. They were instrumental to the validity and defensibility of the standard setting process.

The standard setting for the Algebra I, Biology, and Literature Keystone Exams was conducted by Data Recognition Corporation (DRC) using a Bookmark procedure (Lewis, Mitzel, & Green, 1996) during a workshop held in Harrisburg, Pennsylvania, June 23–24, 2011. After the standard setting event, the descriptors were finalized. Along with the recommended cut scores, final PLDs for each of the Keystone Exams were submitted to the Pennsylvania Board of Education for final approval.

Below is a summary of the process that was used to guide the development of the Keystone Exams PLDs and a summary of the methodology and results of the standard setting workshop. Additional details about the standard setting event can be found in the *Keystone Standard Setting Technical Report* (Pennsylvania Department of Education, 2011).

DEVELOPMENT OVERVIEW FOR THE PERFORMANCE LEVEL DESCRIPTORS

The Keystone Exams PLDs were developed by Pennsylvania educators during two meetings. The goal of the first meeting was to have Pennsylvania educators review the general Pennsylvania Policy Definitions that describe, at a high level, performance expected for each level and complete an in-depth analysis of the Keystone Exams Assessment Anchors and Eligible Content in order to create a bulleted list describing, in detail, what students are expected to know and be able to do at each performance level. The goal of the second meeting was to have committees of Pennsylvania educators review the Pennsylvania Policy Definitions again and draft general descriptors (paragraphs) that build upon and/or summarize the information from the bulleted lists of what students are expected to know and be able to do at each performance level.

Guiding documents were prepared for each meeting. The guiding documents included the following:

- PowerPoint training presentations
- Meeting agendas
- Assessment Anchors and Eligible Content documents
- Policy definitions
- Other relevant materials as needed to help guide the work of the committees

All meeting materials were submitted to PDE/QRT for review and approval before each Keystone Exams meeting following an agreed-upon development schedule. The following section provides specific information concerning each meeting.

¹ The PDE/QRT includes the representatives from the Pennsylvania Department of Education, members of the Quality Review Team, and/or others appointed by the Quality Review Team.

ROLE OF FACILITATORS AND OBSERVERS FOR THE MEETINGS

The role of the facilitators was to ensure that a fair and orderly consensus process was followed for each meeting, that the committee members' work was adequately documented, and that the process stayed on schedule. The facilitators developed the agenda, prepared all meeting materials such as the PowerPoint training presentations and the task-guiding documents, and provided the initial training on the development of the specific descriptors (meeting 1) and the general descriptors (meeting 2). PDE/QRT members supported the facilitation process and/or served as observers of the process.

The facilitators also served as a resource, answering questions pertaining to the content of the standards (Assessment Anchor Content Standards and Eligible Content) and the documents developed to guide the process. Facilitators also summarized the results of each meeting, finalized the results, and prepared the specific descriptors/bulleted lists (meeting 1) and the general descriptors (meeting 2) for PDE/QRT review and approval.

PERFORMANCE LEVEL DESCRIPTORS MEETING 1

CREATING SPECIFIC LISTS DESCRIBING WHAT STUDENTS SHOULD KNOW AND BE ABLE TO DO AT EACH PERFORMANCE LEVEL

The first PLD meeting for Algebra I, Biology, and Literature Exams was held May 18–19, 2010, in Harrisburg, Pennsylvania. The purpose of the first meeting was to guide Pennsylvania educators in understanding the Assessment Anchors and Eligible Content for Algebra I, Biology, and Literature for what the Commonwealth of Pennsylvania determined students should know and be able to do for a given Keystone Exam subject. Committee members applied this understanding to the development of a bulleted list of specific determinations as to the level of knowledge and skills deemed necessary for each performance level. The section below describes the process used in the first meeting.

TRAINING

Pennsylvania educators received general training on how to develop specific PLDs, including training on how to describe student performance in relation to the Keystone Exams Assessment Anchors and Eligible Content. The training also provided educators with a general overview of the Standards Aligned System (SAS) and the high-level plan for the Keystone Exams. Definitions of key terms (e.g., Assessment Anchor Content Standards, Eligible Content, Performance Level Descriptors) were provided along with information on the background and purpose of the Keystone Exams. Keystone Exams content-specific materials (e.g., Assessment Anchor Content Standards, Eligible Content, other guiding documents) were distributed. The PDE/QRT also provided information on the policy definitions for existing Pennsylvania assessments.

ANALYZING THE ASSESSMENT ANCHORS AND ELIGIBLE CONTENT AND THE GENERAL POLICY DEFINITIONS FOR PENNSYLVANIA ASSESSMENTS

Following the introductory training, educators were divided into groups according to each Keystone Exam. Each group focused specifically on the task at hand—developing the specific PLDs for a given Keystone Exam. Committee members were informed of the format of the specific descriptors (bulleted list) and the number of proposed performance levels for each Keystone Exam (Below Basic, Basic, Proficient, and Advanced). Committee members were then given time to familiarize themselves with the policy definitions and the Assessment Anchors and the Eligible Content for a given Keystone Exam. They were provided with PDE/QRT-approved guiding documents to facilitate the process. Beginning with Proficient, committee members were asked to draft, in bulleted-list format, each performance level for Basic, Proficient, and Advanced, making sure to consider the knowledge and skills required or deemed necessary for each performance level. Note: Educators were not asked to create a specific descriptor for Below Basic.

DRAFTING SPECIFIC DESCRIPTORS

Outlined below is the sequence of steps taken to develop specific descriptors. The sequence was not always followed exactly. For example, some steps occurred simultaneously; other steps were repeated as needed or reordered as necessary.

1. The committee began with the development of the bulleted list for Proficient to serve as a model for the work during the remainder of the development process. As a formative first task using the Assessment Anchors and Eligible Content and the Pennsylvania Policy Definition for Proficient, the committee was asked to discuss, deliberate, and reach consensus on its initial bulleted list of the knowledge and skills needed to be considered Proficient. During this process, members were encouraged to consult all available resources and guiding documents. Particular emphasis was placed on the alignment of the knowledge and skills necessary for Proficient performance with what students are expected to know and be able to do as defined by the Assessment Anchors and Eligible Content for Algebra I, Biology, and Literature.
2. Once the committee drafted a bulleted list of the knowledge and skills needed to describe Proficient performance based upon the Assessment Anchors and Eligible Content, a group discussion took place. In reviewing the bulleted list for Proficient, the educators were specifically asked to determine whether all members agreed that the list included the appropriate knowledge and skills from the Assessment Anchors and Eligible Content to describe the Proficient performance level and that all Assessment Anchors and Eligible Content were sufficiently addressed.
3. The results of the discussion were summarized, and suggested revisions were made. The summary feedback was presented to the committee for additional consideration. An open discussion followed. Committee consensus was reached.
4. Following development of the bulleted list of the knowledge and skills needed for the Proficient performance level as determined by the committee, the committee began the development of the bulleted lists describing the specific knowledge and skills needed for Basic and Advanced. To complete the task, the committee members followed the procedures analogous to those used to develop the specific bulleted list for the Proficient performance level. These procedures included, as a formative first task, the committee's use of the Assessment Anchors and Eligible Content and the Pennsylvania Policy Definitions (e.g., Basic, Advanced) to discuss, deliberate, and reach consensus on its initial bulleted list of the knowledge and skills needed for Basic and then Advanced. This order of development—Proficient first, followed by Basic and then Advanced—was followed throughout the remainder of the process.

Once the initial drafts of the bulleted lists for Basic, Proficient, and Advanced were developed, a group discussion took place. To guide the discussion, the following questions were used to evaluate each specific descriptor (bulleted list) for a given performance level (Basic, Proficient, Advanced):

- Is the description of the performance level appropriate? If not, what revisions need to be made?
 - Is the description of the specific Keystone Exam inappropriate because the list of knowledge and skills included in the description of the performance level is too demanding? If so, what revisions need to be made?
 - Is the description inappropriate because the knowledge and skills included in the description of the performance level is inconsistent with the expectation of the high standards as reflected in the Policy Definition? If so, what revisions need to be made?
 - Is the description inappropriate because the knowledge and skills included in the description of the performance level might be too easy? If so, what revisions need to be made?
5. The results of the discussion were summarized, and suggested revisions were listed. The summary feedback was presented to the committee for additional consideration. An open discussion followed. Depending upon the degree of concurrence, the facilitators proposed revisions based on the committee members' feedback to the specific descriptors (bulleted lists) for each descriptor. Committee consensus was reached.

6. Once consensus was reached, the bulleted lists or specific descriptions for each performance level were reviewed once again to confirm that all Assessment Anchors and Eligible Content were sufficiently addressed for each performance level and that the lists showed a clear progression from one performance level to the next. The results of the discussion were summarized, and suggested revisions were listed. The summary feedback was presented to the committee for additional consideration. An open discussion followed. Depending upon the degree of concurrence, the facilitators proposed revisions to the lists for each descriptor based on the committee members' feedback. Committee consensus was reached.
7. Following completion of the committee's work, the specific PLDs or bulleted lists of the knowledge and skills needed for each descriptor were collected. The bulleted lists were prepared for final review by the PDE/QRT. Upon approval by the PDE/QRT, the bulleted lists of the knowledge and skills describing each performance level were posted on the PDE website for additional review and feedback.

PERFORMANCE LEVEL DESCRIPTORS MEETING 2

CREATING GENERAL DESCRIPTIVE PARAGRAPHS DESCRIBING WHAT STUDENTS SHOULD KNOW AND BE ABLE TO DO AT EACH PERFORMANCE LEVEL

The second meeting for Algebra I, Biology, and Literature Exams took place in Harrisburg, Pennsylvania, on April 27–28, 2011. The second meeting built upon the work completed at the first meeting. The purpose of the second meeting was to guide the committee of Pennsylvania educators in developing general PLDs (paragraphs) for each of the performance levels (Basic, Proficient, and Advanced). These paragraphs were clearly written to ensure all stakeholders have a common understanding of what describes expected performance at the various levels. The paragraphs were not to be as specific as the bulleted lists but were to be aligned to the bulleted lists. In order to complete the task, the educators reviewed the Pennsylvania Policy Definitions for the performance levels.

Table 13–1. Pennsylvania Policy Definitions

Level	Description
Advanced	The Advanced Level reflects superior academic performance. Advanced work indicates an in-depth understanding and exemplary display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content.
Proficient	The Proficient Level reflects satisfactory academic performance. Proficient work indicates a solid understanding and adequate display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content.
Basic	The Basic Level reflects marginal academic performance. Basic work indicates a partial understanding and limited display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content. This work is approaching satisfactory performance, but has not been reached. There is a need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.
Below Basic	The Below Basic Level reflects inadequate academic performance. Below Basic work indicates little understanding and minimal display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.

The committee members then reviewed the specific bulleted list describing the knowledge and skills for Proficient based upon the Assessment Anchors and Eligible Content to determine whether the list of knowledge and skills provided in the bulleted list was still in alignment with the Policy Definition for Proficient. This review by the committee also included an in-depth analysis of the Assessment Anchors and Eligible Content. The section below describes, in detail, the process used in the second meeting.

TRAINING

Pennsylvania educators received general training on how to develop general descriptors (paragraphs) that describe performance at the various levels, including training on how to describe student performance in relation to the Keystone Exams Assessment Anchors and Eligible Content. The training also included providing Pennsylvania educators with a general overview of the SAS and the high-level plan for the Keystone Exams. Definitions of key terms (e.g., Assessment Anchor Content Standard, Eligible Content, specific and general Performance Level Descriptors) were provided along with information on the background and purpose of the Keystone Exams. A review of the Pennsylvania Policy Definitions was also included in the training, including a discussion of how the policy definition for Proficient relates to what it means to be Proficient on a given Keystone Exam. Content-specific materials (e.g., Policy Definitions, Assessment Anchor Content Standards and Eligible Content, specific descriptors

or bulleted lists from the first meeting, other guiding documents) were also distributed.

ANALYZING THE ASSESSMENT ANCHORS AND ELIGIBLE CONTENT AND THE POLICY DEFINITION FOR PROFICIENT

Following the introductory training, Pennsylvania educators were divided into groups according to Keystone Exam. Each group focused specifically on the task at hand—developing the general PLD paragraphs (Basic, Proficient, and Advanced) for a given Keystone Exam. To begin the process, educators reviewed the Pennsylvania Policy Definition for Proficient.

DRAFTING GENERAL DESCRIPTOR PARAGRAPHS

Once the committee reviewed the bulleted list for alignment to the Policy Definition for Proficient, committee members were asked to describe, in general terms, the knowledge and skills deemed necessary for each performance level (Basic, Proficient, and Advanced), beginning with Proficient. As a formative first task, committee members were instructed to refer to the bulleted list of the knowledge and skills required or deemed necessary for each performance level. Outlined below is the sequence of steps for the process used to develop the general PLD paragraphs. The sequence was not always followed exactly. For example, some steps occurred simultaneously; other steps were repeated as needed or reordered as necessary.

The committee began with the development of the general descriptor paragraph for the Proficient performance level. This general descriptor served as a model for the committee’s work during the remainder of the development process. Using the Assessment Anchors and Eligible Content, the specific descriptors (bulleted list) for Proficient, and the Pennsylvania Policy Definition for Proficient, the committee was asked to discuss, deliberate, and reach consensus on a written description of the knowledge and skills needed for Proficient. During the process, members were encouraged to consult all available resources and guiding documents. Particular emphasis was placed on the alignment of the knowledge and skills necessary for the Proficient performance descriptor to the Assessment Anchors and Eligible Content for the given Keystone Exam.

Note: In order to help guide educators in the development of the general descriptor paragraph for Proficient, samples of descriptor paragraphs for Algebra I, Biology, and Literature (e.g., Georgia, North Carolina) were provided. The committee members were encouraged to approach the task by noting how the sample general descriptors must provide the right words to define performance—having a balance between keeping the description of Proficient general enough yet not as specific as the bulleted list. Committee members were also encouraged not to focus too heavily upon style, grammar, and mechanics at this stage. In other words, committee members were not to serve as “wordsmiths.”

1. Once an initial draft paragraph summarizing the knowledge and skills needed to describe Proficient performance was developed, a group discussion took place. Committee members were asked to review the draft paragraph and determine whether the paragraph provided a clear description of what it means to be Proficient on a given Keystone Exam and the Policy Definition for Proficient. The goal of the discussion was to reach consensus.
2. Following development of the general paragraph describing Proficient, the committee began the development of the general paragraph describing Basic performance and the general paragraph describing Advanced performance. To complete the task, the committee members followed the procedures analogous to those used to develop the general paragraph describing Proficient performance on a given Keystone Exam. This process included, as a formative first task, using the Assessment Anchors and Eligible Content, the specific descriptors (bulleted lists), and the Pennsylvania Policy Definitions for a given level (e.g., Basic, Advanced) and discussing, deliberating, and reaching consensus on the knowledge and skills needed for Basic and the knowledge and skills needed for Advanced. This order of development—Proficient first, followed by Basic and then Advanced—was followed throughout the remainder of the process.
3. Once the initial draft paragraphs were developed for the other performance levels, a group discussion took place. In reviewing the state of development of the general PLD paragraphs at this stage, the committee members were asked to consider the following questions:
4. Does each paragraph clearly summarize the knowledge and skills required for a given performance level (Basic, Proficient, and Advanced)? If not, what revisions need to be made?

5. Does each paragraph provide for an appropriate description of the performance level? In other words, does each paragraph provide an overview or summary of the knowledge and skills appropriate for a given performance level? If not, what revisions need to be made?
6. Does any paragraph provide information that should not be included in the description of the performance level? If so, what revisions need to be made?
7. Is there information in any PLD paragraph that does not align well with the Pennsylvania Policy Definitions for a given performance level? If so, what revisions need to be made?
8. Do any paragraphs include information that might be inconsistent with the knowledge and skills defined by the Assessment Anchors and Eligible Content? If so, what revisions need to be made?
9. Does any paragraph include information describing performance that might be too demanding or too easy? If so, what revisions need to be made?
10. The results of the discussion were summarized, and revisions to each general PLD paragraph were made. Committee consensus was reached.
11. Once consensus was reached, the paragraphs describing performance at each level were reviewed again by the committee to confirm the following:
12. The PLD paragraphs show a clear progression from one performance level to the next level.
13. The PLD paragraphs are consistent with the Pennsylvania Policy Definitions.
14. The PLD paragraphs are aligned to the Assessment Anchors and Eligible Content.
15. The results of the discussion were summarized, suggested revisions were made, and committee members' feedback was incorporated into the paragraphs. Committee consensus was reached.
16. Following completion of the committee's work, the general PLD paragraphs were provided to PDE/QRT for final review and feedback. Upon approval by PDE/QRT, the general PLD paragraphs were used to guide the standard setting process.

STANDARD SETTING

A major purpose in the design of the standard setting workshop for the Keystone Exams is to establish procedures to set the performance cuts for the newly developed exams and, at the same time, adhere to the framework required by federal guidelines (USED, 2004) for setting performance levels. Federal guidelines (USED, 2004: Sect 2) specify that the setting of performance standards must involve the following elements:

- Formal adoption of performance categories that comprise at least three levels
- Pluralistic representation by education stakeholders, to include, for example, members of the public, school teachers and administrators, special education teachers, etc.
- Performance standards based primarily on expert judgment regarding content-based expectations of student achievement, but including the consideration of student impact data
- Descriptions of the competencies associated with each performance level

Accordingly, the standard setting workshop is designed to satisfy the following goals:

- A defensible and federally acceptable standard setting methodology that emphasizes a content-based approach for recommending the new performance standards
- The incorporation of PLDs developed by Pennsylvania educators into the standard setting process. (The larger goal around the incorporation of PLDs into the process is to help ensure the alignment of Pennsylvania's content standards to performance expectations as established by the recommended cut scores.)

The panelists were informed that the results from this meeting would be presented to the Board for review and possible adoption.

PANELIST RECRUITMENT

PDE selected committee members for the Algebra I, Biology, and Literature standard setting workshop mostly from members who participated in the May 2010 and April 2011 Performance Level Advisory committees. These committee members were selected as the starting pool because they represented the diversity of the Commonwealth of Pennsylvania, had a mix of teaching and committee experience, and, most importantly, were familiar with the PLDs of the Keystone Exams. From this list, PDE selected a subset of 25 members for Algebra I, 25 members for Biology, and 23 members for Literature to serve as eligible candidates. DRC, in collaboration with PDE and its Technical Advisory Committee (TAC), established a target of 15 to 20 participants for each of the Keystone Exams in Algebra I, Biology, and Literature.

Between March and June 2011, a great effort was made to recruit enough panelists to meet the target number of participants. In accordance with federal guidelines for representative committees and TAC's recommendation of recruiting a few committee members with higher education experience, the following background factors were applied in the recruitment decision:

- Gender
- Ethnicity
- Grade level and higher education experience
- Content expertise
- Geographic location
- Specializations
- Experience in developing state academic standards, state assessments, and other related activities

However, due to the unavailability of and the cancellation by some committee members, a total of 15, 13, and 11 panelists attended the standard setting workshop for Algebra I, Biology, and Literature, respectively. Table 13–2 contains the summary information about the characteristics of the selected panelists for each content area based on their self-reported responses to the Participant Survey. As can be seen from this table, there were committee members who considered themselves minority in the Algebra I and Literature groups. There were also committee members with administration and/or teaching experience in higher education, special education, and/or individualized education plan (IEP); those with experience working in different regions; and those with different lengths of teaching experience.

Table 13–2. Self-Reported Demographic Composition of Panelists by Content Area

Demographic Information	Algebra I Number	Algebra I Percent	Biology Number	Biology Percent	Literature Number	Literature Percent
Gender: Male	9	60.0%	5	38.5%	5	45.5%
Gender: Female	6	40.0%	8	61.5%	6	54.5%
Ethnicity: Asian	1	6.7%	0	0.0%	0	0.0%
Ethnicity: American Indian	0	0.0%	0	0.0%	0	0.0%
Ethnicity: Black	1	6.7%	0	0.0%	1	9.1%
Ethnicity: Latino	0	0.0%	0	0.0%	0	0.0%
Ethnicity: Multi-Race	0	0.0%	0	0.0%	0	0.0%
Ethnicity: White	13	86.7%	13	100.0%	10	90.9%
Role: Classroom Teacher	8	53.3%	9	69.2%	4	36.4%
Role: Educator	3	20.0%	0	0.0%	0	0.0%
Role: Higher Education Educator	3	20.0%	1	7.7%	4	36.4%
Role: Other	1	6.7%	3	23.1%	3	27.3%
Special Education: Yes	7	46.7%	7	53.8%	4	36.4%
Special Education: No	2	13.3%	4	30.8%	3	27.3%
Special Education: N/A	6	40.0%	2	15.4%	4	36.4%
LEP: Yes	4	26.7%	5	38.5%	2	18.2%
LEP: No	4	26.7%	6	46.2%	5	45.5%
LEP: N/A	6	40.0%	2	15.4%	4	36.4%
LEP: Missing	1	6.7%	0	0.0%	0	0.0%
Region: Urban	3	20.0%	2	15.4%	2	18.2%
Region: Suburban	7	46.7%	5	38.5%	5	45.5%
Region: Rural	4	26.7%	6	46.2%	3	27.3%
Region: Other	1	6.7%	0	0.0%	1	9.1%
Experience: Less than 10 years	0	0.0%	5	38.5%	0	0.0%
Experience: 10–20 years	4	26.7%	2	15.4%	4	36.4%
Experience: 20–30 years	8	53.3%	4	30.8%	3	27.3%
Experience: More than 30 years	3	20.0%	2	15.4%	4	36.4%

MATERIALS PREPARATION

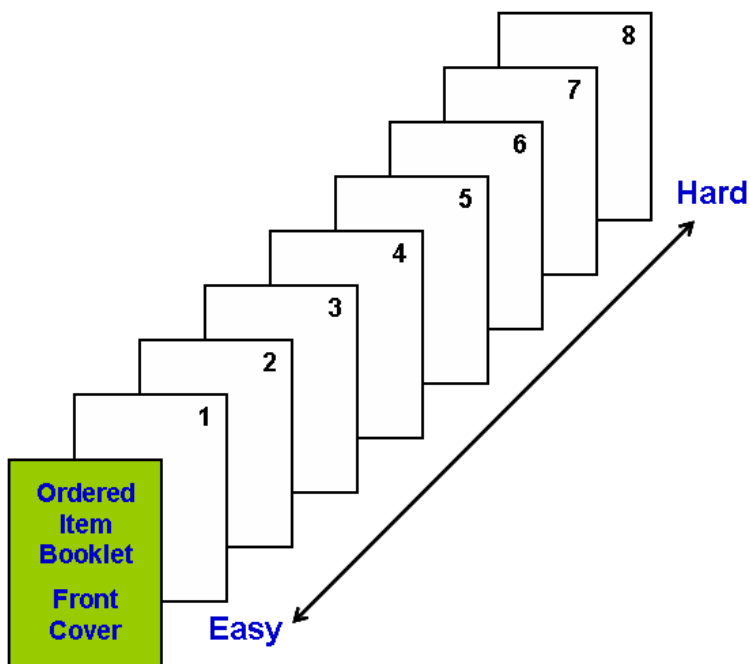
Workshop materials were developed and printed by DRC. The following is a list of materials that were available to panelists during the workshop:

- Item Map
- Item Separation Map
- Ordered Item Booklet (OIB)
- Passages
- Scoring Rubrics
- 2011 Operational Test Form
- PLDs
- Content Standards
- Participant Rating Form
- Participant Survey
- Readiness Form
- Evaluation Form
- Adhesive bookmarks, pens, highlighters, etc.

Item Map. The item map is a summary document displaying relevant information regarding each item. It contains the OIB page number, the original test sequence, item type, key, and content standard. The item map is ordered by difficulty in the same manner as the ordered item booklet. The item separation map is a graphical display of the relative difficulty of each item.

Ordered Item Booklet. The ordered item booklet is composed of all the operational items included in the test given to students in Spring 2011. Items are ordered from the easiest to the hardest. Each page contains an item and a page number. For constructed-response (CR) items, each score point with a sample response has a unique location in the OIB. A visual illustration of the OIB is provided in Figure 13–1.

Figure 13–1. Illustration of Ordered Item Booklet



To ensure there was no item difficulty gap for the items in an OIB, a few field test items were added to the OIBs. Table 13–3 shows the number of items supplemented into the OIBs by content area.

Table 13–3. Number of Score Points in OIB and Number of Items Supplemented

Exam	Number of Score Points in OIB	Number of Items Supplemented
Algebra I	63	3
Biology	69	3
Literature	54	2

Details of all other materials can be found in the *Keystone Standard Setting Technical Report* (Pennsylvania Department of Education, 2011).

DATA PREPARATION

In Bookmark standard setting (Lewis et al., 1996), the locations of items are typically rescaled to produce better alignment with the task of asking panelists what a student should know and be able to do. A probability of 0.67 is often used to find the corresponding item location during rescaling because this probability aligns better with the likelihood panelists use to make their judgment on whether a borderline student should answer the item correctly or receive a score point or higher. For Keystone Exams, the multiple-choice (MC) items were calibrated using the familiar form of the dichotomous Rasch model. The CR items were calibrated using another model in the Rasch family, Master’s partial-credit model (Wright & Masters, 1982). The latter model parameterizes each threshold needed to obtain the maximum score on the task. Consequently, there is one item difficulty parameter for each of the $n - 1$ score transitions (0/1, 1/2, etc.), or thresholds. Using the equated item parameters, the locations of items were rescaled to a response probability of 0.67 (i.e., $RP=0.67$). For MC items, the item locations were found by solving

$$\Phi_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

for the value of β_n that gives $\Phi_{ni} = 0.67$. Φ_{ni} is the probability that person n scores 1 on item i ; β_n is the ability of person n ; and δ_i is the difficulty of item i .

For CR items, the probability of person n scoring x on item i is

$$\pi_{nix} = \frac{\exp \sum_{j=0}^x (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})}, \quad x = 0, 1, \dots, m_i$$

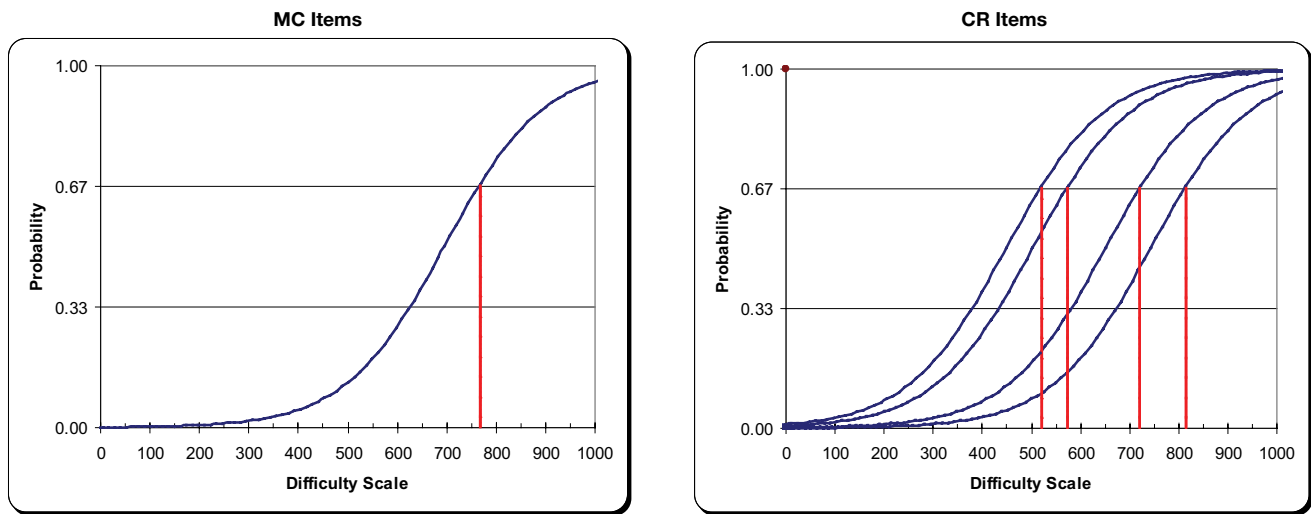
where m_i is the number of thresholds and, for notational convenience,

$$\exp \sum_{j=0}^0 (\beta_n - \delta_j) = 1.$$

This equation expresses the probability of person n scoring x on the m_i threshold of item i as a function of the person’s measure (β_n) and the threshold difficulties (δ_{ij}) of the m_i thresholds for item i . The observation x is a count of the successfully completed item thresholds. The item location for a score point is determined by finding the β_n for the person who has a 0.67 probability of earning this score point or higher.

The figure below shows how the difficulty values of MC items and score values for the CR items were treated in determining their respective OIB placements. For an MC item (left plot), the difficulty is the point on the scale at which the examinees have a 0.67 probability of answering the item correctly. For the CR item (right plot), the four illustrated values (e.g., on a 0–1000 scale) indicate where the examinees have a 0.67 probability of earning a particular score point or higher. The item difficulty for the MC item is 768, and the four threshold values for the CR items are 521, 575, 723, and 815. The value of 521 is the location on the scale where examinees have a 0.67 probability of earning a score of 1 or higher (i.e., 2, 3, or 4). The value of 575 is where examinees have a 0.67 probability of earning a score of 2 or higher (i.e., 3 or 4). The value of 723 is where examinees have a 0.67 probability of earning a score of 3 or 4.

Figure 13–2. Example of Obtaining Item Difficulties for MC and CR Items



TRAINING

The overall training was conducted the first morning of the workshop. Participants were informed that they were to

- be responsible for all secure materials,
- verify their individual placements for each round of judgments, and
- participate in a discussion as a large group.

Content-specific training was conducted after content area groups assembled in different rooms. These training materials included the following:

- Item Map
- Item Separation Map
- OIBs
- Training Rubrics and/or Passages
- PLDs
- Rating Form

Panelists were told that the process includes iterations (rounds) of individual judgments, group discussions, and opportunities to revise judgments. In addition, impacts were presented (percentage of students in each performance level) based on the large groups' results and external data.

BOOKMARK PROCEDURE

DRC utilized a Bookmark method to set the performance standards. Bookmark is one in a broad category of methods commonly referred to as item mapping that focus on items rather than examinees. To begin the process, participants were asked to visualize the knowledge and skills of a student who is at the borderline between two performance levels based on the PLDs. Thereafter, participants were given an ordered item booklet (with items ordered from easiest to most difficult) and asked to assess whether this borderline student has a reasonably high probability of answering each item correctly. "Reasonably high" was defined as 0.67. In addition, an item map was presented that contained the response key, the content objective, and the item sequence in the test booklets. An item separation map was also presented that showed the relative difficulty of each item. Panelists were given a rating form to record their individual placements for all performance levels in each round. Before each round, panelists were asked to fill out a readiness form in order to proceed.

Round 1. The Bookmark procedure proceeded in three rounds. Round 1 began following the review and discussion of PLDs facilitated by a DRC test development specialist. Participants then reviewed the OIBs independently. During this review, they were asked to determine what academic knowledge, skills, and competencies were required for a barely Proficient, Basic, or Advanced student to respond correctly to each successively more difficult item.

Training by the overall psychometric lead during the bookmark placement session emphasized the following points:

- The bookmark represents a judgment of the divide between items that a student at the borderline of a performance level should master and those that are not necessary to master.
- Bookmark placement should not be thought of as separating two items but rather two groups of items. In other words, a placement should not hinge on distinctions drawn for adjacent items without some compelling reason, such as a large gap in content difficulty.
- Students with a scaled score at a given cut score should have approximately a 0.67 probability of correctly responding to a MC item or receiving a certain score point and higher for a CR item at the cut score. These same students should have a higher probability of success on easier items (before the bookmark placement) and a lower probability of success on harder items (after the bookmark placement).

- While placing their bookmarks, panelists should consider what students should know and be able to do in the context of the skills implied by the PLDs and the item content.
- Panelists could start with placing the Basic/Proficient cut point, next the Below Basic/Basic cut point, and finally the Proficient/Advanced cut point.

Panelists were asked to record their bookmark placements on the rating form after they filled out a readiness form, which indicated they had completed the training and understood the standard setting process and their roles. Panelists' judgments were entered into a spreadsheet program. The median ratings of all panelists were calculated. The median placements were treated as the recommended cut scores. In addition, the standard errors associated with the recommended bookmark placements were calculated and associated impact data were determined.

Round 2. Round 2 started with a discussion of Round 1 results. The individual panelists' Round 1 bookmark placements, the median bookmark placements, and the percentage of students in each performance level were presented. Panelists were instructed to verify the ratings entered into the program as correct. A large-group discussion followed. The panelists compared their results with others by considering questions such as why they made their Round 1 placements at the locations where they did and what skills and knowledge were required to answer the items. After that, the impact data, based on the median bookmark placement from Round 1 (using the Spring 2011 operational test score distributions), were provided to help panelists frame the effects of their judgments. During Round 2 discussion, there was no attempt by the facilitators to reach consensus.

After Round 2 discussion, panelists were asked to make a second set of bookmark placements. Before they revised their Round 1 placements, they were asked to fill out the readiness form to make sure they understood how to adjust their placements (if they desired to do so) based on Round 1 information. The judgments were entered into the spreadsheet program to calculate the median cut scores for each table and the full panel. The associated impact data were also calculated.

Round 3. Round 3 began with a discussion of Round 2 results. The process followed in Round 2 was used. More specifically, the individual panelist's Round 2 bookmark placements, the median bookmark placements from Round 2, and the percentage of students in each performance level were presented. Panelists were instructed to verify the ratings entered into the program as correct. A table discussion followed. Panelists compared their results with others by considering questions such as why they made their Round 2 placements at the locations where they did and what skills and knowledge were required to answer the questions. The impact data, based on the median bookmark placement from Round 2, were provided to help panelists frame the effects of their judgments.

The Keystone Exams are one component of Pennsylvania's new system of high school graduation requirements. Because of the high-stakes consequences, the TAC strongly recommended bringing in external impact data to provide panelists with a reference outside of the Keystone Exams. The intent was to achieve reasonableness of results rather than to use the external data in a directive manner. DRC investigated Pennsylvania students' performance on the Pennsylvania System of School Assessment (PSSA), National Assessment of Educational Progress (NAEP), and Student Achievement Test (SAT) and presented external data as shown in Figures 13–3A to 13–3C before panelists made their Round 3 judgments. The panelists were informed of the following points:

- The PSSA and NAEP results were based on students' performance in 2009. The PSSA results were from grades 6–8 and 11. The NAEP results were from grade 8.
- All students in grades 6–8 and 11 in Pennsylvania took the PSSA. A sample that represents the Pennsylvania grade 8 students took the NAEP tests.
- The SAT results were based on the performance of students who took the SAT in 2010 or prior years.
- About 99% of students in the 2010 SAT data file indicated their expected graduation dates were in 2010; most of these students were in grade 11 in 2009. Therefore, the 2010 SAT data and the 2009 PSSA data were matched.
- Based on the matched sample, it was found that students with higher PSSA scores were more likely to take the SAT. To represent the full population in terms of demographics and PSSA scores, the matched sample was weighted by students' demographics and PSSA scores when calculating the impacts.

Figure 13–3A. External Impact Data: Algebra I

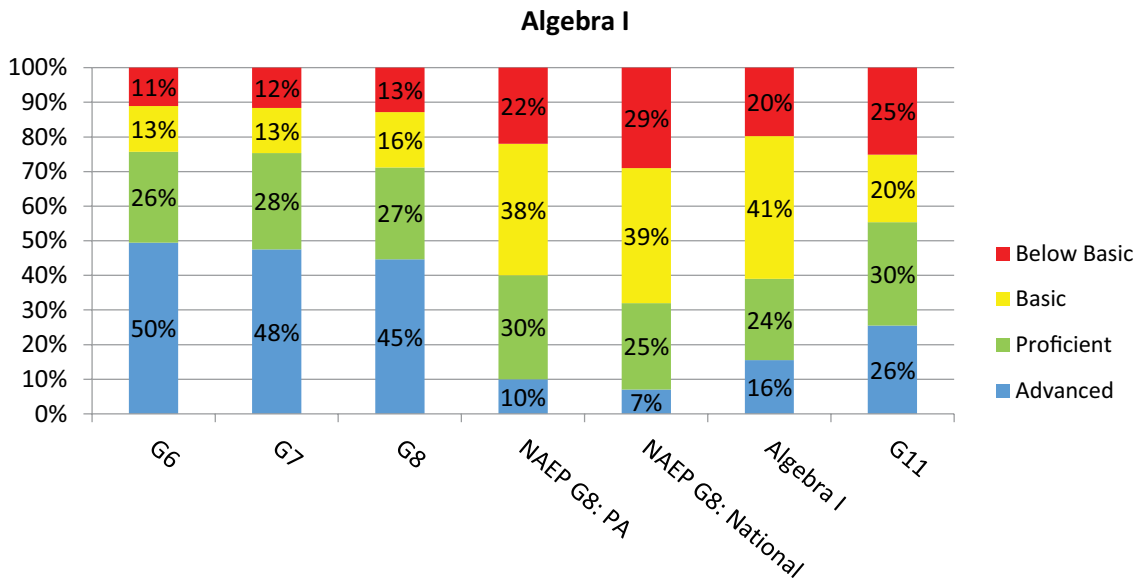


Table 13–4A. Number of Score Points in OIB and Number of Items Supplemented: Algebra I

Performance Level	PSSA G6	PSSA G7	PSSA G8	NAEP G8: PA	NAEP G8: National	KE Alg. I	PSSA G11	College Ready Yes Projected	SAT: College Ready Yes: PA	SAT: College Ready Yes: National
Below Basic	11.1%	11.6%	12.8%	22.0%	29.0%	19.8%	25.1%	0.9%	51.7%	54.0%
Basic	13.2%	13.1%	16.0%	38.0%	39.0%	41.2%	19.5%	7.8%	51.7%	54.0%
Proficient	26.2%	27.8%	26.6%	30.0%	25.0%	23.5%	29.8%	42.2%	51.7%	54.0%
Advanced	49.5%	47.5%	44.7%	10.0%	7.0%	15.5%	25.5%	91.2%	51.7%	54.0%
Below Basic + Basic	24.3%	24.7%	28.8%	60.0%	68.0%	61.0%	44.8%	4.0%	51.7%	54.0%
Proficient + Advanced	75.7%	75.3%	71.3%	40.0%	32.0%	39.0%	55.3%	64.9%	51.7%	54.0%
Total Percentage	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	38.1%	51.7%	54.0%
Total N	128,421	132,803	135,909	3,600	161,700	93,703	135,676	61,118	65,426	N/A

Figure 13–3B. External Impact Data: Biology

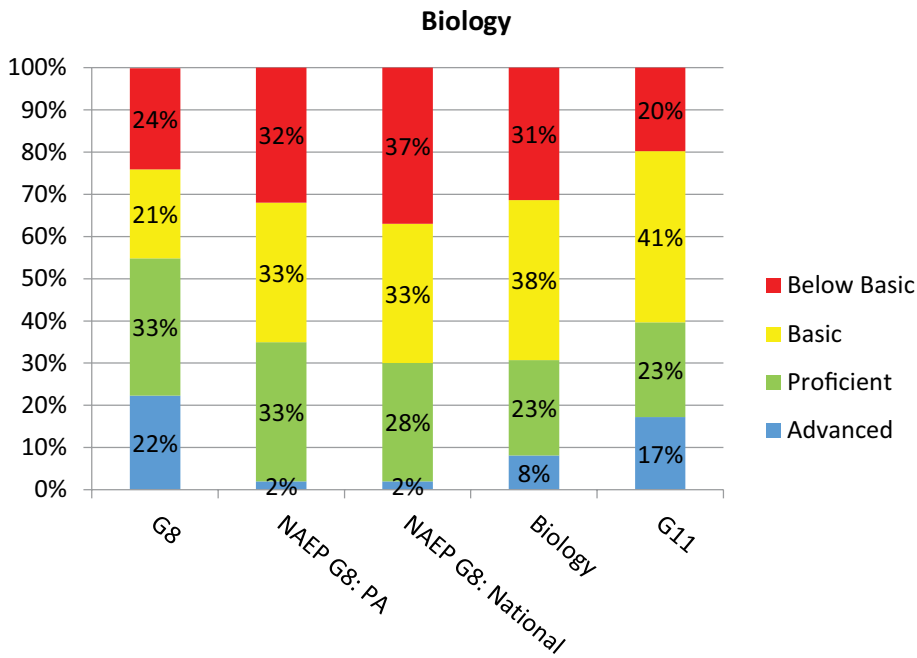


Table 13–4B. Number of Score Points in OIB and Number of Items Supplemented: Biology

Performance Level	PSSA G8	NAEP G8: PA	NAEP G8: National	KE Biology	PSSA G11	College Ready Yes Projected	SAT Math: College Ready Yes: PA	SAT Math: College Ready Yes: National
Below Basic	24.0%	32.0%	37.0%	31.4%	19.8%	1.4%	38.7%	43.0%
Basic	21.1%	33.0%	33.0%	37.9%	40.5%	9.1%	38.7%	43.0%
Proficient	32.5%	33.0%	28.0%	22.6%	22.5%	45.9%	38.7%	43.0%
Advanced	22.3%	2.0%	2.0%	8.1%	17.2%	86.6%	38.7%	43.0%
Below Basic + Basic	45.1%	65.0%	70.0%	69.3%	60.3%	6.6%	38.7%	43.0%
Proficient + Advanced	54.8%	35.0%	30.0%	30.7%	39.7%	63.8%	38.7%	43.0%
Total Percentage	100.0%	100.0%	100.0%	100.0%	100.0%	29.4%	38.7%	43.0%
Total N	134,969	3,600	151,100	46,394	131,534	60,311	65,426	N/A

Figure 13–3L. External Impact Data: Literature

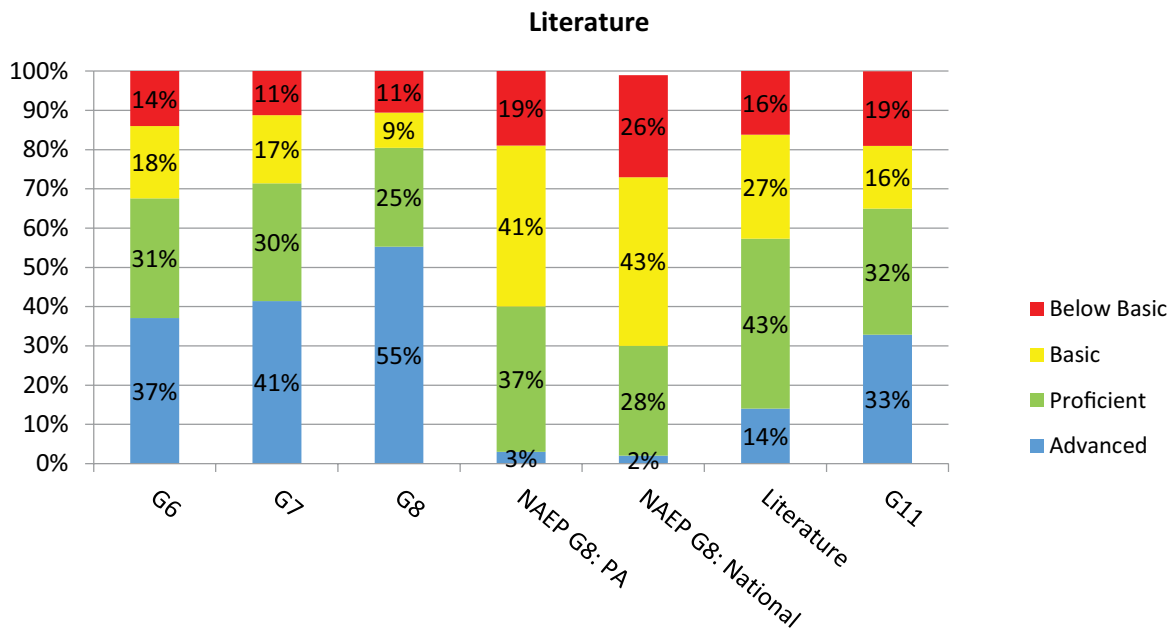


Table 13–4L. Number of Score Points in OIB and Number of Items Supplemented: Literature

Performance Level	PSSA G6	PSSA G7	PSSA G8	NAEP G8: PA	NAEP G8: National	KE Literature	PSSA G11	College Ready Yes: Projected	SAT Math: College Ready Yes: PA	SAT Math: College Ready Yes: National
Below Basic	14.0%	11.2%	10.6%	19.0%	26.0%	16.2%	19.0%	1.2%	46.4%	50.0%
Basic	18.4%	17.4%	8.9%	41.0%	43.0%	26.5%	16.0%	4.7%	46.4%	50.0%
Proficient	30.5%	30.0%	25.2%	37.0%	28.0%	43.3%	32.1%	24.8%	46.4%	50.0%
Advanced	37.1%	41.4%	55.3%	3.0%	2.0%	14.0%	32.9%	76.7%	46.4%	50.0%
Below Basic + Basic	32.4%	28.6%	19.5%	60.0%	69.0%	42.7%	35.1%	2.7%	46.4%	50.0%
Proficient + Advanced	67.6%	71.4%	80.5%	40.0%	30.0%	57.3%	64.9%	51.1%	46.4%	50.0%
Total Percentage	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	34.5%	46.4%	50.0%
Total N	128,284	132,641	135,739	3,500	160,900	42,292	135,470	61,081	65,426	N/A

The Keystone Exams and PSSA results were presented to the panelists first. Panelists were encouraged to compare the impact data and discuss whether the results for the Keystone Exams were reasonable. The NAEP results were added next, and the SAT results were introduced last for comparison and discussion. While discussing the external data, panelists were reminded that all these tests were created for different purposes and might cover different content standards.

Before panelists provided their final judgments, they were instructed to fill out the readiness form to make sure they understood how to adjust their placements (if they desired to do so) based on the Round 2 information and external impact data. After their individual bookmark placements, panelists filled out the evaluation form. The judgments were entered into the spreadsheet program to calculate the median placements for the full panel. The associated impact data were also calculated. The Round 3 results were presented to the panelists for their information after the lunch break.

PANELISTS' RECOMMENDATIONS

Table 13–5 provides a summary of each round’s median, minimum, and maximum ratings (i.e., bookmark page numbers) of the group.

Table 13–5. Summary of Panelists’ Ratings for Each Round

Exam	Round	Bookmark Page Number Median	Bookmark Page Number Min.	Bookmark Page Number Max.	Basic/ Proficient Median	Basic/ Proficient Min.	Basic/ Proficient Max.	Proficient/ Advanced Median	Proficient/ Advanced Min.	Proficient/ Advanced Max.
Alg. I	1	11	6	19	28	14	42	45	33	56
Alg. I	2	11	6	13	26	17	33	42	40	46
Alg. I	3	11	10	12	26	18	30	46	41	46
Bio.	1	9	4	15	26	20	30	56	43	62
Bio.	2	8	7	14	24	21	30	54	50	60
Bio.	3	8	7	12	22	20	30	54	50	60
Lit.	1	8	5	14	27	12	34	47	38	52
Lit.	2	9	8	15	23	15	34	46	38	48
Lit.	3	9	8	15	25	17	34	48	38	48

CUT POINTS AND STANDARD ERRORS

Each bookmark page number is associated with a bookmark difficulty (i.e., logit value). The logit cut is the bookmark difficulty corresponding to the median OIB page number minus one. The logit cut and the standard error (SE) of median logit based on panelists’ Round 1 rating were used to establish the 1 and 2 SE confidence intervals. By bracketing the median cut score by 2 SEs, the 95% confidence interval was identified; the confidence interval can be used to estimate the effects of false positives (passing students who may not actually have sufficient knowledge and skills) or false negatives (failing students who do have sufficient knowledge and skills). PDE can use these standard errors to identify the appropriate cut score by taking into consideration the variance in the human judgments. Table 13–6 summarizes the logit cuts associated with Round 3 median ratings, median +/-1 SE, and median +/-2 SE. The corresponding impacts (percentages in performance level) are provided in this table as well. Note that BB represents Below Basic; B represents Basic; P represents Proficient; and A represents Advanced.

Table 13–6. Summary of Logit Cuts and Impacts

Exam	Stats	Logit Cut BB/B	Logit Cut B/P	Logit Cut P/A	Percentage in Performance Level (%) BB	Percentage in Performance Level (%) B	Percentage in Performance Level (%) P	Percentage in Performance Level (%) A	Percentage in Performance Level (%) P+A
Alg. 1	Median-2SE	-0.7273	0.4291	1.3694	14.7	37.3	30.2	17.8	48.0
Alg. 1	Median-1SE	-0.6181	0.5659	1.5041	17.2	37.8	31.6	13.4	45.0
Alg. 1	Median	-0.5090	0.7027	1.6388	19.8	41.2	27.5	11.5	39.0
Alg. 1	Median+1SE	-0.3999	0.8395	1.7735	22.4	44.4	23.5	9.7	33.2
Alg. 1	Median+2SE	-0.2907	0.9763	1.9082	25.2	47.0	19.7	8.1	27.8
Bio.	Median-2SE	-0.5977	0.3098	1.2500	22.5	38.9	27.9	10.7	38.6
Bio.	Median-1SE	-0.4933	0.3564	1.3205	28.3	33.1	29.3	9.3	38.6
Bio.	Median	-0.3888	0.4029	1.3910	31.4	32.7	27.8	8.1	35.9
Bio.	Median+1SE	-0.2843	0.4494	1.4615	34.5	32.2	25.2	8.1	33.3
Bio.	Median+2SE	-0.1799	0.4960	1.5320	40.7	26.0	26.4	6.9	33.3
Lit.	Median-2SE	-0.6561	0.2338	1.7014	12.4	20.9	46.3	20.4	66.7
Lit.	Median-1SE	-0.5545	0.4116	1.9551	14.2	25.3	46.5	14.0	60.5
Lit.	Median	-0.4530	0.5894	2.2088	16.2	26.5	48.4	8.9	57.3
Lit.	Median+1SE	-0.3515	0.7672	2.4625	18.3	31.4	43.6	6.7	50.3
Lit.	Median+2SE	-0.2499	0.9450	2.7162	20.5	36.5	38.2	4.8	43.0

FINAL RESULTS

After reviewing the results in Table 13–6 and considering panelists’ discussions at the standard setting workshop, PDE recommended using the logit cut scores associated with the median of panelists’ Round 3 ratings for Algebra I and Biology. For Literature, PDE recommended the logits cuts associated with the Round 3 median plus 1 SE.

To avoid negative values on the logit scale, the scaling constants were determined next to linearly convert the logit values to scaled scores. The scaled cut scores for each performance level were obtained by linearly transforming the logit cuts. Details of the scaling process can be found in Chapter Fourteen. A brief description is below.

For Keystone Exams, the linear transformation from logits or Rasch measures to scaled scores was established by anchoring the logit cut for Basic/Proficient to a scaled score 1500 and fixing the slope constant to 50. The intercept constant was calculated next based on the known values 1500, 50, and the logits cut for Basic/Proficient for each content area. In addition, the bottom of the scale was truncated at the lowest obtainable scaled score (LOSS), 1200. The top of the scaled scores was truncated at the highest obtainable scaled score (HOSS), 1800. The recommended scaled score cuts and the corresponding impacts were provided to the Board on July 20, 2011, for approval. Table 13–7 presents the final scaling constants and the Board-approved scaled-score ranges for each performance level.

Table 13–7. Summary of Scaled-Score Ranges and Scaling Constants

Exam	Performance Level Below Basic	Performance Level Basic	Scaling Constants Proficient	Scaling Constants Advanced	Scaling Constants Slope	Scaling Constants Intercept
Algebra I	1200–1438	1439–1499	1500–1545	1546–1800	50	1464.365
Biology	1200–1459	1460–1499	1500–1548	1549–1800	50	1479.355
Literature	1200–1443	1444–1499	1500–1583	1584–1800	50	1461.140

The Keystone Exams are reported by total and modules. Although the panelists made recommendations based on the total test only, the Basic/Proficient cut for the total test is applied directly in setting the passing cut score for each module. In this case, the passing scaled score cut at module level is 1500.

CHAPTER FOURTEEN: SCALING

Scaling is used to transform test score values (i.e., raw scores) onto a scale that can be interpreted by users easily and correctly. Raw scores cannot be used to compare students' achievement across administrations because they depend on the difficulty of the tests. The same student can score higher on an easy test than on a difficult test. To overcome the limitation of raw scores, the scaled scores are introduced to report students' achievement in each Keystone Exam. This chapter describes the two major steps to convert a raw score to a scaled score (SS) and some key considerations for establishing the score scale for Keystone Exams.

RAW SCORES TO RASCH ABILITY ESTIMATES

The pre-equated item parameter estimates for the operational items (further discussed in Chapters Twelve and Fifteen) were used to obtain Rasch person ability estimates and asymptotic standard errors of measurement for each possible raw score value for the overall test, as well as each module. The generation of this raw score-to-Rasch ability was accomplished through application of the fundamental formulas in the Rasch measurement model. The combination of both dichotomously scored multiple-choice (MC) items as well as polytomously scored constructed-response (CR) items requires the use of a partial-credit model (RPCM; Wright & Masters, 1982). The Newton-Raphson iterative procedure is used to obtain precise ability estimates:

$$b_r^{(t+1)} = b_r^t - \frac{r - \sum_i^L \sum_{k=1}^m k P_{rik}^{(t)}}{- \sum_i^L \left[\sum_{k=1}^m k^2 P_{rik}^{(t)} - \left(\sum_{k=1}^m k P_{rik}^{(t)} \right)^2 \right]} \quad r = 1, \dots, M - 1,$$

where b_r^t is the estimated ability of the student with score r after t iterations, k is the number of thresholds, L is the number of items, $M = \sum_i^L m_i$, and $P_{rik}^{(t)}$ is the probability, π_{nix} , defined earlier in Chapter Twelve:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^x (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^x (\beta_n - \delta_{ij})}, \quad x = 0, 1, \dots, m_i$$

The asymptotic standard error of measurement (SEM) was estimated from the denominator of the final iteration:

$$SE(b_r) = \left[\sum_i^L \left[\sum_{k=1}^m k^2 P_{rik}^{(t)} - \left(\sum_{k=1}^m k P_{rik}^{(t)} \right)^2 \right] \right]^{-1/2}.$$

The Rasch ability estimates and the corresponding SEMs are then transformed to scaled scores and SEMs of scaled scores as discussed in the following section.

ZERO AND PERFECT SCORES

A direct ability estimate for zero (no points earned) or perfect (all points earned) raw scores can't be achieved. Thus, a default procedure for estimating such extreme scores was used for the Keystone Exams. Essentially, a fractional raw score (a value less than one, e.g., 0.3) was added to zero scores and subtracted from perfect scores to determine the corresponding logit values for these extreme scores.

RASCH ABILITY ESTIMATES TO SCALED SCORES

Generally, scaled scores are preferred over Rasch ability estimates for reporting purposes. One issue is that Rasch ability estimates are on a scale that includes negative and decimal values. By transforming the Rasch ability estimates to scaled scores, all reported values can become positive integers, which allows for better interpretation by parents and students. Since Rasch ability estimates are comparable after equating to the base administration/year, the transformed scaled scores have a common scale across administrations, even though the corresponding raw scores may differ. Refer to Chapter Fifteen for additional details on equating.

Scaled scores are obtained through a linear transformation of Rasch ability estimates that are associated with each raw score point. Before the linear equation is established for each content area, a few points were considered for the Keystone Exams:

- Avoid scales that might be confused with scores for other types of assessment, for example:
 - Scaled scores ranging from 0 to 100 (because this might be confused with percentage correct scores or percentile ranks)
 - Scaled scores ranging from 200 to 800 (because this might be confused with SAT scores)
 - Scaled scores with similar ranges as the ones for the Pennsylvania System of School Assessment (PSSA) or Classroom Diagnostic Tools (CDT)
- Avoid scales similar to raw scores from a base form.
- Avoid scales that might suggest the scores are more precise than they actually are (i.e., suggesting more precision than can actually be supported by the test scores).
- Avoid scales with negative numbers and decimals.

In terms of industry standard practice, a common perspective is that scaled scores should facilitate score interpretation while at the same time minimize misinterpretation and unwarranted inferences. Often this is done by incorporating some kind of meaning to the scores¹ (Peterson, Kolen, and Hoover, 1989). The incorporation of content meaning is one way to facilitate score interpretation. This might be done in several different ways. For example, the current PSSA scaled scores, like those of many other state assessments, try to input some content meaning by having the PSSA performance level cut scores have known values on the scaled-score metric. Such an approach appears to make good sense given the purposes of a criterion-referenced test like the PSSA.

As a result, a scaled score range of 1200 to 1800 and the Proficient scaled cut-score of 1500 was used to establish the scales for each of the Keystone Exams in Algebra I, Biology, and Literature. It is worth noting that, although careful considerations were given to the selection of these values, they are completely arbitrary based on previously discussed considerations. For example, the label of 1500 could have been called 100 or any other value or letter without affecting any of the relationships among schools, administrations, students, or items. In other words, changing the scale would simply be changing the labels on the axis of a graph without moving any of the points.

LINEAR TRANSFORMATION FORMULAS

The scaled scores for the Keystone Exams are obtained through a linear transformation of the Rasch ability estimates (θ). Specifically,

$$SS = m\theta + b,$$

where m is the slope and b is the intercept. The linear transformation for the Keystone Exams was derived by anchoring the Proficient cut (i.e., Rasch ability estimate) recommended by the panelists at the standard-setting workshop to the scaled score 1499.5 (i.e., 1500 after rounding), and then set the slope of the line. There could be many lines with different slopes going through the anchor point. However, the slope of the line has influence over the variability of the scaled scores. For Keystone Exams, the slope of 50 was chosen because it results in desired scaled score standard deviation. Once the scaled score, slope, and Rasch ability estimate are determined, the

¹ Not everyone agrees with this sentiment. Some have argued the opposite point, that is, any attempt to add meaning to test scores actually predisposes the scores to be misinterpreted (Angoff, 1984).

intercept b can be derived by the equation above. The final slopes and intercepts for deriving scaled scores for the Keystone Exams are provided in Table 14–1.

Table 14–1. Scaling Constants by Content Area

Exam	Scaling Constants Slope	Scaling Constants Intercept
Algebra I	50	1,464.365
Biology	50	1,479.355
Literature	50	1,461.140

ROUNDING

The linearly transformed scaled scores are always rounded to the nearest integer value for reporting purposes. Values greater than or equal to 0.50 are rounded up. Values less than 0.50 are rounded down.

LOWEST OBTAINABLE SCALED SCORES

The Keystone Exams in Algebra I, Biology, and Literature have a lowest obtainable scaled score (LOSS) of 1200. Any derived scaled score less than 1200 is truncated to this minimum value. The selection of a LOSS is mainly based on two considerations: 1) extreme low scaled scores may have an impact on the average of the scaled scores at school/district level and 2) score truncation makes sense from a score precision perspective given measurement errors at the extremes are large. The LOSS value 1200 is established by giving consideration to *chance* performance over the MC items (e.g., if 40 four-option MCs were on a test, approximately 10 points might be earned on guessing alone) and considering the percentage of students who would be awarded the LOSS values.

HIGHEST OBTAINABLE SCALED SCORES

A highest obtainable scale score (HOSS), 1800, is set for the Keystone Exams for the same reasons described for the LOSS value. However, unlike the LOSS value, which is set initially by giving consideration to guessing over MC items, it is somewhat more difficult to determine what rules should be applied to establish the HOSS. Based on the empirical results, the value 1800 corresponds to a logit value (or Rasch ability estimate) that ranged from 6 to 7, and 0 percent of students received this score.

RAW-TO-SCALED-SCORE TABLES

The final raw-to-scaled score conversion tables can be found in Appendix K. Note that only the raw-to-scaled score tables for each single administration were reported. In other words, these tables cannot be used to look for a student's highest total scaled score to date if it is combined from two different administrations. The conditional standard error of measurement (CSEM, see Chapter Eighteen for detailed discussion) and corresponding confidence intervals based on 1 CSEM are also provided in these tables.

CHAPTER FIFTEEN: EQUATING

Equating is a statistical process that is used to adjust scores on test forms so that scores from two different forms can be used interchangeably (Kolen & Brennan, 2004) even though the test forms consist of different items. In large-scale testing programs, it is a common practice to have different item sets appear in different test forms across administrations. Students' raw scores (or number-correct scores) cannot be compared across forms or administrations because the scores depend on the difficulty of the items on a form. The same student can score higher on an easy test than on a difficult test. Although there are various equating methods available for different psychometric paradigms (IRT and CTT), the Keystone utilizes an IRT approach aligned with the assumptions of the Rasch model, the IRT pre-equating method. The first step in any IRT equating method is to conduct scale linking, in which item difficulties from independent calibrations are transformed onto the same scale (Kolen & Brennan, 2014). Once scale linking is conducted, we can proceed with any IRT-based equating methods.

Since the Keystone's inception, a pre-equating design has been implemented due to the many advantages it offers. Specifically, employing a pre-equating method allows for a shortened turn-around time for score reporting due to the use of previously linked item parameters for test construction and development of raw-to-scaled-score tables. In this chapter, we provide a brief comparison of pre- and post-equating, an explanation of the procedure implemented for the Keystone examinations, and a description of the evaluation of pre-equated and post-equated solutions. Summary results are also presented. The equating design and analyses were conducted for all administrations (winter, spring, and summer) of the Keystone examinations.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information. The Pennsylvania Technical Advisory Committee (TAC) provided guidance and direction for processes that were modified. The information provided in this chapter discusses the typical procedures for Keystone, with footnotes that outline any changes for the 2021 processes.

PRE- VS. POST-EQUATING

As with other Pennsylvania assessment programs, the Rasch model is used to guide the test design, form construction, calibration, scaling, and equating of the Keystone Exams. The key step of equating test forms using the Rasch model is to place the item parameters from different administrations on the same scale, also referred to as scale linking. Once the item parameters from different operational test forms are on the same scale, the Newton Raphson procedure can be used to convert number-correct scores to scaled scores as described in Chapter Fourteen. As a result, the scaled scores can be compared and used interchangeably within and across administrations.

As is the case with many K–12 large-scale assessment programs, all scored items are field tested prior to operational use. Once the field-test items' difficulties are placed on the base scale or common metric, in theory, one should not expect the Rasch item difficulties for these items to change (except within a reasonable range of measurement error) after they are administered in an operational test, provided the Rasch model fits the data. Based on this theoretical advantage of using the Rasch model, equating can be conducted using the item parameters that were previously calibrated. This statistical procedure is referred to as pre-equating.

In contrast, post-equating requires data from the current administration to be calibrated. Then, newly estimated item parameters are linked and placed on the same scale as banked item parameters, and scores are equated. With this in mind, pre-equating is advantageous because much of the work is completed before test administration, allowing more time for quality control; whereas post-equating relies on the same given timeframe for calibration, scale linking, equating scores, and implementing quality control procedures.

Although in theory the two equating procedures should provide identical results when the model fits the data, each has its own advantages and disadvantages. The use of pre-equating can facilitate the operational process in terms of rapid score reporting, more time for quality control, and more flexibility in the assessment. One successful application of pre-equating is for computer-adaptive tests (CATs) where test questions are tailored to the student's achievement as the test progresses, such as the Pennsylvania Classroom Diagnostic Tools (CDT). The CDT is

designed to provide diagnostic information about student performance and is available throughout the school year at no cost. CATs require automated scoring for all item types (including constructed-response) and allow for immediate score reporting upon completion of the test. However, a variety of issues need to be considered when using pre-equating in practice. For example, students may not be motivated to take the field tests, especially standalone field tests, which may make the items appear harder in the field test than in the operational test (Eignor, 1985; Eignor & Stocking, 1986; Stocking & Eignor, 1986; Kolen & Harris, 1990). Other concerns for the field-test items include item context, item position, and sample size. In contrast, the use of post-equating, when applicable, does not have the same motivational concerns as with pre-equating. Also, post-equating uses post-administration data and is thought to yield more accurate analysis results, given that the number of students who take the operational tests is usually large. On the other hand, when the reporting window is extremely tight, as is the case with some graduation or end-of-course exams in various states, post-equating must occur within a very short time, allowing less time for the equating analyses and quality control.

EQUATING DESIGN FOR KEYSTONE EXAMS

The Keystone Exams, like many other graduation or end-of-course exams, require quick turnaround of testing results. After the exams are administered, the bulk of the time is consumed by various data-processing steps. As a result, the equating analyses must be produced in a short period of time, which introduces risk of delays in score reporting. In addition, a student's best-score-to-date is calculated as the combination of performance across any administrations (see Chapter Sixteen for details), thus increasing the complexity of equating analyses and score reporting for future administrations. To control the quality and timeliness of post administration processing, pre-equating, one of the most promising applications of Rasch model or item response theory (see Lord, 1980, chap. 13), has been implemented for the Keystone Exams.

To implement the pre-equating model in the Keystone Exams, efforts have been made to enhance the accuracy of pre-equating results based on the findings from the literature. For example, to address the concerns regarding students' motivation to take field tests, stand-alone field tests were not used; rather field-test items were embedded throughout the test so that students would perceive no differences between field-test items and operational items. This approach allows Rasch item difficulty estimates to be used for future pre-equating purposes and is based on the assumption that students should be equally motivated to take the operational and embedded field-test items, especially when they are not aware of which item is a field-test item. To minimize item context and item position effects (i.e., lack of motivation and fatigue), field-test items were interspersed within the operational sections. With this design, students have a smaller chance of knowing the field-test item positions. Fatigue effects due to field-test items being placed in the last section of the operational test can be mitigated in this design as well. To improve the accuracy of the Rasch item difficulties and parameters estimated from the field-test data, DRC scored all MC items and a large sample of CR items given that larger sample sizes can increase the estimation accuracy. The test designs for the operational PSSA mathematics, ELA, and science assessments used multiple test forms that shared several common elements. The operational items were the same on all forms and for all students. Student total raw scores and scaled scores, as well as accountability reporting, were based exclusively on the operational items.

SCALE LINKING

Keystone utilizes a concurrent calibration linking design to obtain item parameters for field-tested items (administered during the spring administration). Results from scale linking are item parameters (Rasch difficulties and thresholds) for field-tested (FT) items that are on the base scale. The chain originates from the scale of measurement defined for each test's base form, which is used as the reference for calibrating all items in the item pool. The base form is usually the form upon which the cut scores were established (see Chapter Thirteen). In the case of the Keystone, scales and cut scores were established for all subjects in 2011; therefore, the examinations are linked to the scales set in 2011.

The Rasch Partial Credit Model (RPCM) is used for the calibrating data, given its flexibility for dichotomously scored (i.e., MC) and polytomously scored (i.e., CR) item types (Masters, 1982). The RPCM is discussed in detail in Chapter Twelve. Without employing scale linking, Rasch difficulties for the field-tested items would not be directly comparable to other items on the base scale. A partially anchored concurrent calibration was employed to estimate all 2021 field-test item parameters for each test on its respective base scale. First, all OP item parameters were evaluated for model-fit to ensure that previously estimated (banked) item parameters were still reasonable and appropriate. Then, OP item parameters were anchored, and FT item parameters were freely estimated for each content area. This allowed for the estimation of FT item parameters on the baseline scale. The following steps are

conducted following each spring administration to estimate the FT item parameters.

1. Calibrate OP and FT items in a partially-anchored concurrent design.
 - a. Anchor OP item parameters to the banked values.
 - b. Include all operational (OP) and FT items¹.
 - c. Include only students who have completed the test and met other criteria described in Chapter Nine.
 - d. FT item parameters are banked for future use².

PRE-EQUATING VERIFICATION

Although extra care has been taken to guarantee the success of pre-equating during the test design, form construction, and calibration of embedded field-test items, DRC also ensured that the pre-equated results had reasonable data-model fit during the pre-equating verification process. Once sufficient data was available, pre-equating verification was conducted to assess data-model fit and allow the parameters of any misfitting items to be freely calibrated. Any misfitting item was identified, and parameters were freely estimated in a subsequent calibration (using a partially anchored design) to improve data-model fit. All calibration was conducted using WINSTEPS (Linacre, 2019). Both sets of item parameters were then used to estimate student abilities, which were then transformed to scaled scores. (Transformation formulas are provided in Chapter Fourteen.) The data and results presented in this section refer to misfitting items as those in which infit mean-square values exceeded a criterion of 1.3. The number of items identified during pre-equating verification for each content area is shown in Table 15–1. No items were identified as misfitting for Algebra in either the winter or spring administration, whereas there was one item identified for each of the Biology winter, Literature winter, and Literature spring administrations. Then, differences were analyzed between fully anchored pre-equated results (hereinafter “pre-equated”) and partially anchored pre-equated results (hereinafter “post-equated”). Pre-equating verification analyses were conducted at the item level, person level, and form level. Additional results from the pre-equating verification analyses can be found in Appendix L.

Table 15–1. Number of Misfitting Items Identified during Pre-Equating Verification by Administration and Content Area

Administration	Algebra I N	Biology N	Literature N
Winter	0	0	1
Spring	0	1	1
Summer	-	-	-

Note. There was no Summer 2021 administration due to the elongated spring testing window from May through September 2021.

At the same time, DRC test development specialists reviewed all misfitting items (infit mean-square values greater than 1.3) and items with large displacement values (absolute value greater than 0.5) to ensure that items were presented in the same manner as their prior administration.

ITEM-LEVEL ANALYSES

Item-level analyses indicate whether the data fit the Rasch model with respect to item-fit statistics and whether the item parameters showed displacement. For item misfit, this included the number of items that had reasonable fit statistics (e.g., greater than 0.7 and less than 1.3) supported by prior literature (Wright & Linacre, 1994). Table 15–2 shows the item fit statistics comparison for each content area and administration. For each administration and content area, both pre-equated and post-equated solutions showed similar fit to the model in terms of infit and outfit statistics. The results support that the data fit the pre-equated solution well.

¹ In 2021, only FT MC items were included because FT OE items were not scored.

² In 2021, FT parameters were banked, however they are considered ineligible for future use given the circumstances surrounding the Covid-19 pandemic and guidance from PA TAC.

Table 15–2. Item Infit and Outfit Mean-Square Statistics by Administration and Content Area

Admin.	Content Area	Method	N	Infit Mean	Infit SD	Infit Min	Infit Max	Infit Range [0.7,1.3]	Outfit Mean	Outfit SD	Outfit Min	Outfit Max	Outfit Range [0.7,1.3]
Winter	Algebra I	Pre	13201	0.99	0.10	0.74	1.22	42/42	1.00	0.14	0.67	1.30	40/42
Winter	Algebra I	Post	13201	0.99	0.10	0.74	1.22	42/42	1.00	0.14	0.67	1.30	40/42
Winter	Biology	Pre	12143	1.00	0.11	0.79	1.23	54/54	1.00	0.17	0.68	1.38	48/54
Winter	Biology	Post	12143	1.00	0.11	0.79	1.23	54/54	1.00	0.17	0.68	1.38	48/54
Winter	Literature	Pre	11714	0.95	0.16	0.60	1.31	37/40	0.97	0.24	0.49	1.47	30/40
Winter	Literature	Post	11714	0.95	0.16	0.60	1.30	37/40	0.97	0.24	0.49	1.49	30/40
Spring	Algebra I	Pre	94861	0.99	0.12	0.69	1.25	41/42	1.01	0.17	0.70	1.37	39/42
Spring	Algebra I	Post	94861	0.99	0.12	0.69	1.25	41/42	1.01	0.17	0.70	1.37	39/42
Spring	Biology	Pre	86599	1.00	0.11	0.79	1.34	53/54	1.01	0.15	0.76	1.41	51/54
Spring	Biology	Post	86599	1.01	0.11	0.79	1.37	53/54	1.02	0.16	0.76	1.41	51/54
Spring	Literature	Pre	84036	1.01	0.15	0.60	1.32	38/40	1.03	0.23	0.59	1.60	34/40
Spring	Literature	Post	84036	1.01	0.15	0.60	1.31	38/40	1.03	0.22	0.59	1.56	34/40
Summer	Algebra I												
Summer	Algebra I												
Summer	Biology												
Summer	Biology												
Summer	Literature												
Summer	Literature												

Note. There was no Summer 2021 administration due to the elongated spring testing window from May through September 2021.

In addition, the results from the fully-anchored pre-equated calibration were assessed to identify any issues with item parameter stability. Displacement refers to the shift in item parameters between anchored values and values that would have been freely estimated in an unanchored calibration. The items with an absolute displacement value greater than or equal to 0.5 were identified and reviewed by DRC test development specialists. Table 15–3 summarizes the outliers flagged by displacement. These items were reviewed by DRC content specialists, but no obvious reasons were found to explain the item difficulty change.

Table 15–3. Count of Items Flagged for Displacement by Administration and Content Area

Administration	Algebra I N	Biology N	Literature N
Winter	0	1	1
Spring	0	0	1
Summer	-	-	-

Note. There was no Summer 2021 administration due to the elongated spring testing window from May through September 2021.

PERSON-LEVEL ANALYSES

The second set of analyses conducted consisted of analyzing person-level fit statistics, which can be another indicator of whether the data fit the model. Table 15–4 summarizes the overall person-level infit and outfit statistics by administration and content area for both the pre-equated and post-equated solutions. The table specifies the mean, standard deviation (SD), minimum (Min), maximum (Max), and proportion of persons who had reasonable fit statistics (e.g., greater than 0.5 and less than 1.5)³ for both infit and outfit statistics. The results in the tables indicate that person-level fit does not vary by equating method.

Furthermore, Appendix L includes the results for the pre-equating verification, including the person infit boxplots for all administrations and content areas for both pre-equated and post-equated solutions. Appendix L also provides boxplots disaggregated by ethnicity, gender, English Learners (ELs), and students with individualized educational programs (IEPs). The person infit plots indicate that the data fits the pre- and post-equated solutions similarly.

Table 15–4. Person Infit and Outfit Mean-Square Statistics by Administration and Content Area

Admin.	Content Area	Method	N	Infit Mean	Infit SD	Infit Min	Infit Max	Infit Range [0.5,1.5]	Outfit Mean	Outfit SD	Outfit Min	Outfit Max	Outfit Range [0.5,1.5] ⁺
Winter	Algebra I	Pre	13201	0.98	0.21	0.17	3.07	97.40%	1.00	0.20	0.05	5.16	97.70%
Winter	Algebra I	Post	13201	0.98	0.21	0.17	3.07	97.40%	1.00	0.20	0.05	5.16	97.70%
Winter	Biology	Pre	12143	0.99	0.15	0.53	1.97	99.60%	1.00	0.15	0.10	2.44	99.10%
Winter	Biology	Post	12143	0.99	0.15	0.53	1.97	99.60%	1.00	0.15	0.10	2.44	99.10%
Winter	Literature	Pre	11714	0.93	0.22	0.32	2.43	97.20%	0.97	0.35	0.14	7.60	91.80%
Winter	Literature	Post	11714	0.93	0.22	0.32	2.43	97.20%	0.97	0.35	0.14	7.63	91.60%
Spring	Algebra I	Pre	94861	0.96	0.20	0.21	3.05	97.80%	1.01	0.23	0.05	9.90	96.40%
Spring	Algebra I	Post	94861	0.96	0.20	0.21	3.05	97.80%	1.01	0.23	0.05	9.90	96.40%
Spring	Biology	Pre	86599	1.00	0.16	0.53	2.53	99.20%	1.01	0.17	0.28	4.00	98.50%
Spring	Biology	Post	86599	1.00	0.16	0.53	2.49	99.20%	1.02	0.17	0.27	4.02	98.50%
Spring	Literature	Pre	84036	0.97	0.23	0.32	3.42	97.00%	1.03	0.34	0.13	7.73	89.60%
Spring	Literature	Post	84036	0.97	0.23	0.32	3.42	97.00%	1.03	0.34	0.13	7.71	89.80%
Summer	Algebra I												
Summer	Algebra I												
Summer	Biology												
Summer	Biology												
Summer	Literature												
Summer	Literature												

Note. There was no Summer 2021 administration due to the elongated spring testing window from May through September 2021.

³ While items and persons are on the same scale, items tend to be more stable. As such, stricter rules are applied to item-fit statistics than person-fit statistics in determining reasonable fit.

NORMALIZED SCALED SCORE DIFFERENCES

On the form level, we evaluated differences between pre-equated and post-equated results. Normalized differences were calculated as the difference between the scaled score divided by the average CSEM of pre- and post-equated results at each raw score point (see Equation below). Table 15–5 displays the descriptive statistics of the normalized scaled score differences as well as the normalized scaled score difference at the proficient cut-score. A difference at the proficient cut-score may indicate a difference in performance level classification. Results indicated that normalized differences were all within reasonable expectations (min = -0.08, max = 0.05), where the largest difference was observed for spring Biology. The plots for normalized scaled score differences are displayed in Appendix L.

$$\text{Normalized Scale Score Difference} = \frac{(SS_{Pre} - SS_{Post})}{((CSEM_{Pre} + CSEM_{Post})/2)}$$

Table 15–5. Normalized Scaled Score Differences Summary by Administration and Content Area

Administration	Content Area	SS Difference Mean	SS Difference Min	SS Difference Max	SS Difference at Prof. Cut
Winter	Algebra I	0	0	0	0
Winter	Biology	0	0	0	0
Winter	Literature	-0.005	-0.067	0	0
Spring	Algebra I	0	0	0	0
Spring	Biology	-0.017	-0.083	0	-0.083
Spring	Literature	0.001	0	0.048	0
Summer	Algebra I				
Summer	Biology				
Summer	Literature				

Note. There was no Summer 2021 administration due to the elongated spring testing window from May through September 2021.

PERFORMANCE LEVEL CLASSIFICATION

Pre-equated solutions were considered reasonable if classification consistency did not change more than 5%. Table 15–6 shows the consistency of classifications with respect to performance levels. The three numeric values within each cell refer to the proportion of students that do not agree at each of the three cuts (Basic, Proficient, and Advanced). If a numeric entry is followed by a negative sign, then pre-equating resulted in a lower percentage of students in the adjacent performance level when compared to post-equating. On the other hand, if the numeric entry is followed by a positive sign, then pre-equating resulted in a higher percentage of students in the adjacent performance level when compared to post-equating.

Performance level classification was identical between the pre- and post-equated solutions for all administrations and content areas except spring Biology, in which 2% of students were classified as Basic when pre-equating was used and Proficient when post-equating was used. Yet these results indicate that the pre-established criterion, in which no more than 5% of performance level classifications changed between the pre- and post-equated solutions, was met. There were no differences observed in performance level classifications for Algebra or Literature. After comparing and evaluating the results, the percentage of students classified differently was less than 5% within each classification. The TAC agreed if classification consistency was less than 5%, then pre-equated solutions should be accepted⁴.

⁴ Given the circumstances surrounding the Covid-19 pandemic in addition to the extended testing window and the dual reporting of scores, the PA TAC supported using the pre-equated solution.

Table 15–6. Performance Level Impact Summary between Pre- and Post-Equated Solutions by Administration and Content Area

Administration	Algebra I	Biology	Literature
Winter	Exact	Exact	Exact
Spring	Exact	(0,2-,0)	Exact
Summer	-	-	-

Note. There was no Summer 2021 administration due to the elongated spring testing window from May through September 2021.

SCALE STABILITY AND MAINTENANCE

Scale stability is a critical component of any testing program. The 2014 Standards of Educational and Psychological Testing state that “Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which the scores are reported” (p.103). Conducting item parameter checks, ensuring that item parameters do not drift over time, and potentially updating operational item parameters are a few ways in which testing programs can maintain scale stability. Although many of these aspects are checked during the pre-equating verification process, it is also important to analyze student performance and scale stability following each administration.

In spring 2021, field-test item statistics and parameters were estimated and banked, but those statistics are not eligible for future forms construction due to the expected impact from the Coronavirus pandemic, including but not limited to the disruption to teaching and learning, the lower participation in statewide summative assessments, and the elongated testing window. However, for consistency purposes, field-test item parameters were estimated for MC items using a partially anchored concurrent calibration, where OP item parameters were anchored on their previously banked values. The final Rasch item parameters can be found in Appendix J. Note that field-test OE items were not scored in spring 2021.

TEST CHARACTERISTIC CURVES AND LOGIT PLOTS

Figures 15–1 and 15–2 help one visualize the across-administration differences in the difficulties of operational items. In Figure 15–1, the test characteristic curves (TCCs) for each administration are presented for each content area. These figures show the winter, spring, and summer TCCs and indicate alignment among the 2021 test forms in terms of difficulty in the logit metric. TCCs that are closely aligned translate into similar raw-score cut points and similar test difficulty across years. The three dotted vertical lines represent the Basic, Proficient, and Advanced cut-scores on the logit (theta) scale. All content areas showed very small differences across administrations.

Figure 15–1. Test Characteristic Curves.

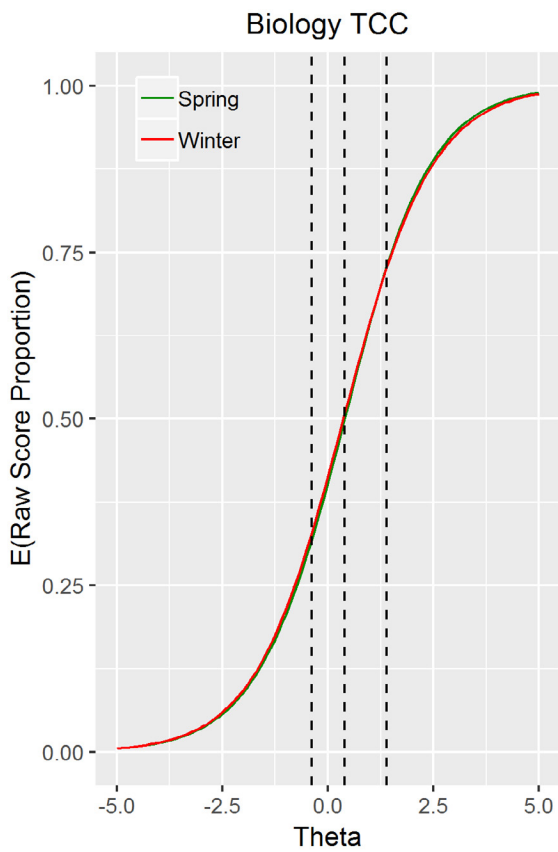
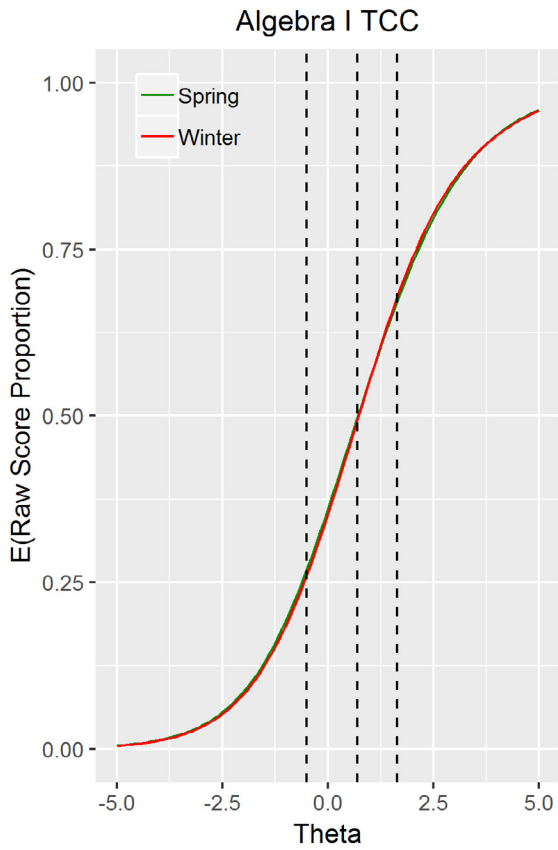


Figure 15–1 (continued). Test Characteristic Curves.

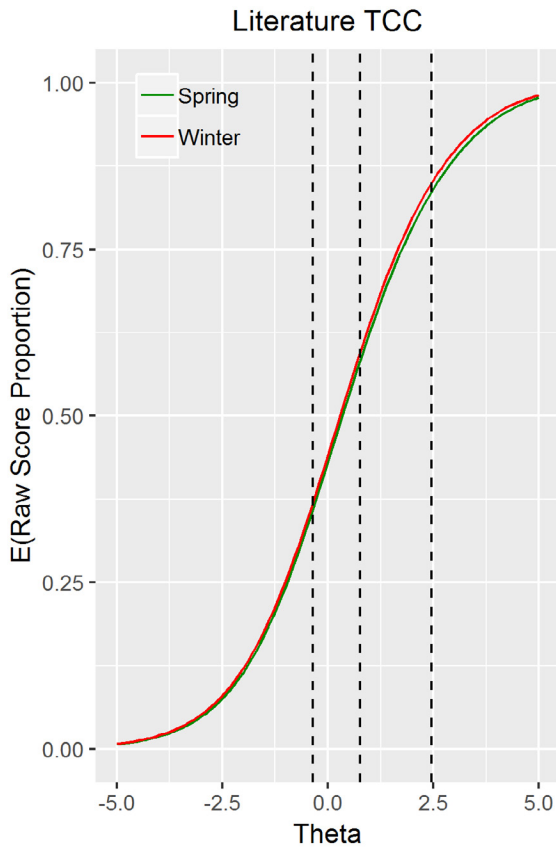


Figure 15–2 displays the relationship between the pre-equated item difficulties (x-axis) and the post-equated item difficulties (y-axis) on the logit (theta) scale. The black line represents the identity line; if points fall on the identity line, it indicates that there is no difference between the pre-equated and post-equated item difficulty. Points that do not fall on the identity line indicate items that were identified as misfitting during the pre-equating verification process and were freely estimated in a subsequent calibration. In all cases, the item difficulties that were freely estimated under the post-equated model were close to the identity line. The plots provide evidence of reasonable across-year stability of item difficulty, meaning the pre-equated item difficulties were similar to the post-equated item difficulties.

Figure 15–2. Logit Plots

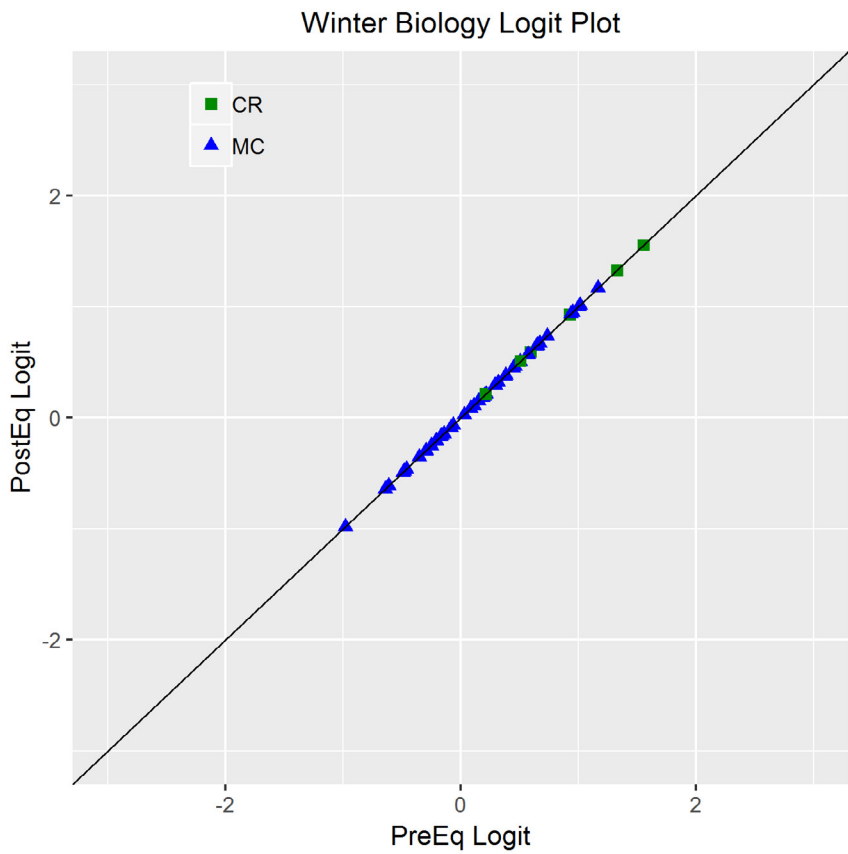
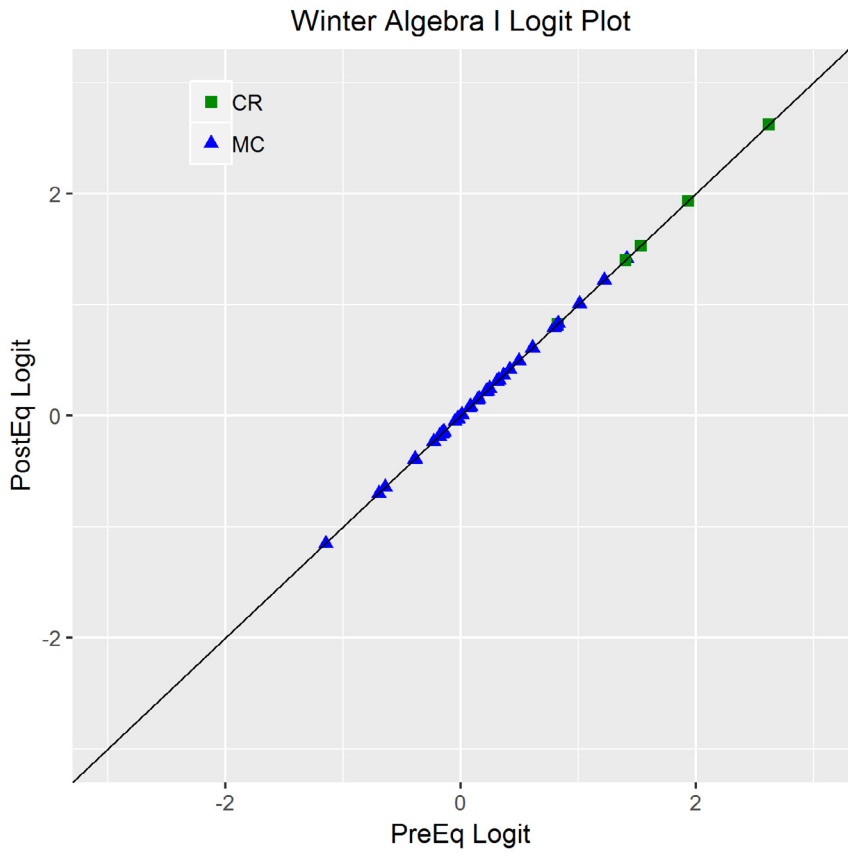


Figure 15–2 (continued). Logit Plots

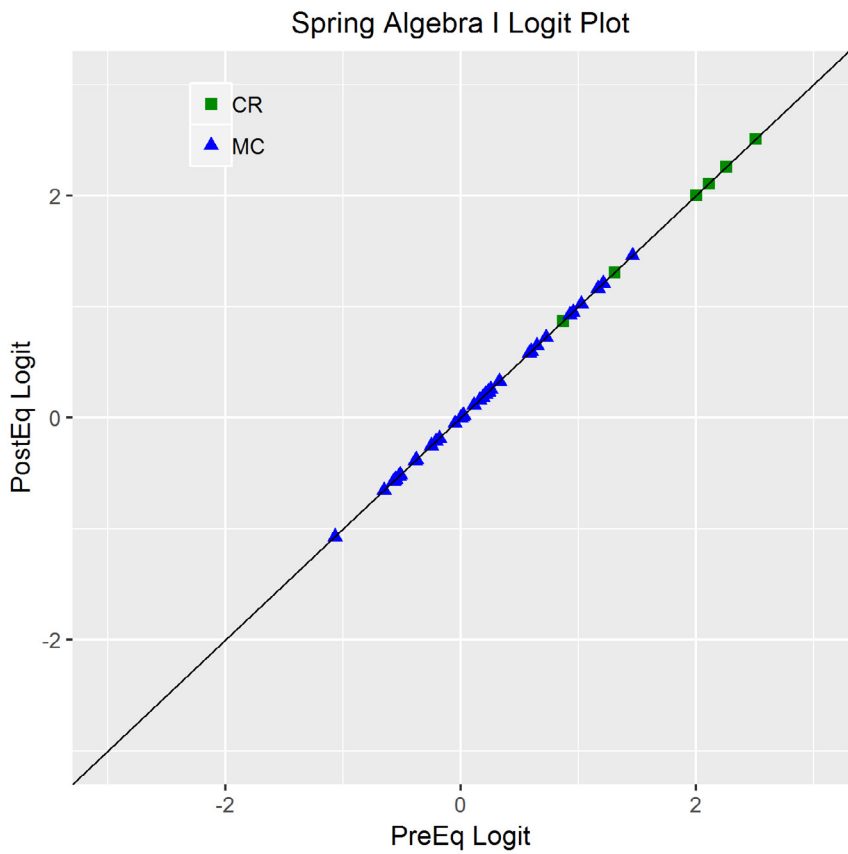
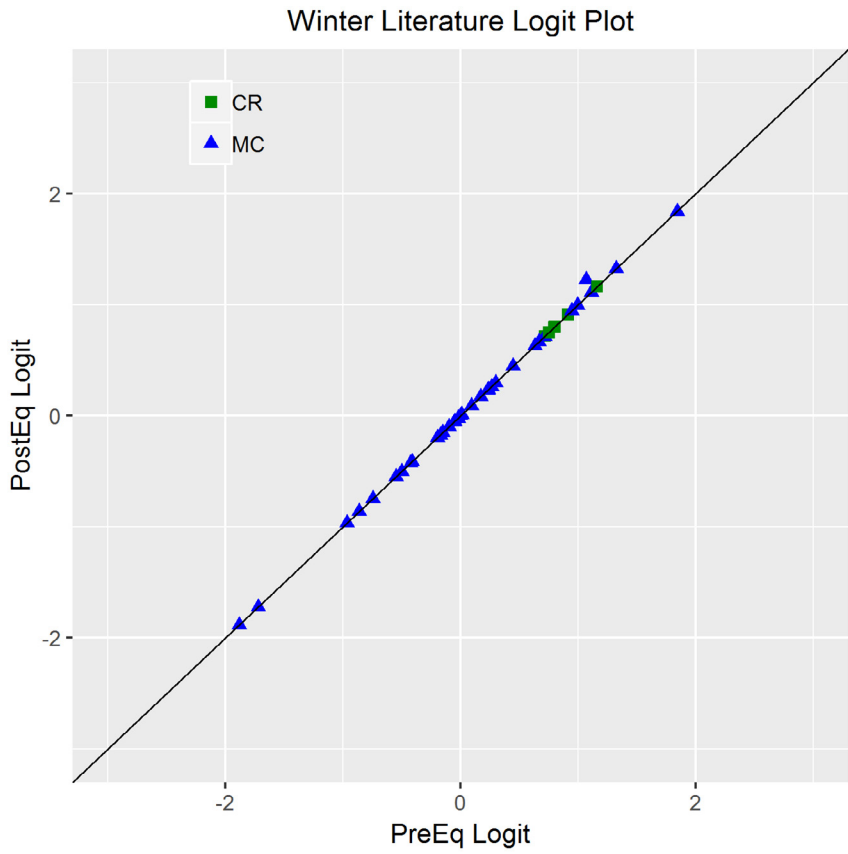
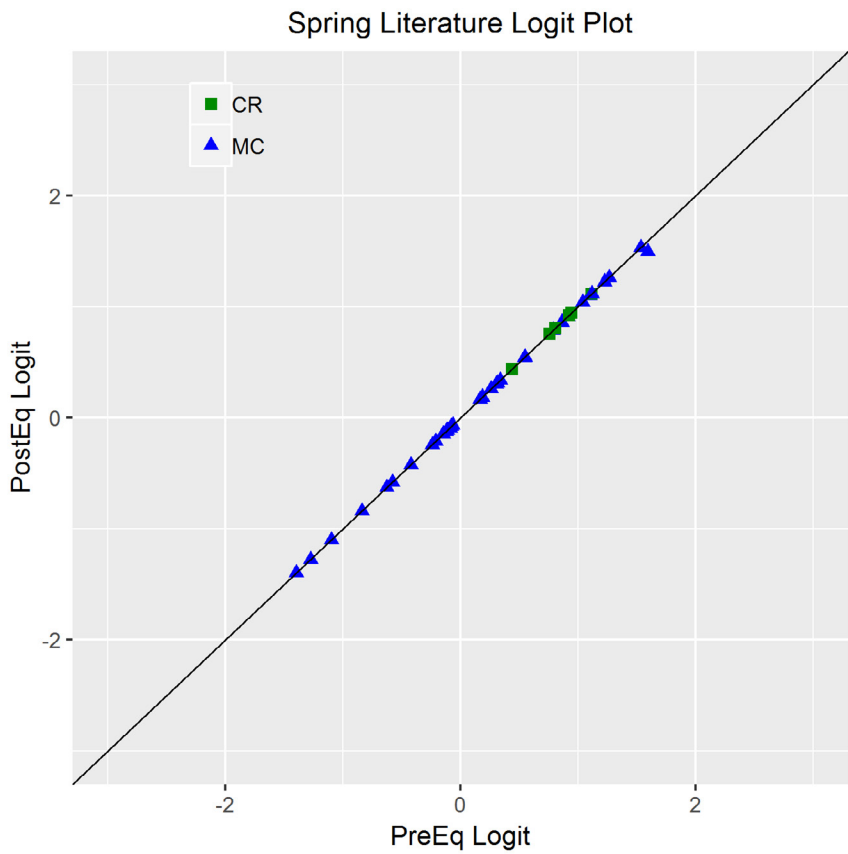
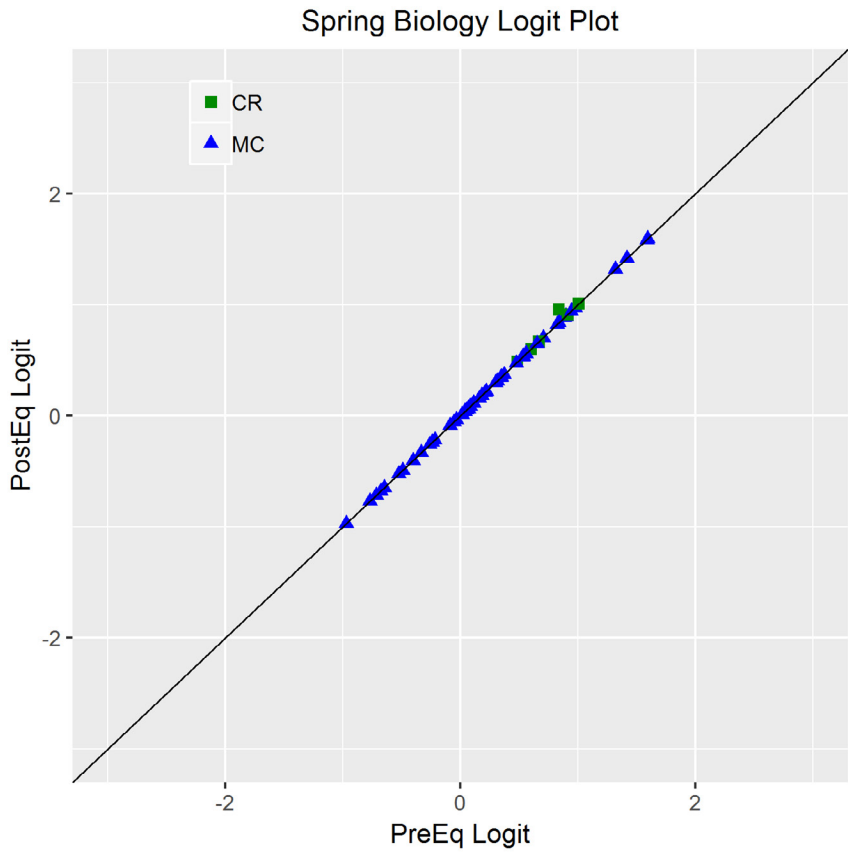


Figure 15–2 (continued). Logit Plots



CHAPTER SIXTEEN: SCORES AND SCORE REPORTS

This chapter provides information about the scores provided for the Pennsylvania Keystone Exams (e.g., scaled scores, performance levels, and module scores), how the scores are presented on score reports, and appropriate and inappropriate uses of the scores.

SCORING

Keystone Exams items include both multiple-choice (MC) and constructed-response (CR) items. Each correct response to an MC item receives a score of 1. Incorrect responses receive a score of 0. Scores on CR items range from 0 to 4, depending on the content area. Table 16–1 summarizes the types of items used in each content-area exam.

Table 16–1. Item Types Used by Content Area

Exam	Item Type MC (1 point)	Item Type CR (3 point)	Item Type CR (4 point)
Algebra I	■		■
Biology	■	■	
Literature	■	■	

DESCRIPTION OF TOTAL-TEST SCORES

Different types of scores have been developed for Keystone Exams reporting. Since the underlying properties of these scores are not necessarily the same, the particular scores used depend on the purposes for which the test has been given. The following types of scores are provided for reporting overall performance on each Keystone Exam:

- Raw scores
- Scaled scores
- Highest total test scaled score to date
- Performance levels

RAW SCORES

A raw score (or number-correct score) is the number of points a student earned over all the operational MC and CR items. By itself, the raw score has very limited utility. One limitation is that it can only be interpreted with reference to the total number of items on a specific exam (e.g., a raw score of 15 on a 20-item exam is different from a raw score of 15 on a 30-item exam). In addition, raw scores depend on the difficulty of test items across test forms (e.g., a raw score of 15 on a test with 20 easy items is different from a raw score of 15 on a test with 20 difficult items). Because the difficulty of the items on a test can change from administration to administration, raw scores should not be compared across administrations.

SCALED SCORES

Scaled scores were introduced in Chapter Fourteen. In the simplest sense, a scaled score is a transformed number-correct score. The specifics of the transformation processes for the Keystone Exams were also discussed in Chapter Fourteen. When all students take the same test items, as with the operational items on the Keystone Exams, the more points the student earns, the higher the associated scaled score will be.

The value of switching to the more abstract scaled-score metric is that it produces more general, interpretable, and equitable results. As noted above, a raw score of 30 is meaningless unless the maximum raw score is known. The difficulty of the test items was also mentioned as an additional challenge with interpreting raw scores. Number-correct scores are transformed to scaled scores to remove the effects of test length and item difficulty. (Strictly speaking, transformation of number-correct scores to percent-correct scores would also remove the effect of test length, but it would do nothing to adjust for the difficulty of the items.)

Another advantage of scaled scores is that they lend themselves to interpretations at what is referred to as an interval level, while raw scores do not. Interval-level scales allow an interpretation of a scaled score difference of 5 points to be the same whether the scores are 1295 vs. 1300 or 1445 vs. 1450. Raw-score differences, in this context, cannot be interpreted in this manner and are thus neither generalizable nor equitable.

A scaled score of 1500—or any other value for a particular content-area exam, such as Algebra I—should have the same absolute meaning in the current administration as it had in previous administrations when test scores are properly equated across administrations. More importantly, a significant increase in the scaled score from the previous administration to the current administration means that student performance improved¹; it does not say anything about whether this administration's exam is easier or harder than last administration's exam. To make these interpretations requires no information about the length or the difficulty of the exam in either administration, although these variables are essential for the process of deriving the scaled scores.

There is considerable auxiliary information presented in this report that might aid in further contextualizing Keystone Exams scaled scores:

- Chapter Fourteen provides information on the development of the Keystone Exams scaled-score system, including transformation formulas, rounding rules, and general scale characteristics (e.g., minimum values).
- Chapter Seventeen provides total-test score statistics. In particular, Table 17–2 lists the scaled-score means and standard deviations for the testing results.

HIGHEST TOTAL TEST SCALED SCORE TO DATE

After an administration, every test-taker earns an administration-specific scaled score based on their overall performance, as well two scaled scores based on their performance on Module 1 and Module 2. As previously discussed, each examination consists of two modules that measure distinct topics within each content area. For example, for Literature Module 1 represents fiction-based content and Module 2 represents non-fiction-based content. If a student takes a Keystone Exam multiple times, they also receive a highest total test score to date that represents the best combination of performance on Module 1 and Module 2. The advantage of this calculation is to combine Module-level performance from different administrations that may represent a higher total score and performance level (i.e., proficient or advanced). Both the administration-specific scaled scores (Module 1, Module 2, and total) and the highest total test scaled score are included on the score reports (see Figures 16–1 and 16–2).

In this technical report, both administration-specific and highest test scaled score to date information is presented. In general, information about raw scores always represents the administration-specific performance. Meaning, results represent the raw scores earned on that specific administration. On the other hand, information about scaled score performance represents the highest scaled score to date calculation unless otherwise specified. Moreover, information about performance level classifications also represents the performance level that corresponds to the highest total test scaled score to date. Meaning when scaled scores and performance levels are presented, results represent the best combination of Module 1 and Module 2 that yield the highest overall score and performance level classification.

PERFORMANCE LEVELS

Keystone Exams results are also reported using four performance levels: Below Basic, Basic, Proficient, and Advanced. The cut scores on the scaled-score metric (i.e., the lowest possible scaled score to enter the Basic, Proficient, and Advanced levels) were presented earlier in this report. However, the information is repeated below (Table 16–2) for convenience.

¹ This example is not an endorsement of conducting a trend analysis with just two years of results. Further, small differences may not be statistically or practically significant.

Table 16–2. Scaled Score Cuts for Each Performance Level by Content Area

Exam	Min	Scaled Score Cuts BB/B	Scaled Score Cuts B/P	Scaled Score Cuts P/A	Max
Algebra I	1,200	1,439	1,500	1,546	1,800
Biology	1,200	1,460	1,500	1,549	1,800
Literature	1,200	1,444	1,500	1,584	1,800

Note. BB = Below Basic; B = Basic; P = Proficient; and A = Advanced

Performance level descriptors (PLDs) are another way to attach meaning to the scaled-score metric. They associate precise quantitative ranges of scaled scores with verbal, qualitative descriptions of student status. While much less precise, the qualitative description of the levels is one way for parents and teachers to interpret the student scores. They are also useful in assessing the status of the school. The Pennsylvania General PLDs developed by Pennsylvania Department of Education (PDE) and teacher panels are given below. These are also included on student score reports.

- **Advanced:** Superior academic performance indicating an in-depth understanding and exemplary display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content.
- **Proficient:** Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content.
- **Basic:** Marginal academic performance indicating work approaching, but not yet reaching, satisfactory performance. Performance indicates a partial understanding and limited display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content. The student may need additional opportunities and/or increased student academic commitment to achieve the Proficient level.
- **Below Basic:** Inadequate academic performance indicating little understanding and minimal display of the skills included in the Keystone Exams Assessment Anchors and Eligible Content. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient level.

DESCRIPTION OF MODULE SCORES

Each of the Keystone Exams in Algebra I, Biology, and Literature contains two modules. A module score describes performance of a student, school, or district on a particular module (content standard defined in the exam). The following types of scores are provided for Keystone Exams at module level:

- Raw scores
- Scaled scores
- Performance levels

MODULE RAW SCORES

Raw scores at module and assessment anchor levels were reported in different summary reports. As described earlier, a raw score is the number of points a student earned over all the operational MC and CR items; it depends on the difficulty and length of the test form; and it should not be compared across administrations. In the summary reports, the school, district, and/or state median points earned were reported at module and assessment anchor levels. These raw scores can provide some diagnostic information when they are compared with the minimum estimated points needed to pass. The latter is calculated by summing the probabilities of a barely proficient student answering the items included in a module or assessment anchor correctly. The sum is rounded up to the nearest integer. The probability is derived using the Rasch models discussed in Chapter Twelve.

MODULE SCALED SCORES

The module scaled scores were provided in the individual student report. For the Keystone Exams, the module scaled score represents a student's achievement on each module. They can be compared across administrations because they are statistically equated. However, it is not advisable to compare scores across modules because each module contains varying item content and difficulty. This variation is also the reason the total scaled score is not the average of the two modules' scaled scores.

MODULE PERFORMANCE LEVELS

Based on the testing results at the module level, students can be classified as Passed or Not Passed. The derived scaled score cut is 1500 for both modules. This cut score is determined by panelists' recommendations for the proficient cut of the corresponding total test. Note that a student who does not pass a module can still be Proficient or above on the total test if the student performs very well on the other module. If a student is not proficient on the total test but passes one module, although it is recommended that this student take both modules during retesting, the student can choose to take just the non-passed module because the final score is based on the highest combination of module scores.

APPROPRIATE SCORE USE

INDIVIDUAL STUDENTS

Scaled scores on the Keystone Exams indicate a student's achievement with respect to the Keystone Exams Assessment Anchors and Eligible Content. Scaled scores are primarily used to determine student performance level classifications (i.e., a criterion-referenced inference). Scaled scores that are based on Rasch models are typically assumed to be of the interval type, so comparisons may be made on differences in scaled scores. If this assumption holds, then it would be safe to infer for Algebra I that the ability difference between 1410 and 1420 represents the same ability difference that separates 1550 and 1560. Scaled scores can also be used to compare the performance of an individual student to the performance of a similar demographic or subgroup at a school or district. Test score standard errors (discussed in Chapter Eighteen) should be considered.

GROUPS OF STUDENTS

Test results can be used to evaluate performance over time. Mean scaled scores can be compared across administrations within the same content area to indicate whether a student's performance is improving across years. Generally, such trend analyses benefit from using mean results from as many test administrations as possible. Different cohorts of students are used (i.e., the same student or students are not tracked across grade levels). All scores can be analyzed within the same content area for any single administration to determine which demographic or program group had, for example, the highest average performance or the highest percentage of students at or above Proficient.

Module scores can help evaluate academic areas for relative strengths or weaknesses. These module scores provide information to identify areas where further diagnosis is warranted. Generalizations from test results may be made to the specific content domain represented by the academic standards measured in the Keystone Exams. However, all instruction and program evaluations should include as much information from other sources as possible to provide a complete picture of performance.

CAUTIONS FOR SCORE USE

EXTREME ERROR FOR EXTREME SCORES

Student scores toward the minimum or maximum ends of the score range have very large standard errors of measurement (SEM) and, therefore, should be interpreted very cautiously. The maximum scaled score only provides a very rough estimate of a student's ability. For example, if a student achieved the maximum scaled score (i.e., 1800), it cannot be determined whether this student could have achieved an even higher scaled score. If the test were 10 items longer, a different estimate might have been obtained. Similarly, if the items on a new test are more difficult than the items on a previous administration, the maximum scaled score would likely be higher on the new test because it would take a greater level of achievement to answer the items correctly. In this manner, extreme scaled scores may vary from one administration to the next even if the number of test items does not change. The fluctuation of extreme scaled scores complicates the comparisons of students with scaled scores at the extreme ends of the score distribution. To minimize confusion and potential misinterpretation, the minimum and maximum

scaled scores possible on the Keystone Exams have been fixed (see Table 16–2) so they do not change between administrations.

UNIQUE SCALE FOR EACH CONTENT AREA

Scaling was conducted for each content-area exam separately. Therefore, the scaled scores should be interpreted only within each content area. The scaled scores are not status indicators in the same sense as percentile ranks (or scales that are essentially transformations of percentile ranks) and therefore cannot be used to profile relative strengths and weaknesses across content areas. As an example, the scaled scores of 1450 in Algebra I and 1400 in Biology gained by a student do not necessarily imply that the student performed better in Algebra I than in Biology.

USING KEYSTONE EXAMS RESULTS FOR OTHER PURPOSES

Other uses or inferences based on Keystone Exams results may or may not be valid as the validity evidence and arguments provided in Chapter Nineteen may not necessarily support other score uses and interpretations. According to the *Standards for Educational and Psychological Tests* (AERA, APA, & NCME, 2014), if a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary. Finally, a universal caveat for any test's result is that it should not be used for placement and educational planning alone. Instead, other information about the student (e.g., other test performance data) should be included.

REPORT DEVELOPMENT

Several months prior to the first release of reports for the Keystone Exams, PDE and DRC conducted focus groups with Pennsylvania educators and parents/guardians. In the focus groups, educators and parents/guardians provided feedback on sample report for the Keystone Exams. Feedback from the focus groups was used to inform the design and content of the Keystone individual and summary reports. The focus groups targeted educator and parent/guardian constituencies in three geographic regions of the state—the Pittsburgh area, the Harrisburg area, and the Philadelphia area.

Two preliminary educator groups were convened in Harrisburg on November 15 and 17, 2010. These groups, totaling 34 educators, reviewed the student report and provided feedback using both a survey and group discussion. Substantive changes to the individual student report were made as a result of these meetings, with two different versions of the report emerging from these reviews. These two groups did not review the summary reports.

A second set of focus groups were conducted in December 2010 to review the updated reports. For the December meetings there were 35 panelists (22 educators & 13 parents) for six focus groups in Pittsburgh (December 3), Harrisburg (December 6), and King of Prussia (December 7). The three educator groups reviewed the two versions of the student report and the one version of the school summary report. The three parent groups reviewed the two versions of the student report.

Feedback from these two focus groups was taken into consideration during final report development. For more information about the focus groups, please refer to the *Keystone Exams Score Report Focus Group Findings* (Pennsylvania Department of Education, 2011).

REPORTS

The following score reports are provided to students, schools, and districts for the Keystone Exams in Algebra I, Biology, and Literature:

- Parent Letter
- Individual student report
- School summary report
- District summary report
- State summary report
- Report interpretation guide

PARENT LETTER

Parent letters were delivered to Pennsylvania districts when district files were posted after each Keystone administration. This score report provided parents and students with their first glimpse of performance on the Keystone Exams. This report provides results at the student level.

INDIVIDUAL STUDENT REPORT

A student report is provided for all students who took the Keystone Exams. Two copies of the individual student report for all Keystone Exams were sent to each school district and charter school for distribution to parents, teachers, guidance counselors, and/or principals. School districts and charter schools may publish the results of the Keystone Exams school-level reports. This report is a two-page color document that provides the types of scores explained earlier in this chapter. Screenshots of the two pages from a sample individual student report are provided in Figures 16–1 and 16–2. Figure 16–2 displays the second page of the Individual Student Report, including scale score on each module, and the highest total test scale score to date.

Figure 16–1. Page 1 of the Individual Student Report



Student Report

Student Name: SAMPLE STUDENT 1


PA Student ID:

School:

District:

Test Date:

Grade:

Content Area:  Biology

Student's Keystone Exam Result			
		Goal Range	
Below Basic	Basic	Proficient	Advanced
		✓	

Dear Parent/Guardian:

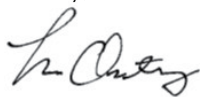
On behalf of the Pennsylvania Department of Education, I would like to thank you for taking the time to review this report. Academic success begins at home. Children with parents/guardians who are engaged in their child's academics are more likely to enjoy the learning process and succeed.

This report provides information about your child's recent performance on a Pennsylvania test known as the Keystone Exam. On this page, you can see your child's overall performance – below basic, basic, proficient or advanced.

Within this report, you will find specific information about your child's performance on the Biology Keystone Exam. It displays your child's Highest Total Test Scale Score to Date for Module 1 and Module 2. Module 1 assesses Cells and Cell Processes, and Module 2 assesses Continuity and Unity of Life.

For detailed information about the Keystone exams, please visit the Pennsylvania Department of Education's Standards Aligned System website at www.pdesas.org, or contact your child's school.

Sincerely,



Noe Ortega
Acting Secretary of Education

About the Keystone Exams

The Keystone Exams are end-of-course assessments designed to evaluate student performance on academic content. The purpose of the Algebra I, Biology, and Literature Keystone Exams is to measure student, educator, and school accountability. Keystone Exams are designed to be administered to students at or near the end of a Keystone-related course. Students' results are banked until their junior year for accountability purposes. Keystone Exams are one component of Pennsylvania's system of high school graduation requirements affecting students in the class of 2023 and beyond.

These tests were developed by Pennsylvania educators and were aligned to the standards adopted by the Pennsylvania State Board of Education. The results help students, parents, and educators understand how well rigorous expectations for student achievement in core subject areas are being met.

A Report Interpretation Guide is available at www.education.pa.gov. Type "Keystone Interpretation Guide" in the search field or consult the local school district or school.

www.pdesas.org

Biology



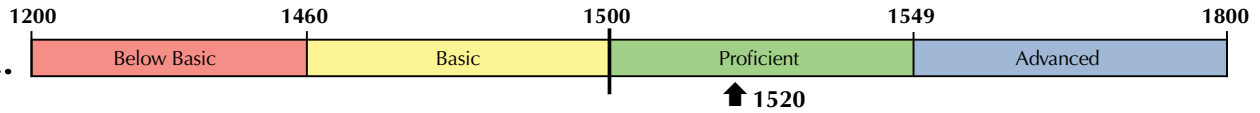
pennsylvania

DEPARTMENT OF EDUCATION

Figure 16–2. Page 2 of the Individual Student Report

Performance Level on Total Test

Student’s highest total test scale score to date is indicated by the (↑). If this student were to test again under similar circumstances, his/her score would likely remain in the following range: 1507-1533.



Inadequate academic performance that indicates little understanding and minimal display of the skills included in the Keystone Exams Assessment Anchors & Eligible Content. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient level.

Marginal academic performance, work approaching, but not yet reaching, satisfactory performance. Performance indicates a partial understanding and limited display of the skills included in the Keystone Exams Assessment Anchors & Eligible Content. The student may need additional instructional opportunities and/or increased student academic commitment to achieve the Proficient level.

Satisfactory academic performance indicating a solid understanding and adequate display of the skills included in the Keystone Exams Assessment Anchors & Eligible Content.

Superior academic performance indicating an in-depth understanding and exemplary display of the skills included in the Keystone Exams Assessment Anchors & Eligible Content.

Biology



SAMPLE STUDENT 1

	Module 1 Cells and Cell Processes			Module 2 Continuity and Unity of Life			Total Test	
	Result	Scale Score	Test Date	Result	Scale Score	Test Date	Scale Score	Performance Level
Highest Total Test Scale Score to Date ¹	Passed	1542		Passed	1501		1520	Proficient
Scores of Three Most Recent Test Events	Passed	1542		Passed	1501			

¹The highest total test scale score to date is the highest score computed from all possible combinations of module 1 and module 2. The total scale score is not the simple average of the module scale scores. Rather, it is weighted by the relative difficulty of questions from each module.

SUMMARY REPORTS

Summary reports are provided at the school, district, and state levels. These reports contain summary information about the percentage of students in each of the four performance levels. Raw scores are also provided by assessment anchor to allow schools or districts to identify strengths and weaknesses at the content-strand level.

REPORT INTERPRETATION GUIDE

A report interpretation guide is provided to help parents and other Keystone Exams stakeholders better understand test-result information presented in the individual student report. The report interpretation guide can be found on the PDE website at www.education.pa.gov.

CHAPTER SEVENTEEN: OPERATIONAL TEST STATISTICS

This chapter presents various summary statistics for the total-test scores based on the final data file described in Chapter Nine. Related information covered elsewhere in this report includes the item-level statistics that were presented in Chapters Eleven (classical item statistics) and Twelve (Rasch item statistics). Please refer to these chapters for additional consideration as item difficulty distributions can affect total score distributions.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

PERFORMANCE LEVEL STATISTICS

Table 17–1 presents performance level percentages by test administration, content area, and student type. Performance level classifications are based on the highest total scaled score to date. As can be seen from the table, the overall percentage in each performance level varied from administration to administration, depending on the ratio of the first-time testers and retesters. In general, retesters had a lower percentage of students in the Proficient and Advanced levels than first-time testers.

Table 17–1A. Performance Level Percentages: All Testers

Administration	Content Area	N	Below Basic (%)	Basic (%)	Proficient (%)	Advanced (%)
Winter	Algebra I	13,200	19.5	45.5	22.8	12.3
Winter	Biology	12,144	21.5	29.5	28.8	20.2
Winter	Literature	11,722	12.3	25.7	52.7	9.4
Spring	Algebra I	109,710	25.6	38.5	22.4	13.5
Spring	Biology	100,976	24.4	30.8	26.7	18.0
Spring	Literature	97,928	14.7	28.1	46.7	10.5
Summer						
Summer						
Summer						

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Table 17–1B. Performance Level Percentages: First-time Testers

Administration	Content Area	N	Below Basic (%)	Basic (%)	Proficient (%)	Advanced (%)
Winter	Algebra I	10,746	20.8	40.3	24.2	14.7
Winter	Biology	11,203	21.0	26.9	30.2	21.9
Winter	Literature	11,042	11.8	23.8	54.4	10.0
Spring	Algebra I	103,295	25.7	37.2	22.9	14.2
Spring	Biology	98,419	24.2	30.3	27.1	18.5
Spring	Literature	96,409	14.5	27.8	47.1	10.6
Summer						
Summer						
Summer						

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Table 17–1C. Performance Level Percentages: Retesters

Administration	Content Area	N	Below Basic (%)	Basic (%)	Proficient (%)	Advanced (%)
Winter	Algebra I	2,454	13.9	68.1	16.6	1.4
Winter	Biology	941	27.1	60.0	12.1	0.7
Winter	Literature	680	20.0	55.1	24.4	0.4
Spring	Algebra I	6,415	22.6	60.5	14.8	2.1
Spring	Biology	2,557	34.6	51.3	12.6	1.4
Spring	Literature	1,519	25.2	50.3	23.2	1.3
Summer						
Summer						
Summer						

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

SCALED SCORES

Table 17–2 provides the scaled score means and standard deviations by test administration, content area, and student type (all testers, first-time testers and retesters). The descriptive statistics are based on the highest total scaled score to date following each administration. As can be seen from the table, first-time testers earned higher scaled scores than retesters.

Table 17–2. Means and Standard Deviations of Scaled Scores

Administration	Content Area	All Testers Mean	All Testers SD	First-Time Testers Mean	First-Time Testers SD	Retesters Mean	Retesters SD
Winter	Algebra I	1484.5	52.4	1486.9	55.8	1473.8	31.4
Winter	Biology	1503.1	51.3	1505.5	52.3	1475.4	24.7
Winter	Literature	1513.5	57.4	1515.8	57.7	1476.9	36.9
Spring	Algebra I	1480.1	58.6	1480.9	59.6	1467.6	36.8
Spring	Biology	1498.6	52.5	1499.3	52.8	1472.9	28.9
Spring	Literature	1509.7	60.9	1510.3	60.9	1473.5	44.3
Summer							
Summer							
Summer							

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

RAW SCORES

SUMMARY STATISTICS

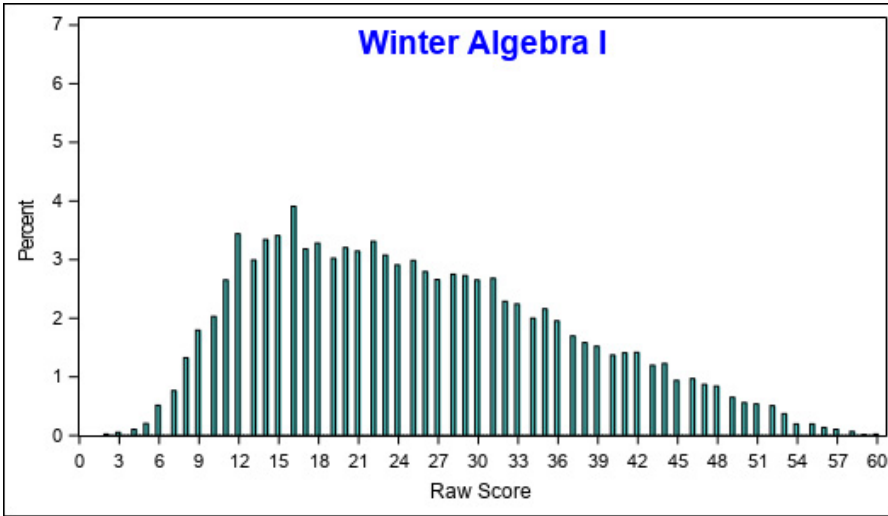
The reader is referred to Appendix M to review summary statistics for the operational raw scores. The statistics reported include number of points possible (Pts.), number of items (Len.), number of students tested (*N*), mean raw score (Mean), standard deviation of raw score (SD), reliability (*r*), and traditional standard error of measurement (SEM).

SCORE DISTRIBUTIONS

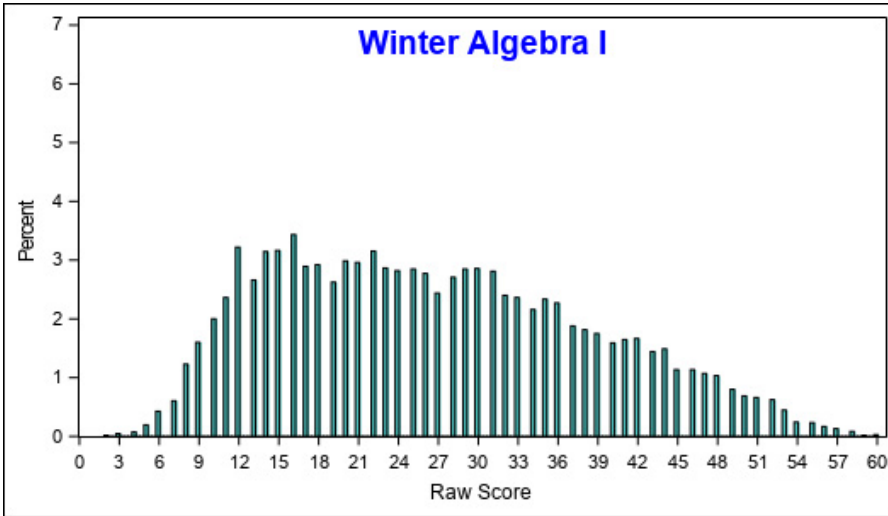
Raw score distributions for each administration and content area are provide in Figure 17–1. Graphs are shown for all test-takers, first-time test-takers, and retesters . As can be seen from the graphs, the retesters tended to score lower than the first-time testers. For all testers, the distribution of raw scores for Algebra I and Biology are positively skewed while the distributions of raw scores for Literature are negatively skewed. The summer raw score distributions are not presented due to the extended spring 2021 testing windows and cancellation of summer 2021 Keystone exams.

Figure 17–1. Raw Score Distributions

All Testers



First-Time Testers



Retesters

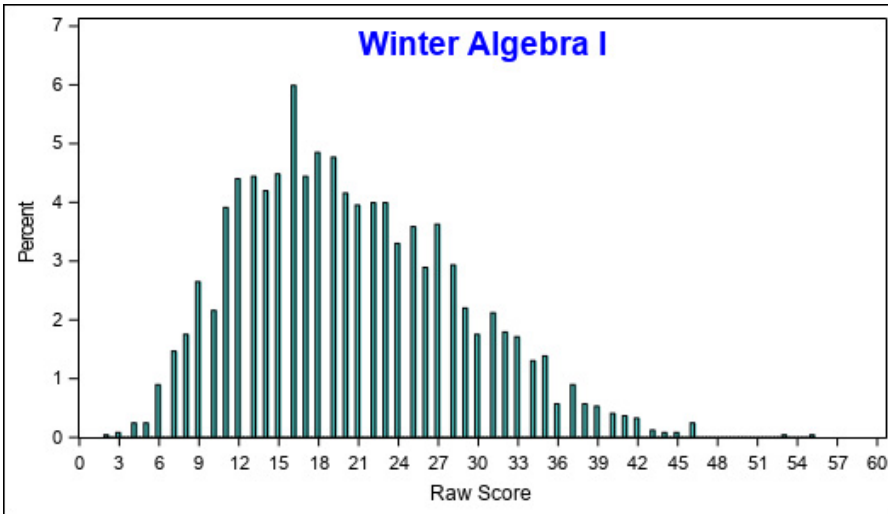
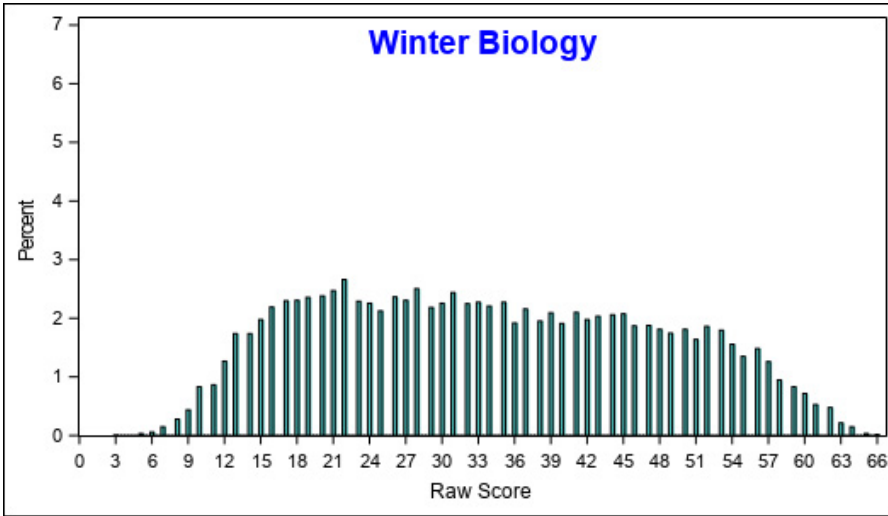
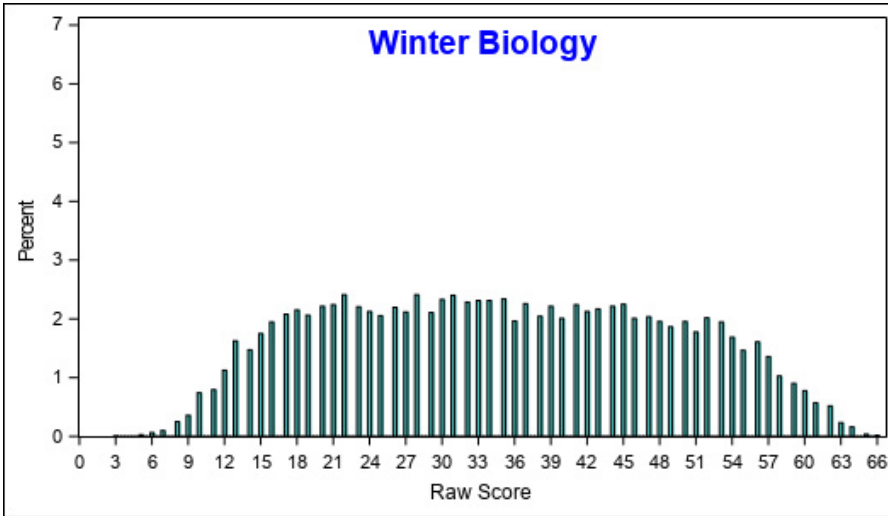


Figure 17–1 (continued). Raw Score Distributions

All Testers



First-Time Testers



Retesters

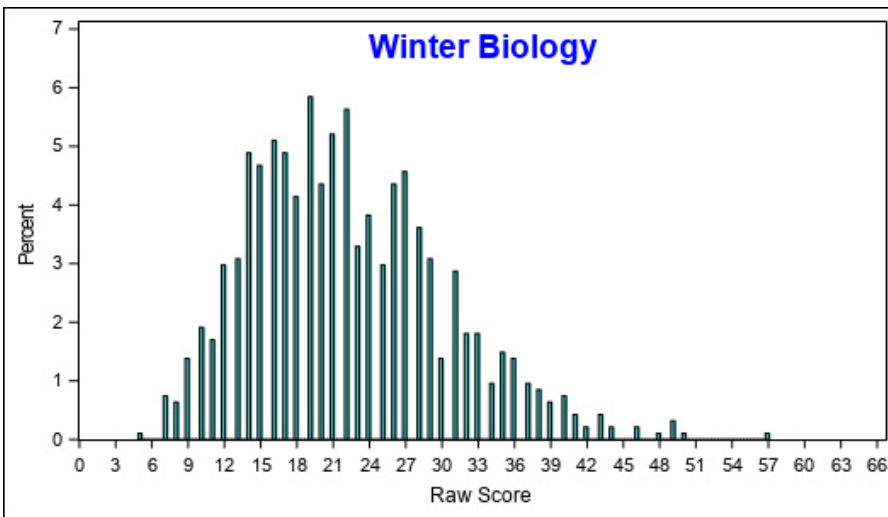
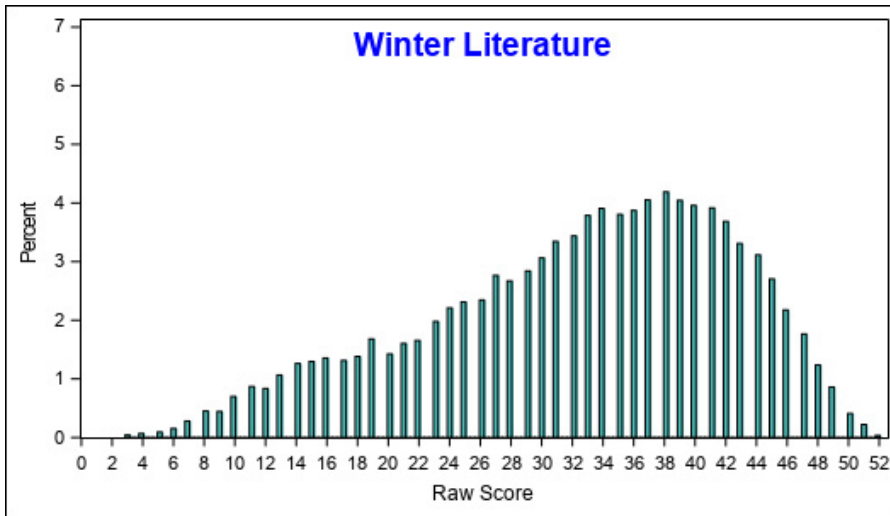
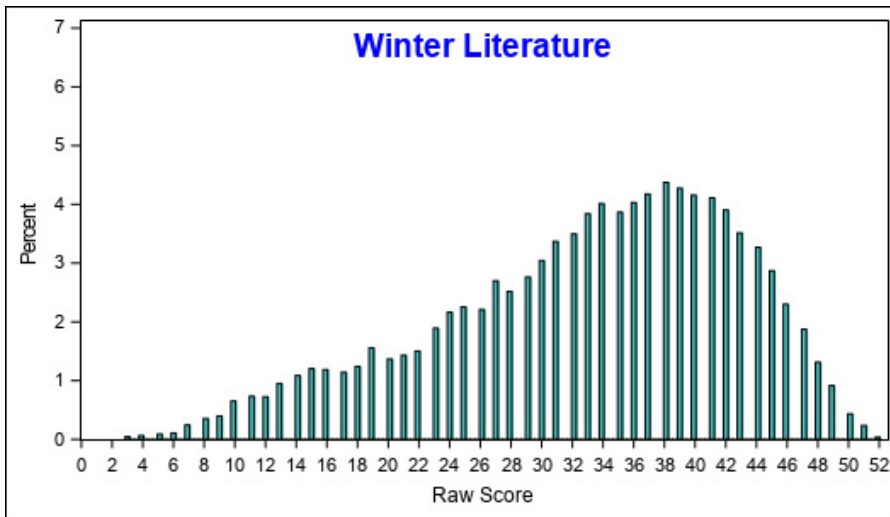


Figure 17–1 (continued). Raw Score Distributions

All Testers



First-Time Testers



Retesters

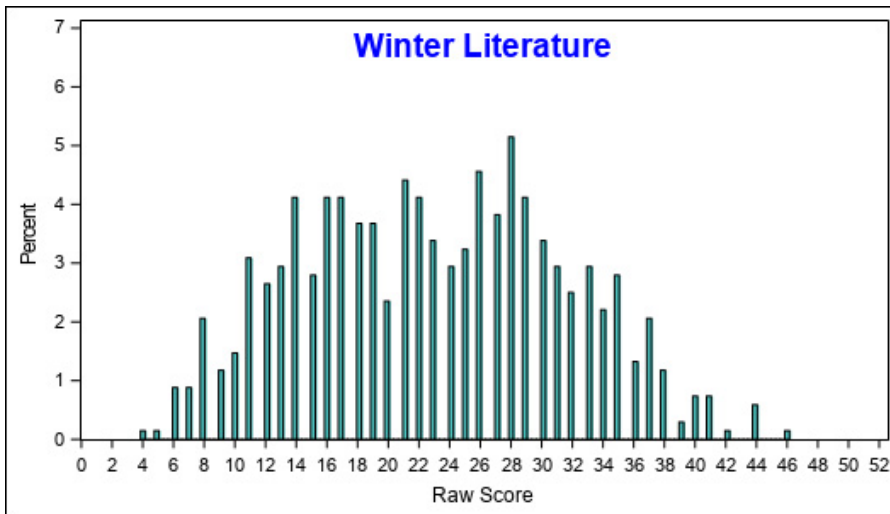
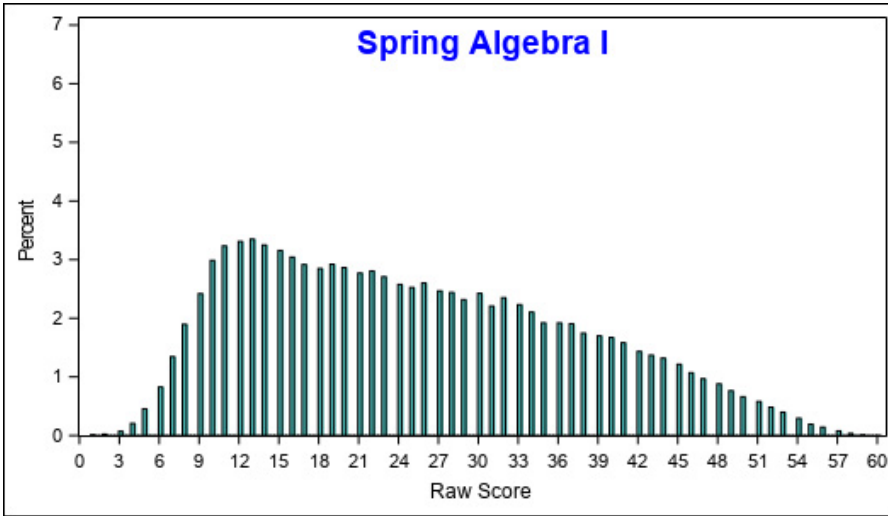
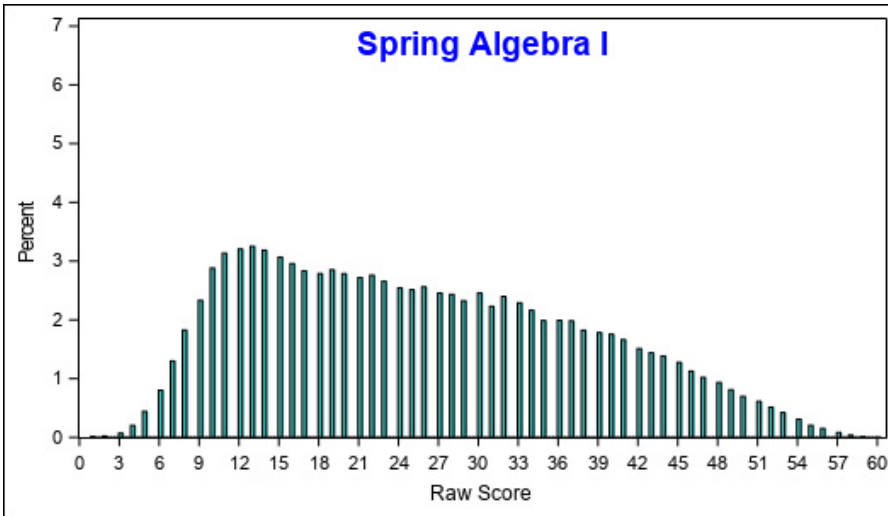


Figure 17–1 (continued). Raw Score Distributions

All Testers



First-Time Testers



Retesters

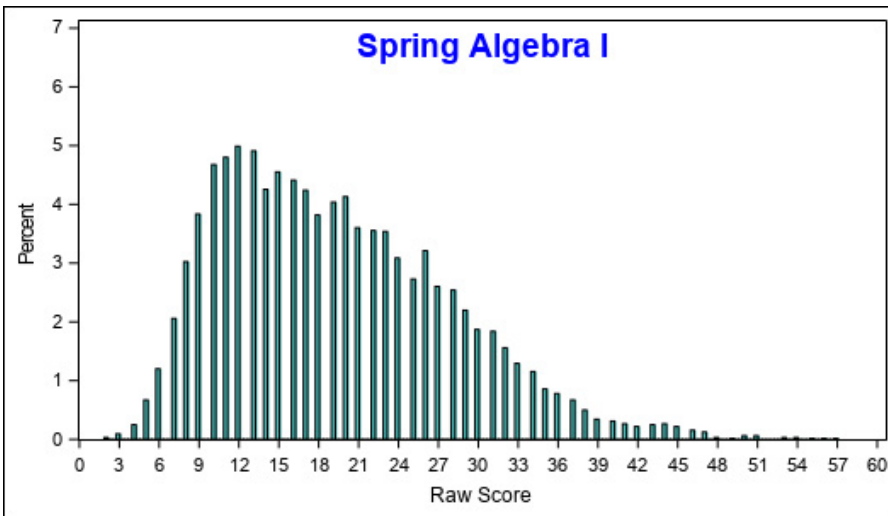
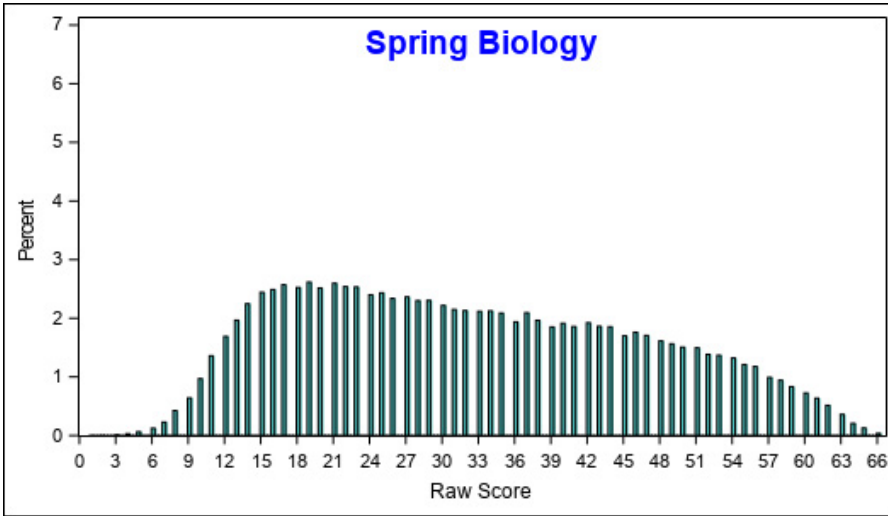
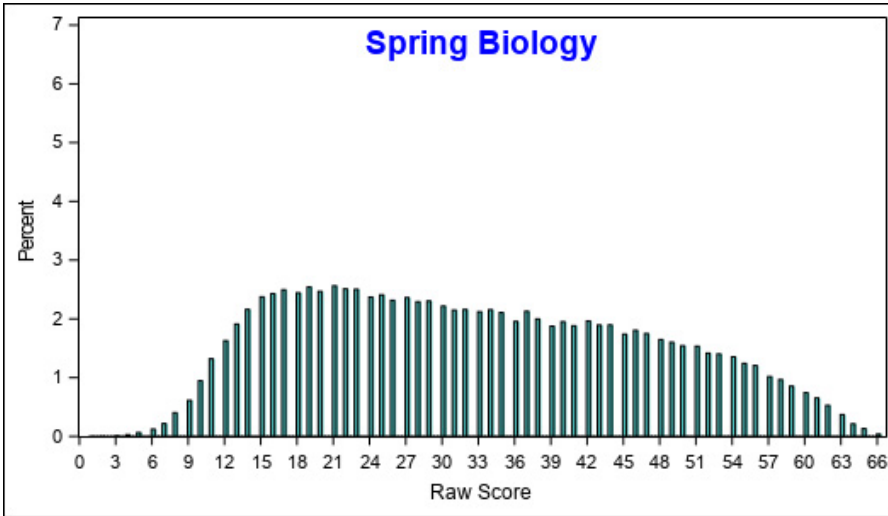


Figure 17–1 (continued). Raw Score Distributions

All Testers



First-Time Testers



Retesters

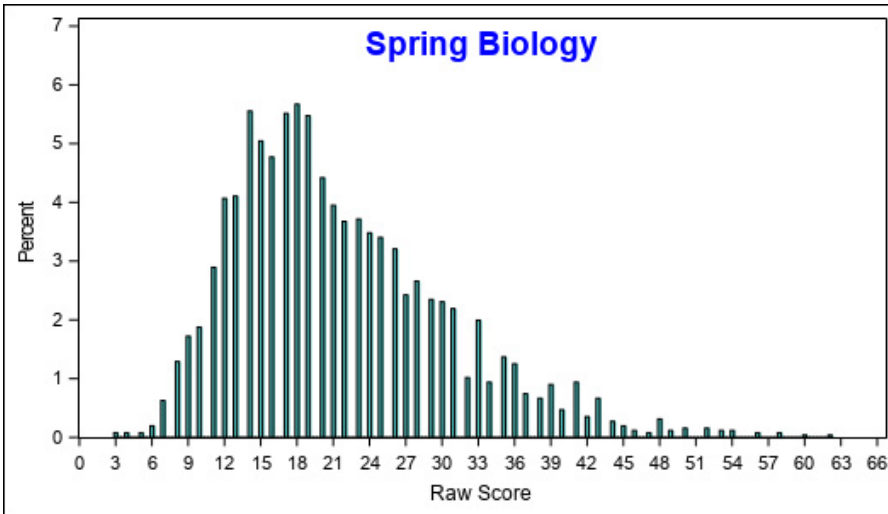
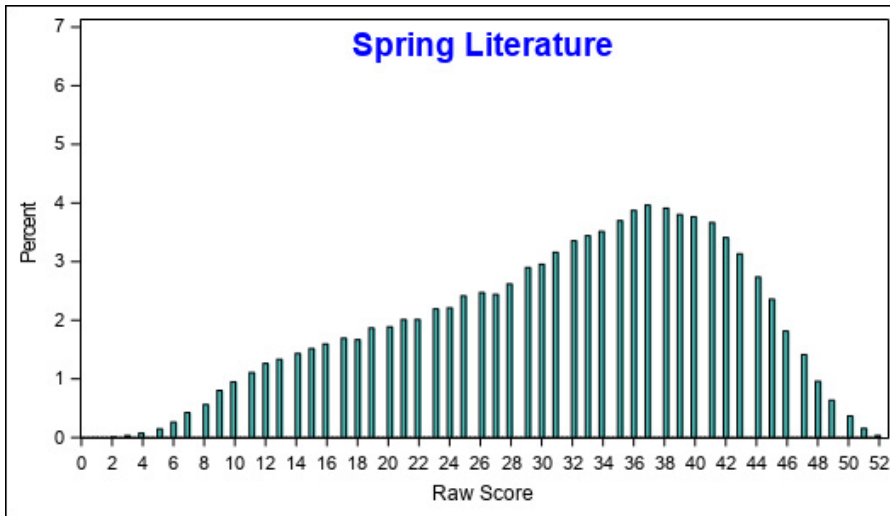
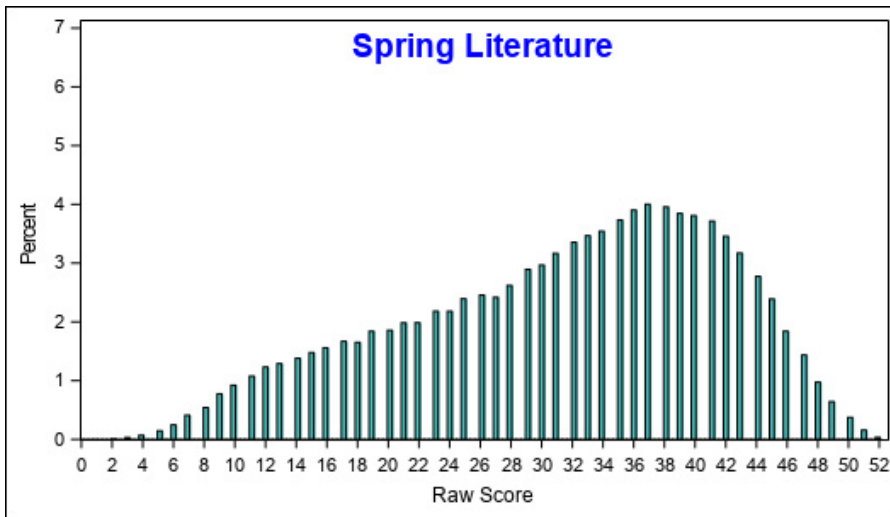


Figure 17–1 (continued). Raw Score Distributions

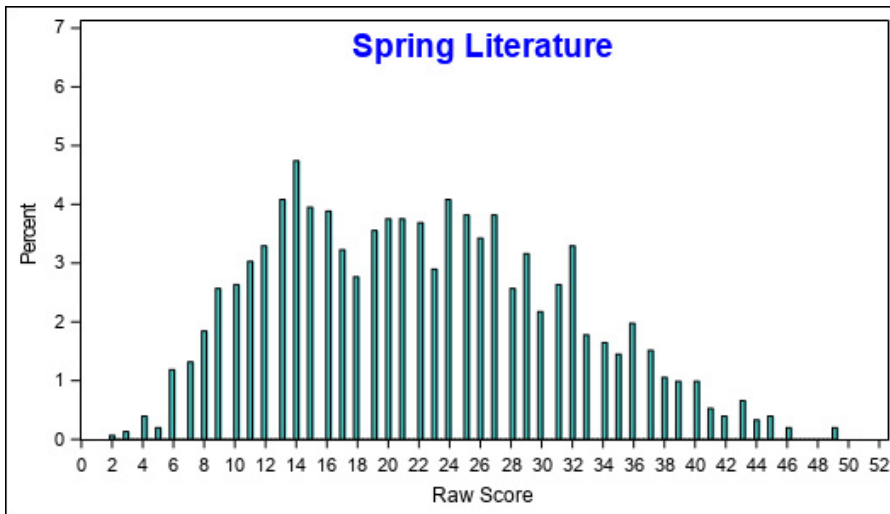
All Testers



First-Time Testers



Retesters



CHAPTER EIGHTEEN: RELIABILITY

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

This chapter addresses the reliability of Pennsylvania Keystone Exams test scores. According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), reliability refers to

the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group (p. 222).

Frisbie (2005) highlighted several elements of this definition. First, reliability is a property of test scores, not of a test itself. Many may appreciate this distinction, but in casual usage, individuals frequently make reference to a reliable test. While reliability concerns test scores (and not the test specifically), it's important to appreciate the fact that test scores can be affected by characteristics of the instrument. For example, all other things being equal, tests with more items/points tend to be more reliable than tests with fewer items/points. Second, reliability coefficients are group specific. Reliabilities tend to be higher in populations that are more heterogeneous and lower in populations that are more homogeneous. Consequently, both test length and population heterogeneity should be considered when evaluating reliability.

There are other reliability considerations that may be less evident from the definition above yet are still important for test users to understand. While freedom from measurement error is highlighted in the definition, reliability is specifically concerned with random sources of error. Indeed, the degree of inconsistency due to random error sources is what determines reliability: less consistency is associated with lower reliability and more consistency is associated with higher reliability. Of course, systematic error sources also exist. These can artificially increase reliability and decrease validity. (Validity is further discussed in Chapter Nineteen.)

Another noteworthy issue is that multiple sources of error exist (e.g., the day of testing, the items used, the raters who score the items). However, most widely used reliability indices only reflect a single type of error. Consequently, it is important for test users to understand which specific type of error is being considered in a reliability study, and equally, if not more importantly, which types are not.

Understanding the distinction between relative error and absolute error is important because many reliability indices only reflect relative error. Relative error is of interest whenever the relative ordering of individuals with respect to their test performance is of interest. When specific score values are considered important (e.g., if cut scores are used), then absolute error is of interest, too. Generally, there is more error variance when considering the absolute scores of examinees, which, in turn, suggests lower reliability. Understanding examinee rank-order stability is also important; however, such stability might be well achieved even when the specific score values are considerably different.

As the above discussion suggests, reliability is a complex, nonunitary notion that cannot be adequately represented by a single number. There are several reliability indices available, and these may not provide the same results (Frisbie, 2005). The remainder of this chapter covers the following:

- Reliability coefficients and their interpretation
- Unconditional and conditional standard errors of measurement
- Decision consistency
- Rater agreement

RELIABILITY INDICES

As shown below, the reliability coefficient expresses the consistency of test scores as the ratio of true score variance to total score variance. The total variance contains two components: variance in true scores and variance due to the imperfections in the measurement process. Put differently, total variance equals true score variance plus error variance.¹

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Reliability coefficients indicate the degree to which differences in test scores reflect true differences in the attribute being tested rather than random fluctuations. Total test score variance (i.e., individual differences) is partly due to real differences in the attribute (true variance) and partly due to random error in the measurement process (error variance).

Reliability coefficients range from 0.0 to 1.0. If all test score variance were true, the index would equal 1.0. The index would be 0.0 if none of the test score variance were true. Such scores would be pure random noise—that is, all measurement error. If the index had a value of 1.0, scores would be perfectly consistent—that is, contain no measurement error. Although values of 1.0 are never achieved in practice, it is clear that larger coefficients are more desirable as they indicate that test scores are less influenced by random error. (How big is big enough and how small is too small are issues considered in a later section.)

As noted in the introduction, there are several different indices that can be used to estimate this ratio. One approach is referred to as internal consistency, which is derived from analyzing the performance consistency of individuals over the items within a test. As discussed below, these internal consistency indices do not take into account other sources of error, such as day-to-day variations (e.g., student health, testing environment) or rater inconsistency.

COEFFICIENT ALPHA

Although a number of reliability indices exist, perhaps the most frequently reported for achievement tests is coefficient alpha. Consequently, this index is the one reported for the Keystone Exams (see the column with title “r” in Appendix M). Alpha indicates the internal consistency over the responses to a set of items measuring an underlying trait, in this case, academic achievement, in content areas such as Algebra I, Biology, and Literature.

Alpha is an internal consistency index. It can be conceptualized as the extent to which an exchangeable set of items from the same domain would result in a similar rank ordering of students. Note that relative error is reflected in this index. Variation in student performance from one sample of items to the next should be of particular concern for any achievement test user. Consider two hypothetical vocabulary tests intended for the same group of students. Each test contains different sets of unique words that are believed to be randomly equivalent, perhaps like the ones shown below:

Table 18–1. Two Hypothetical Vocabulary Tests

Test One	Test Two
Abase	Abate
Boon	Bilk
Capricious	Circuitous
Deface	Debase
....
Zealous	Zenith

¹ A Covariance term is not required as true scores and error are assumed to be uncorrelated in classical test theory.

If a representative group of students could take both of these tests, the correlation between the scores obtained would represent the parallel-forms reliability of the test scores. However, such data-collection designs are impractical in large-scale settings, and fatigue and practice effects are likely to affect the results. Internal-consistency reliability indices arose in part to provide reliability measures using the data from just a single test administration. So, if students only took Test One and the coefficient alpha index for those test scores were high, this would suggest that Test Two would provide a very similar rank ordering of the students if they had taken it instead. If coefficient alpha were low, dissimilar rank orderings would likely be observed—again, relative-error variance is reflected in alpha.

FORMULA

Consider the following data matrix representing the scores of persons (rows) on items (columns):

Table 18–2. Person × Item Score (X_{pi}) Infinite (Population-Universe) Matrix

Person	Item 1	Item 2	... l	... k
1	Y_{11}	Y_{12}	... Y_{1i}	... X_{1k}
2	Y_{21}	Y_{22}	... Y_{2i}	... X_{2k}
.....				
P	Y_{p1}	Y_{p2}	... Y_{pi}	... X_{pk}
.....				
N	Y_{N1}	Y_{N2}	... Y_{Ni}	... X_{Nk}

Note. Adapted from Cronbach and Shavelson (2004).

Then, a general computational formula for alpha is as follows:

$$\alpha = \frac{N}{N - 1} \left(1 - \frac{\sum_{i=1}^N \sigma_{Y_i}^2}{\sigma_X^2} \right)$$

where N is the number of parts (items or testlets), σ_X^2 is the variance of the observed total test scores, and $\sigma_{Y_i}^2$ is the variance of part i .

FURTHER INTERPRETATIONS

RULES OF THUMB

Which reliability values are considered high enough? Which values are considered too low? Although frequently asked for, any rules of thumb for interpreting the magnitude of reliability indices are mostly arbitrary. Another approach is to research the reliabilities from similar testing instruments to see what values are commonly observed. For the Keystone Exams, comparisons to tests of similar lengths that were administered to similar student populations from other large-scale assessment programs would be relevant. For many other state assessment programs, reliabilities in the low 0.90s are usually the highest ever observed, and reliabilities in the high 0.80s are very common.

The lower a given reliability coefficient, the greater the potential for over-interpretation of the associated results. As suggested earlier, there is no firm guideline regarding how low is too low. However, as an informative point of reference, a reliability coefficient of 0.50 would mean that there is as much error variance as true-score variance in the scores.

IS ALPHA A LOWER LIMIT TO RELIABILITY?

According to Brennan (1998), the conventional wisdom that coefficient alpha is a lower limit to reliability is based largely on a misunderstanding. In reflecting on the 50th anniversary of his seminal 1951 article, Cronbach—in Cronbach and Shavelson (2004)—expressed similar misgivings about this conventional wisdom:

one could argue that alpha was almost an unbiased estimate of the desired reliability. . . the *almost* in the preceding sentence refers to a small mathematical detail that causes the alpha coefficient to run a trifle lower than the desired value. This detail is of no consequence and does not support the statement made frequently in textbooks or in articles that alpha is a lower value to the reliability coefficient. That statement is justified by reasoning that starts with the definition of the desired coefficient as the expected consistency among measurements that had a higher degree of parallelism than the random parallel concept implied.

The assumptions for three common parallelism models are presented in Table 18–3. Alpha’s assumptions come from the Essentially-Tau-Equivalent model, which does not require equal means or equal variances across test parts. Based on this, Brennan (1998) asserts that the lower-limit issue, as conceptualized by many, provides an answer to a question that is of minimal importance. Reframed differently, the goal of selecting a reliability coefficient is not to find the one that provides the highest coefficient, but the one that most accurately reflects the test data under study.

It is important to note that there are factors encountered in practice that may legitimately make coefficient alpha an underestimate of reliability. However, there are also factors that might make coefficient alpha an overestimate of reliability. Both possibilities are discussed further below and generally arise when the Essentially-Tau-Equivalent assumptions are strained.

Table 18–3. Summary of Expectations/Observable Relationships for Different Parallelism Models

Relationship	Degree of Measurement Parallelism* Classically Parallel	Degree of Measurement Parallelism* Essentially-Tau Equivalent	Degree of Measurement Parallelism* Congeneric
Content Similarity	Yes	Yes	Yes
Equal Means across Parts	Yes	No	No
Equal Variances across Parts	Yes	No	No
Equal Covariances across Parts	Yes	Yes	No
Equal Covariances with other Variables	Yes	Yes	No

*Note. Other models exist but are not considered here due to their limited application in practice.

BIASES THAT MIGHT MAKE ALPHA AN UNDERESTIMATE OF RELIABILITY

There are factors that might negatively bias coefficient alpha, making the apparent reliability lower than it may actually be. In practice, two situations frequently encountered that might cause this include tests that are composed of mixed item types (e.g., MC and CR items) and tests that include a planned stratification of the test items according to topics or subdomains.

Although both situations strictly violate the assumptions used in deriving the coefficient alpha (i.e., the tests are not based on equal part lengths in the former case and are not randomly parallel in the latter case), neither necessarily guarantees that the reliability will be markedly lower. In the latter case, reliability will be underestimated only when strand items are homogeneous enough for the average covariance within strata to exceed the average covariance between strata. Although both are potential influences for the Keystone Exams, the total test score reliabilities (i.e., r) reported in Appendix M ranged from 0.89 to 0.92, indicating highly consistent test scores for these instruments.

BIASES THAT MIGHT MAKE ALPHA AN OVERESTIMATE OF RELIABILITY

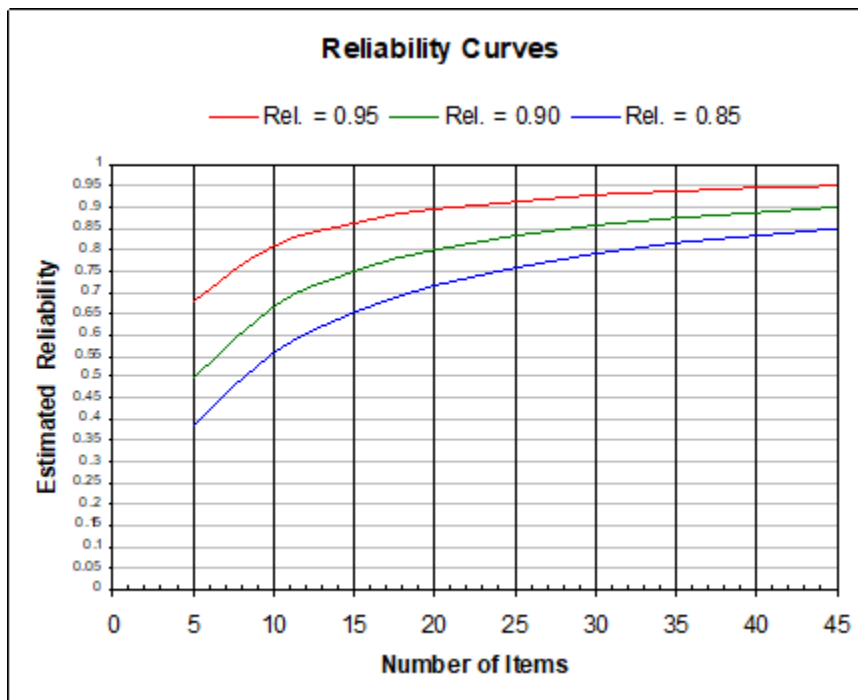
As emphasized in earlier sections, coefficient alpha only takes into account measurement error that arises from the selection of items used on a particular test form. There are other sources of random inaccuracy. One is due to the occasion of testing. Examples of other various random conditions that might affect students on any particular testing occasions include illness, fatigue, and anxiety. Also, when a test includes CR items, another source of random fluctuation can be the CR item scorers. In a sense, alpha may be positively biased because it does not take into account these other important sources of random error. Actually, any internal consistency reliability index might understate the overall problem of measurement error because they all ignore such sources of random error.

Another positive bias can occur when items are associated (clustered) with a common stimulus. Item bundles and testlets are other frequently used terms for this situation. One concrete example is when multiple reading comprehension items are associated with a common passage selection. Again, such a situation does not guarantee that the reliability estimate will be markedly affected, but the potential exists.

MODULE SCORE RELIABILITY

As noted in the introduction, reliabilities tend to go up in value with an increase in test length and go down in value with a decrease in test length. Figure 18–1 illustrates this relationship for a hypothetical 45-point test with three total score reliabilities: 0.95, 0.90, and 0.85. As an example, the curve for reliability equal to 0.90 suggests that a 15-item module would be expected to have a score reliability of 0.75. The use of the Spearman-Brown prophecy formula assumes all items are exchangeable, which in practice they may not be. While such a chart may not perfectly model actual module correlations, the intent is only to illustrate the substantial impact that limited numbers of items impact that module-score reliability. One should not be surprised that module scores with more points tend to show higher reliability coefficients and those with fewer points tend to show lower reliability coefficients.

Figure 18–1. Example of the Relationship Between Test Length and Reliability



As can be seen in Appendix M, the reliability coefficients at the module level were always less than the total score reliabilities. The reliability of module scores were greater than 0.83, which is likely because the number of items at the module level is half of the number of items in the total test.

STANDARD ERROR OF MEASUREMENT

The reliability coefficient is a unit-free indicator that reflects the degree to which scores are free of measurement error. It always ranges between 0.0 and 1.0 regardless of the test's scale. Reliability coefficients best reflect the extent to which measurement inconsistencies may be present or absent in a group. However, they are not that useful for helping users interpret test scores. The standard error of measurement (SEM) is another indicator of test score precision that is better suited for determining the effect of measurement inconsistencies on the scores obtained by individual examinees. This is particularly so for conditional SEMs (CSEM) discussed further below.

TRADITIONAL STANDARD ERROR OF MEASUREMENT

A precise, theoretical interpretation of the SEM (see Appendix M) is somewhat unwieldy. A beginning point for understanding the concept is as follows. If everyone being tested had the same true score,² there would still be some variation in observed scores due to imperfections in the measurement process, such as random differences in attention during instruction or concentration during testing or the sampling of test items. The standard error is defined as the standard deviation³ of the distribution of observed scores for students with identical true scores. Because the SEM is an index of the random variability in test scores in actual score units, it represents very important information for test score users.

The SEM formula is provided below.

$$SEM = SD\sqrt{1-reliability}$$

It indicates that the value of the SEM depends on both the reliability coefficient and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value), the SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value), the SEM would be 0.0. In other words, a perfectly reliable test has no measurement error (Harvill, 1991). Additionally, the value of the SEM takes the group variation (i.e., score standard deviation) into account. Consider that an SEM of 3.0 on a 10-point test would be very different from an SEM of 3.0 on a 100-point test.

TRADITIONAL SEM CONFIDENCE INTERVALS

The SEM is an index of the random variability in test scores in actual score units, which is why it has such great utility for test score users. SEMs allow statements regarding the precision of individual tests scores. SEMs help place reasonable limits (Gulliksen, 1950) around observed scores through construction of an approximate score band. Often referred to as confidence intervals, these bands are constructed by taking the observed scores, X , and adding and subtracting a multiplicative factor of the SEM. As an example, students with a given true score will have observed scores that fall between ± 1 SEM about two-thirds of the time.⁴ For ± 2 SEM confidence intervals, the percentage increases to about 95 percent.

FURTHER INTERPRETATIONS

ONE SEM FOR ALL TEST SCORES

The SEM approach described above only provides a single numerical estimate for constructing the confidence intervals for examinees regardless of their score levels. In reality, however, such confidence intervals vary according to one's score. Consequently, care should be taken when using the SEM for students with extreme scores. An alternate approach that conditions the SEM on a student's score estimate is described in the next sections.

GROUP SPECIFIC

As noted in the introduction, reliabilities are group specific. The same is true for SEMs because both score reliabilities and score standard deviations vary across groups.

² True score is the score the person would receive if the measurement process were perfect.

³ The standard deviation of a distribution is a measure of the dispersion of the observations. For the normal distribution, about 16 percent of the observations are more than one standard deviation above the mean.

⁴ Some prefer the following interpretation: if a student were tested an infinite number of times, the ± 1 SEM confidence intervals constructed for each score would capture the student's true score 68 percent of the time.

RAW SCORE METRIC

The SEM approach is calculated using raw scores, and as such, the resulting confidence interval bands are on the raw score metric. Error bands on the scaled score metric are considered in the next section.

TYPE OF ERROR REFLECTED

The interpretation of the SEM should be driven by the type of score reliability that underpins it. So, the Keystone Exams SEMs involve the same source of error relevant to internal consistency indices. As noted earlier, a precise technical explanation of the SEM (and resulting confidence intervals) can be unwieldy. Because of this, score users are often provided less complex interpretations.

One simpler description sometimes used is that a confidence interval represents the possible score range that one would observe if a student could be tested twice with the same instrument. Taking the same test on a different day implies the only source of random error being considered is related to the occasion of testing—such as a student might be sleepier one day than another, might be sick, or might not have eaten a good breakfast. There is a reliability index that captures this source of random error, and it is referred to as the test-retest reliability coefficient. This is not the type of reliability computed for the Keystone Exams. When internal consistency reliability estimates are used, such an explanation blurs the fact that random error based on the occasion of testing is not considered.

When SEMs are derived from internal consistency reliability estimates, a better approach is to describe the confidence interval as providing reasonable bounds for the range of scores that a student might receive if he or she took an equivalent version of the test. (That is, the student took a test that covered exactly the same content but included a different set of items.) As an example, if the Algebra I score was 1450 and the SEM band was 1435 to 1465, then a student would be likely to receive a score somewhere between 1435 and 1465 if he or she took a different version of the test.

RESULTS AND OBSERVATIONS

Coefficient alpha results and associated (traditional) SEMs for various Keystone Exam scores are documented in Appendix M. Values were derived using the final data file (see Chapter Nine). The results are organized by administration and then content area. Each table also breaks out the modules and groups of interest such as the total student population (overall), gender, ethnicity, English learner (EL), students with an individualized education plan (IEP), and the economically disadvantaged (ED). The statistics reported include the number of points possible (Pts.), number of items (Len.), number of students tested (N), mean number of score points received (Mean), standard deviation of test scores (SD), reliability (r), and traditional standard error of measurement (SEM).

Note that these tables report the standard deviations of observed scores. Assuming normally distributed scores, one would expect about two-thirds of the observations to be within one standard deviation of the mean. An estimate of the standard deviation of the true scores can be computed as

$$\hat{\sigma}_T = \sqrt{\hat{\sigma}_x^2 - \hat{\sigma}_x^2(1 - \hat{\rho}_x)}$$

The overall test score reliability values are high (with a value of 0.91 or higher) for Algebra I, Biology, and Literature. The reliabilities at the module level tended to be lower, likely due to the fact that each module contains half as many items as the entire test. It was also noted that reliabilities tend to go up in value with an increase in population heterogeneity and go down in value with a decrease in more homogeneous populations. Once again, there is no firm guideline regarding how low is too low. The lower a given reliability coefficient, the greater the potential for over-interpretation. As a point of reference, a reliability coefficient of 0.50 would suggest that there is as much error variance as true-score variance in the scores. It should be noted that the reliability of group mean scores (e.g., school or district means) tends to be higher than that of individual scores, suggesting interpretation of module scores at these aggregate levels is likely reasonable.

RASCH CONDITIONAL STANDARD ERRORS OF MEASUREMENT

The CSEM also indicates the degree of measurement error but does so in scaled-score units and varies as a function of a student's actual scaled score. Therefore, the CSEM may be especially useful in characterizing measurement precision in the neighborhood of a score level used for decision-making—such as cut scores for identifying students who meet a performance standard.

Technically, when a Rasch model is applied, the CSEM at any given point on the ability continuum is defined as the reciprocal of the square root of the test information function derived from the Rasch scaling model:

$$CSEM(\hat{\beta}_n) = \frac{1}{\sqrt{I(\hat{\beta}_n)}}$$

where $CSEM(\hat{\beta}_n)$ is conditional standard error of measurement and $I(\hat{\beta}_n)$ is test information function. Test information depends on the sum of the corresponding information functions for the test items. Item information depends on each item's difficulty and conditional item score variance. The formula above utilizes the Rasch ability β_n metric. The conditional standard error on the scaled-score (SS) metric is determined simply by multiplying the $CSEM(\hat{\beta}_n)$ by the slope (multiplicative constant, m) of the linear transformation equation used to convert the Rasch ability estimates to scaled scores.

$$CSEM(SS) = CSEM(\hat{\beta}_n) * m$$

Chapter Fourteen provides the linear transformation formulas for each of the Keystone Exams.

RASCH CSEM CONFIDENCE INTERVALS

CSEMs also allow statements regarding the precision of individual tests scores. And like SEMs, they help place reasonable limits around observed scaled scores through construction of an approximate score band. The confidence intervals are constructed by adding and subtracting a multiplicative factor of the CSEM and may be interpreted as described in the earlier section.

FURTHER INTERPRETATIONS

DIFFERENT CSEMS FOR DIFFERENT TEST SCORES

The CSEM approach provides different numerical estimates for constructing the confidence intervals for examinees depending on their specific score levels. The magnitude of the CSEM values is U-shaped, with larger CSEM values associated with lower and higher scores.

GROUP SPECIFIC

Assuming reasonable model-data fit—as explored in Chapter Twelve—the Rasch-based CSEMs (conditioned on score level) should not vary across groups.

SCALED-SCORE METRIC

The CSEM and associated confidence interval bands are on the scaled-score metric.

TYPE OF ERROR REFLECTED

The CSEMs documented on the Keystone Exams score reports are the Rasch-based conditional standard errors of measurement described above. These are provided by the program WINSTEPS described in Chapter Twelve. As noted earlier, these CSEMs are based on the concept of statistical information. For the purpose of providing a simpler explanation of CSEMs to test score users, the earlier description of SEMs framed using the idea of internal consistency reliability was provided in the Keystone Exams score report interpretive guide.⁵ Score report content is considered in greater detail in Chapter Sixteen.

⁵ Because Rasch CSEMs are based on statistical information, it is questionable whether they account for error variance due to items. However, it seems difficult to construct a simple explanation of Rasch CSEMs for the general public.

RESULTS AND OBSERVATIONS

Figure 18–2 shows the Rasch CSEMs associated with each scaled-score level. (This information is also provided in tabular form in Appendix K.) Values were derived using the calibration data file described in Chapter Nine. The scaled scores and associated CSEMs represent the administration-specific scores. The values are fairly consistent across a noticeably large range of the scaled scores, as demonstrated by the relatively flat bottoms of most plots. The values increase at both extremes (i.e., at smaller and larger scaled scores) give these figures their typical U-shaped pattern. The three red-dashed lines represent the Basic, Proficient, and Advanced scaled cut-scores, respectively, moving from lower to higher scaled score values. CSEM values at the cut score lines are associated with smaller values, indicating more precise measurement occurs at these cuts.

Figure 18–2. Conditional Standard Error Plots for Each Administration and Content Area

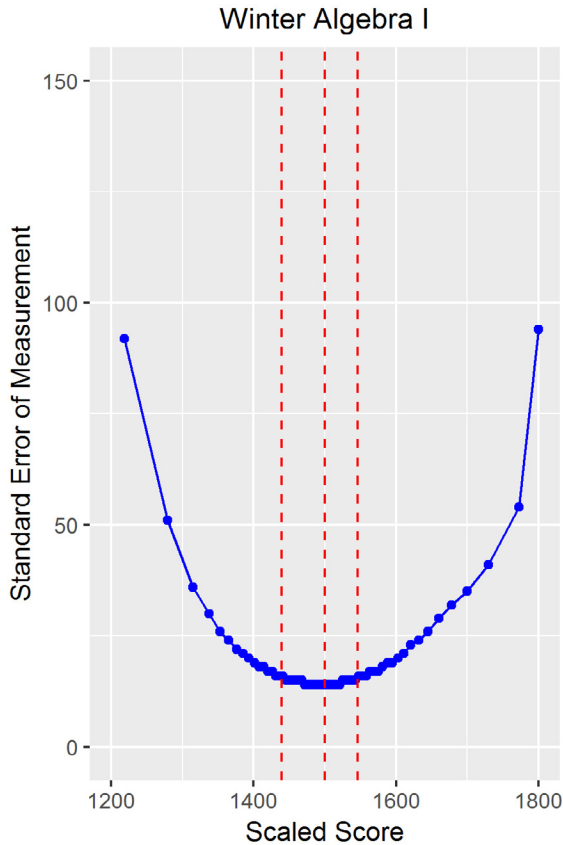


Figure 18–2 (continued). Conditional Standard Error Plots for Each Administration and Content Area

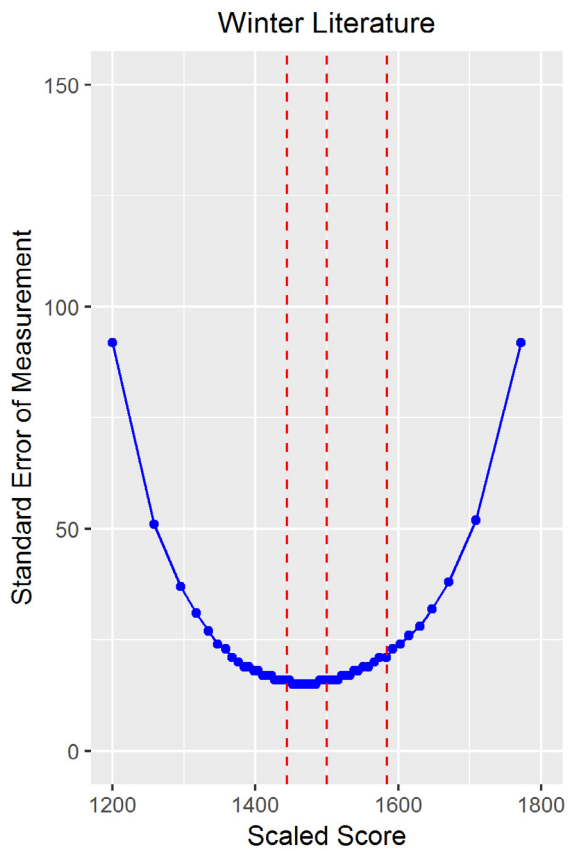
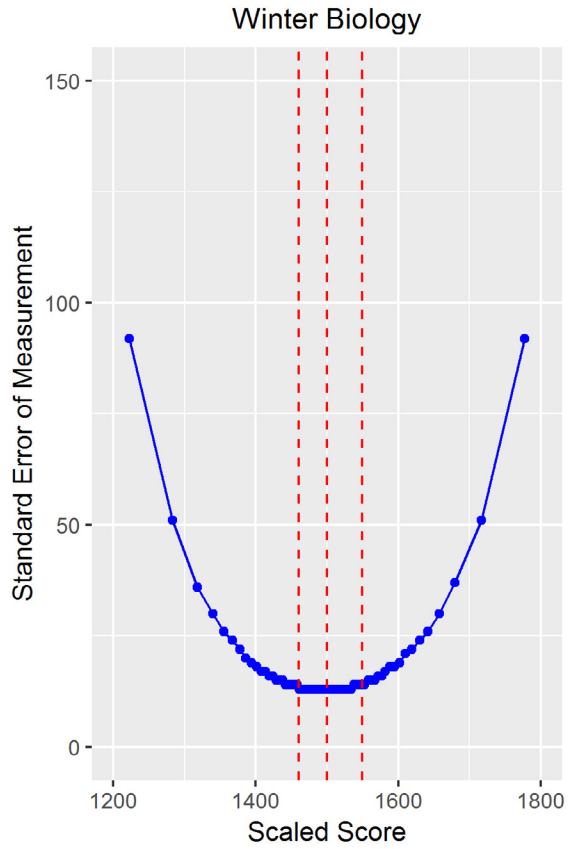


Figure 18–2 (continued). Conditional Standard Error Plots for Each Administration and Content Area

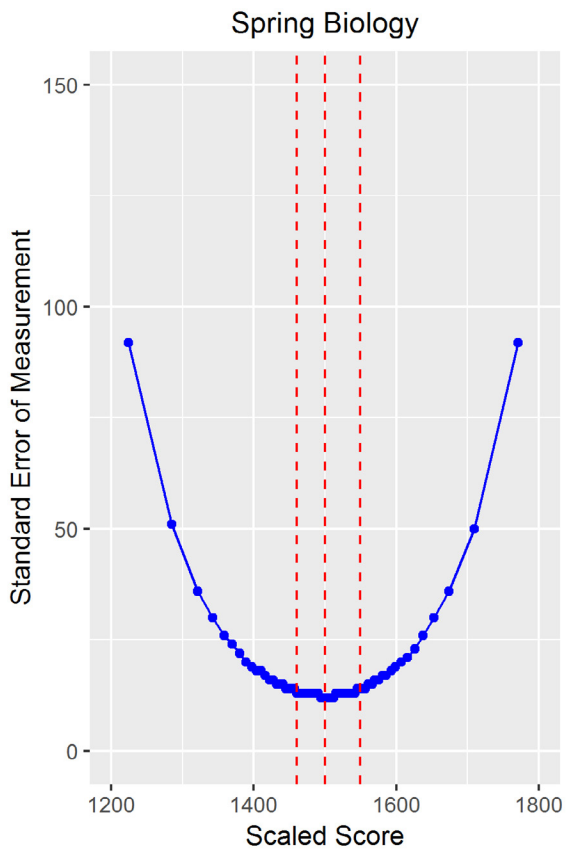
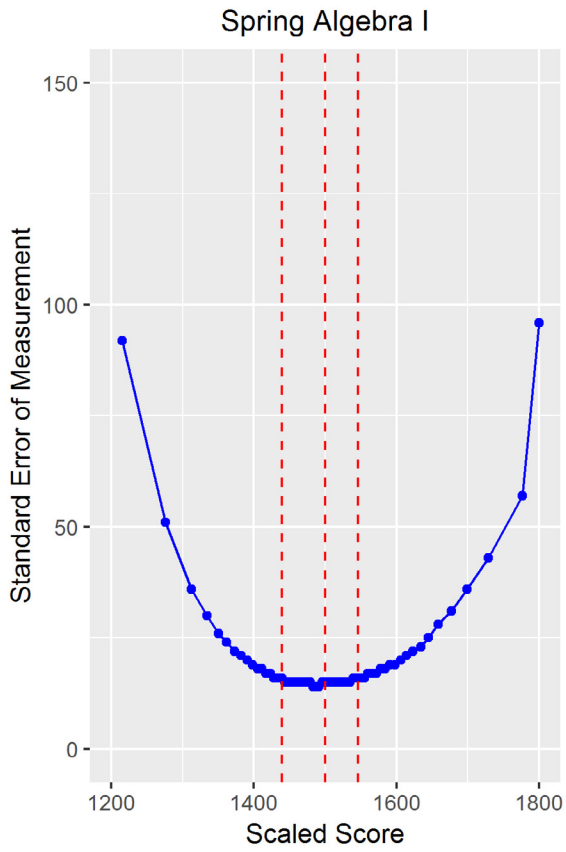
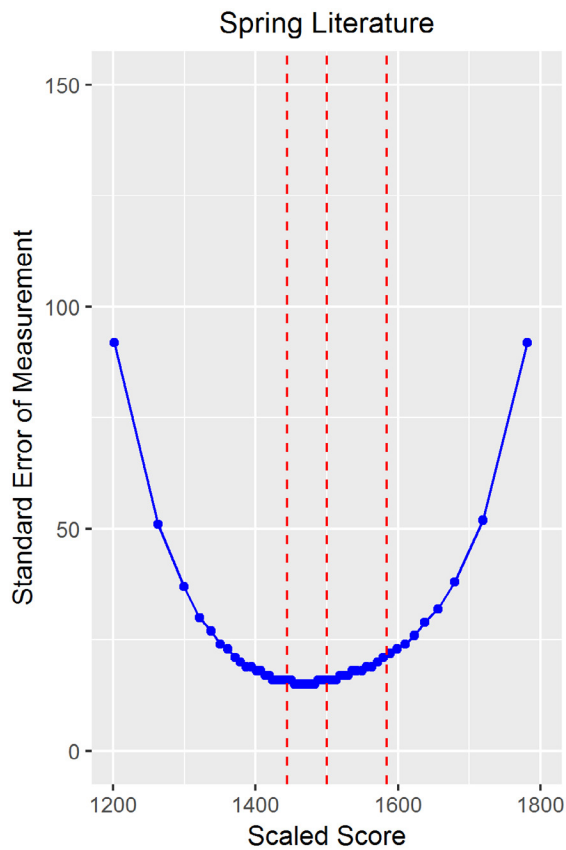


Figure 18–2 (continued). Conditional Standard Error Plots for Each Administration and Content Area



RELIABILITY OF PERFORMANCE LEVEL CLASSIFICATION DECISIONS

Student performance on the Keystone Exams is classified into one of four achievement levels using the cut scores described in Chapter Thirteen. The reliability of the classification decisions can be assessed by two statistics: decision accuracy and decision consistency.

DECISION ACCURACY

Decision accuracy describes the extent to which performance level classification decisions based on the administered test form would agree with the decisions that would be made on the basis of a perfectly reliable test (i.e., if it was possible to know each examinee's true score). Decision accuracy answers the question: How does the actual classification of test takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known?

DECISION CONSISTENCY

Decision consistency describes the extent to which classification decisions based on the administered test form would agree with the decisions made if a parallel alternate form had been administered. Decision consistency answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test?

Since the true scores are unknown and it is not feasible to repeat the Keystone Exams in order to estimate the proportion of students who would be reclassified in the same performance levels, a statistical model needs to be imposed on the data in order to project the consistency of classifications solely using data from the available administration (Hambleton and Novick, 1973). Although a number of procedures are available, two well-known methods were developed by Hanson and Brennan (1990) and Livingston and Lewis (1995), utilizing specific true score models. These approaches are fairly complex, and the cited sources contain details regarding the statistical models used to calculate the decision accuracy and consistency from a single administration.

For Keystone Exams, given the two approaches provide similar results, true scores and single-form scores on forms parallel to the one actually given are estimated following the Livingston and Lewis (1995) method. The decision accuracy is estimated using an estimated joint distribution of reported performance-level classifications on the current form of the exam and the performance-level classifications based on the true score. Decision consistency is estimated using an estimated joint distribution of reported performance-level classifications on the current form of the exam and performance-level classifications on the parallel alternate form. In each case, the proportion of performance-level classifications with exact agreement is the sum of the entries in the diagonal of the contingency table representing the joint distribution. Reliability of classification at each performance-level cut score is estimated by collapsing the joint distribution at the passing score boundary into a 2-by-2 table and summing the two entries.

Several factors might affect the classification decision accuracy and consistency. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications. Another factor is the location of the cut score in the score distribution. More consistent classifications are observed when the cut scores are located away from the mass of the score distribution. For example, when scores are close to being normally distributed, the mass is concentrated in the middle of the distribution, and thus, classifications tend to become more consistent when cut scores go up from 70 percent to 80 percent, or, alternatively, go down from 30 percent to 20 percent. The number of performance levels is also a consideration. Consistency indices for four performance levels should be lower than for those based on two categories. This is not surprising since classification using four levels would allow more opportunity to change achievement levels. Hence, there would be more classification errors with four achievement levels, resulting in lower consistency indices.

The results—derived using the program *BB-Class* (Brennan, 2004)—for the overall accuracy and consistency across all four performance levels as well as for the dichotomies created by the three cut scores are presented in Table 18–4.

Across all administrations and content areas, the overall decision accuracy ranged from 0.79 to 0.81 and the decision consistency ranged from 0.71 to 0.73. Dichotomous decisions have the higher accuracy and consistency values than the overall. The decision accuracy of the Basic/Proficient cut scores ranged from 0.92 to 0.93 and the decision consistency ranged from 0.87 to 0.94. The decision accuracy results indicate that at least 92% of students meeting or exceeding the Proficient cut score would receive the same classification if their true scores were known. If a parallel test were administered, at least 87% of students meeting or exceeding the Proficient cut score would be classified in the same way.

Table 18–4. Reliability of Performance-Level Classification Decisions

Administration	Content Area	Statistic	Overall	Below Basic/ Basic	Basic/Proficient	Proficient/ Advanced
Winter	Algebra I	Accuracy	0.79	0.91	0.92	0.96
Winter	Algebra I	Consistency	0.71	0.87	0.89	0.94
Winter	Biology	Accuracy	0.80	0.93	0.93	0.94
Winter	Biology	Consistency	0.72	0.90	0.90	0.92
Winter	Literature	Accuracy	0.81	0.96	0.92	0.93
Winter	Literature	Consistency	0.73	0.94	0.89	0.91
Spring	Algebra I	Accuracy	0.80	0.92	0.93	0.96
Spring	Algebra I	Consistency	0.73	0.89	0.90	0.94
Spring	Biology	Accuracy	0.80	0.92	0.93	0.95
Spring	Biology	Consistency	0.72	0.89	0.90	0.93
Spring	Literature	Accuracy	0.80	0.95	0.92	0.93
Spring	Literature	Consistency	0.72	0.93	0.89	0.90
Summer	Algebra I	Accuracy				
Summer	Algebra I	Consistency				
Summer	Biology	Accuracy				
Summer	Biology	Consistency				
Summer	Literature	Accuracy				
Summer	Literature	Consistency				

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

RATER AGREEMENT

Because CR items are included on the Keystone Exams, another source of random error is related to the scorers of those items. Frisbie (2005) noted that “test score reliability differs from scorer reliability” and that “the need for one kind of estimate cannot be satisfied by the other.” Additionally, the data most easily obtainable that captures this information comes from the “10 percent read behinds” collected during the scoring process. Partly because of the way these data are obtained and reported (i.e., it’s **not** a ratio of true score variance over observed score variance), the term *rater agreement* is used here, not *rater reliability* or *inter-rater reliability* as these terms are somewhat misleading.

The rater agreements for the Keystone Exams are presented in Tables 18–5 to 18–7. In addition, the percentages awarded to each score point are also presented in these tables. As the table shows, the exact inter-rater agreement percentages ranged from 79 to 100 percent for Winter and 82 to 100 percent for Spring. Overall, Algebra I had the highest exact agreements while Literature tended to have lowest exact agreement. The percentages of exact and adjacent agreement for all content areas were nearly 100.

Table 18–5. Inter-Rater Agreement and Percentage Awarded for Each Score Point for CR Items: Winter

Content Area	Item	Inter-Rater Agreement % Exact	Inter-Rater Agreement % Adjacent	% Exact + Adjacent Agreement	% Awarded for Score Point 0	% Awarded for Score Point 1	% Awarded for Score Point 2	% Awarded for Score Point 3	% Awarded for Score Point 4	% Awarded for Score Point B/NS
Algebra 1	1A	99	1	100	44	36	NA	NA	NA	20
Algebra 1	1B	99	1	100	52	12	17	NA	NA	20
Algebra 1	1C	100	0	100	78	3	NA	NA	NA	20
Algebra 1	2	97	3	100	40	28	7	3	1	20
Algebra 1	3A	100	0	100	30	50	NA	NA	NA	21
Algebra 1	3B	99	1	100	62	17	NA	NA	NA	21
Algebra 1	3C	99	1	100	46	33	NA	NA	NA	21
Algebra 1	3D	99	1	100	55	25	NA	NA	NA	21
Algebra 1	4	95	5	100	44	23	7	5	1	20
Algebra 1	5A	100	0	100	37	42	NA	NA	NA	20
Algebra 1	5B	100	0	100	60	19	NA	NA	NA	20
Algebra 1	5C	100	0	100	74	5	NA	NA	NA	20
Algebra 1	5D	100	0	100	74	5	NA	NA	NA	20
Algebra 1	6	89	11	100	10	28	23	16	1	22
Biology	1	91	9	100	43	18	13	10	NA	15
Biology	2	93	7	100	10	37	37	2	NA	14
Biology	3	92	7	99	16	30	28	12	NA	14
Biology	4	82	17	99	30	33	18	4	NA	14
Biology	5	89	11	100	26	20	18	21	NA	14
Biology	6	79	20	99	16	22	26	22	NA	14
Literature	1	82	18	100	4	35	37	10	NA	14
Literature	2	80	20	100	5	32	33	15	NA	14
Literature	3	85	15	100	6	32	34	11	NA	17
Literature	4	90	10	100	3	30	41	11	NA	14
Literature	5	82	18	100	10	30	33	11	NA	16
Literature	6	92	8	100	7	25	39	11	NA	17

Note. Some of the Algebra I CR items were scored by part. For example, 1A in the second column means part A of item 1. B/NS in the last column represents blank/non-scorable. NA means not applicable.

Table 18–6. Inter-Rater Agreement and Percentage Awarded for Each Score Point for CR Items: Spring

Content Area	Item	Inter-Rater Agreement % Exact	Inter-Rater Agreement % Adjacent	% Exact + Adjacent Agreement	% Awarded for Score Point 0	% Awarded for Score Point 1	% Awarded for Score Point 2	% Awarded for Score Point 3	% Awarded for Score Point 4	% Awarded for Score Point B/NS
Algebra 1	1	90	10	100	25	37	16	7	2	14
Algebra 1	2	95	5	100	42	35	5	2	0	16
Algebra 1	3A	100	0	100	15	70	NA	NA	NA	15
Algebra 1	3B	98	2	100	34	51	NA	NA	NA	15
Algebra 1	3C	99	1	100	78	7	NA	NA	NA	15
Algebra 1	3D	100	0	100	69	16	NA	NA	NA	15
Algebra 1	4A	100	0	100	50	37	NA	NA	NA	14
Algebra 1	4B	100	0	100	39	47	NA	NA	NA	14
Algebra 1	4C	100	0	100	62	24	NA	NA	NA	14
Algebra 1	4D	100	0	100	86	1	NA	NA	NA	14
Algebra 1	5A	99	1	100	55	30	NA	NA	NA	14
Algebra 1	5B	100	0	100	77	9	NA	NA	NA	14
Algebra 1	5C	99	1	100	74	12	NA	NA	NA	14
Algebra 1	5D	99	1	100	74	12	NA	NA	NA	14
Algebra 1	6	90	10	100	33	24	14	9	4	16
Biology	1	85	14	99	23	24	22	17	NA	13
Biology	2	90	10	100	26	22	21	17	NA	14
Biology	3	96	4	100	29	27	19	12	NA	14
Biology	4	92	8	100	25	22	20	13	NA	19
Biology	5	92	8	100	26	28	20	12	NA	13
Biology	6	85	15	100	19	30	23	11	NA	16
Literature	1	82	18	100	3	22	44	16	NA	14
Literature	2	83	17	100	9	29	38	10	NA	14
Literature	3	83	17	100	7	27	38	11	NA	17
Literature	4	87	13	100	5	36	38	5	NA	15
Literature	5	87	13	100	9	25	43	8	NA	15
Literature	6	82	18	100	8	24	38	13	NA	18

Note. Some of the Algebra I CR items were scored by part. For example, 1A in the second column means part A of item 1. B/NS in the last column represents blank/non-scorable. NA means not applicable.

Table 18–7. Inter-Rater Agreement and Percentage Awarded for Each Score Point for CR Items: Summer

Content Area	Item	Inter-Rater Agreement % Exact	Inter-Rater Agreement % Adjacent	% Exact + Adjacent Agreement	% Awarded for Score Point 0	% Awarded for Score Point 1	% Awarded for Score Point 2	% Awarded for Score Point 3	% Awarded for Score Point 4	% Awarded for Score Point B/NS
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Algebra 1										
Biology										
Biology										
Biology										
Biology										
Biology										
Biology										
Literature										
Literature										
Literature										
Literature										
Literature										
Literature										

Note. Some of the Algebra I CR items were scored by part. For example, 1A in the second column means part A of item 1. B/NS in the last column represents blank/non-scorable. NA means not applicable.

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

CHAPTER NINETEEN: VALIDITY

As defined in the *Standards for Educational and Psychological Testing*, validity is “the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests” (AERA, APA, & NCME, 2014, p.11). The *Standards* provide a framework for describing the sources of evidence that should be considered when evaluating validity. These sources include evidence based on test content, response processes, the internal structure of the test, the relationships between test scores and other variables, and the consequences of testing. In addition, when Rasch models are used to analyze assessment data, validity considerations related to those processes should also be explored.

The validity process involves the collection of a variety of evidence to support the proposed test score interpretations and uses. The entire technical report describes the technical aspects of the Keystone Exams in support of their score interpretations and uses. Each of the previous chapters contributes important components that pertain to score validation: test development, test administration, test scoring, item analysis, Rasch calibration, scaling, equating, score reporting, and reliability. This chapter summarizes and synthesizes the evidence based on the framework of the *Standards*. The purposes and intended uses of the Keystone Exams are reviewed first, and then each type of validity evidence is addressed in turn.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

PURPOSES AND INTENDED USES OF THE KEYSTONE EXAMS

The *Standards* emphasize that validity pertains to how test scores are used. To help contextualize the evidence that are presented in this chapter, the purposes of the Keystone Exams will be reviewed first. The Keystone Exams, which began in 2010–2011 for Algebra I, Biology, and Literature, are one component of Pennsylvania’s system of high school graduation requirements. Students take the exams upon completing specific courses. The Keystone Exams results help schools and districts guide students toward meeting state standards. Students who do not score Proficient or above on a Keystone Exam module may choose to complete a project-based assessment for that module, provided that they meet the requirements detailed below.

- The student has taken the course.
- The student was unsuccessful in achieving a score of Proficient or Advanced on the Keystone Exam after at least two attempts.
- The student has met the district’s attendance requirements for the course.
- The student has participated in a satisfactory manner in supplemental instructional services.

EVIDENCE BASED ON TEST CONTENT

Test content validity evidence for the Keystone Exams rests greatly on establishing a link between each piece of the assessment (i.e., the items) and what students should know and be able to do as prescribed by the Keystone Exams Assessment Anchors and Eligible Content. The Keystone Exams are intended to measure the knowledge and skills described in the Assessment Anchors and Eligible Content for Algebra I, Biology, and Literature.

Lane (1999) suggests taking the following steps to support the content validity of an assessment. In the case of Keystone Exams, one should

- Evaluate the degree to which the test specifications represent and align with the knowledge and skills described in the Assessment Anchors and Eligible Content for Algebra I, Biology, and Literature.
- Evaluate the alignment between the Keystone Exams items and test specifications to ensure representativeness.
- Evaluate the extent to which the curriculum aligns with the Assessment Anchors and Eligible Content.

- Conduct content reviews of the Keystone Exams items using a panel of content experts to see whether items measure the intended construct or are the sources of construct-irrelevant variance.
- Conduct fairness reviews of the items to avoid issues related to a specific subpopulation.
- Evaluate procedures for administration and scoring such as the appropriateness of instructions to examinees, time limit for the assessment, and training of raters.
- Submit operational tests to third-party independent reviews.

Chapters Two through Eight of this report present a considerable amount of evidence related to test content. As described in these chapters, all the items were developed and aligned with the Keystone Exams Assessment Anchors and Eligible Content for Algebra I, Biology, and Literature. After development, items underwent multiple rounds of content and bias reviews. After being field tested, items were reviewed with respect to their statistical properties. Items selected for the operational assessment had to pass content, psychometric, and PDE reviews. Tests were administered according to standardized procedures with allowable accommodations.

Some of the efforts made to ensure content validity are summarized below.

- DRC used Webb’s (1999) Depth of Knowledge (DOK) model to ensure the Keystone Exams items aligned with the Assessment Anchors and Eligible Content and the Academic Content Standards in terms of both content and cognitive levels.
- DRC established detailed test and item/passage development specifications and ensured the items were sufficient in number and adequately distributed across content, levels of cognitive complexity, and levels of difficulty.
- DRC selected qualified item writers and provided training to help ensure they wrote high-quality items.
- All newly developed items were first reviewed by content specialists and editors at DRC to make sure they measured the intended Assessment Anchors and Eligible Content for Algebra I, Biology, and Literature. Appropriateness for the intended students was also considered, as well as DOK, graphics, grammar/punctuation, language demand, and distractor reasonableness.
- Prior to field testing, the test items were submitted to content committees (composed of Pennsylvania educators) for review with respect to the following categories:
 - Overall quality and clarity
 - Anchor, Eligible Content, and/or standard alignment
 - Grade-level appropriateness
 - Difficulty level
 - Domain of knowledge (DOK)
 - Appropriate sources of challenge (e.g., unintended content and skills)
 - Correct answer
 - Quality of distractors
 - Graphics
 - Appropriate language demand
 - Freedom from bias
- The items were also submitted to a Bias, Fairness, and Sensitivity Committee for review. This committee reviewed items for issues related to diversity, gender, and other pertinent factors.

- Items passing all the prior hurdles were tried out as embedded field-test items in the operational test. Several statistical analyses were conducted on the field-test data including classical item analyses, distractor analyses, and differential item functioning (DIF). Items were again carefully reviewed by DRC staff and a committee of Pennsylvania teachers with respect to their statistical characteristics. DIF was used to detect test items that might bias test scores for particular groups. Empirical investigation of DIF strengthens the validity evidence related to score interpretations for students in particular groups by eliminating potential sources of construct-irrelevant variance.
- The Keystone Exams were administered according to standardized procedures with allowable accommodations. Students were given ample time to complete the tests (i.e., there were no speediness issues).
- As described in Chapter Eight, the raters for constructed-response (CR) items were carefully recruited and well trained. Their scoring was monitored throughout the scoring session to ensure that an acceptable level of scoring accuracy was maintained.

EVIDENCE BASED ON RESPONSE PROCESS

Response-process evidence is used to examine the extent to which the cognitive skills and processes employed by students match those identified in the test developer’s defined construct domains for all students and for each subgroup. Think-aloud procedures or cognitive labs can be used to collect this type of evidence. In addition, when an assessment includes CR items, an examination of the extent to which the raters interpret and apply the scoring criteria accurately when assigning scores to students’ responses on CR items also adds response-process validity evidence.

For the operational Keystone Exams offered in winter, spring, and summer, no cognitive lab studies were conducted to collect the response-process evidence. Rather, for all the Keystone Exams, well-organized scorer training and subsequent monitoring of rating accuracy helped ensure that raters strictly followed the scoring criteria and that no features unrelated to the rubric significantly affected their scoring.

EVIDENCE BASED ON INTERNAL STRUCTURE

As described in the *Standards* (2014), internal-structure evidence refers to the degree to which the relationships between test items and test components conform to the construct on which the proposed test interpretations are based. For each Keystone Exam, one total test score as well as module scores were reported (see Chapter Sixteen for more information about the Keystone Exams scores). Several dimensionality studies were conducted in order to provide internal-structure evidence relating to the use of both types of scores.

ITEM-TEST CORRELATIONS

Item-test correlations are provided in Appendix J and discussed in Chapter Eleven. All item correlations were greater than 0, indicating a positive association with the total score.

DIFFERENTIAL ITEM FUNCTIONING (DIF)

DIF analyses with respect to gender, ethnicity, and administration mode address construct-irrelevant variance, which represents an important threat to the validity of achievement tests. As noted in Chapter Five, newly field-test items were screened and reviewed for DIF. Only items approved by teacher committees were eligible for operational use. After operational use, DIF analyses were conducted on the operational items again to investigate whether the DIF code changed between field testing and operational use. Table 19–1 shows the number of items in which the DIF codes changed between field testing and operational testing, either from A/B to C-level DIF, or vice versa. There were very few items in which the bias code changed from A/B to C, and few items in which the bias code changed from C to A/B. Items with changed DIF code tended to be in Keystone Literature.

Table 19–1. Summary of Bias Code Change from Field Test to Operational Test

Administration	Content Area	Change	Male/Female	White/Black	PPT/CBT
Winter	Algebra I	C → A/B	0	0	0
Winter	Algebra I	A/B → C	0	0	0
Winter	Biology	C → A/B	0	0	0
Winter	Biology	A/B → C	0	0	0
Winter	Literature	C → A/B	1	0	1
Winter	Literature	A/B → C	2	0	1
Spring	Algebra I	C → A/B	0	0	0
Spring	Algebra I	A/B → C	0	0	0
Spring	Biology	C → A/B	0	0	0
Spring	Biology	A/B → C	0	0	0
Spring	Literature	C → A/B	0	0	0
Spring	Literature	A/B → C	1	0	0
Summer	Algebra I	C → A/B			
Summer	Algebra I	A/B → C			
Summer	Biology	C → A/B			
Summer	Biology	A/B → C			
Summer	Literature	C → A/B			
Summer	Literature	A/B → C			

Note. PPT represents the paper-and-pencil-based test, and CBT represents the computer-based test.

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

MODULE CORRELATIONS

Correlations and disattenuated correlations among module scores for the Keystone Exams are presented below. Values were derived from the Keystone final data files (see Chapter Nine). These data can also provide information on score dimensionality that is part of internal-structure evidence. All Keystone Exams have two modules, which represent unique, non-overlapping topics. The intercorrelations between the modules within the content areas were positive and ranged from 0.83 to 0.85. The intercorrelations between modules in different content areas ranged from 0.60 to 0.76.

Table 19–2. Correlations among Algebra I, Biology, and Literature Modules

Administration	Content Area	Module	Algebra I Module 1	Algebra I Module 2	Biology Module 1	Biology Module 2	Literature Module 1	Literature Module 2
Winter	Algebra I	Module 1	-					
Winter	Algebra I	Module 2	0.83	-				
Winter	Biology	Module 1	0.71	0.69	-			
Winter	Biology	Module 2	0.70	0.68	0.85	-		
Winter	Literature	Module 1	0.60	0.60	0.73	0.76	-	
Winter	Literature	Module 2	0.62	0.63	0.74	0.77	0.83	-
Spring	Algebra I	Module 1	-					
Spring	Algebra I	Module 2	0.85	-				
Spring	Biology	Module 1	0.68	0.69	-			
Spring	Biology	Module 2	0.70	0.72	0.85	-		
Spring	Literature	Module 1	0.61	0.60	0.70	0.72	-	
Spring	Literature	Module 2	0.63	0.63	0.71	0.75	0.84	-
Summer	Algebra I	Module 1						
Summer	Algebra I	Module 2						
Summer	Biology	Module 1						
Summer	Biology	Module 2						
Summer	Literature	Module 1						
Summer	Literature	Module 2						

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

The correlations in Table 19–2 are based on the observed module scores. These observed-score correlations are weakened by existing measurement error contained within each module score. As a result, disattenuated correlations could provide an estimate of the relationships among modules if there were no measurement error. (An important caveat is explained further below.) The disattenuated correlation coefficients R_{12} can be computed by using the formula (Spearman, 1904; Spearman, 1910) below:

$$R_{12} = \frac{r_{12}}{\sqrt{r_{11}r_{22}}},$$

where r_{12} is the observed correlation, and r_{11} and r_{22} are the reliabilities for Module 1 and Module 2. Disattenuated correlations very near 1.00 suggest that the same or very similar constructs are being measured. Values somewhat less than 1.00 suggest that different modules are measuring slightly different aspects of the same construct. Values markedly less than 1.00 suggest the modules reflect different constructs.

Table 19–3 shows the corresponding disattenuated correlations for each Keystone Exam. Given that none of these modules had perfect reliabilities (see Chapter Eighteen), the disattenuated module correlations are higher than their observed score counterparts.

Table 19–3. Disattenuated Correlations among Algebra I, Biology, and Literature Modules

Administration	Content Area	Module	Algebra I Module 1	Algebra I Module 2	Biology Module 1	Biology Module 2	Literature Module 1	Literature Module 2
Winter	Algebra I	Module 1	-					
Winter	Algebra I	Module 2	1.00	-				
Winter	Biology	Module 1	0.83	0.81	-			
Winter	Biology	Module 2	0.83	0.81	0.99	-		
Winter	Literature	Module 1	0.72	0.71	0.85	0.90	-	
Winter	Literature	Module 2	0.74	0.76	0.86	0.91	1.00	-
Spring	Algebra I	Module 1	-					
Spring	Algebra I	Module 2	1.00	-				
Spring	Biology	Module 1	0.79	0.80	-			
Spring	Biology	Module 2	0.81	0.83	0.99	-		
Spring	Literature	Module 1	0.73	0.71	0.82	0.86	-	
Spring	Literature	Module 2	0.74	0.74	0.83	0.87	0.99	-
Summer								
Summer								
Summer								
Summer								
Summer								
Summer								-

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

The within-content area disattenuated correlations were high (e.g., above 0.99), suggesting that the within-content area modules might be measuring essentially the same construct. This, in turn, suggests that the within content module scores might not provide unique information about the strengths or weaknesses of many of the students.

On a fairly consistent basis, the disattenuated correlations among the modules within each content area are higher than the correlations among modules across different content areas. In general, within-content area module correlations greater than or equal to 0.99, while across-content area module disattenuated correlations range from 0.71 to 0.87.

It should be noted that some caution is needed when interpreting the disattenuated results because the reliabilities used to calculate the disattenuated correlations are subject to both upward and downward biases. Consequently, some of the values in the table above may be higher or lower than they should be, depending on which bias prevails for any given pair of module scores. When the reliabilities are lower than they should be, the disattenuated correlations will be inflated and, in some instances, can appear higher than the theoretical correlation maximum value of 1.00.

Overall, the internal structure evidence presented supports the assumption that related elements of each of the Keystone Exams are correlated in the intended manner. The modules *within* each content area have stronger relationships than the *across* content area modules, which further supports reporting a total score for each different content area.

The module scores present more of a mixed message. Since the modules in each content area were designed to measure distinct components of the content area, it is reasonable to expect that the inter-content module correlations should be positive and strong but, ideally, not extremely high. However, the disattenuated correlations imply that some modules are essentially measuring the same constructs for most of the students. Consequently, there may be less support for providing results for some module scores beyond the total score. While there is content rationale underlying the creation of the module scores, the empirical correlations illustrate that caution is required when using the module scores to identify a student’s strengths and weaknesses. Certainly, instructional

programs should not be based on module score information alone, but rather in conjunction with other sources of evidence available (e.g., teacher observations, other exam performance).

EVIDENCE BASED ON RELATIONSHIPS WITH OTHER VARIABLES

As described in the *Standards*, “Evidence based on relationships with other variables provides evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretations” (AERA, APA, & NCME, 2014, p. 16). This category of evidence refers to external structure evidence and has been classified as three types of evidence: *convergent*, *discriminant*, and *criterion-related*. Convergent evidence is provided by relationships among students’ performances on different assessments intended to measure a similar construct. Discriminant evidence is provided by relationships among students’ performances on different tests intended to measure different constructs. Criterion-related evidence, either predictive or concurrent, is provided by relationships between students’ test scores and their performances on a criterion measure (Cronbach, 1971; Messick, 1989).

The correlations among students’ test scores on different Keystone Exams including Algebra I, Biology, and Literature are shown in Table 19–4 to provide some discriminant validity evidence. In this table, both the observed correlations and the disattenuated correlations (in the parentheses) are reported. Each Keystone Exam measures different constructs, so the correlations among them were not expected to be extremely high. The values in this table are consistent with this expectation. As can be seen, the correlations among the Keystone Exams ranged from 0.69 to 0.81.

Table 19–4. Correlations Among Student Performance

Administration	Content Area	Algebra I	Biology
Winter	Biology	0.77 (0.84)	
Winter	Literature	0.69 (0.75)	0.81 (0.88)
Spring	Biology	0.77 (0.84)	
Spring	Literature	0.70 (0.76)	0.79 (0.86)
Summer	Biology		
Summer	Literature		

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

External evidence for the Keystone Exams is examined by using students’ scores on the 2021 Pennsylvania System of School Assessment (PSSA) as the external criteria. The final 2021 Algebra I, Biology and Literature data files were merged with the PSSA mathematics, science, and ELA data by students’ PA secure IDs. Then the correlations between students’ scores on the Keystone Exams and on the PSSA were calculated as one component of external evidence. This analysis was attempted for all Keystone administrations. However, only enough students were obtained for the spring administration. Table 19–5 summarizes the sample sizes and correlations by grade and content area after the file merging of the spring Keystone Exams and the PSSA.

Table 19–5. Correlation between Spring Keystone Exam Scores and PSSA Scores by Grade

Keystone Content Area/ PSSA Subject	Grade 7 N	Grade 7 Correlation	Grade 8 N	Grade 8 Correlation
Algebra I/ Mathematics	5,391	0.78	21,758	0.83
Biology/ Science	NA	NA	223	0.71
Literature/ ELA	NA	NA	19	0.59

The correlations within the same content area ranged from 0.59 to 0.83. These results suggest the Keystone Exams measured something similar but not identical to the corresponding PSSA tests. The correlations between PSSA ELA and Keystone Literature examinations are based on a small sample, and therefore should not be overinterpreted. The results also provide external evidence in support of the Keystone Exams as a valid measure of students' achievement.

The collection of external evidence relating to the Keystone Exams is an ongoing process once the data are collected in the future. Other criterion-related evidence can be evaluated by the relationships between the Keystone Exams and criterion variables such as the Scholastic Aptitude Test (SAT), the American College Testing (ACT), or students' Grade Point Average (GPA) in their first college course.

EVIDENCE BASED ON CONSEQUENCES OF TESTS

Based on the *Standards* (AERA, APA, & NCME, 2014), evidence of the consequences of implementing an assessment program is an additional source of validity information. Both positive and negative (intended and unintended) consequences of score-based inferences must be investigated to fully evaluate the pool of validity evidence.

Lane and Stone (2002) summarized the general *intended* consequences for state assessments and accountability programs:

- Student, teacher, and administrator motivation and effort
- Curriculum and instruction practices (including content and strategies)
- Improved learning for all students
- Content and format of classroom assessments
- Professional development support
- Use and nature of test-preparation activities
- Student, teacher, administrator, and public awareness and beliefs about the assessment, criteria for judging performance, and the use of assessment results

Evidence for the improvement of student learning can be seen by examining the change in the percentage of students who scored Proficient or Advanced from the inception of the Keystone program in 2011 to the most recent administrations. Table 19–6 provides the percentages of students who scored Proficient or Advanced by administration and content. Values are derived from the first-time test-takers during each spring administration and represent administration-specific performance level classification.

Table 19–6. Percentages of Students earning Proficient or Advanced Performance Level Classifications

Administration	Algebra I Number	Algebra I Percent	Biology Number	Biology Percent	Literature Number	Literature Percent
Spring 2011	94,697	38.6	46,979	35.7	42,808	49.9
Spring 2013	157,811	47.6	134,995	47.2	117,830	63.1
Spring 2014	124,954	51.5	119,274	52.9	113,477	60.9
Spring 2015	121,255	50.1	115,936	58.0	114,387	67.2
Spring 2016	120,104	50.6	116,345	56.8	112,361	65.0
Spring 2017	118,704	51.7	115,473	56.6	110,966	64.5
Spring 2018	117,345	50.6	114,704	57.2	109,387	64.7
Spring 2019	117,677	48.4	113,834	54.7	110,030	63.3
Spring 2020						
Spring 2021	103,296	37.1	98,419	45.6	96,409	57.7

Note. Spring 2020 data is not available due to school closures and cancellation of Keystone testing in Spring 2020.

Lane and Stone (2002) also summarized the possible unintended outcomes:

- Narrowing of curriculum and instruction to focus only on the specific standards assessed and ignoring the broader construct reflected in the specified standards
- Use of test-preparation materials that are closely linked to the assessment without making changes to instruction
- Use of unethical test-preparation materials or administration procedures
- Differential performance gains for subgroups of students
- Inappropriate or unfair uses of test scores, such as questionable practices in reassignment of teachers or principals
- For some students, decreased confidence and motivation to learn and to perform well on the assessment because of past experiences with assessments

As noted above, one important piece of consequential evidence pertains to the use of assessment results. As shown in Chapter Sixteen, there are several different types of scores and score reports used for the Keystone Exams. The extent to which various groups of users (e.g., students, teachers) interpret these scores and reports appropriately affects the validity of subsequent uses of these results. Chapter Sixteen is intended to provide accurate and clear test score and report information with the hope that this will help users avoid unintended uses and interpretations of the Keystone Exams results. Nevertheless, evidence pertaining to other consequences of the Keystone Exams needs continued research.

EVIDENCE RELATED TO THE USE OF RASCH MODEL

Since the Rasch model is the basis of all calibration, scaling, and equating analyses associated with the Keystone Exams, the validity of the inferences from these results depends on the degree to which the assumptions of the model are met, as well as the fit between the model and the test data. As discussed in Chapter Twelve, the underlying assumptions of Rasch models were essentially met for all the Keystone Exams data, indicating the appropriateness of using the Rasch models to analyze the Keystone Exams data.

VALIDITY EVIDENCE SUMMARY

Validity evidence related to test content was reviewed earlier in this chapter. The early chapters of this technical report show that a strong link can be established between each item on the Keystone Exam and its associated Eligible Content. Details regarding how the operational Keystone Exams were assembled to reflect the state content standards and detailed information regarding educator reviews (including content, bias, and sensitivity reviews) are presented in Chapter Six.

Module score intercorrelations were also presented in this chapter. In general, within-content area module scores (e.g., Algebra I Module 1 and Algebra Module 2) were correlated more highly than across-content module scores (e.g., Literature Module 1 and Algebra Module 1). Consequently, this provides some favorable evidence regarding the internal and external relationships between the test components.

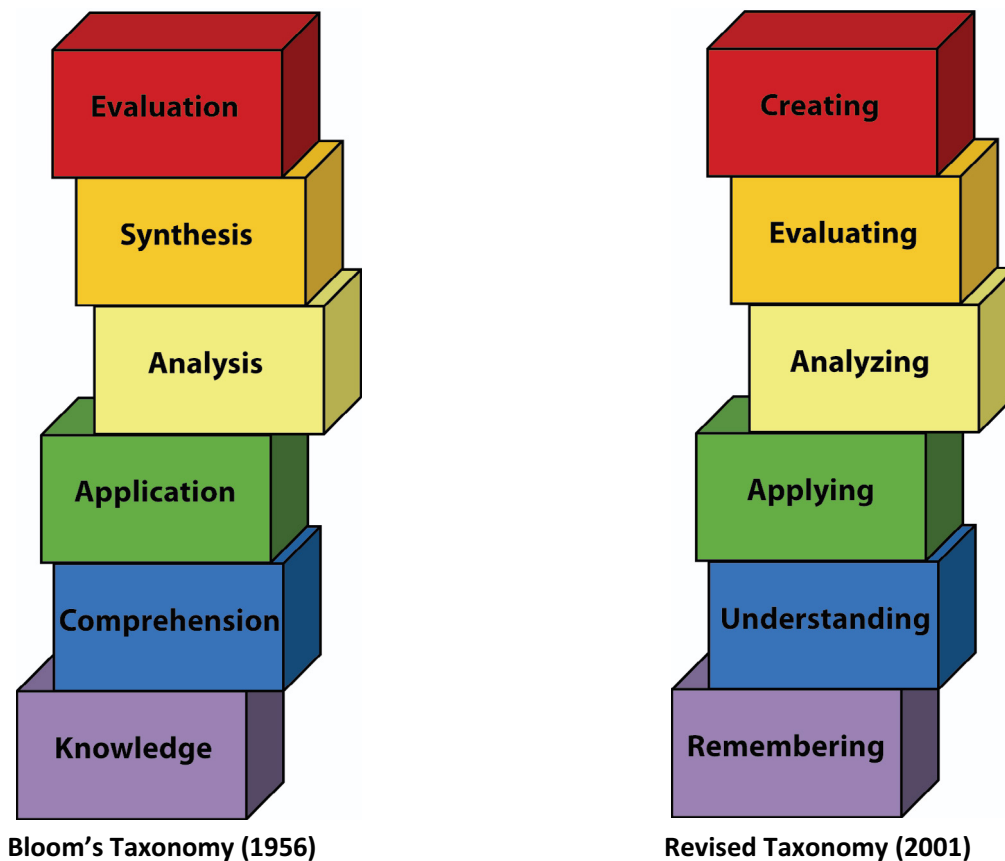
Moreover, validity of score inferences is bolstered when test scores are consistent. Here, the reliabilities of the total test scores presented in Chapter Eighteen were high.

As reported in Chapter Five, DIF with respect to gender, ethnicity, and test administration mode helps address construct-irrelevant variance, which represents an important threat to the validity of inferences made from achievement test scores. As noted in that chapter, field test items are screened and reviewed for DIF. Only items approved by data review committees are eligible for operational use.

APPENDIX A: UNDERSTANDING DEPTH OF KNOWLEDGE AND COGNITIVE COMPLEXITY

One of the steps in the item review process involves Pennsylvania educators' review of items for cognitive complexity or the nature of thinking. One model for classifying thinking into cognitive levels of complexity is Bloom's Taxonomy. Bloom's Taxonomy was first presented in 1956 through the publication, *The Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain*. This taxonomy identifies six levels within the cognitive domain, from the simple recall or recognition of facts, at the lowest level, through increasingly more complex levels, to the highest level which is classified as evaluation.

During the late 1990s, the original Bloom's Taxonomy was revised (Anderson and Krathwohl, 2001). In the 2001 version of Bloom's Taxonomy, the names of the six major cognitive process categories or levels were revised to indicate action (verbs) rather than non-action (nouns) as noted in the graphic below.



More recently, Webb's Depth-of-Knowledge Levels have also been used in the review of items for cognitive demand. Webb's Depth of Knowledge was created by Norman Webb from the Wisconsin Center for Education Research. Webb's definition of depth of knowledge is the degree or complexity of knowledge that the content curriculum standards and expectations require. Therefore, when reviewing items for depth of knowledge, the item is reviewed to determine whether or not it is as demanding cognitively as what the actual content curriculum standard expects. In the case of the Pennsylvania Keystone items, the item meets the criterion if the depth of knowledge of the item is in alignment with the depth of knowledge of the Assessment Anchor as defined by the Eligible Content.

Webb's Depth of Knowledge includes four levels, from the lowest (basic recall) to the highest (extended thinking). Verb examples that represent each level in Webb's Depth of Knowledge can be found in the information that follows. However, verbs alone do not describe the depth of knowledge. Rather, depth of knowledge also focuses upon how well the students need to know the content before they can respond to a given item.

Because Bloom's Taxonomy (1956) is very familiar to many teachers, information comparing Bloom's Taxonomy and Webb's Depth of Knowledge is provided to Pennsylvania educators during the review of the Keystone items. The comparison serves as a "bridge" for teachers to understand Webb's Depth of Knowledge as compared to Bloom's Taxonomy.

ALGEBRA I DEPTH OF KNOWLEDGE

DEPTH OF KNOWLEDGE GUIDELINES FOR REVIEW OF ALGEBRA I, ALGEBRA II, AND GEOMETRY ITEMS

Committees of Pennsylvania educators review each Keystone Exam item, not only to determine whether or not the item measures what it is intended to measure, but also to determine whether or not the item aligns with the cognitive level or depth of knowledge of the Assessment Anchor as defined by the Eligible Content. The information below provides a definition of the four depth-of-knowledge levels. The charts at the end of the section also provide a comparison between Bloom's Taxonomy and Webb's Depth of Knowledge for mathematics (Algebra I, Algebra II, and Geometry). Included are examples of verbs (i.e., the action). Using this information as well as the charts, Pennsylvania educators are asked to determine the depth of knowledge of each item and to verify that the depth of knowledge of each item is in alignment with the depth of knowledge of the Assessment Anchor as defined by the Eligible Content.

DEFINITIONS OF WEBB'S DEPTH OF KNOWLEDGE

Level 1 (Recall) requires the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify Level 1 include "identify," "recall," "recognize," "use," and "measure." Verbs such as "describe" and "explain" could be classified at different levels, depending on what is to be described and explained.

Level 2 (Skill/Concept) requires the engagement of some mental processing beyond a habitual response. A Level 2 item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of objects or phenomena and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph, or reading information from the graph, are also at Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited only to number skills, but may involve visualization skills and probability skills. Other Level 2 activities include noticing or describing non-trivial patterns; explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and deciding which concepts to apply in order to solve a complex problem.

Level 4 (Extended Thinking) requires complex reasoning, planning, developing, and thinking most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2. However, if the student is to conduct a river study that requires

taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas *within* the content area or *among* content areas—and have to select one approach among many alternatives on how the situation should be solved, in order to be at this highest level. Level 4 activities include designing *and* conducting experiments and projects; developing and proving conjectures; making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

Note: Multiple-choice and constructed-response items can be written at a depth-of-knowledge Level 4; however, to design an item in this format is difficult, as it would require research, investigation, and application, often over an extended period of time (e.g., performance-based tasks; portfolios; research studies/projects).

(Webb, N. 1997, 1999, 2002, 2005, 2006)

Table A–1. Bloom’s Taxonomy – Algebra I

Categories (1956)	Definition	Examples of Action Words*
Knowledge	Student remembers, or recalls appropriate previously learned information.	define; identify; name; select; state; order; (involves a one-step problem)
Comprehension	Student translates, comprehends, or interprets information based on prior learning.	convert; estimate; explain; express; factor; generalize; give example; identify; indicate; locate; picture; (involves two or more steps)
Application	Student selects, transfers, and uses data and principles to complete a task or problem with minimum directions.	apply; choose; compute; employ; interpret; graph; modify; operate; plot; practice; solve; use; (involves three or more steps)
Analysis	Student distinguishes, classifies, and relates assumptions, hypotheses, evidence, or structure of a statement or question.	compare; contrast; correlate; differentiate; discriminate; examine; infer; maximize; minimize; prioritize; subdivide; test
Synthesis	Student originates, integrates, and combines ideas into a product, plan, or proposal that is new to him or her.	arrange; collect; construct; design; develop; formulate; organize; set up; prepare; plan; propose; create experiment and record data
Evaluation	Student appraises, assesses, or critiques on a basis of specific standards and criteria.	appraise; assess; defend estimate; evaluate; judge; predict; rate; validate; verify

Table A–2. Webb’s Depth of Knowledge – Algebra I

Categories	Definition	Example of Action Words*
Recall	Student recalls facts, information, procedures, or definitions.	define; identify; name; select; state; order; one step
Basic Application of Skill/ Concept	Student uses information, conceptual knowledge, and procedures.	apply; choose; compute; employ; interpret; graph; modify; operate; plot; practice; solve; use; two or more steps
Strategic Thinking	Student uses reasoning and develops a plan or sequence of steps; process has some complexity.	compare; contrast; correlate; differentiate; discriminate; examine; infer; maximize; minimize; prioritize; subdivide; test
Extended Thinking	Student conducts an investigation, needs time to think and process multiple conditions of the problem or task. (The item/task generally requires several days or weeks to complete.)	arrange; collect; construct; design; develop; formulate; organize; set up; prepare; plan; propose; create experiment and record data

*Some action words (verbs) can be classified at different depth-of-knowledge levels depending on the context of the item and the complexity of the action.

BIOLOGY DEPTH OF KNOWLEDGE

Note: “Knowledge” can refer both to content knowledge and knowledge of *scientific processes*. This meaning of knowledge is consistent with the National Science Education Standards (NSES), which terms “Science as Inquiry” as its first Content Standard.

Committees of Pennsylvania educators review each Keystone Exam item, not only to determine whether or not the item measures what it is intended to measure, but also to determine whether or not the item aligns with the cognitive level or depth of knowledge of the Assessment Anchor as defined by the Eligible Content. The information below provides a definition of the four depth-of-knowledge levels. The charts at the end of the section also provide a comparison between Bloom’s Taxonomy and Webb’s Depth of Knowledge for biology. Included are examples of verbs (i.e., the action). Using this information as well as the charts, Pennsylvania educators are asked to determine the depth of knowledge of each item and to verify that the depth of knowledge of each item is in alignment with the depth of knowledge of the Assessment Anchor as defined by the Eligible Content.

DEFINITIONS OF WEBB’S DEPTH OF KNOWLEDGE

Level 1 (Recall) requires the recall of information, such as a fact, definition, term, or a simple procedure, as well as performance of a simple science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A “simple” procedure is well defined and typically involves only one step. Verbs such as “identify,” “recall,” “recognize,” “use,” “calculate,” and “measure” generally represent cognitive work at the recall level. Simple word problems that can be directly translated into and solved by a formula are considered Level 1. Verbs such as “describe” and “explain” could be classified at different depth-of-knowledge levels, depending on the complexity of what is to be described and explained.

A student answering a Level 1 item either knows the answer or does not: that is, the item does not need to be “figured out” or “solved.” In other words, if the knowledge necessary to answer an item automatically provides the answer to it, then the item is at Level 1. If the knowledge needed to answer the item is not automatically provided in the stem, the item is at least at Level 2. Some examples that represent but do not constitute all Level 1 performance are as follows:

- Recall or recognize a fact, term, or property.
- Represent in words or diagrams a scientific concept or relationship.
- Provide or recognize a standard scientific representation for simple phenomenon.
- Perform a routine procedure, such as measuring length.

Level 2 (Skills and Concepts) requires the engagement of some mental processing beyond recalling. The content knowledge or process involved is **more complex** than in Level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply **more than one step**. For example, to compare data requires first identifying characteristics of the objects or phenomena and then grouping or ordering the objects. Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts. Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different depth-of-knowledge levels, depending on the complexity of the action. For example, interpreting information from a simple graph, which requires reading information from the graph, is a Level 2. An item that requires interpretation from a complex graph, such as making decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated, is at Level 3. Some examples that represent but do not constitute all of Level 2 performance are as follows:

- Specify and explain the relationship between facts, terms, properties, or variables.
- Describe and explain examples and non-examples of science concepts.
- **Select a procedure according to specified criteria and perform it.**

- Formulate a routine problem, given data and conditions.
- Organize, represent, and interpret data.

Level 3 (Strategic Thinking) requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity does not result only from the fact that there could be multiple answers, a possibility for both Levels 1 and 2, but because the multi-step task requires more demanding reasoning. In most instances, requiring students to explain their thinking is at Level 3; requiring a very simple explanation or a word or two should be at Level 2. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Experimental designs in Level 3 typically involve more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems. Some examples that represent but do not constitute all Level 3 performance are as follows:

- Identify research questions and design investigations for a scientific problem.
- Solve non-routine problems.
- Develop a scientific model for a complex situation.
- Form conclusions from experimental data.

Level 4 (Extended Thinking) requires high cognitive demands and complexity. Students are required to make several connections—relate ideas within the content area or among content areas—and have to select or devise one approach among many alternatives to solve the problem. Many on-demand assessment instruments will not include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. “Develop generalizations of the results obtained and the strategies used and apply them to new problem situations,” is an example of a grade 8 objective that is a Level 4. Many, but not all, performance assessments and open-ended assessment activities requiring significant thought will be Level 4.

Level 4 involves complex reasoning, experimental design and planning, and probably will require an extended period of time either for the science investigation required by an objective, or for carrying out the multiple steps of an assessment item. However, the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student is asked to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4. Some examples that represent but do not constitute all Level 4 performance are as follows:

- Based on data provided from a complex experiment that is novel to the student, deduct the fundamental relationship between several controlled variables.
- Conduct an investigation, from specifying a problem to designing and carrying out an experiment, to analyzing its data and forming conclusions.

Note: Multiple-choice and constructed-response items can be written at a depth-of-knowledge Level 4; however, to design an item in this format is difficult, as it would require research, investigation, and application, often over an extended period of time (e.g. performance-based tasks, portfolios, research studies/projects).

(Webb, N. 1997, 1999, 2002, 2005, 2006)

Table A–3. Bloom’s Taxonomy – Biology

Categories (1956)	Definition	Examples of Action Words*
Knowledge	Student remembers or recalls appropriate previously learned information.	identify; recall; observe; recognize; use; calculate; measure; order
Comprehension	Student translates, comprehends, or interprets information based on prior learning.	explain; interpret; describe; classify; identify; recognize; predict
Application	Student selects, transfers, and uses data and principles to complete a task or problem with minimum directions.	apply; classify; experiment; interpret; use; order; calculate
Analysis	Student distinguishes, classifies, and relates assumptions, hypotheses, evidence, or structure of a statement or question.	analyze; order; explain; classify; arrange; compare; contrast; infer; calculate; categorize; examine; experiment; question; test
Synthesis	Student originates, integrates, and combines ideas into a product, plan, or proposal that is new to him or her.	combine; arrange; rearrange; modify; invent; design; construct; organize; predict; infer; conclude; create; experiment and record data
Evaluation	Student appraises, assesses, or critiques on a basis of specific standards and criteria.	evaluate; measure; explain; compare; summarize; predict; test decide; rate; conclude

Table A–4. Webb’s Depth of Knowledge – Biology

Categories	Definition	Examples of Action Words*
Recall	Student recalls facts, information, procedures, or definitions.	identify; recall; observe; recognize; use; calculate; measure; order
Basic Application of Skill/ Concept	Student uses information, conceptual knowledge, and procedures.	explain; interpret; describe; classify; identify; order; recognize; predict; apply; use; calculate; organize; estimate; observe; collect; and display data
Strategic Thinking	Student uses reasoning and develops a plan or sequence of steps; process has some complexity.	analyze; order; explain; classify; arrange; compare; contrast; infer; interpret; calculate; categorize; examine; experiment; question; predict; evaluate; test
Extended Thinking	Student conducts an investigation, needs time to think and process multiple conditions of the problem or task. (The item/task generally requires several days or weeks to complete.)	combine; arrange; rearrange; propose; evaluate; modify; invent; design; construct; organize; predict; infer; conclude; evaluate; create; experiment and record data

*Some action words (verbs) can be classified at different depth-of-knowledge levels depending on the context of the item and the complexity of the action.

LITERATURE DEPTH OF KNOWLEDGE

Note: *The levels are based on Valencia and Wixson (2000, pp. 909–935).*

Committees of Pennsylvania educators review each Keystone Exam item, not only to determine whether or not the item measures what it is intended to measure, but also to determine whether or not the item aligns with the cognitive level or depth of knowledge of the Assessment Anchor as defined by the Eligible Content. The information below provides a definition of the four depth-of-knowledge levels. The charts at the end of the section also provide a comparison between Bloom’s Taxonomy and Webb’s Depth of Knowledge for literature. Included are examples of verbs (i.e., the action). Using this information as well as the charts, Pennsylvania educators are asked to determine the depth of knowledge of each item and to verify that the depth of knowledge of each item is in alignment with the depth of knowledge of the Assessment Anchor as defined by the Eligible Content.

DEFINITIONS OF WEBB’S DEPTH OF KNOWLEDGE

Level 1 requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text, as well as basic comprehension of a text, is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from text, slight paraphrasing of specific details from the text, or simple understanding of a single word or phrase. Some examples that represent but do not constitute all Level 1 performance are as follows:

- Support ideas by reference to verbatim or only slightly paraphrased details from the text.
- Use a dictionary to find the meanings of words.
- Recognize figurative language in a reading passage.

Level 2 requires the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Inter-sentence analysis of inference is required. Some important concepts are covered, but not in a complex way. Content curriculum standards and items at this level may include words such as summarize, interpret, infer, classify, organize, collect, display, compare, and determine whether fact or opinion. Literal main ideas are stressed. A Level 2 item may require students to apply skills and concepts that are covered in Level 1. However, items require closer understanding of text, possibly through the item’s paraphrasing of both the question and the answer. Some examples that represent but do not constitute all Level 2 performance are as follows:

- Use context cues to identify the meaning of unfamiliar words, phrases, and expressions that could otherwise have multiple meanings.
- Predict a logical outcome based on information in a selection.
- Identify and summarize the major events in a narrative.

Level 3 requires deeper knowledge. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Content curriculum standards and items (Assessment Anchors as defined by the Eligible Content) at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students’ application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent but do not constitute all Level 3 performance are as follows:

- Explain or recognize how the author’s purpose affects the interpretation of a selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Level 4 requires higher-order thinking and deep knowledge. The content curriculum standard or item at this level will probably require an extended activity, with extended time provided for completing it. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. Students take information from

at least one passage of a text and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent but do not constitute all Level 4 performance are as follows:

- Analyze and synthesize information from more than one source.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

Note: Multiple-choice and constructed-response items can be written at a depth-of-knowledge Level 4; however, to design an item in this format is difficult, as it would require research, investigation, and application, often over an extended period of time (e.g. performance-based tasks, portfolios, research studies/projects).

(Webb, N. 2005; Valencia and Wixson, 2000)

Table A-5. Bloom’s Taxonomy – Literature

Categories (1956)	Definition	Examples of Action Words*
Knowledge	Student remembers or recalls appropriate previously learned information.	define; identify; name; recall; recognize; select; tell
Comprehension	Student translates, comprehends, or interprets information based on prior learning.	describe; distinguish; explain; identify; indicate; interpret; locate; recognize; restate; summarize
Application	Student selects, transfers, and uses data and principles to complete a task or problem with minimum directions.	apply; choose; demonstrate; determine; interpret; inform; select; show; use
Analysis	Student distinguishes, classifies, and relates assumptions, hypotheses, evidence, or structure of a statement or question.	analyze; characterize; compare; contrast; discriminate; distinguish; explain; infer
Synthesis	Student originates, integrates, and combines ideas into a product, plan, or proposal that is new to him or her.	compose; create; develop; formulate; generalize; organize
Evaluation	Student appraises, assesses, or critiques on a basis of specific standards and criteria.	assess; conclude; convince; defend; evaluate; explain; justify; predict; prove; support

Table A-6. Webb’s Depth of Knowledge – Literature

Categories	Definition	Examples of Action Words*
Recall	Student recalls facts, information, procedures, or definitions.	define; identify; locate; name; recall; recognize; sequence; tell
Basic Application of Skill/Concept	Student uses information, conceptual knowledge, and procedures.	apply; compare; comprehend; identify; describe; determine; infer; interpret; predict; summarize; use
Strategic Thinking	Student uses reasoning and develops a plan or sequence of steps; process has some complexity.	analyze; cite evidence; compare; contrast; draw conclusions; explain; generalize; infer; interpret; evaluate; recognize; summarize; support
Extended Thinking	Student conducts an investigation, needs time to think and process multiple conditions of the problem or task. (The item/task generally requires several days or weeks to complete.)	describe and illustrate; evaluate; examine and explain; analyze; synthesize

*Some action words (verbs) can be classified at different depth-of-knowledge levels depending on the context of the item and the complexity of the action.

APPENDIX B: GENERAL SCORING GUIDELINES

ALGEBRA I

4 Points

- The response demonstrates a *thorough* understanding of the mathematical concepts and procedures required by the task.
- The response provides correct answer(s) with clear and complete mathematical procedures shown and a correct explanation, as required by the task. Response may contain a minor “blemish” or omission in work or explanation that does not detract from demonstrating a thorough understanding.

3 Points

- The response demonstrates a *general* understanding of the mathematical concepts and procedures required by the task.
- The response and explanation (as required by the task) are mostly complete and correct. The response may have minor errors or omissions that do not detract from demonstrating a general understanding.

2 Points

- The response demonstrates a *partial* understanding of the mathematical concepts and procedures required by the task.
- The response is somewhat correct with partial understanding of the required mathematical concepts and/or procedures demonstrated and/or explained. The response may contain some work that is incomplete or unclear.

1 Point

- The response demonstrates a *minimal* understanding of the mathematical concepts and procedures required by the task.

0 Points

- The response has no correct answer and *insufficient* evidence to demonstrate any understanding of the mathematical concepts and procedures required by the task.

Special Categories within zero reported separately:

Blank	Blank, entirely erased, entirely crossed out, or consists entirely of whitespace
Refusal	Refusal to respond to the task
Off Task	Makes no reference to the item but is not an intentional refusal
Foreign Language	Written entirely in a language other than English
Illegible	Illegible or incoherent

COPYRIGHT© PA DEPARTMENT OF EDUCATION. DO NOT DUPLICATE.

BIOLOGY

3 Points

- The response demonstrates a *thorough* understanding of the scientific content, concepts, and/or procedures required by the task(s).
- The response provides a clear, complete, and correct response as required by the task(s). The response may contain a minor blemish or omission in work or explanation that does not detract from demonstrating a thorough understanding.

2 Points

- The response demonstrates a *partial* understanding of the scientific content, concepts, and/or procedures required by the task(s).
- The response is somewhat correct with partial understanding of the required scientific content, concepts, and/or procedures demonstrated and/or explained. The response may contain some work that is incomplete or unclear.

1 Point

- The response demonstrates a *minimal* understanding of the scientific content, concepts, and/or procedures required by the task(s).
- The response is somewhat correct with minimal understanding of the required scientific content, concepts, and/or procedures demonstrated and/or explained. The response may contain some work that is incomplete or unclear.

0 Points

- The response provides *insufficient* evidence to demonstrate any understanding of the scientific content, concepts, and/or procedures as required by the task(s).
- The response may show only information copied or rephrased from the question or insufficient correct information to receive a score of 1.

Special Categories within zero reported separately:

Blank.....Blank, entirely erased, entirely crossed out, or consists entirely of whitespace

Refusal.....Refusal to respond to the task

Off Task.....Makes no reference to the item but is not an intentional refusal

Foreign Language.....Written entirely in a language other than English

IllegibleIllegible or incoherent

COPYRIGHT© PA DEPARTMENT OF EDUCATION. DO NOT DUPLICATE.

LITERATURE

3 Points

- The response provides a clear, complete, and accurate answer to the task.
- The response provides relevant and specific information from the passage.

2 Points

- The response provides a partial answer to the task.
- The response provides limited information from the passage and may include inaccuracies.

1 Point

- The response provides a minimal answer to the task.
- The response provides little or no information from the passage and may include inaccuracies.

OR

- The response relates minimally to the task.

0 Points

- The response is totally incorrect or irrelevant or contains insufficient information to demonstrate comprehension.

Special Categories within zero reported separately:

Blank.....Blank, entirely erased, entirely crossed out, or consists entirely of white space

Refusal.....Refusal to respond to the task

Off Task.....Makes no reference to the item but is not an intentional refusal

Foreign Language.....Written entirely in a language other than English

IllegibleIllegible or incoherent

COPYRIGHT© PA DEPARTMENT OF EDUCATION. DO NOT DUPLICATE.

APPENDIX C: ITEM AND TEST DEVELOPMENT PROCESS FOR THE KEYSTONE EXAMS

Table C–1. Item and Test Development Process for the Keystone Exams

Step	Description
1. Review Guiding Documentation	Each year item and test development specialists meet internally to review all guiding documentation related to the Keystone Exams. Documentation reviewed includes the test design blueprints, the Keystone Assessment Anchors and Eligible Content, the test item specifications, the test style specifications (style guide), and all test content descriptions.
2. Meet with PDE to Confirm Understanding of Program	The goal of the meeting each year is to ensure that item and test development teams have a clear understanding of PDE’s vision for test development. A successful development cycle requires a clear understanding of Pennsylvania’s content-area test specifications and of any unique interpretations of the Keystone Assessment Anchors (if any).
3. Create Preliminary Test Item Development Plan	Item and test development specialists generate a preliminary development plan which includes an overview of the program, the internal and external (PDE) review and approval processes, a projected schedule for development of test items—including the number of test items to be developed for review by PDE and subsequent review by the committees of Pennsylvania educators. Item and test development specialists also generate strategies for securing passages and developing passage-based items, etc.
4. Meet with PDE to Finalize Test Item Development Plan	Over the course of the meeting, item and test development specialists verify all steps in the development process including timelines and schedules for test item/test development.
5. Analyze Item Bank	Existing test items in the current Keystone Exams Item Bank are reviewed for technical psychometric quality as well as for their match to the Assessment Anchors. During this phase, test development specialists also make a tally of the test items by Assessment Anchor—including test development specialists’ best thinking regarding the number of usable test items in the existing item bank. A tally is also made of the number of usable passages, as well as other stimulus prompts in the bank, including science scenarios.
6. Refine Test Item Development Plan to Include Writers and Subcontractors	Item and test development specialists identify the writers who will write the test items (test development specialists or other professional item writers, subcontractors, etc.), the estimated number of writers needed, the qualifications of writers, and the approximate number of test items to be submitted by each source.
7. Train Item Writers	Item and test development specialists train item writers, as needed. Item writers who have written for the Keystone Exams in the past receive updated information, as needed.
8. Write and Review Items	Test items are written by item writers after training is complete, and feedback is provided by the item and test development specialists to item writers on a regular basis. As test items are written, they are reviewed and edited in a series of internal reviews. Item and test development specialists review and edit items to include, but not limited to, the following: match to Assessment Anchor/Eligible Content, relevance to purpose, accuracy of content, item difficulty, interest level, depth of knowledge and cognitive complexity, adherence to the principles of Universal Design, and freedom from issues of bias/fairness/sensitivity. At the same time, the process of procuring permissions also begins, including securing permissions for passages, art, etc.
9. Enter Test Items into Database	Upon acceptance from item writers, test items are entered into the item management system, IDEAS (Item Development and Educational Assessment System). Item data stored in the system database includes, but is not limited to, the following: readability, cognitive level, estimated level of difficulty, alignment to assessment anchors, and correlation to stimulus passages.
10. Prepare Item Set for Sample Item Review by PDE	Item and test development specialists prepare a subset of the items for review by PDE.
11. PDE Conducts Sample Item Review	After a subset of the items is submitted to PDE for review, PDE reviews the items and provides feedback to item and test development teams via a conference call. Items are revised per PDE feedback.

Table C–1 (continued). Item and Test Development Process for the Keystone Exams

Step	Description
12. Continue to Write and Review Items	The remaining items are written, and feedback is provided by the item and test development specialists to item writers on a regular basis. Items are entered into the item management system, IDEAS (Item Development and Educational Assessment System) (See step 8 and step 9).
13. Review Items Prior to Test Item Review and Validation Sessions	Prior to New Item Content Review, all items are submitted to PDE for review. Item and test development specialists incorporate all PDE feedback, and PDE-requested edits to items are made.
14. Prepare for Test Item Review Sessions (the New Item Content Review and the Bias, Fairness, and Sensitivity Review)	Item and test development specialists prepare all items and stimulus passages for review by the New Item Content Review Committee (consisting of Pennsylvania educators) and by the separate Bias, Fairness, and Sensitivity Committee (consisting of a panel of experts). Item and test development specialists also prepare training materials needed for training committee members to review items for content or for bias, fairness, and sensitivity issues. All training materials and other ancillary materials (e.g. agendas, presentations, etc.) are also developed and then submitted to PDE for review and approval. Invitations are also sent to Pennsylvania educators and national experts from PDE-approved committee lists.
15. Conduct Test Item Review Sessions (the New Item Content Review and the Bias, Fairness, and Sensitivity Review)	Committees of Pennsylvania educators and national experts review items in two meetings: one addressing item content and quality, the other addressing bias, fairness, and sensitivity. PDE, with support from item and test development specialists, presents training on how to review new test items for content considerations or bias/fairness/sensitivity issues. At the New Item Content Review, suggested edits to test items are made and/or replacement test items are written during the actual item review so that both the committee and the PDE are able to observe changes to the test items and approve the test items during the committee review process. At the Bias, Fairness, and Sensitivity Review, experts in bias, fairness, and sensitivity review all test items and passages and come to a consensus about any issues that are noted. At both meetings the results are carefully documented.
16. Conduct Item Review Resolution and Cleanup	Following the conclusion of the New Item Content Review Committee meetings, PDE re-examines the consensus changes suggested by the committee members during the New Item Content Review Committee meetings. DRC item and test development specialists then record all of PDE's follow-up decisions and changes. During this cleanup process, PDE either accepts the changes as requested by the committee, or PDE rejects the decision of the committee. If a committee decision is rejected, PDE provides an alternate decision for DRC to implement. During this cleanup process, PDE also interprets the report from the Bias, Fairness, and Sensitivity Committee meetings and subsequently applies changes to test items and passages. DRC item and test development specialists then apply the changes to the test items and passages per PDE's decisions.
17. Submit Field Test Items for Final Sign-Off	PDE-approved changes are applied to the items, non-permissioned passages, prompts, etc. (Changes reflect PDE's arbitration of the committee decisions.) Once all revisions to the items, non-permissioned passage text, and/or the art used by test items and passages are completed, the test items are submitted to PDE for final review and sign-off. (Changes requested to permissioned passages are sought from the publisher of record, and, if approved by the copyright holders, changes are implemented.) [PDE's approval process for field test items generally occurs simultaneously with PDE's approval of the core test forms. See step 25.]
	To follow the path for new field test items, skip to step 22, or to follow the chronological test development path, continue with step 18.
18. Review Results of the Field Test	Following the administration of a field test form and the subsequent rangefinding and field test scoring processes for field test items, performance data for all field test items are analyzed by DRC psychometricians and test development specialists. Test item performance data that meet certain triggering criteria are flagged for additional reviews by test development specialists. Flagged field test items with extreme performance data are considered psychometrically unusable and are removed from future operational consideration. Field test items with marginal performance data are prepared for the Field Test Item Data Review meeting.

Table C–1 (continued). Item and Test Development Process for the Keystone Exams






Step	Description
19. Prepare for Field Test Item Data Review	Test development specialists prepare all items and stimulus passages for review by the Field Test Item Data Review Committee (which consists of Pennsylvania educators). Psychometricians also prepare training materials needed for training committee members to review items for their performance. All training materials and other ancillary materials (e.g. agendas, presentations, etc.) are submitted to PDE for review and approval. Invitations are also sent to Pennsylvania educators from PDE-approved committee lists.
20. Conduct Field Test Item Data Review	Committees of Pennsylvania educators review the performance data of flagged field test items. Psychometricians present training on how to review field test items based on their performance data. At the Item Data Review, committee members examine the performance of the items and determine whether the field test item is technically sound and appropriate for use on an operational Keystone Exams test. Since test items cannot be modified at the Field Test Item Data Review, the committee can either accept an item as is or the committee can reject the item.
21. Conduct Field Test Item Data Review Reconciliation	Following the conclusion of the Field Test Item Data Review Committee meetings, PDE re-examines the consensus decisions (accept or reject) suggested by the committee members during the Field Test Item Data Review Committee meetings. Test development specialists record all of PDE's follow-up decisions and changes. During this cleanup process, PDE either accepts the decisions of the data review committee, or PDE rejects the decisions of the data review committee. If a committee decision is not accepted, PDE provides an alternate decision for test development specialists to implement. All PDE-approved changes to the test items status (accepted or rejected) are incorporated into the Item Development and Educational Assessment System, IDEAS.
22. Select Items to Fill Core, Field Test, and Equating Block Positions in Core and Field Test Forms	After the PDE-approved changes to the new field test items is completed AND the results of the prior field test have been finalized following data review, test development specialists collaborate with psychometricians to follow the Test Design Blueprints and build requirements to make the initial selection of items for core and field test positions for all test forms. In later administrations, core-to-core linking items will also be selected during this step.
23. Review Core and Equating Block Selections	After test content and psychometric requirements have been achieved for core, the core items are provided to PDE for review and approval. Any changes to the content of the core requested by PDE are balanced with psychometric requirements until all core positions are approved by PDE, test development specialists, and psychometricians.
24. Construct Test Forms	Items, passages, and test components are assembled into forms using the form construction and typesetting function of DRC's Item Development and Educational Assessment System, IDEAS. Forms are reviewed internally for style and formatting requirements.
25. Review Typeset Forms	After forms are constructed in IDEAS, draft hard copies of the forms are produced and presented to PDE for review and approval. Any changes to the content of the core requested by PDE are balanced with psychometric requirements until all core positions are approved by PDE, test development specialists, and psychometricians. PDE also re-reviews all field test items appearing in the test forms. DRC applies changes to the field test items as required.
26. Print Test Forms	Following PDE's approval of the test forms, DRC completes a series of final proofing of all test forms. Final forms (along with ancillary materials) are then approved for printing.
27. Assemble Documentation of Test Materials	Metadata for each test item and form is documented and proofed, including: grade, form, session/section, item sequence, reporting category, Assessment Anchor, Descriptor, Eligible Content, number of points, item type, number of answer options, item usage, stimulus ID, etc.
28. Prepare Online Forms	Following approval of the print forms, all online forms are prepared. Forms are rendered in form sets, and items and forms are compared for continuity with the print forms as well as to ensure that all tools and features are functioning as expected.
	To follow the path for new field test items, return to step 18.

APPENDIX D: ITEM AND DATA REVIEW CARD EXAMPLES

ITEM REVIEW CARD EXAMPLE

<p>Standard: Explain how factors such as pH, temperature, and concentration levels can affect enzyme function.</p>	PA Keystone Item Card
<p>1. <i>[Blurred text]</i></p> <p>A. <i>[Blurred text]</i></p> <p>B. <i>[Blurred text]</i></p> <p>C. <i>[Blurred text]</i></p> <p>D. <i>[Blurred text]</i></p>	
	Item ID
	Content Area
	Science
	Course
	Biology
	Scenario ID
	Scenario Title
	Grade
	HS
	KAACS Standards
	BIO.A.2.3.2
	Item Type
	Multiple Choice
	Points
	1
	Depth of Knowledge
	2
	Est Difficulty
	Medium
	Key

DATA REVIEW CARD EXAMPLE

Standard: Compare and/or order any real numbers (rational and irrational may be mixed).		PA Keystone Data Card
<p>1. </p> <p>A. </p> <p>B. </p> <p>C. </p> <p>D. </p>		
		Item ID
		Content Area
		Mathematics
		Course
		Algebra I
		Passage ID
		Passage Title
		Grade
		HS
		Standards
		KAACS: A1.1.1.1.1
		Item Type
		Multiple Choice
		Points
		1
		Calculator
		Yes
		Depth of Knowledge
		2
		Est Difficulty
		High
		Key
		Focus

Administration

Form Name	Use Function	Rptg Flag	Seq	Period	Year	Session	Calc	Mode/Ext	Grade
						1	Yes		HS

Traditional Statistics

N	P-Val	Mean	Item Total Corr
	0.34		0.10

Fit Statistics

Outfit t	Infit t	Outfit MnSq	Infit MnSq	Chi-sq	Deg Free	Item Fit	Fit
9.9	9.9	1.28	1.18				

IRT Statistics

Label	Final	Final S.E.	Preliminary	Preliminary S.E.
Location	1.39	0.02		

Distractor/Step Specific

Label	Proportion	Corr	Avg Meas	Threshold
A*	0.34	0.10		
B	0.24	0.11		
C	0.25	-0.22		
D	0.17	0.01		
MULTS	0.00			
OMITS	0.00			

DIF Analysis

Category	Bias Code	Num Value	N - Ref	N - Focal
MALEFEMALE	A-	-0.13	4709	4550
PAPERONLINE	A+	0.15	8242	1029
WHITEBLACK	A-	-0.23	6812	1245
WHITEHISPANIC	A-	-0.16	6812	726

The purpose of this form is to provide guidelines to the item review process in terms of item characteristics that are essential in building a fair and balanced assessment. Use these guidelines in conjunction with the Item Rating Sheet when recording your feedback on individual items.

	Content Alignment	Options
Standards, Anchors, Eligible Content	Does the content of the item align with the Standard/Anchor/Eligible Content? Each item was written to assess a particular Standard/Anchor/Eligible Content statement which is indicated on the individual Item Card. Consider the degree to which the item is, in fact, aligned with the indicated eligible content. In making this judgment, it is important to consider whether the content is aligned (e.g., do the eligible content and the item both deal with fractions) and whether the required performance is aligned (e.g., if the eligible content calls for a comparison to be made, is this reflected in the item).	HIGHER —Aligns to the higher level of the EC LOWER —Aligns to the lower level of the EC NONE —No alignment with EC

	Rigor Level Alignment	Options
Grade	Is the item grade-level appropriate? Is the content consistent with the experiences of a student at the grade level assessed? Is the challenge level appropriate for the grade?	ABOVE Grade Level AT Grade Level BELOW Grade Level
Difficulty	Do you agree with the item’s difficulty rating? Item Difficulty is indicated as Low, Medium, and High. Is your rating in agreement with the difficulty rating on the Item Form?	HIGH MEDIUM LOW
Depth of Knowledge	Depth of Knowledge is based on the alignment work of Norman Webb. Rate each item based on the cognitive demand, using the following levels: <ol style="list-style-type: none"> 1. Recall – Recall of a fact, information, or procedure. 2. Basic Application of Skill or Concept – Use of information, conceptual knowledge, procedures, two or more steps, etc. 3. Strategic Thinking – Requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer. 4. Extended Thinking – Requires an investigation, time to think and process multiple conditions of the problem or task, and more than 10 minutes to do non-routine manipulations. (This level is generally not assessed in on-demand assessments.) 	4 = Extended Thinking 3 = Strategic Thinking 2 = Basic Application 1 = Recall

Rigor Level Alignment		Options
Source of Challenge	Is the source of challenge appropriately targeted to the content? The hardest part of the item (i.e., source of challenge) should be the content that is targeted. For example, in mathematics, the mathematics should be the major source of challenge rather than the wording or graphic. Students should not give an incorrect answer to a mathematics item because the reading level is too high or a graphic is flawed. Conversely, students should not give correct answers for reasons such as prior knowledge that make the answer to the question obvious (e.g., if the question asks which country has the largest population and students are to read a graph that includes China, there is no need to read the graph to answer the question).	Y = Yes N = No

Technical Design		Options
Correct Answer	Is there one clear, correct answer option? There should be no other answer that “could” be correct. CAUTION: This does not mean that “good” distractors are unfair.	Y = Yes N = No
Distractors	Are distractors fair and appropriate? Distractors that are appropriate offer students reasonable choices that can be arrived at by making common errors. There should be no distractors that make no sense at all. It should be possible to examine each option and to reason how a student with some deficiency in knowledge or skill could choose it. The distractors should be formatted according to acceptable standards of test construction (e.g., a phrase that is common to each distractor should be in the stem).	Y = Yes N = No N/A = OE items do not have distractors
Graphics	Are the graphics clear and accurate?	Y = Yes N = No N/A = No graphic

Universal Design		Options
Language Demand	Is language clear, well-formatted, and precise? Does the item use correct terminology for the content area? In order for all students to enter into the questions of the assessment, they must be able to understand them. If the items are formatted poorly, use unnecessarily complex words or phrases, or use figures or layouts that are difficult to understand, some students will give incorrect answers due to these factors rather than the content that is being assessed.	Y = Yes N = No
Bias	Is the item free of bias? All students will not be able to enter into the assessment if bias considerations are not resolved. Does the item contain clear bias problems? A <i>thorough, independent bias review</i> (separate from this meeting) <i>will be completed for all items</i> .	Y = Yes N = No

Status		Options
Acceptance Status	This is an overall judgment about the item. Based on the consensus of the committee, indicate whether the item was approved without revision to the content of the item or whether the item was accepted by the committee after revision of the content of the item. If there is a dissenting view (opposed to the committee consensus), record a	—Approved as is —Accepted with suggested revisions

NOTES:

- If you leave a box blank on the Item Rating Sheet, it will be recorded to indicate that you did not have any specific feedback for that item or issue.
- If you object to the consensus of the committee, please note this on the item rating sheet and then record a brief explanation of the dissenting view on the back of the Item Rating Sheet.
- **Do NOT remove any items from the item binder at any time.**
- You must sign your item rating sheet.

APPENDIX F: TALLY SHEETS

ALGEBRA I-WINTER 2020/2021													
Keystone Exam					Algebra I								
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items				Points				
					Number of Core Items				Core Points				
					MC	SCR	ECR	Total	MC	SCR	ECR	Total	
A1.1: Operations and Linear Equations & Inequalities	1			Operations with Real Numbers and Expressions		1		1		4		4	
	1	1	1	Compare and/or order any real numbers.	1			1	1			1	
	1	1	2	Simplify square roots.	1			1	1			1	
	1	2	1	Find the Greatest Common Factor (GCF) and/or the Least Common Multiple (LCM) for sets of monomials.									
	1	3	1	Simplify/evaluate expressions involving properties/laws of exponents, roots, and/or absolute values to solve problems.									
	1	4	1	Use estimation to solve problems.	1			1	1			1	
	1	5	1	Add, subtract, and/or multiply polynomial expressions (express answers in simplest form).	1			1	1			1	
	1	5	2	Factor algebraic expressions, including difference of squares and trinomials.	1			1	1			1	
	1	5	3	Simplify/reduce a rational algebraic expression.	1			1	1			1	
	Total For Assessment Anchor A1.1.1					6	1		7	6	4		10
	2			Linear Equations		1		1		4		4	
	2	1	1	Write, solve, and/or apply a linear equation.	1			1	1			1	
	2	1	2	Use and/or identify an algebraic property to justify any step in an equation-solving process.	1			1	1			1	
	2	1	3	Interpret solutions to problems in the context of the problem situation.	1			1	1			1	
	2	2	1	Write and/or solve a system of linear equations (including problem situations) using graphing, substitution, and/or elimination.	2			2	2			2	
	2	2	2	Interpret solutions to problems in the context of the problem situation.	1			1	1			1	
	Total For Assessment Anchor A1.1.2					6	1		7	6	4		10
	3			Linear Inequalities			1	1			4	4	
	3	1	1	Write or solve compound inequalities and/or graph their solution sets on a number line .	1			1	1			1	
	3	1	2	Identify or graph the solution set to a linear inequality on a number line.	1			1	1			1	
	3	1	3	Interpret solutions to problems in the context of the problem situation.	1			1	1			1	
	3	2	1	Write and/or solve a system of linear inequalities using graphing.	1			1	1			1	
	3	2	2	Interpret solutions to problems in the context of the problem situation.	2			2	2			2	
	Total For Assessment Anchor A1.1.3					6		1	7	6		4	10
Total For Reporting Category A1.1					18	2	1	21	18	8	4	30	

Keystone Exam

Algebra I

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items				Points				
					Number of Core Items				Core Points				
					MC	SCR	ECR	Total	MC	SCR	ECR	Total	
A1.2: Linear Functions & Data Organizations	1			Functions			1	1			4	4	
	1	1	1	Analyze a set of data for the existence of a pattern and represent the pattern algebraically and/or graphically.	1			1	1			1	
	1	1	2	Determine whether a relation is a function, given a set of points or a graph.	2			2	2			2	
	1	1	3	Identify the domain or range of a relation.	1			1	1			1	
	1	2	1	Create, interpret, and/or use the equation, graph, or table of a linear function.	1			1	1			1	
	1	2	2	Translate from one representation of a linear function to another.	1			1	1			1	
	Total For Assessment Anchor A1.2.1					6		1	7	6		4	10
	2			Coordinate Geometry			1	1			4	4	
	2	1	1	Identify, describe, and/or use constant rates of change.	1			1	1			1	
	2	1	2	Apply the concept of linear rate of change (slope) to solve problems.	1			1	1			1	
	2	1	3	Write or identify a linear equation when given...	1			1	1			1	
	2	1	4	Determine the slope and/or y-intercept represented by a linear equation or graph.	2			2	2			2	
	2	2	1	Draw, identify, find, and/or write an equation for a line of best fit for a scatter plot.	1			1	1			1	
	Total For Assessment Anchor A1.2.2					6		1	7	6		4	10
	3			Data Analysis		1		1		4		4	
	3	1	1	Calculate and/or interpret the range, quartiles, and interquartile range of data.	1			1	1			1	
	3	2	1	Estimate or calculate to make predictions based on a circle, line, bar graph, measures of central tendency, or other representations.	2			2	2			2	
	3	2	2	Analyze data, make predictions, and/or answer questions based on displayed data.	1			1	1			1	
	3	2	3	Make predictions using the equations or graphs of best-fit lines of scatter plots.	1			1	1			1	
	3	3	1	Find probabilities for compound events.	1			1	1			1	
	Total For Assessment Anchor A1.2.3					6	1		7	6	4		10
Total For Reporting Category A1.2					18	1	2	21	18	4	8	30	

ALGEBRA I-SPRING 2021

Keystone Exam

Algebra I

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items				Points				
					Number of Core Items				Core Points				
					MC	SCR	ECR	Total	MC	SCR	ECR	Total	
A1.1: Operations and Linear Equations & Inequalities	1			Operations with Real Numbers and Expressions			1	1			4	4	
	1	1	1	Compare and/or order any real numbers.									
	1	1	2	Simplify square roots.	1			1	1			1	
	1	2	1	Find the Greatest Common Factor (GCF) and/or the Least Common Multiple (LCM) for sets of monomials.	1			1	1			1	
	1	3	1	Simplify/evaluate expressions involving properties/laws of exponents, roots, and/or absolute values to solve problems.	1			1	1			1	
	1	4	1	Use estimation to solve problems.	1			1	1			1	
	1	5	1	Add, subtract, and/or multiply polynomial expressions (express answers in simplest form).	1			1	1			1	
	1	5	2	Factor algebraic expressions, including difference of squares and trinomials.									
	1	5	3	Simplify/reduce a rational algebraic expression.	1			1	1			1	
	Total For Assessment Anchor A1.1.1					6		1	7	6		4	10
	2			Linear Equations		1		1		4			4
	2	1	1	Write, solve, and/or apply a linear equation.	1			1	1				1
	2	1	2	Use and/or identify an algebraic property to justify any step in an equation-solving process.	1			1	1				1
	2	1	3	Interpret solutions to problems in the context of the problem situation.	2			2	2				2
	2	2	1	Write and/or solve a system of linear equations (including problem situations) using graphing, substitution, and/or elimination.	1			1	1				1
	2	2	2	Interpret solutions to problems in the context of the problem situation.	1			1	1				1
	Total For Assessment Anchor A1.1.2					6	1		7	6	4		10
	3			Linear Inequalities			1	1			4		4
	3	1	1	Write or solve compound inequalities and/or graph their solution sets on a number line .	2			2	2				2
	3	1	2	Identify or graph the solution set to a linear inequality on a number line.	1			1	1				1
	3	1	3	Interpret solutions to problems in the context of the problem situation.	1			1	1				1
	3	2	1	Write and/or solve a system of linear inequalities using graphing.	1			1	1				1
	3	2	2	Interpret solutions to problems in the context of the problem situation.	1			1	1				1
	Total For Assessment Anchor A1.1.3					6		1	7	6		4	10
	Total For Reporting Category A1.1					18	1	2	21	18	4	8	30

Keystone Exam

Algebra I

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items				Points				
					Number of Core Items				Core Points				
					MC	SCR	ECR	Total	MC	SCR	ECR	Total	
A1.2: Linear Functions & Data Organizations	1			Functions		1		1		4		4	
	1	1	1	Analyze a set of data for the existence of a pattern and represent the pattern algebraically and/or graphically.	1			1	1			1	
	1	1	2	Determine whether a relation is a function, given a set of points or a graph.	1			1	1			1	
	1	1	3	Identify the domain or range of a relation.	2			2	2			2	
	1	2	1	Create, interpret, and/or use the equation, graph, or table of a linear function.	1			1	1			1	
	1	2	2	Translate from one representation of a linear function to another.	1			1	1			1	
	Total For Assessment Anchor A1.2.1					6	1		7	6	4		10
	2			Coordinate Geometry			1	1			4		4
	2	1	1	Identify, describe, and/or use constant rates of change.	1			1	1				1
	2	1	2	Apply the concept of linear rate of change (slope) to solve problems.	1			1	1				1
	2	1	3	Write or identify a linear equation when given...	2			2	2				2
	2	1	4	Determine the slope and/or y-intercept represented by a linear equation or graph.	1			1	1				1
	2	2	1	Draw, identify, find, and/or write an equation for a line of best fit for a scatter plot.	1			1	1				1
	Total For Assessment Anchor A1.2.2					6		1	7	6		4	10
	3			Data Analysis		1		1		4			4
	3	1	1	Calculate and/or interpret the range, quartiles, and interquartile range of data.	2			2	2				2
	3	2	1	Estimate or calculate to make predictions based on a circle, line, bar graph, measures of central tendency, or other representations.	1			1	1				1
	3	2	2	Analyze data, make predictions, and/or answer questions based on displayed data.	1			1	1				1
	3	2	3	Make predictions using the equations or graphs of best-fit lines of scatter plots.	1			1	1				1
	3	3	1	Find probabilities for compound events.	1			1	1				1
	Total For Assessment Anchor A1.2.3					6	1		7	6	4		10
Understand measurable attributes and units, systems, processes of measurement													
Total For Reporting Category A1.2					18	2	1	21	18	8	4	30	

ALGEBRA I-SUMMER 2021

Keystone Exam

Algebra I

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items				Points				
					Number of Core Items				Core Points				
					MC	SCR	ECR	Total	MC	SCR	ECR	Total	
A1.1: Operations and Linear Equations & Inequalities	1			Operations with Real Numbers and Expressions		1		1		4		4	
	1	1	1	Compare and/or order any real numbers.	1			1	1			1	
	1	1	2	Simplify square roots.	1			1	1			1	
	1	2	1	Find the Greatest Common Factor (GCF) and/or the Least Common Multiple (LCM) for sets of monomials.	1			1	1			1	
	1	3	1	Simplify/evaluate expressions involving properties/laws of exponents, roots, and/or absolute values to solve problems.	1			1	1			1	
	1	4	1	Use estimation to solve problems.									
	1	5	1	Add, subtract, and/or multiply polynomial expressions (express answers in simplest form).									
	1	5	2	Factor algebraic expressions, including difference of squares and trinomials.	1			1	1			1	
	1	5	3	Simplify/reduce a rational algebraic expression.	1			1	1			1	
	Total For Assessment Anchor A1.1.1					6	1		7	6	4		10
	2			Linear Equations			1	1			4		4
	2	1	1	Write, solve, and/or apply a linear equation.	2			2	2				2
	2	1	2	Use and/or identify an algebraic property to justify any step in an equation-solving process.	1			1	1				1
	2	1	3	Interpret solutions to problems in the context of the problem situation.	1			1	1				1
	2	2	1	Write and/or solve a system of linear equations (including problem situations) using graphing, substitution, and/or elimination.	1			1	1				1
	2	2	2	Interpret solutions to problems in the context of the problem situation.	1			1	1				1
	Total For Assessment Anchor A1.1.2					6		1	7	6		4	10
	3			Linear Inequalities		1		1		4			4
	3	1	1	Write or solve compound inequalities and/or graph their solution sets on a number line .	1			1	1				1
	3	1	2	Identify or graph the solution set to a linear inequality on a number line.	1			1	1				1
	3	1	3	Interpret solutions to problems in the context of the problem situation.	2			2	2				2
	3	2	1	Write and/or solve a system of linear inequalities using graphing.	1			1	1				1
	3	2	2	Interpret solutions to problems in the context of the problem situation.	1			1	1				1
	Total For Assessment Anchor A1.1.3					6	1		7	6	4		10
	Total For Reporting Category A1.1					18	2	1	21	18	8	4	30

Keystone Exam					Algebra I								
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items				Points				
					Number of Core Items				Core Points				
					MC	SCR	ECR	Total	MC	SCR	ECR	Total	
A1.2: Linear Functions & Data Organizations	1			Functions		1		1			4		4
	1	1	1	Analyze a set of data for the existence of a pattern and represent the pattern algebraically and/or graphically.	1			1	1				1
	1	1	2	Determine whether a relation is a function, given a set of points or a graph.	1			1	1				1
	1	1	3	Identify the domain or range of a relation.	1			1	1				1
	1	2	1	Create, interpret, and/or use the equation, graph, or table of a linear function.	2			2	2				2
	1	2	2	Translate from one representation of a linear function to another.	1			1	1				1
	Total For Assessment Anchor A1.2.1					6	1		7	6	4		10
	2			Coordinate Geometry		1		1			4		4
	2	1	1	Identify, describe, and/or use constant rates of change.	1			1	1				1
	2	1	2	Apply the concept of linear rate of change (slope) to solve problems.	1			1	1				1
	2	1	3	Write or identify a linear equation when given...	1			1	1				1
	2	1	4	Determine the slope and/or y-intercept represented by a linear equation or graph.	1			1	1				1
	2	2	1	Draw, identify, find, and/or write an equation for a line of best fit for a scatter plot.	2			2	2				2
	Total For Assessment Anchor A1.2.2					6	1		7	6	4		10
	3			Data Analysis			1	1				4	4
	3	1	1	Calculate and/or interpret the range, quartiles, and interquartile range of data.	1			1	1				1
	3	2	1	Estimate or calculate to make predictions based on a circle, line, bar graph, measures of central tendency, or other representations.	1			1	1				1
	3	2	2	Analyze data, make predictions, and/or answer questions based on displayed data.	2			2	2				2
	3	2	3	Make predictions using the equations or graphs of best-fit lines of scatter plots.	1			1	1				1
	3	3	1	Find probabilities for compound events.	1			1	1				1
	Total For Assessment Anchor A1.2.3					6		1	7	6		4	10
	Total For Reporting Category A1.2					18	2	1	21	18	8	4	30

Keystone Exam					Biology							
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points				
					Number of Core Items			Core Points				
					MC	CR	Total	MC	CR	Total		
BIO.A: Basic Biological Principles	1			Basic Biological Principles								
	1	1	1	Describe the characteristics of life shared by all prokaryotic and eukaryotic organisms.	2		2	2		2		
	1	2	1	Compare cellular structures and their functions in prokaryotic and eukaryotic cells.	1	1	2	1	3	4		
	1	2	2	Describe and interpret relationships between structure and function at various levels of biological organization.	2		2	2		2		
	Total For Assessment Anchor BIO.A.1					5	1	6	5	3	8	
	2			The Chemical Basis for Life								
	2	1	1	Describe the unique properties of water and how these properties support life on Earth.	2		2	2		2		
	2	2	1	Explain how carbon is uniquely suited to form biological macromolecules.	1		1	1		1		
	2	2	2	Describe how biological macromolecules form from monomers.		1	1		3	3		
	2	2	3	Compare the structure and function of carbohydrates, lipids, proteins, and nucleic acids in organisms.	1		1	1		1		
	2	3	1	Describe the role of an enzyme as a catalyst in regulating a specific biochemical reaction.	1		1	1		1		
	2	3	2	Explain how factors such as pH, temperature, and concentration levels can affect enzyme function.	1		1	1		1		
	Total For Assessment Anchor BIO.A.2					6	1	7	6	3	9	
	3			Bioenergetics								
	3	1	1	Describe the fundamental roles of plastids (e.g., chloroplasts) and mitochondria in energy transformations.	3		3	3		3		
	3	2	1	Compare the basic transformation of energy during photosynthesis and cellular respiration.	1		1	1		1		
	3	2	2	Describe the role of ATP in biochemical reactions.	2		2	2		2		
	Total For Assessment Anchor BIO.A.3					6		6	6		6	
	4			Homeostasis and Transport								
	4	1	1	Describe how the structure of the plasma membrane allows it to function as a regulatory structure and/or protective barrier for a cell.	2		2	2		2		
	4	1	2	Compare the mechanisms that transport materials across the plasma membrane.	1	1	2	1	3	4		
	4	1	3	Describe how membrane-bound cellular organelles.	2		2	2		2		
	4	2	1	Explain how organisms maintain homeostasis.	2		2	2		2		
	Total For Assessment Anchor BIO.A.4					7	1	8	7	3	10	
Total For Reporting Category BIO.A					24	3	27	24	9	33		

Keystone Exam

Biology

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
BIO.B: Cell Growth and Reproduction	1			Cell Growth and Reproduction							
	1	1	1	Describe the events that occur during the cell cycle: interphase, nuclear division.	1	1	2	1	3	4	
	1	1	2	Compare the processes and outcomes of mitotic and meiotic nuclear divisions.	1		1	1		1	
	1	2	1	Describe how the process of DNA replication results in the transmission and/or conservation of genetic information.	1		1	1		1	
	1	2	2	Explain the functional relationships between DNA, genes, alleles, and chromosomes and their roles in inheritance.	1		1	1		1	
	Total For Assessment Anchor BIO.B.1					4	1	5	4	3	7
	2				Genetics						
	2	1	1	Describe and/or predict observed patterns of inheritance.							
	2	1	2	Describe processes that can alter composition or number of chromosomes.	1		1	1		1	
	2	2	1	Describe how the processes of transcription and translation are similar in all organisms.	1		1	1		1	
	2	2	2	Describe the role of ribosomes, endoplasmic reticulum, Golgi apparatus, and the nucleus in the production of specific types of proteins.	1		1	1		1	
	2	3	1	Describe how genetic mutations alter the DNA sequence and may or may not affect phenotype.	2		2	2		2	
	2	4	1	Explain how genetic engineering has impacted the fields of medicine, forensics, and agriculture.	1	1	2	1	3	4	
	Total For Assessment Anchor BIO.B.2					6	1	7	6	3	9
	3				Theory of Evolution						
	3	1	1	Explain how natural selection can impact allele frequencies of a population.	1		1	1		1	
	3	1	2	Describe the factors that can contribute to the development of new species.	1		1	1		1	
	3	1	3	Explain how genetic mutations may result in genotypic and phenotypic variations within a population.	1		1	1		1	
	3	2	1	Interpret evidence supporting the theory of evolution.	1		1	1		1	
	3	3	1	Distinguish between the scientific terms: hypothesis, inference, law, theory, principle, fact, and observation.	2		2	2		2	
	Total For Assessment Anchor BIO.B.3					6		6	6		6
	4				Ecology						
	4	1	1	Describe the levels of ecological organization.	1		1	1		1	
	4	1	2	Describe characteristic biotic and abiotic components of aquatic and terrestrial ecosystems.	2		2	2		2	
	4	2	1	Describe how energy flows through an ecosystem.	2		2	2		2	
	4	2	2	Describe biotic interactions in an ecosystem.	1		1	1		1	
4	2	3	Describe how matter recycles through an ecosystem.	1		1	1		1		
4	2	4	Describe how ecosystems change in response to natural and human disturbances.	1		1	1		1		
4	2	5	Describe the effects of limiting factors on population dynamics and potential species extinction.		1	1		3	3		
Total For Assessment Anchor BIO.B.4					8	1	9	8	3	11	
Total For Assessment Anchor BIO.B					24	3	27	24	9	33	

Keystone Exam

Biology

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
BIO.A: Basic Biological Principles	1			Basic Biological Principles							
	1	1	1	Describe the characteristics of life shared by all prokaryotic and eukaryotic organisms.	1		1	1		1	
	1	2	1	Compare cellular structures and their functions in prokaryotic and eukaryotic cells.	2	1	3	2	3	5	
	1	2	2	Describe and interpret relationships between structure and function at various levels of biological organization.	1		1	1		1	
	Total For Assessment Anchor BIO.A.1					4	1	5	4	3	7
	2				The Chemical Basis for Life						
	2	1	1	Describe the unique properties of water and how these properties support life on Earth.	2		2	2		2	
	2	2	1	Explain how carbon is uniquely suited to form biological macromolecules.	1		1	1		1	
	2	2	2	Describe how biological macromolecules form from monomers.	1		1	1		1	
	2	2	3	Compare the structure and function of carbohydrates, lipids, proteins, and nucleic acids in organisms.	1		1	1		1	
	2	3	1	Describe the role of an enzyme as a catalyst in regulating a specific biochemical reaction.	1		1	1		1	
	2	3	2	Explain how factors such as pH, temperature, and concentration levels can affect enzyme function.	1	1	2	1	3	4	
	Total For Assessment Anchor BIO.A.2					7	1	8	7	3	10
	3				Bioenergetics						
	3	1	1	Describe the fundamental roles of plastids (e.g., chloroplasts) and mitochondria in energy transformations.	1	1	2	1	3	4	
	3	2	1	Compare the basic transformation of energy during photosynthesis and cellular respiration.	2		2	2		2	
	3	2	2	Describe the role of ATP in biochemical reactions.	2		2	2		2	
	Total For Assessment Anchor BIO.A.3					5	1	6	5	3	8
	4				Homeostasis and Transport						
	4	1	1	Describe how the structure of the plasma membrane allows it to function as a regulatory structure and/or protective barrier for a cell.	1		1	1		1	
	4	1	2	Compare the mechanisms that transport materials across the plasma membrane.	2		2	2		2	
	4	1	3	Describe how membrane-bound cellular organelles.	3		3	3		3	
	4	2	1	Explain how organisms maintain homeostasis.	2		2	2		2	
	Total For Assessment Anchor BIO.A.4					8		8	8		8
Total For Reporting Category BIO.A					24	3	27	24	9	33	

Keystone Exam

Biology

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
BIO.B: Cell Growth and Reproduction	1			Cell Growth and Reproduction							
	1	1	1	Describe the events that occur during the cell cycle: interphase, nuclear division.	1		1	1		1	
	1	1	2	Compare the processes and outcomes of mitotic and meiotic nuclear divisions.	1		1	1		1	
	1	2	1	Describe how the process of DNA replication results in the transmission and/or conservation of genetic information.	2		2	2		2	
	1	2	2	Explain the functional relationships between DNA, genes, alleles, and chromosomes and their roles in inheritance.	2		2	2		2	
	Total For Assessment Anchor BIO.B.1					6		6	6		6
	2			Genetics							
	2	1	1	Describe and/or predict observed patterns of inheritance.	1		1	1		1	
	2	1	2	Describe processes that can alter composition or number of chromosomes.	2		2	2		2	
	2	2	1	Describe how the processes of transcription and translation are similar in all organisms.	1		1	1		1	
	2	2	2	Describe the role of ribosomes, endoplasmic reticulum, Golgi apparatus, and the nucleus in the production of specific types of proteins.	1		1	1		1	
	2	3	1	Describe how genetic mutations alter the DNA sequence and may or may not affect phenotype.		1	1		3	3	
	2	4	1	Explain how genetic engineering has impacted the fields of medicine, forensics, and agriculture.	1		1	1		1	
	Total For Assessment Anchor BIO.B.2					6	1	7	6	3	9
	3			Theory of Evolution							
	3	1	1	Explain how natural selection can impact allele frequencies of a population.		1	1		3	3	
	3	1	2	Describe the factors that can contribute to the development of new species.	1		1	1		1	
	3	1	3	Explain how genetic mutations may result in genotypic and phenotypic variations within a population.	1		1	1		1	
	3	2	1	Interpret evidence supporting the theory of evolution.	2		2	2		2	
	3	3	1	Distinguish between the scientific terms: hypothesis, inference, law, theory, principle, fact, and observation.	1		1	1		1	
	Total For Assessment Anchor BIO.B.3					5	1	6	5	3	8
	4			Ecology							
	4	1	1	Describe the levels of ecological organization.	1		1	1		1	
	4	1	2	Describe characteristic biotic and abiotic components of aquatic and terrestrial ecosystems.	1		1	1		1	
4	2	1	Describe how energy flows through an ecosystem.	1	1	2	1	3	4		
4	2	2	Describe biotic interactions in an ecosystem.	1		1	1		1		
4	2	3	Describe how matter recycles through an ecosystem.	1		1	1		1		
4	2	4	Describe how ecosystems change in response to natural and human disturbances.	1		1	1		1		
4	2	5	Describe the effects of limiting factors on population dynamics and potential species extinction.	1		1	1		1		
Total For Assessment Anchor BIO.B.4					7	1	8	7	3	10	
Total For Assessment Anchor BIO.B					24	3	27	24	9	33	

Keystone Exam					Biology						
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
BIO.A: Basic Biological Principles	1			Basic Biological Principles							
	1	1	1	Describe the characteristics of life shared by all prokaryotic and eukaryotic organisms.	1	1	2	1	3	4	
	1	2	1	Compare cellular structures and their functions in prokaryotic and eukaryotic cells.	2		2	2		2	
	1	2	2	Describe and interpret relationships between structure and function at various levels of biological organization.	2		2	2		2	
	Total For Assessment Anchor BIO.A.1					5	1	6	5	3	8
	2			The Chemical Basis for Life							
	2	1	1	Describe the unique properties of water and how these properties support life on Earth.	1		1	1		1	
	2	2	1	Explain how carbon is uniquely suited to form biological macromolecules.	2		2	2		2	
	2	2	2	Describe how biological macromolecules form from monomers.	2		2	2		2	
	2	2	3	Compare the structure and function of carbohydrates, lipids, proteins, and nucleic acids in organisms.	1		1	1		1	
	2	3	1	Describe the role of an enzyme as a catalyst in regulating a specific biochemical reaction.	1		1	1		1	
	2	3	2	Explain how factors such as pH, temperature, and concentration levels can affect enzyme function.		1	1		3	3	
	Total For Assessment Anchor BIO.A.2					7	1	8	7	3	10
	3			Bioenergetics							
	3	1	1	Describe the fundamental roles of plastids (e.g., chloroplasts) and mitochondria in energy transformations.	2		2	2		2	
	3	2	1	Compare the basic transformation of energy during photosynthesis and cellular respiration.	2		2	2		2	
	3	2	2	Describe the role of ATP in biochemical reactions.	1	1	2	1	3	4	
	Total For Assessment Anchor BIO.A.3					5	1	6	5	3	8
	4			Homeostasis and Transport							
	4	1	1	Describe how the structure of the plasma membrane allows it to function as a regulatory structure and/or protective barrier for a cell.	2		2	2		2	
	4	1	2	Compare the mechanisms that transport materials across the plasma membrane.	1		1	1		1	
	4	1	3	Describe how membrane-bound cellular organelles.	2		2	2		2	
	4	2	1	Explain how organisms maintain homeostasis.	2		2	2		2	
	Total For Assessment Anchor BIO.A.4					7		7	7		7
Total For Reporting Category BIO.A					24	3	27	24	9	33	

Keystone Exam

Biology

Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
BIO.B: Cell Growth and Reproduction	1			Cell Growth and Reproduction							
	1	1	1	Describe the events that occur during the cell cycle: interphase, nuclear division.	1	1	2	1	3	4	
	1	1	2	Compare the processes and outcomes of mitotic and meiotic nuclear divisions.	1		1	1		1	
	1	2	1	Describe how the process of DNA replication results in the transmission and/or conservation of genetic information.	1		1	1		1	
	1	2	2	Explain the functional relationships between DNA, genes, alleles, and chromosomes and their roles in inheritance.	1		1	1		1	
	Total For Assessment Anchor BIO.B.1					4	1	5	4	3	7
	2				Genetics						
	2	1	1	Describe and/or predict observed patterns of inheritance.	2		2	2		2	
	2	1	2	Describe processes that can alter composition or number of chromosomes.	1		1	1		1	
	2	2	1	Describe how the processes of transcription and translation are similar in all organisms.	1		1	1		1	
	2	2	2	Describe the role of ribosomes, endoplasmic reticulum, Golgi apparatus, and the nucleus in the production of specific types of proteins.	1		1	1		1	
	2	3	1	Describe how genetic mutations alter the DNA sequence and may or may not affect phenotype.	2		2	2		2	
	2	4	1	Explain how genetic engineering has impacted the fields of medicine, forensics, and agriculture.	1		1	1		1	
	Total For Assessment Anchor BIO.B.2					8		8	8		8
	3				Theory of Evolution						
	3	1	1	Explain how natural selection can impact allele frequencies of a population.	1	1	2	1	3	4	
	3	1	2	Describe the factors that can contribute to the development of new species.	1		1	1		1	
	3	1	3	Explain how genetic mutations may result in genotypic and phenotypic variations within a population.	1		1	1		1	
	3	2	1	Interpret evidence supporting the theory of evolution.	1		1	1		1	
	3	3	1	Distinguish between the scientific terms: hypothesis, inference, law, theory, principle, fact, and observation.	1		1	1		1	
	Total For Assessment Anchor BIO.B.3					5	1	6	5	3	8
	4				Ecology						
	4	1	1	Describe the levels of ecological organization.	1		1	1		1	
	4	1	2	Describe characteristic biotic and abiotic components of aquatic and terrestrial ecosystems.	1		1	1		1	
	4	2	1	Describe how energy flows through an ecosystem.	1		1	1		1	
	4	2	2	Describe biotic interactions in an ecosystem.	1	1	2	1	3	4	
	4	2	3	Describe how matter recycles through an ecosystem.	1		1	1		1	
4	2	4	Describe how ecosystems change in response to natural and human disturbances.	1		1	1		1		
4	2	5	Describe the effects of limiting factors on population dynamics and potential species extinction.	1		1	1		1		
Total For Assessment Anchor BIO.B.4					7	1	8	7	3	10	
Total For Assessment Anchor BIO.B					24	3	27	24	9	33	

Keystone Exam					Literature						
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
L.F: Fiction	1			Reading for Meaning—Fiction							
	1	1	1	Identify and/or analyze the author's intended purpose of a text.							
	1	1	2	Explain, describe, and/or analyze examples of a text that support the author's intended purpose.	1		1	1		1	
	1	1	3	Analyze, interpret, and evaluate how authors use techniques and elements of fiction to effectively communicate an idea or concept.	1		1	1		1	
	1	2	1	Identify and/or apply a synonym or antonym of a word used in a text.							
	1	2	2	Identify how the meaning of a word is changed when an affix is added; identify the meaning of a word with an affix from a text.	1		1	1		1	
	1	2	3	Use context clues to determine or clarify the meaning of unfamiliar, multiple-meaning, or ambiguous words.							
	1	2	4	Draw conclusions about connotations of words.	2		2	2		2	
	1	3	1	Identify and/or explain stated or implied main ideas and relevant supporting details from a text.	1		1	1		1	
	1	3	2	Summarize the key details and events of a fictional text, in part or as a whole.							
	Total For Assessment Anchor L.F.1					6		6	6		6
	2				Analyzing and Interpreting Literature—Fiction						
	2	1	1	1	Make inferences and/or draw conclusions based on analysis of a text.	1		1	1		1
	2	1	2	2	Cite evidence from a text to support generalizations.	2		2	2		2
	2	2	1	1	Analyze how literary form relates to and/or influences meaning of a text.						
	2	2	2	2	Compare and evaluate the characteristics that distinguish fiction from literary nonfiction.						
	2	2	3	3	Explain, interpret, compare, describe, analyze, and/or evaluate connections between texts.	1	2	3	1	6	7
	2	2	4	4	Compare and evaluate the characteristics that distinguish narrative, poetry, and drama.						
	2	3	1	1	Explain, interpret, compare, describe, analyze, and/or evaluate character in a variety of fiction:	1	1	2	1	3	4
	2	3	2	2	Explain, interpret, compare, describe, analyze, and/or evaluate setting in a variety of fiction:						
	2	3	3	3	Explain, interpret, compare, describe, analyze, and/or evaluate plot in a variety of fiction:						
	2	3	4	4	Explain, interpret, compare, describe, analyze, and/or evaluate theme in a variety of fiction:	1		1	1		1
	2	3	5	5	Explain, interpret, compare, describe, analyze, and/or evaluate tone, style, and/or mood in a variety of fiction:	1		1	1		1
	2	3	6	6	Explain, interpret, compare, describe, analyze, and/or evaluate point of view in a variety of fiction:	1		1	1		1
	2	4	1	1	Interpret and analyze works from a variety of genres for literary, historical, and/or cultural significance.						
	2	5	1	1	Identify, explain, interpret, describe, and/or analyze the effects of personification, simile, metaphor, hyperbole, satire, foreshadowing, flashback, imagery, allegory, symbolism, dialect, allusion, and irony in a text.	2		2	2		2
	2	5	2	2	Identify, explain, and analyze the structure of poems and sound devices.	1		1	1		1
	2	5	3	3	Identify and analyze how stage directions, monologue, dialogue, soliloquy, and dialect support dramatic script.						
Total For Assessment Anchor L.F.2					11	3	14	11	9	20	
Total For Reporting Category L.F					17	3	20	17	9	26	

Keystone Exam				Literature								
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points				
					Number of Core Items			Core Points				
					MC	CR	Total	MC	CR	Total		
L.N: Nonfiction	1			Reading for Meaning—Nonfiction								
	1	1	1	Identify and/or analyze the author's intended purpose of a text.	1		1	1		1		
	1	1	2	Explain, describe, and/or analyze examples of a text that support the author's intended purpose.	1		1	1		1		
	1	1	3	Analyze, interpret, and evaluate how authors use techniques and elements of nonfiction to effectively communicate an idea or concept.								
	1	1	4	Explain how an author's use of key words or phrases in text informs and influences the reader.	1		1	1		1		
	1	2	1	Identify and/or apply a synonym or antonym of a word used in a text.								
	1	2	2	Identify how the meaning of a word is changed when an affix is added; identify the meaning of a word with an affix from a text.								
	1	2	3	Use context clues to determine or clarify the meaning of unfamiliar, multiple-meaning, or ambiguous words.	2		2	2		2		
	1	2	4	Draw conclusions about connotations of words.	1		1	1		1		
	1	3	1	Identify and/or explain stated or implied main ideas and relevant supporting details from a text.								
	1	3	2	Summarize the key details and events of a nonfictional text, in part or as a whole.	1		1	1		1		
	1	3	3	Analyze the interrelationships of ideas and events in text to determine how one idea or event may interact and influence another.	1		1	1		1		
	Total For Assessment Anchor L.N.1					8		8	8		8	
	2				Data Analysis							
	2	1	1		Make inferences and/or draw conclusions based on analysis of a text.							
	2	1	2		Cite evidence from a text to support generalizations.							
	2	2	1		Analyze how literary form relates to and/or influences meaning of a text.	1		1	1		1	
	2	2	2		Compare and evaluate the characteristics that distinguish fiction from literary nonfiction.							
	2	2	3		Explain, interpret, compare, describe, analyze, and/or evaluate connections between texts.							
	2	3	1		Explain, interpret, compare, describe, analyze, and/or evaluate character in a variety of nonfiction:							
	2	3	2		Explain, interpret, compare, describe, analyze, and/or evaluate setting in a variety of nonfiction:							
	2	3	3		Explain, interpret, compare, describe, analyze, and/or evaluate plot in a variety of nonfiction:							
	2	3	4		Explain, interpret, compare, describe, analyze, and/or evaluate theme in a variety of nonfiction:							
	2	3	5		Explain, interpret, compare, describe, analyze, and/or evaluate tone, style, and/or mood in a variety of nonfiction:		1	1		3	3	
	2	3	6		Explain, interpret, compare, describe, analyze, and/or evaluate point of view in a variety of nonfiction:	1		1	1		1	
	2	4	1		Identify, analyze, and evaluate the structure and format of complex informational texts.	1		1	1		1	
	2	4	2		Identify, explain, compare, interpret, describe, and/or analyze the sequence of steps in a list of directions.							
	2	4	3		Explain, interpret, and/or analyze the effect of text organization, including headings, graphics, and charts.							
	2	4	4		Make connections between a text and the content of graphics and charts.							
	2	4	5		Analyze and evaluate how graphics and charts clarify, simplify, and organize complex informational texts.							
	2	5	1		Differentiate between fact and opinion.							
	2	5	2		Explain, interpret, describe, and/or analyze the use of facts and opinions in a text.	1	1	2	1	3	4	
2	5	3		Distinguish essential from nonessential information.								
2	5	4		Identify, explain, and/or interpret bias and propaganda techniques in nonfictional text.	2		2	2		2		
2	5	5		Explain, describe, and/or analyze the effectiveness of bias (explicit and implicit) and propaganda techniques in nonfictional text.	1		1	1		1		
2	5	6		Explain, interpret, describe, and/or analyze the author's defense of a claim to make a point or construct an argument in nonfictional text.	2	1	3	2	3	5		
Total For Assessment Anchor L.N.2					9	3	12	9	9	18		
Total For Reporting Category L.N					17	3	20	17	9	26		

LITERATURE-SPRING 2021

Keystone Exam						Literature					
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
L.F: Fiction	1			Reading for Meaning—Fiction							
	1	1	1	Identify and/or analyze the author's intended purpose of a text.							
	1	1	2	Explain, describe, and/or analyze examples of a text that support the author's intended purpose.	1		1	1			1
	1	1	3	Analyze, interpret, and evaluate how authors use techniques and elements of fiction to effectively communicate an idea or concept.	1		1	1			1
	1	2	1	Identify and/or apply a synonym or antonym of a word used in a text.							
	1	2	2	Identify how the meaning of a word is changed when an affix is added; identify the meaning of a word with an affix from a text.	1		1	1			1
	1	2	3	Use context clues to determine or clarify the meaning of unfamiliar, multiple-meaning, or ambiguous words.	2		2	2			2
	1	2	4	Draw conclusions about connotations of words.							
	1	3	1	Identify and/or explain stated or implied main ideas and relevant supporting details from a text.	1		1	1			1
	1	3	2	Summarize the key details and events of a fictional text, in part or as a whole.	1		1	1			1
	Total For Assessment Anchor L.F.1					7	7	7	7	7	7
	2				Analyzing and Interpreting Literature—Fiction						
	2	1	1	1	Make inferences and/or draw conclusions based on analysis of a text.	1		1	1		1
	2	1	2	2	Cite evidence from a text to support generalizations.	1		1	1		1
	2	2	1	1	Analyze how literary form relates to and/or influences meaning of a text.						
	2	2	2	2	Compare and evaluate the characteristics that distinguish fiction from literary nonfiction.	1		1	1		1
	2	2	3	3	Explain, interpret, compare, describe, analyze, and/or evaluate connections between texts.						
	2	2	4	4	Compare and evaluate the characteristics that distinguish narrative, poetry, and drama.						
	2	3	1	1	Explain, interpret, compare, describe, analyze, and/or evaluate character in a variety of fiction:	1	1	2	1	3	4
	2	3	2	2	Explain, interpret, compare, describe, analyze, and/or evaluate setting in a variety of fiction:						
	2	3	3	3	Explain, interpret, compare, describe, analyze, and/or evaluate plot in a variety of fiction:	1		1	1		1
	2	3	4	4	Explain, interpret, compare, describe, analyze, and/or evaluate theme in a variety of fiction:	1		1	1		1
	2	3	5	5	Explain, interpret, compare, describe, analyze, and/or evaluate tone, style, and/or mood in a variety of fiction:		1	1		3	3
	2	3	6	6	Explain, interpret, compare, describe, analyze, and/or evaluate point of view in a variety of fiction:	1		1	1		1
	2	4	1	1	Interpret and analyze works from a variety of genres for literary, historical, and/or cultural significance.						
	2	5	1	1	Identify, explain, interpret, describe, and/or analyze the effects of personification, simile, metaphor, hyperbole, satire, foreshadowing, flashback, imagery, allegory, symbolism, dialect, allusion, and irony in a text.	2	1	3	2	3	5
	2	5	2	2	Identify, explain, and analyze the structure of poems and sound devices.						
	2	5	3	3	Identify and analyze how stage directions, monologue, dialogue, soliloquy, and dialect support dramatic script.	1		1	1		1
Total For Assessment Anchor L.F.2					10	3	13	10	9	19	
Total For Reporting Category L.F					17	3	20	17	9	26	

Keystone Exam				Literature							
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
L.N: Nonfiction	1			Reading for Meaning—Nonfiction							
	1	1	1	Identify and/or analyze the author's intended purpose of a text.							
	1	1	2	Explain, describe, and/or analyze examples of a text that support the author's intended purpose.		1	1		3	3	
	1	1	3	Analyze, interpret, and evaluate how authors use techniques and elements of nonfiction to effectively communicate an idea or concept.							
	1	1	4	Explain how an author's use of key words or phrases in text informs and influences the reader.							
	1	2	1	Identify and/or apply a synonym or antonym of a word used in a text.							
	1	2	2	Identify how the meaning of a word is changed when an affix is added; identify the meaning of a word with an affix from a text.	1		1	1		1	
	1	2	3	Use context clues to determine or clarify the meaning of unfamiliar, multiple-meaning, or ambiguous words.	1		1	1		1	
	1	2	4	Draw conclusions about connotations of words.							
	1	3	1	Identify and/or explain stated or implied main ideas and relevant supporting details from a text.	1		1	1		1	
	1	3	2	Summarize the key details and events of a nonfictional text, in part or as a whole.							
	1	3	3	Analyze the interrelationships of ideas and events in text to determine how one idea or event may interact and influence another.	2		2	2		2	
	Total For Assessment Anchor L.N.1					5	1	6	5	3	8
	2				Data Analysis						
	2	1	1	Make inferences and/or draw conclusions based on analysis of a text.	1		1	1		1	
	2	1	2	Cite evidence from a text to support generalizations.							
	2	2	1	Analyze how literary form relates to and/or influences meaning of a text.							
	2	2	2	Compare and evaluate the characteristics that distinguish fiction from literary nonfiction.							
	2	2	3	Explain, interpret, compare, describe, analyze, and/or evaluate connections between texts.	1	2	3	1	6	7	
	2	3	1	Explain, interpret, compare, describe, analyze, and/or evaluate character in a variety of nonfiction:	2		2	2		2	
	2	3	2	Explain, interpret, compare, describe, analyze, and/or evaluate setting in a variety of nonfiction:	1		1	1		1	
	2	3	3	Explain, interpret, compare, describe, analyze, and/or evaluate plot in a variety of nonfiction:							
	2	3	4	Explain, interpret, compare, describe, analyze, and/or evaluate theme in a variety of nonfiction:							
	2	3	5	Explain, interpret, compare, describe, analyze, and/or evaluate tone, style, and/or mood in a variety of nonfiction:	1		1	1		1	
	2	3	6	Explain, interpret, compare, describe, analyze, and/or evaluate point of view in a variety of nonfiction:	2		2	2		2	
	2	4	1	Identify, analyze, and evaluate the structure and format of complex informational texts.	1		1	1		1	
	2	4	2	Identify, explain, compare, interpret, describe, and/or analyze the sequence of steps in a list of directions.							
	2	4	3	Explain, interpret, and/or analyze the effect of text organization, including headings, graphics, and charts.							
	2	4	4	Make connections between a text and the content of graphics and charts.							
	2	4	5	Analyze and evaluate how graphics and charts clarify, simplify, and organize complex informational texts.							
	2	5	1	Differentiate between fact and opinion.							
	2	5	2	Explain, interpret, describe, and/or analyze the use of facts and opinions in a text.							
2	5	3	Distinguish essential from nonessential information.	1		1	1		1		
2	5	4	Identify, explain, and/or interpret bias and propaganda techniques in nonfictional text.								
2	5	5	Explain, describe, and/or analyze the effectiveness of bias (explicit and implicit) and propaganda techniques in nonfictional text.	1		1	1		1		
2	5	6	Explain, interpret, describe, and/or analyze the author's defense of a claim to make a point or construct an argument in nonfictional text.	1		1	1		1		
Total For Assessment Anchor L.N.2					12	2	14	12	6	18	
Total For Reporting Category L.N					17	3	20	17	9	26	

LITERATURE-SUMMER 2021

Keystone Exam				Literature							
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points			
					Number of Core Items			Core Points			
					MC	CR	Total	MC	CR	Total	
L.F: Fiction	1			Reading for Meaning—Fiction							
	1	1	1	Identify and/or analyze the author's intended purpose of a text.	1		1	1		1	
	1	1	2	Explain, describe, and/or analyze examples of a text that support the author's intended purpose.							
	1	1	3	Analyze, interpret, and evaluate how authors use techniques and elements of fiction to effectively communicate an idea or concept.	1		1	1		1	
	1	2	1	Identify and/or apply a synonym or antonym of a word used in a text.	1		1	1		1	
	1	2	2	Identify how the meaning of a word is changed when an affix is added; identify the meaning of a word with an affix from a text.							
	1	2	3	Use context clues to determine or clarify the meaning of unfamiliar, multiple-meaning, or ambiguous words.	1		1	1		1	
	1	2	4	Draw conclusions about connotations of words.	1		1	1		1	
	1	3	1	Identify and/or explain stated or implied main ideas and relevant supporting details from a text.							
	1	3	2	Summarize the key details and events of a fictional text, in part or as a whole.	1		1	1		1	
	Total For Assessment Anchor L.F.1					6	6	6	6	6	6
	2				Analyzing and Interpreting Literature—Fiction						
	2	1	1		Make inferences and/or draw conclusions based on analysis of a text.	1		1	1		1
	2	1	2		Cite evidence from a text to support generalizations.	1		1	1		1
	2	2	1		Analyze how literary form relates to and/or influences meaning of a text.						
	2	2	2		Compare and evaluate the characteristics that distinguish fiction from literary nonfiction.	1		1	1		1
	2	2	3		Explain, interpret, compare, describe, analyze, and/or evaluate connections between texts.						
	2	2	4		Compare and evaluate the characteristics that distinguish narrative, poetry, and drama.						
	2	3	1		Explain, interpret, compare, describe, analyze, and/or evaluate character in a variety of fiction:	1	1	2	1	3	4
	2	3	2		Explain, interpret, compare, describe, analyze, and/or evaluate setting in a variety of fiction:	1	1	2	1	3	4
	2	3	3		Explain, interpret, compare, describe, analyze, and/or evaluate plot in a variety of fiction:	1		1	1		1
	2	3	4		Explain, interpret, compare, describe, analyze, and/or evaluate theme in a variety of fiction:						
	2	3	5		Explain, interpret, compare, describe, analyze, and/or evaluate tone, style, and/or mood in a variety of fiction:	1		1	1		1
	2	3	6		Explain, interpret, compare, describe, analyze, and/or evaluate point of view in a variety of fiction:	1	1	2	1	3	4
	2	4	1		Interpret and analyze works from a variety of genres for literary, historical, and/or cultural significance.						
	2	5	1		Identify, explain, interpret, describe, and/or analyze the effects of personification, simile, metaphor, hyperbole, satire, foreshadowing, flashback, imagery, allegory, symbolism, dialect, allusion, and irony in a text.	3		3	3		3
	2	5	2		Identify, explain, and analyze the structure of poems and sound devices.						
2	5	3		Identify and analyze how stage directions, monologue, dialogue, soliloquy, and dialect support dramatic script.							
Total For Assessment Anchor L.F.2					11	3	14	11	9	20	
Total For Reporting Category L.F					17	3	20	17	9	26	

Keystone Exam					Literature							
Reporting Category	Assessment Anchor	Descriptor (Sub-anchor)	Eligible Content	Focus	Items			Points				
					Number of Core Items			Core Points				
					MC	CR	Total	MC	CR	Total		
L.N: Nonfiction	1			Reading for Meaning—Nonfiction								
	1	1	1	Identify and/or analyze the author's intended purpose of a text.	1		1	1		1		
	1	1	2	Explain, describe, and/or analyze examples of a text that support the author's intended purpose.	1		1	1		1		
	1	1	3	Analyze, interpret, and evaluate how authors use techniques and elements of nonfiction to effectively communicate an idea or concept.	1		1	1		1		
	1	1	4	Explain how an author's use of key words or phrases in text informs and influences the reader.								
	1	2	1	Identify and/or apply a synonym or antonym of a word used in a text.								
	1	2	2	Identify how the meaning of a word is changed when an affix is added; identify the meaning of a word with an affix from a text.	1		1	1		1		
	1	2	3	Use context clues to determine or clarify the meaning of unfamiliar, multiple-meaning, or ambiguous words.								
	1	2	4	Draw conclusions about connotations of words.								
	1	3	1	Identify and/or explain stated or implied main ideas and relevant supporting details from a text.								
	1	3	2	Summarize the key details and events of a nonfictional text, in part or as a whole.								
	1	3	3	Analyze the interrelationships of ideas and events in text to determine how one idea or event may interact and influence another.	2		2	2		2		
	Total For Assessment Anchor L.N.1					6		6	6		6	
	2				Data Analysis							
	2	1	1		Make inferences and/or draw conclusions based on analysis of a text.		1	1		3	3	
	2	1	2		Cite evidence from a text to support generalizations.	2		2	2		2	
	2	2	1		Analyze how literary form relates to and/or influences meaning of a text.	1		1	1		1	
	2	2	2		Compare and evaluate the characteristics that distinguish fiction from literary nonfiction.	1		1	1		1	
	2	2	3		Explain, interpret, compare, describe, analyze, and/or evaluate connections between texts.							
	2	3	1		Explain, interpret, compare, describe, analyze, and/or evaluate character in a variety of nonfiction:		1	1		3	3	
	2	3	2		Explain, interpret, compare, describe, analyze, and/or evaluate setting in a variety of nonfiction:							
	2	3	3		Explain, interpret, compare, describe, analyze, and/or evaluate plot in a variety of nonfiction:							
	2	3	4		Explain, interpret, compare, describe, analyze, and/or evaluate theme in a variety of nonfiction:							
	2	3	5		Explain, interpret, compare, describe, analyze, and/or evaluate tone, style, and/or mood in a variety of nonfiction:	1		1	1		1	
	2	3	6		Explain, interpret, compare, describe, analyze, and/or evaluate point of view in a variety of nonfiction:							
	2	4	1		Identify, analyze, and evaluate the structure and format of complex informational texts.	1		1	1		1	
	2	4	2		Identify, explain, compare, interpret, describe, and/or analyze the sequence of steps in a list of directions.	1		1	1		1	
	2	4	3		Explain, interpret, and/or analyze the effect of text organization, including headings, graphics, and charts.	2		2	2		2	
	2	4	4		Make connections between a text and the content of graphics and charts.							
	2	4	5		Analyze and evaluate how graphics and charts clarify, simplify, and organize complex informational texts.							
	2	5	1		Differentiate between fact and opinion.							
	2	5	2		Explain, interpret, describe, and/or analyze the use of facts and opinions in a text.							
2	5	3		Distinguish essential from nonessential information.								
2	5	4		Identify, explain, and/or interpret bias and propaganda techniques in nonfictional text.	1		1	1		1		
2	5	5		Explain, describe, and/or analyze the effectiveness of bias (explicit and implicit) and propaganda techniques in nonfictional text.								
2	5	6		Explain, interpret, describe, and/or analyze the author's defense of a claim to make a point or construct an argument in nonfictional text.	1	1	2	1	3	4		
Total For Assessment Anchor L.N.2					11	3	14	11	9	20		
Total For Reporting Category L.N					17	3	20	17	9	26		

APPENDIX G: KEYSTONE EXAMS MODULE LAYOUT PLANS

Table G–1A. Winter 2020/2021, Spring 2021 and Summer 2021 Algebra I Keystone Exams Section Layout Plan

Module	Number of MC	Estimated MC Item Breakdown	Number of CR	Estimated CR Item Breakdown	Testing Time	Administration Time
1	23	18—Operational (Core) Items; 5—Embedded Field Test Items	4	3—Operational (Core) Items; 1— Embedded Field Test Items	75	85–90
2	23	18—Operational (Core) Items; 5—Embedded Field Test Items	4	3—Operational (Core) Items; 1— Embedded Field Test Items	75	85–90

Table G–1B. Winter 2020/2021, Spring 2021 and Summer 2021 Biology Keystone Exams Section Layout Plan

Module	Number of MC	Estimated MC Item Breakdown	Number of CR	Estimated CR Item Breakdown	Testing Time	Administration Time
1	32	24—Operational (Core) Items; 8—Embedded Field Test Items	4	3—Operational (Core) Items; 1— Embedded Field Test Items	72	82–87
2	32	24—Operational (Core) Items; 8—Embedded Field Test Items	4	3—Operational (Core) Items; 1— Embedded Field Test Items	72	82–87

Table G–1L. Winter 2020/2021, Spring 2021 and Summer 2021 Literature Keystone Exams Section Layout Plan

Module	Number of MC	Estimated MC Item Breakdown	Number of CR	Estimated CR Item Breakdown	Testing Time	Administration Time
1	23	17—Operational (Core) Items; 6—Embedded Field Test Items	4	3—Operational (Core) Items; 1— Embedded Field Test Items	73	83–88
2	23	17—Operational (Core) Items; 6—Embedded Field Test Items	4	3—Operational (Core) Items; 1— Embedded Field Test Items	73	83–88

APPENDIX H: MEAN RAW SCORES BY FORM

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

Tables in this appendix present the mean raw scores for all forms by content area and is based on the final data (see Chapter Nine). In addition, accommodated forms (i.e., Spanish, Braille, large print, VSL) were removed to best represent comparisons among similar groups of test-takers.

Table H-1. Mean Raw Scores by Form

Column Heading	Definition
Form	Form
<i>N</i>	Number of students
L	Length
Pts	Points possible
Min	Minimum
Max	Maximum
Mean	Mean
Med	Median
<i>SD</i>	Standard deviation

ALGEBRA I: SPRING

Table H-2. Algebra I Mean Raw Scores by Form Table

Form	N	L	Pts	Min	Max	Mean	Med	SD
All	101761	42	60	1	60	25.8	24	12.2
1	5089	42	60	1	59	25.6	24	12.0
2	5092	42	60	1	59	25.6	24	12.2
3	5094	42	60	1	59	26.0	25	12.4
4	5141	42	60	1	58	25.7	24	12.1
5	5075	42	60	1	60	25.7	24	12.3
6	5103	42	60	3	60	25.9	25	12.1
7	5069	42	60	3	59	25.8	24	12.1
8	5097	42	60	2	59	25.8	25	12.3
9	5094	42	60	1	59	25.8	24	12.4
10	5074	42	60	1	58	25.9	24	12.2
11	5116	42	60	1	58	26.1	25	12.1
12	5075	42	60	3	59	25.5	24	12.3
13	5087	42	60	3	58	25.8	25	12.3
14	5075	42	60	1	60	25.9	25	12.1
15	5057	42	60	2	58	25.8	24	12.2
16	5090	42	60	1	57	25.7	24	12.0
17	5045	42	60	3	58	25.8	24	12.3
18	5105	42	60	1	60	26.0	25	12.2
19	5114	42	60	1	58	25.8	24	12.1
20	5069	42	60	2	59	25.9	25	12.2

BIOLOGY: SPRING

Table H-3. Biology Mean Raw Scores by Form Table

Form	N	L	Pts	Min	Max	Mean	Med	SD
All	96858	54	66	1	66	32.9	32	14.1
1	4806	54	66	4	66	32.8	32	14.2
2	4836	54	66	2	65	33.0	32	14.1
3	4837	54	66	4	66	32.9	32	14.3
4	4839	54	66	1	66	32.9	32	13.9
5	4815	54	66	3	66	32.9	32	14.1
6	4872	54	66	3	66	32.7	31	14.1
7	4804	54	66	2	66	33.2	32	14.1
8	4794	54	66	3	66	33.0	32	14.0
9	4822	54	66	1	66	32.9	32	14.0
10	4832	54	66	5	66	33.0	32	14.0
11	4821	54	66	4	65	32.9	31	14.0
12	4808	54	66	3	66	33.1	33	14.1
13	4896	54	66	5	66	32.9	32	14.1
14	4882	54	66	2	66	32.8	32	14.0
15	4862	54	66	1	65	33.2	32	14.1
16	4913	54	66	4	66	32.4	31	14.0
17	4846	54	66	3	66	33.1	32	14.1
18	4897	54	66	4	66	32.7	31	14.1
19	4854	54	66	3	66	32.7	32	14.2
20	4822	54	66	4	66	33.0	31	14.3

LITERATURE: SPRING

Table H-4. Literature Mean Raw Scores by Form Table

Form	N	L	Pts	Min	Max	Mean	Med	SD
All	96365	40	52	0	52	31.2	33	10.4
1	4839	40	52	3	52	31.1	33	10.3
2	4823	40	52	4	52	31.1	33	10.2
3	4804	40	52	3	52	31.5	33	10.3
4	4844	40	52	4	52	31.3	33	10.5
5	4809	40	52	2	52	31.1	33	10.6
6	4851	40	52	2	52	31.0	33	10.5
7	4802	40	52	3	52	31.0	33	10.4
8	4807	40	52	2	52	31.3	33	10.4
9	4849	40	52	0	52	31.2	33	10.3
10	4814	40	52	4	52	31.5	33	10.3
11	4817	40	52	2	51	31.1	33	10.3
12	4852	40	52	3	52	31.1	32	10.3
13	4802	40	52	1	51	31.4	33	10.3
14	4808	40	52	1	51	31.4	33	10.4
15	4853	40	52	2	52	31.4	33	10.3
16	4789	40	52	2	52	31.4	33	10.4
17	4812	40	52	3	51	31.2	33	10.4
18	4796	40	52	3	52	31.4	33	10.3
19	4810	40	52	4	52	31.0	32	10.3
20	4784	40	52	2	52	31.3	33	10.5

APPENDIX I: DEMOGRAPHIC AND ACCOMMODATION DATA

WINTER

Table I-1. Students Assessed on the Winter Keystone: Algebra I

Description	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)	2	10	39	439	3,510	4,235	2,160	86	10,481
Total number of CBT processed (Number)	0	0	128	128	1,390	1,970	835	17	4,468
Total number of tests processed (Number)	2	10	167	567	4,900	6,205	2,995	103	14,949
Total number of tests processed with a score (Number)	0	10	167	562	4,558	5,497	2,195	76	13,065
Total number of tests processed with a score (Percent)	0	100	100	99.1	93	88.6	73.3	73.8	87.4
Total number of tests processed without a score (Number)	2	0	0	5	342	708	800	27	1,884
Total number of tests processed without a score (Percent)	100	0	0	.9	7	11.4	26.7	26.2	12.6

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table I-2. Students Assessed on the Winter Keystone: Biology

Description	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)	1	0	1,378	4,910	1,468	121	7,878
Total number of CBT processed (Number)	0	1	665	3,841	1,168	20	5,695
Total number of tests processed (Number)	1	1	2,043	8,751	2,636	141	13,573
Total number of tests processed with a score (Number)	1	1	1,580	8,063	2,281	111	12,037
Total number of tests processed with a score (Percent)	100	100	77.3	92.1	86.5	78.7	88.7
Total number of tests processed without a score (Number)	0	0	463	688	355	30	1,536
Total number of tests processed without a score (Percent)	0	0	22.7	7.9	13.5	21.3	11.3

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I-3. Students Assessed on the Winter Keystone: Literature

Description	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)	0	0	171	5,896	2,034	149	8,250
Total number of CBT processed (Number)	0	0	86	3,278	1,392	18	4,774
Total number of tests processed (Number)	0	0	257	9,174	3,426	167	13,024
Total number of tests processed with a score (Number)	0	0	199	8,191	3,047	135	11,572
Total number of tests processed with a score (Percent)	0	0	77.4	89.3	88.9	80.8	88.9
Total number of tests processed without a score (Number)	0	0	58	983	379	32	1,452
Total number of tests processed without a score (Percent)	0	0	22.6	10.7	11.1	19.2	11.1

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I-4. Counts of Students without Scores on the Winter Keystone: Algebra I

Reason for Non-Assessment	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)	0	0	0	1	132	201	468	18	820
Extended absence from school (Percent)	0	0	0	20	38.6	28.4	58.5	66.7	43.5
Non-attempt (Number)	0	0	0	0	80	275	70	4	429
Non-attempt (Percent)	0	0	0	0	23.4	38.8	8.8	14.8	22.8
Medical emergency (Number)	0	0	0	0	7	14	12	0	33
Medical emergency (Percent)	0	0	0	0	2	2	1.5	0	1.8
Parental request - Chapter 4 (Number)	0	0	0	0	9	10	146	2	167
Parental request - Chapter 4 (Percent)	0	0	0	0	2.6	1.4	18.3	7.4	8.9
Parental request - Other reasons (Number)	0	0	0	2	69	136	52	1	260
Parental request - Other reasons (Percent)	0	0	0	40	20.2	19.2	6.5	3.7	13.8
Other reasons (Number)	2	0	0	2	45	72	52	2	175
Other reasons (Percent)	100	0	0	40	13.2	10.2	6.5	7.4	9.3
Total not assessed (Number)	2	0	0	5	342	708	800	27	1,884

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table I-5. Counts of Students without Scores on the Winter Keystone: Biology

Reason for Non-Assessment	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)	0	0	63	200	190	15	468
Extended absence from school (Percent)	0	0	13.6	29.1	53.5	50	30.5
Non-attempt (Number)	0	0	336	144	39	0	519
Non-attempt (Percent)	0	0	72.6	20.9	11	0	33.8
Medical emergency (Number)	0	0	2	19	4	0	25
Medical emergency (Percent)	0	0	.4	2.8	1.1	0	1.6
Parental request - Chapter 4 (Number)	0	0	1	14	27	3	45
Parental request - Chapter 4 (Percent)	0	0	.2	2	7.6	10	2.9
Parental request - Other reasons (Number)	0	0	51	251	63	4	369
Parental request - Other reasons (Percent)	0	0	11	36.5	17.7	13.3	24
Other reasons (Number)	0	0	10	60	32	8	110
Other reasons (Percent)	0	0	2.2	8.7	9	26.7	7.2
Total not assessed (Number)	0	0	463	688	355	30	1,536

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I-6. Counts of Students without Scores on the Winter Keystone: Literature

Reason for Non-Assessment	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)	0	0	38	318	156	19	531
Extended absence from school (Percent)	0	0	65.5	32.3	41.2	59.4	36.6
Non-attempt (Number)	0	0	11	418	94	6	529
Non-attempt (Percent)	0	0	19	42.5	24.8	18.8	36.4
EL in first year in U.S. schools (Number)	0	0	0	0	1	0	1
EL in first year in U.S. schools (Percent)	0	0	0	0	.3	0	.1
Medical emergency (Number)	0	0	1	8	5	0	14
Medical emergency (Percent)	0	0	1.7	.8	1.3	0	1
Parental request - Chapter 4 (Number)	0	0	0	13	6	3	22
Parental request - Chapter 4 (Percent)	0	0	0	1.3	1.6	9.4	1.5
Parental request - Other reasons (Number)	0	0	4	146	77	0	227
Parental request - Other reasons (Percent)	0	0	6.9	14.9	20.3	0	15.6
Other reasons (Number)	0	0	4	80	40	4	128
Other reasons (Percent)	0	0	6.9	8.1	10.6	12.5	8.8
Total not assessed (Number)	0	0	58	983	379	32	1,452

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–7. Demographic Characteristics of Students taking the Winter Keystone: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)	0	4	65	277	2,376	2,675	1,065	25	6,487
Female (Percent)	0	40	38.9	49.3	52.1	48.7	48.5	32.9	49.7
Male (Number)	0	6	102	285	2,182	2,822	1,130	51	6,578
Male (Percent)	0	60	61.1	50.7	47.9	51.3	51.5	67.1	50.3
American Indian/Alaskan Native (not Hispanic) (Number)	0	0	0	2	11	7	6	0	26
American Indian/Alaskan Native (not Hispanic) (Percent)	0	0	0	.4	.2	.1	.3	0	.2
Asian (not Hispanic) (Number)	0	1	41	36	201	119	31	1	430
Asian (not Hispanic) (Percent)	0	10	24.6	6.4	4.4	2.2	1.4	1.3	3.3
Black or African American (not Hispanic) (Number)	0	0	1	21	229	535	244	9	1,039
Black or African American (not Hispanic) (Percent)	0	0	.6	3.7	5	9.7	11.1	11.8	8
Hispanic (any race) (Number)	0	1	2	29	375	822	291	7	1,527
Hispanic (any race) (Percent)	0	10	1.2	5.2	8.2	15	13.3	9.2	11.7
Multi-Racial (not Hispanic) (Number)	0	1	5	27	163	194	57	3	450
Multi-Racial (not Hispanic) (Percent)	0	10	3	4.8	3.6	3.5	2.6	3.9	3.4
White (not Hispanic) (Number)	0	7	118	445	3,575	3,816	1,565	56	9,582
White (not Hispanic) (Percent)	0	70	70.7	79.2	78.4	69.4	71.3	73.7	73.3
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)	0	0	0	2	4	4	1	0	11
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)	0	0	0	.4	.1	.1	0	0	.1
IEP (not gifted) (Number)	0	0	2	14	294	902	582	33	1,827
IEP (not gifted) (Percent)	0	0	1.2	2.5	6.5	16.4	26.5	43.4	14
Student exited IEP in last 2 years (Number)	0	0	4	6	76	118	44	2	250
Student exited IEP in last 2 years (Percent)	0	0	2.4	1.1	1.7	2.1	2	2.6	1.9
Title I (Number)	0	0	0	13	375	906	523	26	1,843
Title I (Percent)	0	0	0	2.3	8.2	16.5	23.8	34.2	14.1
Title III served (Number)	0	0	0	0	42	170	76	5	293
Title III served (Percent)	0	0	0	0	.9	3.1	3.5	6.6	2.2
Title III not served (Number)	0	0	0	0	0	0	0	0	0
Title III not served (Percent)	0	0	0	0	0	0	0	0	0
Migrant student (Number)	0	0	0	0	1	9	5	0	15
Migrant student (Percent)	0	0	0	0	0	.2	.2	0	.1
EL enrolled first year (Number)	0	0	0	0	3	8	5	0	16
EL enrolled first year (Percent)	0	0	0	0	.1	.1	.2	0	.1
EL enrolled not first year (Number)	0	0	0	2	40	166	72	5	285
EL enrolled not first year (Percent)	0	0	0	.4	.9	3	3.3	6.6	2.2
Exited ESL/bilingual program and in first year of monitoring (Number)	0	0	0	0	5	18	3	0	26
Exited ESL/bilingual program and in first year of monitoring (Percent)	0	0	0	0	.1	.3	.1	0	.2

Table I-7 (continued). Demographic Characteristics of Students taking the Winter Keystone: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in 2nd year of monitoring (Number)	0	0	0	4	9	17	4	0	34
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)	0	0	0	.7	.2	.3	.2	0	.3
Former EL no longer monitored (Number)	0	0	6	14	99	131	23	0	273
Former EL no longer monitored (Percent)	0	0	3.6	2.5	2.2	2.4	1	0	2.1
LIFE first year (Number)	0	0	0	0	0	0	0	0	0
LIFE first year (Percent)	0	0	0	0	0	0	0	0	0
LIFE not first year (Number)	0	0	0	0	1	1	1	0	3
LIFE not first year (Percent)	0	0	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Number)	0	0	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Percent)	0	0	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Number)	0	0	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Percent)	0	0	0	0	0	0	0	0	0
Foreign exchange student (Number)	0	0	0	0	0	0	0	0	0
Foreign exchange student (Percent)	0	0	0	0	0	0	0	0	0
Economically disadvantaged (Number)	0	0	17	82	1,353	2,441	1,121	49	5,063
Economically disadvantaged (Percent)	0	0	10.2	14.6	29.7	44.4	51.1	64.5	38.8
Historically Underperforming Subgroup (Number)	0	0	19	95	1,546	2,894	1,381	58	5,993
Historically Underperforming Subgroup (Percent)	0	0	11.4	16.9	33.9	52.6	62.9	76.3	45.9
Enrollment in school of residence after Oct 1 (Number)	0	0	0	2	60	93	77	8	240
Enrollment in school of residence after Oct 1 (Percent)	0	0	0	.4	1.3	1.7	3.5	10.5	1.8
Enrollment in district of residence after Oct 1 (Number)	0	0	0	2	25	62	66	6	161
Enrollment in district of residence after Oct 1 (Percent)	0	0	0	.4	.5	1.1	3	7.9	1.2
Enrollment as PA resident after Oct 1 (Number)	0	0	0	2	9	19	21	0	51
Enrollment as PA resident after Oct 1 (Percent)	0	0	0	.4	.2	.3	1	0	.4
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	10	97	35	2,367	726	228	15	3,478
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	100	58.1	6.2	51.9	13.2	10.4	19.7	26.6
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	3	10	254	283	168	10	728

Table I-7 (continued). Demographic Characteristics of Students taking the Winter Keystone: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	1.8	1.8	5.6	5.1	7.7	13.2	5.6
Military family (Number)	0	0	2	1	37	104	34	0	178
Military family (Percent)	0	0	1.2	.2	.8	1.9	1.5	0	1.4
Homeless (Number)	0	0	0	0	0	0	0	0	0
Homeless (Percent)	0	0	0	0	0	0	0	0	0
Foster (Number)	0	0	0	0	4	27	20	4	55
Foster (Percent)	0	0	0	0	.1	.5	.9	5.3	.4
Home schooled (Number)	0	0	0	0	0	0	0	0	0
Home schooled (Percent)	0	0	0	0	0	0	0	0	0
Court/agency placed (Number)	0	0	0	1	7	26	30	20	84
Court/agency placed (Percent)	0	0	0	.2	.2	.5	1.4	26.3	.6
Number of assessed students (Number)	0	10	167	562	4,558	5,497	2,195	76	13,065

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table I–8. Demographic Characteristics of Students taking the Winter Keystone: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)	0	1	841	3,940	1,073	47	5,902
Female (Percent)	0	100	53.2	48.9	47	42.3	49
Male (Number)	1	0	739	4,122	1,208	64	6,134
Male (Percent)	100	0	46.8	51.1	53	57.7	51
American Indian/Alaskan Native (not Hispanic) (Number)	0	0	2	8	5	0	15
American Indian/Alaskan Native (not Hispanic) (Percent)	0	0	.1	.1	.2	0	.1
Asian (not Hispanic) (Number)	0	0	97	300	42	2	441
Asian (not Hispanic) (Percent)	0	0	6.1	3.7	1.8	1.8	3.7
Black or African American (not Hispanic) (Number)	0	0	131	696	250	11	1,088
Black or African American (not Hispanic) (Percent)	0	0	8.3	8.6	11	9.9	9
Hispanic (any race) (Number)	0	0	266	611	163	3	1,043
Hispanic (any race) (Percent)	0	0	16.8	7.6	7.1	2.7	8.7
Multi-Racial (not Hispanic) (Number)	0	0	55	267	76	8	406
Multi-Racial (not Hispanic) (Percent)	0	0	3.5	3.3	3.3	7.2	3.4
White (not Hispanic) (Number)	1	1	1,027	6,175	1,744	87	9,035
White (not Hispanic) (Percent)	100	100	65	76.6	76.5	78.4	75.1
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)	0	0	2	6	1	0	9
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)	0	0	.1	.1	0	0	.1
IEP (not gifted) (Number)	0	0	150	1,034	504	30	1,718
IEP (not gifted) (Percent)	0	0	9.5	12.8	22.1	27	14.3
Student exited IEP in last 2 years (Number)	0	0	20	139	38	3	200
Student exited IEP in last 2 years (Percent)	0	0	1.3	1.7	1.7	2.7	1.7
Title I (Number)	0	1	63	789	340	25	1,218
Title I (Percent)	0	100	4	9.8	14.9	22.5	10.1
Title III served (Number)	0	0	15	105	50	4	174
Title III served (Percent)	0	0	.9	1.3	2.2	3.6	1.4
Title III not served (Number)	0	0	0	0	0	0	0
Title III not served (Percent)	0	0	0	0	0	0	0
Migrant student (Number)	0	0	2	9	2	0	13
Migrant student (Percent)	0	0	.1	.1	.1	0	.1
EL enrolled first year (Number)	0	0	2	5	2	0	9
EL enrolled first year (Percent)	0	0	.1	.1	.1	0	.1
EL enrolled not first year (Number)	0	0	17	109	55	4	185
EL enrolled not first year (Percent)	0	0	1.1	1.4	2.4	3.6	1.5
Exited ESL/bilingual program and in first year of monitoring (Number)	0	0	4	17	1	0	22
Exited ESL/bilingual program and in first year of monitoring (Percent)	0	0	.3	.2	0	0	.2
Exited ESL/bilingual program and in 2nd year of monitoring (Number)	0	0	6	11	1	0	18
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)	0	0	.4	.1	0	0	.1

Table I-8 (continued). Demographic Characteristics of Students taking the Winter Keystone: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Former EL no longer monitored (Number)	0	0	50	125	22	1	198
Former EL no longer monitored (Percent)	0	0	3.2	1.6	1	.9	1.6
LIFE first year (Number)	0	0	0	0	0	0	0
LIFE first year (Percent)	0	0	0	0	0	0	0
LIFE not first year (Number)	0	0	0	5	1	0	6
LIFE not first year (Percent)	0	0	0	.1	0	0	0
Former EL exited and in 3rd year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Percent)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Percent)	0	0	0	0	0	0	0
Foreign exchange student (Number)	0	0	0	1	0	0	1
Foreign exchange student (Percent)	0	0	0	0	0	0	0
Economically disadvantaged (Number)	1	1	563	2,544	998	55	4,162
Economically disadvantaged (Percent)	100	100	35.6	31.6	43.8	49.5	34.6
Historically Underperforming Subgroup (Number)	1	1	637	3,134	1,228	67	5,068
Historically Underperforming Subgroup (Percent)	100	100	40.3	38.9	53.8	60.4	42.1
Enrollment in school of residence after Oct 1 (Number)	0	0	37	96	74	14	221
Enrollment in school of residence after Oct 1 (Percent)	0	0	2.3	1.2	3.2	12.6	1.8
Enrollment in district of residence after Oct 1 (Number)	0	0	13	48	54	10	125
Enrollment in district of residence after Oct 1 (Percent)	0	0	.8	.6	2.4	9	1
Enrollment as PA resident after Oct 1 (Number)	0	0	8	17	12	1	38
Enrollment as PA resident after Oct 1 (Percent)	0	0	.5	.2	.5	.9	.3
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	1,020	1,088	349	18	2,475
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	64.6	13.5	15.3	16.2	20.6
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	145	256	154	15	570
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	9.2	3.2	6.8	13.5	4.7
Military family (Number)	0	0	6	38	41	2	87
Military family (Percent)	0	0	.4	.5	1.8	1.8	.7
Homeless (Number)	0	0	0	0	0	0	0
Homeless (Percent)	0	0	0	0	0	0	0
Foster (Number)	0	0	3	29	20	3	55
Foster (Percent)	0	0	.2	.4	.9	2.7	.5
Home schooled (Number)	0	0	0	0	0	0	0
Home schooled (Percent)	0	0	0	0	0	0	0
Court/agency placed (Number)	1	0	1	17	37	19	75
Court/agency placed (Percent)	100	0	.1	.2	1.6	17.1	.6
Number of assessed students (Number)	1	1	1,580	8,063	2,281	111	12,037

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–9. Demographic Characteristics of Students taking the Winter Keystone: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)	0	0	100	3,970	1,390	57	5,517
Female (Percent)	0	0	50.3	48.5	45.6	42.2	47.7
Male (Number)	0	0	99	4,218	1,657	78	6,052
Male (Percent)	0	0	49.7	51.5	54.4	57.8	52.3
American Indian/Alaskan Native (not Hispanic) (Number)	0	0	0	11	6	0	17
American Indian/Alaskan Native (not Hispanic) (Percent)	0	0	0	.1	.2	0	.1
Asian (not Hispanic) (Number)	0	0	4	321	56	3	384
Asian (not Hispanic) (Percent)	0	0	2	3.9	1.8	2.2	3.3
Black or African American (not Hispanic) (Number)	0	0	35	554	214	50	853
Black or African American (not Hispanic) (Percent)	0	0	17.6	6.8	7	37	7.4
Hispanic (any race) (Number)	0	0	27	832	183	14	1,056
Hispanic (any race) (Percent)	0	0	13.6	10.2	6	10.4	9.1
Multi-Racial (not Hispanic) (Number)	0	0	9	260	100	4	373
Multi-Racial (not Hispanic) (Percent)	0	0	4.5	3.2	3.3	3	3.2
White (not Hispanic) (Number)	0	0	124	6,208	2,488	64	8,884
White (not Hispanic) (Percent)	0	0	62.3	75.8	81.7	47.4	76.8
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)	0	0	0	3	0	0	3
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)	0	0	0	0	0	0	0
IEP (not gifted) (Number)	0	0	27	982	553	37	1,599
IEP (not gifted) (Percent)	0	0	13.6	12	18.1	27.4	13.8
Student exited IEP in last 2 years (Number)	0	0	5	109	53	6	173
Student exited IEP in last 2 years (Percent)	0	0	2.5	1.3	1.7	4.4	1.5
Title I (Number)	0	0	37	704	350	32	1,123
Title I (Percent)	0	0	18.6	8.6	11.5	23.7	9.7
Title III served (Number)	0	0	4	82	32	5	123
Title III served (Percent)	0	0	2	1	1.1	3.7	1.1
Title III not served (Number)	0	0	0	0	0	0	0
Title III not served (Percent)	0	0	0	0	0	0	0
Migrant student (Number)	0	0	0	7	2	0	9
Migrant student (Percent)	0	0	0	.1	.1	0	.1
EL enrolled first year (Number)	0	0	0	3	3	0	6
EL enrolled first year (Percent)	0	0	0	0	.1	0	.1
EL enrolled not first year (Number)	0	0	6	86	33	6	131
EL enrolled not first year (Percent)	0	0	3	1	1.1	4.4	1.1
Exited ESL/bilingual program and in first year of monitoring (Number)	0	0	0	19	1	0	20
Exited ESL/bilingual program and in first year of monitoring (Percent)	0	0	0	.2	0	0	.2
Exited ESL/bilingual program and in 2nd year of monitoring (Number)	0	0	0	17	3	0	20
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)	0	0	0	.2	.1	0	.2

Table I-9 (continued). Demographic Characteristics of Students taking the Winter Keystone: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Former EL no longer monitored (Number)	0	0	2	167	33	0	202
Former EL no longer monitored (Percent)	0	0	1	2	1.1	0	1.7
LIFE first year (Number)	0	0	0	0	0	0	0
LIFE first year (Percent)	0	0	0	0	0	0	0
LIFE not first year (Number)	0	0	0	0	1	0	1
LIFE not first year (Percent)	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 3rd year of monitoring (Percent)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Number)	0	0	0	0	0	0	0
Former EL exited and in 4th year of monitoring (Percent)	0	0	0	0	0	0	0
Foreign exchange student (Number)	0	0	0	1	0	0	1
Foreign exchange student (Percent)	0	0	0	0	0	0	0
Economically disadvantaged (Number)	0	0	90	2,535	1,175	80	3,880
Economically disadvantaged (Percent)	0	0	45.2	30.9	38.6	59.3	33.5
Historically Underperforming Subgroup (Number)	0	0	102	3,072	1,435	95	4,704
Historically Underperforming Subgroup (Percent)	0	0	51.3	37.5	47.1	70.4	40.6
Enrollment in school of residence after Oct 1 (Number)	0	0	26	121	84	9	240
Enrollment in school of residence after Oct 1 (Percent)	0	0	13.1	1.5	2.8	6.7	2.1
Enrollment in district of residence after Oct 1 (Number)	0	0	6	68	57	7	138
Enrollment in district of residence after Oct 1 (Percent)	0	0	3	.8	1.9	5.2	1.2
Enrollment as PA resident after Oct 1 (Number)	0	0	1	28	6	0	35
Enrollment as PA resident after Oct 1 (Percent)	0	0	.5	.3	.2	0	.3
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	50	1,215	323	33	1,621
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	25.1	14.8	10.6	24.4	14
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)	0	0	25	318	194	23	560
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)	0	0	12.6	3.9	6.4	17	4.8
Military family (Number)	0	0	0	37	49	0	86
Military family (Percent)	0	0	0	.5	1.6	0	.7
Homeless (Number)	0	0	0	0	0	0	0
Homeless (Percent)	0	0	0	0	0	0	0
Foster (Number)	0	0	2	31	18	4	55
Foster (Percent)	0	0	1	.4	.6	3	.5
Home schooled (Number)	0	0	0	0	0	0	0
Home schooled (Percent)	0	0	0	0	0	0	0
Court/agency placed (Number)	0	0	4	32	33	19	88
Court/agency placed (Percent)	0	0	2	.4	1.1	14.1	.8
Number of assessed students (Number)	0	0	199	8,191	3,047	135	11,572

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–10. Incidence of Presentation Accommodations Received on the Winter Keystone: Algebra I

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)	2	N/A	2
Braille format (Percent)	0	N/A	0
Large print format (Number)	4	N/A	4
Large print format (Percent)	0	N/A	0
Computer Assistive Technology (Number)	2	N/A	2
Computer Assistive Technology (Percent)	0	N/A	0
Some test items/questions read aloud (Number)	55	53	108
Some test items/questions read aloud (Percent)	.6	1.4	.8
All test items/questions read aloud (Number)	33	30	63
All test items/questions read aloud (Percent)	.4	.8	.5
Test items/questions signed (Number)	0	0	0
Test items/questions signed (Percent)	0	0	0
Test items/questions interpreted for EL student (Number)	0	0	0
Test items/questions interpreted for EL student (Percent)	0	0	0
Amplification device (Number)	0	1	1
Amplification device (Percent)	0	0	0
Magnification device (Number)	2	1	3
Magnification device (Percent)	0	0	0
Color overlay (Number)	0	N/A	0
Color overlay (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	1	4	5
Other (per Accommodations Guidelines) (Percent)	0	.1	0
Spanish version (Number)	13	N/A	13
Spanish version (Percent)	.1	N/A	.1
Audio (Number)	N/A	154	154
Audio (Percent)	N/A	4.1	1.2
Color Chooser (Number)	N/A	6	6
Color Chooser (Percent)	N/A	.2	0
Contrasting Text Chooser (Number)	N/A	2	2
Contrasting Text Chooser (Percent)	N/A	.1	0
Reverse Contrast (Number)	N/A	1	1
Reverse Contrast (Percent)	N/A	0	0
Refreshable Braille (Number)	N/A	0	0
Refreshable Braille (Percent)	N/A	0	0
Video Sign Language (Number)	N/A	0	0
Video Sign Language (Percent)	N/A	0	0
Number of assessed students (Number)	9,274	3,791	13,065

Table I–11. Incidence of Presentation Accommodations Received on the Winter Keystone: Biology

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)	0	N/A	0
Braille format (Percent)	0	N/A	0
Large print format (Number)	4	N/A	4
Large print format (Percent)	.1	N/A	0
Computer Assistive Technology (Number)	0	N/A	0
Computer Assistive Technology (Percent)	0	N/A	0
Some test items/questions read aloud (Number)	37	72	109
Some test items/questions read aloud (Percent)	.5	1.5	.9
All test items/questions read aloud (Number)	34	64	98
All test items/questions read aloud (Percent)	.5	1.3	.8
Test items/questions signed (Number)	0	0	0
Test items/questions signed (Percent)	0	0	0
Test items/questions interpreted for EL student (Number)	0	0	0
Test items/questions interpreted for EL student (Percent)	0	0	0
Amplification device (Number)	0	1	1
Amplification device (Percent)	0	0	0
Magnification device (Number)	1	0	1
Magnification device (Percent)	0	0	0
Color overlay (Number)	0	N/A	0
Color overlay (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	1	2	3
Other (per Accommodations Guidelines) (Percent)	0	0	0
Spanish version (Number)	20	N/A	20
Spanish version (Percent)	.3	N/A	.2
Audio (Number)	N/A	213	213
Audio (Percent)	N/A	4.4	1.8
Color Chooser (Number)	N/A	25	25
Color Chooser (Percent)	N/A	.5	.2
Contrasting Text Chooser (Number)	N/A	25	25
Contrasting Text Chooser (Percent)	N/A	.5	.2
Reverse Contrast (Number)	N/A	0	0
Reverse Contrast (Percent)	N/A	0	0
Refreshable Braille (Number)	N/A	0	0
Refreshable Braille (Percent)	N/A	0	0
Video Sign Language (Number)	N/A	0	0
Video Sign Language (Percent)	N/A	0	0
Number of assessed students (Number)	7,176	4,861	12,037

Table I–12. Incidence of Presentation Accommodations Received on the Winter Keystone: Literature

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)	1	N/A	1
Braille format (Percent)	0	N/A	0
Large print format (Number)	3	N/A	3
Large print format (Percent)	0	N/A	0
Computer Assistive Technology (Number)	1	N/A	1
Computer Assistive Technology (Percent)	0	N/A	0
Amplification device (Number)	0	2	2
Amplification device (Percent)	0	0	0
Magnification device (Number)	1	1	2
Magnification device (Percent)	0	0	0
Color overlay (Number)	0	N/A	0
Color overlay (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	0	8	8
Other (per Accommodations Guidelines) (Percent)	0	.2	.1
Color Chooser (Number)	N/A	25	25
Color Chooser (Percent)	N/A	.6	.2
Contrasting Text Chooser (Number)	N/A	23	23
Contrasting Text Chooser (Percent)	N/A	.6	.2
Reverse Contrast (Number)	N/A	0	0
Reverse Contrast (Percent)	N/A	0	0
Refreshable Braille (Number)	N/A	0	0
Refreshable Braille (Percent)	N/A	0	0
Number of assessed students (Number)	7,490	4,082	11,572

Table I–13. Incidence of Response Accommodations Received on the Winter Keystone: Algebra I

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student's direction (Number)	3	1	4
Test administrator marked multiple-choice responses at student's direction (Percent)	0	0	0
Test administrator scribed open-ended responses at student's direction (Number)	2	2	4
Test administrator scribed open-ended responses at student's direction (Percent)	0	.1	0
Test administrator transcribed student responses (Number)	6	0	6
Test administrator transcribed student responses (Percent)	.1	0	0
Qualified interpreter translated, transcribed, and/or scribed student's signed responses (Number)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed student's signed responses (Percent)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Number)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Percent)	0	0	0
Keyboard, word processor, or computer (Number)	2	N/A	2
Keyboard, word processor, or computer (Percent)	0	N/A	0
Braille/Notetaker (Number)	3	N/A	3
Braille/Notetaker (Percent)	0	N/A	0
Augmentative communication device (Number)	0	0	0
Augmentative communication device (Percent)	0	0	0
Computer Assistive Technology (Number)	0	N/A	0
Computer Assistive Technology (Percent)	0	N/A	0
Translation dictionary for EL student (Number)	20	1	21
Translation dictionary for EL student (Percent)	.2	0	.2
Other (per Accommodations Guidelines) (Number)	3	0	3
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	9,274	3,791	13,065

Table I–14. Incidence of Response Accommodations Received on the Winter Keystone: Biology

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student's direction (Number)	2	0	2
Test administrator marked multiple-choice responses at student's direction (Percent)	0	0	0
Test administrator scribed open-ended responses at student's direction (Number)	5	0	5
Test administrator scribed open-ended responses at student's direction (Percent)	.1	0	0
Test administrator transcribed student responses (Number)	9	2	11
Test administrator transcribed student responses (Percent)	.1	0	.1
Qualified interpreter translated, transcribed, and/or scribed student's signed responses (Number)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed student's signed responses (Percent)	0	0	0
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Number)	2	2	4
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Percent)	0	0	0
Keyboard, word processor, or computer (Number)	1	N/A	1
Keyboard, word processor, or computer (Percent)	0	N/A	0
Braille/Notetaker (Number)	0	N/A	0
Braille/Notetaker (Percent)	0	N/A	0
Augmentative communication device (Number)	0	0	0
Augmentative communication device (Percent)	0	0	0
Computer Assistive Technology (Number)	0	N/A	0
Computer Assistive Technology (Percent)	0	N/A	0
Translation dictionary for EL student (Number)	3	0	3
Translation dictionary for EL student (Percent)	0	0	0
Other (per Accommodations Guidelines) (Number)	0	0	0
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	7,176	4,861	12,037

Table I–15. Incidence of Response Accommodations Received on the Winter Keystone: Literature

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student's direction (Number)	5	0	5
Test administrator marked multiple-choice responses at student's direction (Percent)	.1	0	0
Test administrator scribed open-ended responses at student's direction (Number)	7	0	7
Test administrator scribed open-ended responses at student's direction (Percent)	.1	0	.1
Test administrator transcribed student responses (Number)	8	0	8
Test administrator transcribed student responses (Percent)	.1	0	.1
Keyboard, word processor, or computer (Number)	9	N/A	9
Keyboard, word processor, or computer (Percent)	.1	N/A	.1
Braille/Notetaker (Number)	0	N/A	0
Braille/Notetaker (Percent)	0	N/A	0
Augmentative communication device (Number)	0	0	0
Augmentative communication device (Percent)	0	0	0
Computer Assistive Technology (Number)	0	N/A	0
Computer Assistive Technology (Percent)	0	N/A	0
Other (per Accommodations Guidelines) (Number)	0	0	0
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	7,490	4,082	11,572

Table I–16. Incidence of Setting Accommodations Received on the Winter Keystone: Algebra I

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)	1	0	1
Hospital/home setting (Percent)	0	0	0
One-on-one setting (Number)	19	2	21
One-on-one setting (Percent)	.2	.1	.2
Small group setting (Number)	641	226	867
Small group setting (Percent)	6.9	6	6.6
Other (per Accommodations Guidelines) (Number)	2	2	4
Other (per Accommodations Guidelines) (Percent)	0	.1	0
Number of assessed students (Number)	9,274	3,791	13,065

Table I–17. Incidence of Setting Accommodations Received on the Winter Keystone: Biology

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)	2	0	2
Hospital/home setting (Percent)	0	0	0
One-on-one setting (Number)	17	3	20
One-on-one setting (Percent)	.2	.1	.2
Small group setting (Number)	593	343	936
Small group setting (Percent)	8.3	7.1	7.8
Other (per Accommodations Guidelines) (Number)	2	1	3
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	7,176	4,861	12,037

Table I–18. Incidence of Setting Accommodations Received on the Winter Keystone: Literature

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)	1	0	1
Hospital/home setting (Percent)	0	0	0
One-on-one setting (Number)	14	4	18
One-on-one setting (Percent)	.2	.1	.2
Small group setting (Number)	600	237	837
Small group setting (Percent)	8	5.8	7.2
Other (per Accommodations Guidelines) (Number)	5	4	9
Other (per Accommodations Guidelines) (Percent)	.1	.1	.1
Number of assessed students (Number)	7,490	4,082	11,572

Table I–19. Incidence of Timing Accommodations Received on the Winter Keystone: Algebra I

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)	585	188	773
Extended time (Percent)	6.3	5	5.9
Frequent breaks (Number)	37	84	121
Frequent breaks (Percent)	.4	2.2	.9
Changed test schedule (Number)	8	0	8
Changed test schedule (Percent)	.1	0	.1
Other (per Accommodations Guidelines) (Number)	3	1	4
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	9,274	3,791	13,065

Table I–20. Incidence of Timing Accommodations Received on the Winter Keystone: Biology

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)	421	149	570
Extended time (Percent)	5.9	3.1	4.7
Frequent breaks (Number)	55	56	111
Frequent breaks (Percent)	.8	1.2	.9
Changed test schedule (Number)	6	1	7
Changed test schedule (Percent)	.1	0	.1
Other (per Accommodations Guidelines) (Number)	0	2	2
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	7,176	4,861	12,037

Table I–21. Incidence of Timing Accommodations Received on the Winter Keystone: Literature

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)	624	251	875
Extended time (Percent)	8.3	6.1	7.6
Frequent breaks (Number)	46	46	92
Frequent breaks (Percent)	.6	1.1	.8
Changed test schedule (Number)	5	1	6
Changed test schedule (Percent)	.1	0	.1
Other (per Accommodations Guidelines) (Number)	3	2	5
Other (per Accommodations Guidelines) (Percent)	0	0	0
Number of assessed students (Number)	7,490	4,082	11,572

Table I–22. Accommodation Rate for Non-IEP and IEP Students on the Winter Keystone Exams: Algebra I

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)	8,029	3,209	11,238
Non-Accommodated (Number)	7,454	3,114	10,568
Non-Accommodated (Percent)	92.8	97	94
Accommodated (Number)	575	95	670
Accommodated (Percent)	7.2	3	6
IEP Students (Number)	1,245	582	1,827
Non-Accommodated (Number)	667	304	971
Non-Accommodated (Percent)	53.6	52.2	53.1
Accommodated (Number)	578	278	856
Accommodated (Percent)	46.4	47.8	46.9

Table I–23. Accommodation Rate for Non-IEP and IEP Students on the Winter Keystone Exams: Biology

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)	6,157	4,162	10,319
Non-Accommodated (Number)	5,790	4,117	9,907
Non-Accommodated (Percent)	94	98.9	96
Accommodated (Number)	367	45	412
Accommodated (Percent)	6	1.1	4
IEP Students (Number)	1,019	699	1,718
Non-Accommodated (Number)	486	326	812
Non-Accommodated (Percent)	47.7	46.6	47.3
Accommodated (Number)	533	373	906
Accommodated (Percent)	52.3	53.4	52.7

Table I–24. Accommodation Rate for Non-IEP and IEP Students on the Winter Keystone Exams: Literature

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)	6,463	3,510	9,973
Non-Accommodated (Number)	5,915	3,378	9,293
Non-Accommodated (Percent)	91.5	96.2	93.2
Accommodated (Number)	548	132	680
Accommodated (Percent)	8.5	3.8	6.8
IEP Students (Number)	1,027	572	1,599
Non-Accommodated (Number)	500	326	826
Non-Accommodated (Percent)	48.7	57	51.7
Accommodated (Number)	527	246	773
Accommodated (Percent)	51.3	43	48.3

Table I–25. Incidence of IEP and EL Students Receiving Accommodations on the Winter Keystone: Algebra I

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Some test items/questions read aloud (Number)	1	1	4	49
PPT - Some test items/questions read aloud (Percent)	3.6	.5	.1	4
PPT - All test items/questions read aloud (Number)	1	1	6	25
PPT - All test items/questions read aloud (Percent)	3.6	.5	.1	2.1
PPT - Small group setting (Number)	8	16	101	516
PPT - Small group setting (Percent)	28.6	7.3	1.3	42.4
PPT - Extended time (Number)	5	21	432	127
PPT - Extended time (Percent)	17.9	9.6	5.5	10.4
PPT - Frequent breaks (Number)	0	0	5	32
PPT - Frequent breaks (Percent)	0	0	.1	2.6
PPT - Number assessed (Number)	28	219	7,810	1,217
CBT - Some test items/questions read aloud (Number)	0	0	2	51
CBT - Some test items/questions read aloud (Percent)	0	0	.1	8.9
CBT - All test items/questions read aloud (Number)	0	0	0	30
CBT - All test items/questions read aloud (Percent)	0	0	0	5.2
CBT - Small group setting (Number)	2	0	13	211
CBT - Small group setting (Percent)	28.6	0	.4	36.7
CBT - Extended time (Number)	0	0	84	104
CBT - Extended time (Percent)	0	0	2.7	18.1
CBT - Frequent breaks (Number)	1	0	6	77
CBT - Frequent breaks (Percent)	14.3	0	.2	13.4
CBT - Number assessed (Number)	7	47	3,162	575
Total - Some test items/questions read aloud (Number)	1	1	6	100
Total - Some test items/questions read aloud (Percent)	2.9	.4	.1	5.6
Total - All test items/questions read aloud (Number)	1	1	6	55
Total - All test items/questions read aloud (Percent)	2.9	.4	.1	3.1
Total - Small group setting (Number)	10	16	114	727
Total - Small group setting (Percent)	28.6	6	1	40.6
Total - Extended time (Number)	5	21	516	231
Total - Extended time (Percent)	14.3	7.9	4.7	12.9
Total - Frequent breaks (Number)	1	0	11	109
Total - Frequent breaks (Percent)	2.9	0	.1	6.1
Total - Number assessed (Number)	35	266	10,972	1,792

Table I–26. Incidence of IEP and EL Students Receiving Accommodations on the Winter Keystone: Biology

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Some test items/questions read aloud (Number)	0	0	2	35
PPT - Some test items/questions read aloud (Percent)	0	0	0	3.5
PPT - All test items/questions read aloud (Number)	0	0	0	34
PPT - All test items/questions read aloud (Percent)	0	0	0	3.4
PPT - Small group setting (Number)	7	7	89	490
PPT - Small group setting (Percent)	50	7.8	1.5	48.8
PPT - Extended time (Number)	3	18	254	146
PPT - Extended time (Percent)	21.4	20	4.2	14.5
PPT - Frequent breaks (Number)	0	0	5	50
PPT - Frequent breaks (Percent)	0	0	.1	5
PPT - Number assessed (Number)	14	90	6,067	1,005
CBT - Some test items/questions read aloud (Number)	3	0	0	69
CBT - Some test items/questions read aloud (Percent)	27.3	0	0	10
CBT - All test items/questions read aloud (Number)	0	0	2	62
CBT - All test items/questions read aloud (Percent)	0	0	0	9
CBT - Small group setting (Number)	6	3	24	310
CBT - Small group setting (Percent)	54.5	3.8	.6	45.1
CBT - Extended time (Number)	4	0	22	123
CBT - Extended time (Percent)	36.4	0	.5	17.9
CBT - Frequent breaks (Number)	0	0	2	54
CBT - Frequent breaks (Percent)	0	0	0	7.8
CBT - Number assessed (Number)	11	79	4,083	688
Total - Some test items/questions read aloud (Number)	3	0	2	104
Total - Some test items/questions read aloud (Percent)	12	0	0	6.1
Total - All test items/questions read aloud (Number)	0	0	2	96
Total - All test items/questions read aloud (Percent)	0	0	0	5.7
Total - Small group setting (Number)	13	10	113	800
Total - Small group setting (Percent)	52	5.9	1.1	47.3
Total - Extended time (Number)	7	18	276	269
Total - Extended time (Percent)	28	10.7	2.7	15.9
Total - Frequent breaks (Number)	0	0	7	104
Total - Frequent breaks (Percent)	0	0	.1	6.1
Total - Number assessed (Number)	25	169	10,150	1,693

Table I–27. Incidence of IEP and EL Students Receiving Accommodations on the Winter Keystone: Literature

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Small group setting (Number)	9	14	96	481
PPT - Small group setting (Percent)	52.9	21.2	1.5	47.6
PPT - Extended time (Number)	2	2	455	165
PPT - Extended time (Percent)	11.8	3	7.1	16.3
PPT - Frequent breaks (Number)	0	0	3	43
PPT - Frequent breaks (Percent)	0	0	0	4.3
PPT - Number assessed (Number)	17	66	6,397	1,010
CBT - Small group setting (Number)	5	4	18	210
CBT - Small group setting (Percent)	62.5	8.7	.5	37.2
CBT - Extended time (Number)	5	1	113	132
CBT - Extended time (Percent)	62.5	2.2	3.3	23.4
CBT - Frequent breaks (Number)	1	0	3	42
CBT - Frequent breaks (Percent)	12.5	0	.1	7.4
CBT - Number assessed (Number)	8	46	3,464	564
Total - Small group setting (Number)	14	18	114	691
Total - Small group setting (Percent)	56	16.1	1.2	43.9
Total - Extended time (Number)	7	3	568	297
Total - Extended time (Percent)	28	2.7	5.8	18.9
Total - Frequent breaks (Number)	1	0	6	85
Total - Frequent breaks (Percent)	4	0	.1	5.4
Total - Number assessed (Number)	25	112	9,861	1,574

SUMMER

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

Table I–28. Students Assessed on the Summer Keystone: Algebra I

Description	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)									
Total number of CBT processed (Number)									
Total number of tests processed (Number)									
Total number of tests processed with a score (Number)									
Total number of tests processed with a score (Percent)									
Total number of tests processed without a score (Number)									
Total number of tests processed without a score (Percent)									

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table I–29. Students Assessed on the Summer Keystone: Biology

Description	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)							
Total number of CBT processed (Number)							
Total number of tests processed (Number)							
Total number of tests processed with a score (Number)							
Total number of tests processed with a score (Percent)							
Total number of tests processed without a score (Number)							
Total number of tests processed without a score (Percent)							

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–30. Students Assessed on the Summer Keystone: Literature

Description	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Total number of PPT processed (Number)							
Total number of CBT processed (Number)							
Total number of tests processed (Number)							
Total number of tests processed with a score (Number)							
Total number of tests processed with a score (Percent)							
Total number of tests processed without a score (Number)							
Total number of tests processed without a score (Percent)							

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–31. Counts of Students without Scores on the Summer Keystone: Algebra I

Reason for Non-Assessment	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)									
Extended absence from school (Percent)									
Non-attempt (Number)									
Non-attempt (Percent)									
Medical emergency (Number)									
Medical emergency (Percent)									
Parental request - Chapter 4 (Number)									
Parental request - Chapter 4 (Percent)									
Parental request - Other reasons (Number)									
Parental request - Other reasons (Percent)									
Other reasons (Number)									
Other reasons (Percent)									
Total not assessed (Number)									

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table I–32. Counts of Students without Scores on the Summer Keystone: Biology

Reason for Non-Assessment	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)							
Extended absence from school (Percent)							
Non-attempt (Number)							
Non-attempt (Percent)							
Medical emergency (Number)							
Medical emergency (Percent)							
Parental request - Chapter 4 (Number)							
Parental request - Chapter 4 (Percent)							
Parental request - Other reasons (Number)							
Parental request - Other reasons (Percent)							
Other reasons (Number)							
Other reasons (Percent)							
Total not assessed (Number)							

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–33. Counts of Students without Scores on the Summer Keystone: Literature

Reason for Non-Assessment	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Extended absence from school (Number)							
Extended absence from school (Percent)							
Non-attempt (Number)							
Non-attempt (Percent)							
EL in first year in U.S. schools (Number)							
EL in first year in U.S. schools (Percent)							
Medical emergency (Number)							
Medical emergency (Percent)							
Parental request - Chapter 4 (Number)							
Parental request - Chapter 4 (Percent)							
Parental request - Other reasons (Number)							
Parental request - Other reasons (Percent)							
Other reasons (Number)							
Other reasons (Percent)							
Total not assessed (Number)							

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–34. Demographic Characteristics of Students taking the Summer Keystone: Algebra I

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)									
Female (Percent)									
Male (Number)									
Male (Percent)									
American Indian/Alaskan Native (not Hispanic) (Number)									
American Indian/Alaskan Native (not Hispanic) (Percent)									
Asian (not Hispanic) (Number)									
Asian (not Hispanic) (Percent)									
Black or African American (not Hispanic) (Number)									
Black or African American (not Hispanic) (Percent)									
Hispanic (any race) (Number)									
Hispanic (any race) (Percent)									
Multi-Racial (not Hispanic) (Number)									
Multi-Racial (not Hispanic) (Percent)									
White (not Hispanic) (Number)									
White (not Hispanic) (Percent)									
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)									
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)									
IEP (not gifted) (Number)									
IEP (not gifted) (Percent)									
Student exited IEP in last 2 years (Number)									
Student exited IEP in last 2 years (Percent)									
Title I (Number)									
Title I (Percent)									
Title III served (Number)									
Title III served (Percent)									
Title III not served (Number)									
Title III not served (Percent)									
Migrant student (Number)									
Migrant student (Percent)									
EL enrolled first year (Number)									
EL enrolled first year (Percent)									
EL enrolled not first year (Number)									
EL enrolled not first year (Percent)									
Exited ESL/bilingual program and in first year of monitoring (Number)									
Exited ESL/bilingual program and in first year of monitoring (Percent)									

Table I–34 (continued). Demographic Characteristics of Students taking the Summer Keystone: Algebra

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in 2nd year of monitoring (Number)									
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)									
Former EL no longer monitored (Number)									
Former EL no longer monitored (Percent)									
LIFE first year (Number)									
LIFE first year (Percent)									
LIFE not first year (Number)									
LIFE not first year (Percent)									
Former EL exited and in 3rd year of monitoring (Number)									
Former EL exited and in 3rd year of monitoring (Percent)									
Former EL exited and in 4th year of monitoring (Number)									
Former EL exited and in 4th year of monitoring (Percent)									
Foreign exchange student (Number)									
Foreign exchange student (Percent)									
Economically disadvantaged (Number)									
Economically disadvantaged (Percent)									
Historically Underperforming Subgroup (Number)									
Historically Underperforming Subgroup (Percent)									
Enrollment in school of residence after Oct 1 (Number)									
Enrollment in school of residence after Oct 1 (Percent)									
Enrollment in district of residence after Oct 1 (Number)									
Enrollment in district of residence after Oct 1 (Percent)									
Enrollment as PA resident after Oct 1 (Number)									
Enrollment as PA resident after Oct 1 (Percent)									
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)									
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)									
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)									
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)									
Military family (Number)									
Military family (Percent)									
Homeless (Number)									

Table I–34 (continued). Demographic Characteristics of Students taking the Summer Keystone: Algebra

Demographic or Educational Characteristic	Other*	Gr.6	Gr.7	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Homeless (Percent)									
Foster (Number)									
Foster (Percent)									
Home schooled (Number)									
Home schooled (Percent)									
Court/agency placed (Number)									
Court/agency placed (Percent)									
Number of assessed students (Number)									

*Other combines students coded as (1) below Grade 6, (2) ungraded, or (3) without a coded grade

Table I–35. Demographic Characteristics of Students taking the Summer Keystone: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)							
Female (Percent)							
Male (Number)							
Male (Percent)							
American Indian/Alaskan Native (not Hispanic) (Number)							
American Indian/Alaskan Native (not Hispanic) (Percent)							
Asian (not Hispanic) (Number)							
Asian (not Hispanic) (Percent)							
Black or African American (not Hispanic) (Number)							
Black or African American (not Hispanic) (Percent)							
Hispanic (any race) (Number)							
Hispanic (any race) (Percent)							
Multi-Racial (not Hispanic) (Number)							
Multi-Racial (not Hispanic) (Percent)							
White (not Hispanic) (Number)							
White (not Hispanic) (Percent)							
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)							
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)							
IEP (not gifted) (Number)							
IEP (not gifted) (Percent)							
Student exited IEP in last 2 years (Number)							
Student exited IEP in last 2 years (Percent)							
Title I (Number)							
Title I (Percent)							
Title III served (Number)							
Title III served (Percent)							
Title III not served (Number)							
Title III not served (Percent)							
Migrant student (Number)							
Migrant student (Percent)							
EL enrolled first year (Number)							
EL enrolled first year (Percent)							
EL enrolled not first year (Number)							
EL enrolled not first year (Percent)							
Exited ESL/bilingual program and in first year of monitoring (Number)							
Exited ESL/bilingual program and in first year of monitoring (Percent)							
Exited ESL/bilingual program and in 2nd year of monitoring (Number)							

Table I–35 (continued). Demographic Characteristics of Students taking the Summer Keystone: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)							
Former EL no longer monitored (Number)							
Former EL no longer monitored (Percent)							
LIFE first year (Number)							
LIFE first year (Percent)							
LIFE not first year (Number)							
LIFE not first year (Percent)							
Former EL exited and in 3rd year of monitoring (Number)							
Former EL exited and in 3rd year of monitoring (Percent)							
Former EL exited and in 4th year of monitoring (Number)							
Former EL exited and in 4th year of monitoring (Percent)							
Foreign exchange student (Number)							
Foreign exchange student (Percent)							
Economically disadvantaged (Number)							
Economically disadvantaged (Percent)							
Historically Underperforming Subgroup (Number)							
Historically Underperforming Subgroup (Percent)							
Enrollment in school of residence after Oct 1 (Number)							
Enrollment in school of residence after Oct 1 (Percent)							
Enrollment in district of residence after Oct 1 (Number)							
Enrollment in district of residence after Oct 1 (Percent)							
Enrollment as PA resident after Oct 1 (Number)							
Enrollment as PA resident after Oct 1 (Percent)							
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)							
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)							
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)							
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)							
Military family (Number)							
Military family (Percent)							
Homeless (Number)							
Homeless (Percent)							
Foster (Number)							
Foster (Percent)							

Table I–35 (continued). Demographic Characteristics of Students taking the Summer Keystone: Biology

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Home schooled (Number)							
Home schooled (Percent)							
Court/agency placed (Number)							
Court/agency placed (Percent)							
Number of assessed students (Number)							

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–36. Demographic Characteristics of Students taking the Summer Keystone: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Female (Number)							
Female (Percent)							
Male (Number)							
Male (Percent)							
American Indian/Alaskan Native (not Hispanic) (Number)							
American Indian/Alaskan Native (not Hispanic) (Percent)							
Asian (not Hispanic) (Number)							
Asian (not Hispanic) (Percent)							
Black or African American (not Hispanic) (Number)							
Black or African American (not Hispanic) (Percent)							
Hispanic (any race) (Number)							
Hispanic (any race) (Percent)							
Multi-Racial (not Hispanic) (Number)							
Multi-Racial (not Hispanic) (Percent)							
White (not Hispanic) (Number)							
White (not Hispanic) (Percent)							
Native Hawaiian or Other Pacific Islander (not Hispanic) (Number)							
Native Hawaiian or Other Pacific Islander (not Hispanic) (Percent)							
IEP (not gifted) (Number)							
IEP (not gifted) (Percent)							
Student exited IEP in last 2 years (Number)							
Student exited IEP in last 2 years (Percent)							
Title I (Number)							
Title I (Percent)							
Title III served (Number)							
Title III served (Percent)							
Title III not served (Number)							
Title III not served (Percent)							
Migrant student (Number)							
Migrant student (Percent)							
EL enrolled first year (Number)							
EL enrolled first year (Percent)							
EL enrolled not first year (Number)							
EL enrolled not first year (Percent)							
Exited ESL/bilingual program and in first year of monitoring (Number)							
Exited ESL/bilingual program and in first year of monitoring (Percent)							

Table I–36 (continued). Demographic Characteristics of Students taking the Summer Keystone: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Exited ESL/bilingual program and in 2nd year of monitoring (Number)							
Exited ESL/bilingual program and in 2nd year of monitoring (Percent)							
Former EL no longer monitored (Number)							
Former EL no longer monitored (Percent)							
LIFE first year (Number)							
LIFE first year (Percent)							
LIFE not first year (Number)							
LIFE not first year (Percent)							
Former EL exited and in 3rd year of monitoring (Number)							
Former EL exited and in 3rd year of monitoring (Percent)							
Former EL exited and in 4th year of monitoring (Number)							
Former EL exited and in 4th year of monitoring (Percent)							
Foreign exchange student (Number)							
Foreign exchange student (Percent)							
Economically disadvantaged (Number)							
Economically disadvantaged (Percent)							
Historically Underperforming Subgroup (Number)							
Historically Underperforming Subgroup (Percent)							
Enrollment in school of residence after Oct 1 (Number)							
Enrollment in school of residence after Oct 1 (Percent)							
Enrollment in district of residence after Oct 1 (Number)							
Enrollment in district of residence after Oct 1 (Percent)							
Enrollment as PA resident after Oct 1 (Number)							
Enrollment as PA resident after Oct 1 (Percent)							
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Number)							
Enrollment in school of residence after previous Oct 1 but on/before current Oct 1 (Percent)							
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Number)							
Enrollment in district of residence after previous Oct 1 but on/before current Oct 1 (Percent)							
Military family (Number)							
Military family (Percent)							
Homeless (Number)							

Table I–36 (continued). Demographic Characteristics of Students taking the Summer Keystone: Literature

Demographic or Educational Characteristic	Other*	Gr.8	Gr.9	Gr.10	Gr.11	Gr.12	Total
Homeless (Percent)							
Foster (Number)							
Foster (Percent)							
Home schooled (Number)							
Home schooled (Percent)							
Court/agency placed (Number)							
Court/agency placed (Percent)							
Number of assessed students (Number)							

*Other combines students coded as (1) below Grade 8, (2) ungraded, or (3) without a coded grade

Table I–37. Incidence of Presentation Accommodations Received on the Summer Keystone: Algebra I

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)			
Braille format (Percent)			
Large print format (Number)			
Large print format (Percent)			
Computer Assistive Technology (Number)			
Computer Assistive Technology (Percent)			
Some test items/questions read aloud (Number)			
Some test items/questions read aloud (Percent)			
All test items/questions read aloud (Number)			
All test items/questions read aloud (Percent)			
Test items/questions signed (Number)			
Test items/questions signed (Percent)			
Test items/questions interpreted for EL student (Number)			
Test items/questions interpreted for EL student (Percent)			
Amplification device (Number)			
Amplification device (Percent)			
Magnification device (Number)			
Magnification device (Percent)			
Color overlay (Number)			
Color overlay (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Spanish version (Number)			
Spanish version (Percent)			
Audio (Number)			
Audio (Percent)			
Color Chooser (Number)			
Color Chooser (Percent)			
Contrasting Text Chooser (Number)			
Contrasting Text Chooser (Percent)			
Reverse Contrast (Number)			
Reverse Contrast (Percent)			
Refreshable Braille (Number)			
Refreshable Braille (Percent)			
Video Sign Language (Number)			
Video Sign Language (Percent)			
Number of assessed students (Number)			

Table I–38. Incidence of Presentation Accommodations Received on the Summer Keystone: Biology

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)			
Braille format (Percent)			
Large print format (Number)			
Large print format (Percent)			
Computer Assistive Technology (Number)			
Computer Assistive Technology (Percent)			
Some test items/questions read aloud (Number)			
Some test items/questions read aloud (Percent)			
All test items/questions read aloud (Number)			
All test items/questions read aloud (Percent)			
Test items/questions signed (Number)			
Test items/questions signed (Percent)			
Test items/questions interpreted for EL student (Number)			
Test items/questions interpreted for EL student (Percent)			
Amplification device (Number)			
Amplification device (Percent)			
Magnification device (Number)			
Magnification device (Percent)			
Color overlay (Number)			
Color overlay (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Spanish version (Number)			
Spanish version (Percent)			
Audio (Number)			
Audio (Percent)			
Color Chooser (Number)			
Color Chooser (Percent)			
Contrasting Text Chooser (Number)			
Contrasting Text Chooser (Percent)			
Reverse Contrast (Number)			
Reverse Contrast (Percent)			
Refreshable Braille (Number)			
Refreshable Braille (Percent)			
Video Sign Language (Number)			
Video Sign Language (Percent)			
Number of assessed students (Number)			

Table I–39. Incidence of Presentation Accommodations Received on the Summer Keystone: Literature

Type of Presentation Accommodation	PPT	CBT	Total
Braille format (Number)			
Braille format (Percent)			
Large print format (Number)			
Large print format (Percent)			
Computer Assistive Technology (Number)			
Computer Assistive Technology (Percent)			
Amplification device (Number)			
Amplification device (Percent)			
Magnification device (Number)			
Magnification device (Percent)			
Color overlay (Number)			
Color overlay (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Color Chooser (Number)			
Color Chooser (Percent)			
Contrasting Text Chooser (Number)			
Contrasting Text Chooser (Percent)			
Reverse Contrast (Number)			
Reverse Contrast (Percent)			
Refreshable Braille (Number)			
Refreshable Braille (Percent)			
Number of assessed students (Number)			

Table I–40. Incidence of Response Accommodations Received on the Summer Keystone: Algebra I

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student’s direction (Number)			
Test administrator marked multiple-choice responses at student’s direction (Percent)			
Test administrator scribed open-ended responses at student’s direction (Number)			
Test administrator scribed open-ended responses at student’s direction (Percent)			
Test administrator transcribed student responses (Number)			
Test administrator transcribed student responses (Percent)			
Qualified interpreter translated, transcribed, and/or scribed student’s signed responses (Number)			
Qualified interpreter translated, transcribed, and/or scribed student’s signed responses (Percent)			
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Number)			
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Percent)			
Keyboard, word processor, or computer (Number)			
Keyboard, word processor, or computer (Percent)			
Braille/Notetaker (Number)			
Braille/Notetaker (Percent)			
Augmentative communication device (Number)			
Augmentative communication device (Percent)			
Computer Assistive Technology (Number)			
Computer Assistive Technology (Percent)			
Translation dictionary for EL student (Number)			
Translation dictionary for EL student (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–41. Incidence of Response Accommodations Received on the Summer Keystone: Biology

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student’s direction (Number)			
Test administrator marked multiple-choice responses at student’s direction (Percent)			
Test administrator scribed open-ended responses at student’s direction (Number)			
Test administrator scribed open-ended responses at student’s direction (Percent)			
Test administrator transcribed student responses (Number)			
Test administrator transcribed student responses (Percent)			
Qualified interpreter translated, transcribed, and/or scribed student’s signed responses (Number)			
Qualified interpreter translated, transcribed, and/or scribed student’s signed responses (Percent)			
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Number)			
Qualified interpreter translated, transcribed, and/or scribed EL student responses (Percent)			
Keyboard, word processor, or computer (Number)			
Keyboard, word processor, or computer (Percent)			
Braille/Notetaker (Number)			
Braille/Notetaker (Percent)			
Augmentative communication device (Number)			
Augmentative communication device (Percent)			
Computer Assistive Technology (Number)			
Computer Assistive Technology (Percent)			
Translation dictionary for EL student (Number)			
Translation dictionary for EL student (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–42. Incidence of Response Accommodations Received on the Summer Keystone: Literature

Type of Response Accommodation	PPT	CBT	Total
Test administrator marked multiple-choice responses at student’s direction (Number)			
Test administrator marked multiple-choice responses at student’s direction (Percent)			
Test administrator scribed open-ended responses at student’s direction (Number)			
Test administrator scribed open-ended responses at student’s direction (Percent)			
Test administrator transcribed student responses (Number)			
Test administrator transcribed student responses (Percent)			
Keyboard, word processor, or computer (Number)			
Keyboard, word processor, or computer (Percent)			
Braille/Notetaker (Number)			
Braille/Notetaker (Percent)			
Augmentative communication device (Number)			
Augmentative communication device (Percent)			
Computer Assistive Technology (Number)			
Computer Assistive Technology (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–43. Incidence of Setting Accommodations Received on the Summer Keystone: Algebra I

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)			
Hospital/home setting (Percent)			
One-on-one setting (Number)			
One-on-one setting (Percent)			
Small group setting (Number)			
Small group setting (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–44. Incidence of Setting Accommodations Received on the Summer Keystone: Biology

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)			
Hospital/home setting (Percent)			
One-on-one setting (Number)			
One-on-one setting (Percent)			
Small group setting (Number)			
Small group setting (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–45. Incidence of Setting Accommodations Received on the Summer Keystone: Literature

Type of Setting Accommodation	PPT	CBT	Total
Hospital/home setting (Number)			
Hospital/home setting (Percent)			
One-on-one setting (Number)			
One-on-one setting (Percent)			
Small group setting (Number)			
Small group setting (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–46. Incidence of Timing Accommodations Received on the Summer Keystone: Algebra I

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)			
Extended time (Percent)			
Frequent breaks (Number)			
Frequent breaks (Percent)			
Changed test schedule (Number)			
Changed test schedule (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–47. Incidence of Timing Accommodations Received on the Summer Keystone: Biology

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)			
Extended time (Percent)			
Frequent breaks (Number)			
Frequent breaks (Percent)			
Changed test schedule (Number)			
Changed test schedule (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–48. Incidence of Timing Accommodations Received on the Summer Keystone: Literature

Type of Timing Accommodation	PPT	CBT	Total
Extended time (Number)			
Extended time (Percent)			
Frequent breaks (Number)			
Frequent breaks (Percent)			
Changed test schedule (Number)			
Changed test schedule (Percent)			
Other (per Accommodations Guidelines) (Number)			
Other (per Accommodations Guidelines) (Percent)			
Number of assessed students (Number)			

Table I–49. Accommodation Rate for Non-IEP and IEP Students on the Summer Keystone Exams: Algebra I

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)			
Non-Accommodated (Number)			
Non-Accommodated (Percent)			
Accommodated (Number)			
Accommodated (Percent)			
IEP Students (Number)			
Non-Accommodated (Number)			
Non-Accommodated (Percent)			
Accommodated (Number)			
Accommodated (Percent)			

Table I–50. Accommodation Rate for Non-IEP and IEP Students on the Summer Keystone Exams: Biology

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)			
Non-Accommodated (Number)			
Non-Accommodated (Percent)			
Accommodated (Number)			
Accommodated (Percent)			
IEP Students (Number)			
Non-Accommodated (Number)			
Non-Accommodated (Percent)			
Accommodated (Number)			
Accommodated (Percent)			

Table I–51. Accommodation Rate for Non-IEP and IEP Students on the Summer Keystone Exams: Literature

Student Subgroup Tested	PPT	CBT	Total
Non-IEP Students (Number)			
Non-Accommodated (Number)			
Non-Accommodated (Percent)			
Accommodated (Number)			
Accommodated (Percent)			
IEP Students (Number)			
Non-Accommodated (Number)			
Non-Accommodated (Percent)			
Accommodated (Number)			
Accommodated (Percent)			

Table I–52. Incidence of IEP and EL Students Receiving Accommodations on the Summer Keystone: Algebra I

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Some test items/questions read aloud (Number)				
PPT - Some test items/questions read aloud (Percent)				
PPT - All test items/questions read aloud (Number)				
PPT - All test items/questions read aloud (Percent)				
PPT - Small group setting (Number)				
PPT - Small group setting (Percent)				
PPT - Extended time (Number)				
PPT - Extended time (Percent)				
PPT - Frequent breaks (Number)				
PPT - Frequent breaks (Percent)				
PPT - Number assessed (Number)				
CBT - Some test items/questions read aloud (Number)				
CBT - Some test items/questions read aloud (Percent)				
CBT - All test items/questions read aloud (Number)				
CBT - All test items/questions read aloud (Percent)				
CBT - Small group setting (Number)				
CBT - Small group setting (Percent)				
CBT - Extended time (Number)				
CBT - Extended time (Percent)				
CBT - Frequent breaks (Number)				
CBT - Frequent breaks (Percent)				
CBT - Number assessed (Number)				
Total - Some test items/questions read aloud (Number)				
Total - Some test items/questions read aloud (Percent)				
Total - All test items/questions read aloud (Number)				
Total - All test items/questions read aloud (Percent)				
Total - Small group setting (Number)				
Total - Small group setting (Percent)				
Total - Extended time (Number)				
Total - Extended time (Percent)				
Total - Frequent breaks (Number)				
Total - Frequent breaks (Percent)				
Total - Number assessed (Number)				

Table I–53. Incidence of IEP and EL Students Receiving Accommodations on the Summer Keystone: Biology

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Some test items/questions read aloud (Number)				
PPT - Some test items/questions read aloud (Percent)				
PPT - All test items/questions read aloud (Number)				
PPT - All test items/questions read aloud (Percent)				
PPT - Small group setting (Number)				
PPT - Small group setting (Percent)				
PPT - Extended time (Number)				
PPT - Extended time (Percent)				
PPT - Frequent breaks (Number)				
PPT - Frequent breaks (Percent)				
PPT - Number assessed (Number)				
CBT - Some test items/questions read aloud (Number)				
CBT - Some test items/questions read aloud (Percent)				
CBT - All test items/questions read aloud (Number)				
CBT - All test items/questions read aloud (Percent)				
CBT - Small group setting (Number)				
CBT - Small group setting (Percent)				
CBT - Extended time (Number)				
CBT - Extended time (Percent)				
CBT - Frequent breaks (Number)				
CBT - Frequent breaks (Percent)				
CBT - Number assessed (Number)				
Total - Some test items/questions read aloud (Number)				
Total - Some test items/questions read aloud (Percent)				
Total - All test items/questions read aloud (Number)				
Total - All test items/questions read aloud (Percent)				
Total - Small group setting (Number)				
Total - Small group setting (Percent)				
Total - Extended time (Number)				
Total - Extended time (Percent)				
Total - Frequent breaks (Number)				
Total - Frequent breaks (Percent)				
Total - Number assessed (Number)				

Table I–54. Incidence of IEP and EL Students Receiving Accommodations on the Summer Keystone: Literature

Accommodation Received by Administration Mode	Both IEP and EL	EL and non-IEP	General Education (non-IEP or EL)	IEP and non-EL
PPT - Small group setting (Number)				
PPT - Small group setting (Percent)				
PPT - Extended time (Number)				
PPT - Extended time (Percent)				
PPT - Frequent breaks (Number)				
PPT - Frequent breaks (Percent)				
PPT - Number assessed (Number)				
CBT - Small group setting (Number)				
CBT - Small group setting (Percent)				
CBT - Extended time (Number)				
CBT - Extended time (Percent)				
CBT - Frequent breaks (Number)				
CBT - Frequent breaks (Percent)				
CBT - Number assessed (Number)				
Total - Small group setting (Number)				
Total - Small group setting (Percent)				
Total - Extended time (Number)				
Total - Extended time (Percent)				
Total - Frequent breaks (Number)				
Total - Frequent breaks (Percent)				
Total - Number assessed (Number)				

APPENDIX J: ITEM STATISTICS

Appendix contains item statistics for each item type (multiple-choice, and constructed-response) by content area and administration (winter, spring, and summer).

The spring administration of the Keystone exams was cancelled due to Coronavirus (COVID-19) mitigation efforts. Due to the cancelled spring administration, test materials were not delivered and there are no test results for analysis. Consequently, tables and graphs that usually display Spring Keystone test data will not be populated within this section of the 2020 Keystone Exams Technical Report. This includes any form-level or item-level information in order to save items and/or forms for future use. Additional data analyses will be conducted in 2021 and will be included in the 2021 Keystone Exams Technical Report. Refer to the Preface for additional information.

Table J-1. Item Statistics

Column Heading	Definition
Ref	Reference line number
PubID	Item ID
Form	Test form
<i>N</i>	Number of students
PVal	<i>P</i> -Value
P()	Proportion selecting given response ('-' represents omitted responses and '**' represents multiple responses)
ITCorr	Item total correlation
Corr()	Correlation of options/points and total test score ('-' represents omitted responses and '**' represents multiple responses)
Meas	Rasch item difficulty measure estimate
SEM	Standard error of Rasch item difficulty measure estimate
z-Infit	z infit statistic
MS-Infit	Mean square infit statistic
z-Outfit	z outfit statistic
MS-Outfit	Mean square outfit statistic
M/F	Male/Female DIF code
W/B	White/Black DIF code
W/H	White/Hispanic DIF code
O/P	Online computer-based test/paper-pencil-based test DIF code

MULTIPLE-CHOICE ITEMS

Table J–2. Algebra I Multiple-Choice Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
1	818801	418786	0	13200	0.52	0.11	0.14	0.23	0.52	0.00	0.00	0.47	-0.17	-0.17	-0.28	0.47	0.22	0.02	-9.90	0.93	-9.25	0.89
2	819097	419082	0	13200	0.63	0.08	0.63	0.08	0.22	0.00	0.00	0.27	-0.24	0.27	-0.19	-0.04	0.14	0.02	9.90	1.10	9.90	1.13
3	818258	418243	0	13200	0.46	0.20	0.08	0.46	0.27	0.00	0.00	0.41	-0.04	-0.12	0.41	-0.35	0.61	0.02	0.02	1.00	0.25	1.00
4	700779	300764	0	13200	0.76	0.76	0.07	0.07	0.11	0.00	0.00	0.34	0.34	-0.18	-0.21	-0.15	-0.64	0.02	-9.90	0.87	-9.90	0.82
5	736768	336753	0	13200	0.51	0.22	0.12	0.51	0.14	0.00	0.00	0.32	-0.31	-0.11	0.32	0.02	0.31	0.02	9.90	1.08	9.90	1.13
6	984152	584137	0	13200	0.67	0.11	0.11	0.11	0.67	0.00	0.00	0.34	-0.07	-0.24	-0.20	0.34	-0.15	0.02	-4.09	0.97	-0.60	0.99
7	897482	497467	0	13200	0.60	0.07	0.20	0.12	0.60	0.00	0.00	0.44	-0.22	-0.27	-0.16	0.44	-0.23	0.02	-8.49	0.94	-6.98	0.90
8	736798	336783	0	13200	0.30	0.30	0.19	0.27	0.23	0.00	0.00	0.31	0.31	-0.11	-0.17	-0.05	1.41	0.02	9.90	1.12	9.90	1.25
9	984149	584134	0	13200	0.57	0.19	0.57	0.11	0.12	0.00	0.00	0.40	-0.19	0.40	-0.25	-0.12	0.23	0.02	-1.12	0.99	-2.10	0.98
10	818287	418272	0	13200	0.64	0.15	0.64	0.10	0.11	0.01	0.00	0.33	-0.11	0.33	-0.22	-0.15	-0.39	0.02	2.82	1.02	2.47	1.04
11	736797	336782	0	13200	0.29	0.17	0.19	0.34	0.29	0.01	0.00	0.28	-0.02	-0.14	-0.13	0.28	1.01	0.02	4.06	1.04	5.48	1.07
12	993818	593803	0	13200	0.61	0.14	0.08	0.61	0.16	0.01	0.00	0.32	-0.14	-0.22	0.32	-0.10	-0.16	0.02	6.52	1.05	6.44	1.09
13	819629	419614	0	13200	0.44	0.21	0.14	0.21	0.44	0.01	0.00	0.33	-0.05	-0.21	-0.16	0.33	0.80	0.02	9.90	1.11	9.90	1.15
14	816623	416608	0	13200	0.45	0.10	0.38	0.45	0.06	0.01	0.00	0.44	-0.14	-0.30	0.44	-0.10	0.08	0.02	-0.46	1.00	-0.66	0.99
15	896425	496410	0	13200	0.48	0.48	0.27	0.13	0.11	0.01	0.00	0.45	0.45	-0.17	-0.22	-0.21	-0.02	0.02	-0.54	1.00	-0.97	0.99
16	818261	418246	0	13200	0.47	0.47	0.27	0.12	0.12	0.01	0.00	0.49	0.49	-0.18	-0.23	-0.25	0.42	0.02	-9.90	0.90	-9.90	0.88
17	819205	419190	0	13200	0.37	0.30	0.37	0.16	0.16	0.01	0.00	0.31	0.00	0.31	-0.23	-0.15	0.50	0.02	9.90	1.08	9.55	1.11
18	800166	400151	0	13200	0.77	0.08	0.11	0.77	0.04	0.01	0.00	0.41	-0.25	-0.19	0.41	-0.19	-1.15	0.02	-8.17	0.91	-7.40	0.84
19	966691	566676	0	13200	0.56	0.11	0.56	0.22	0.11	0.00	0.00	0.40	-0.15	0.40	-0.25	-0.16	-0.03	0.02	-3.00	0.98	0.73	1.01
20	817160	417145	0	13200	0.55	0.55	0.15	0.11	0.19	0.00	0.00	0.22	0.22	-0.16	-0.16	-0.01	0.37	0.02	9.90	1.20	9.90	1.23
21	816457	416442	0	13200	0.52	0.17	0.52	0.13	0.17	0.00	0.00	0.28	-0.22	0.28	-0.09	-0.07	0.15	0.02	9.90	1.12	9.90	1.16
22	902459	502444	0	13200	0.49	0.49	0.14	0.18	0.18	0.00	0.00	0.42	0.42	-0.23	-0.16	-0.17	0.22	0.02	-3.12	0.98	-1.16	0.99
23	984290	584275	0	13200	0.39	0.23	0.26	0.12	0.39	0.00	0.00	0.36	-0.09	-0.19	-0.17	0.36	0.61	0.02	2.55	1.02	2.76	1.03
24	985794	585779	0	13200	0.57	0.16	0.21	0.57	0.05	0.00	0.00	0.31	-0.12	-0.17	0.31	-0.17	0.07	0.02	9.90	1.08	5.82	1.07
25	896219	496204	0	13200	0.61	0.61	0.07	0.07	0.25	0.00	0.00	0.53	0.53	-0.25	-0.23	-0.32	-0.18	0.02	-9.90	0.83	-9.90	0.76
26	993302	593287	0	13200	0.34	0.20	0.24	0.22	0.34	0.00	0.00	0.36	-0.08	-0.06	-0.26	0.36	0.81	0.02	-0.32	1.00	2.27	1.03
27	820457	420442	0	13200	0.50	0.25	0.12	0.50	0.12	0.00	0.00	0.39	-0.22	-0.20	0.39	-0.10	0.33	0.02	1.89	1.01	1.43	1.02
28	983367	583352	0	13200	0.64	0.13	0.13	0.10	0.64	0.00	0.00	0.24	-0.10	-0.14	-0.10	0.24	-0.16	0.02	9.90	1.10	9.90	1.16
29	979846	579831	0	13200	0.59	0.14	0.16	0.59	0.11	0.00	0.00	0.34	-0.20	-0.09	0.34	-0.19	-0.14	0.02	4.18	1.03	8.25	1.12
30	817732	417717	0	13200	0.63	0.17	0.63	0.12	0.08	0.00	0.00	0.36	-0.17	0.36	-0.16	-0.21	-0.05	0.02	-2.50	0.98	-1.80	0.98
31	702470	302455	0	13200	0.45	0.45	0.16	0.21	0.18	0.00	0.00	0.45	0.45	-0.18	-0.21	-0.17	0.25	0.02	-5.34	0.96	-4.80	0.94

Table J-2 (continued). Algebra I Multiple-Choice Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
32	985797	585782	0	13200	0.68	0.08	0.10	0.68	0.13	0.00	0.00	0.48	-0.21	-0.28	0.48	-0.23	-0.39	0.02	-9.90	0.83	-9.90	0.77
33	983837	583822	0	13200	0.36	0.18	0.28	0.18	0.36	0.00	0.00	0.39	-0.20	-0.10	-0.16	0.39	1.22	0.02	6.26	1.06	7.57	1.11
34	820456	420441	0	13200	0.44	0.19	0.44	0.13	0.24	0.00	0.00	0.24	-0.12	0.24	-0.11	-0.06	0.83	0.02	9.90	1.22	9.90	1.30
35	713850	313835	0	13200	0.74	0.06	0.09	0.74	0.12	0.00	0.00	0.30	-0.11	-0.16	0.30	-0.19	-0.70	0.02	-5.48	0.95	-1.59	0.97
36	969916	569901	0	13200	0.53	0.53	0.20	0.16	0.12	0.00	0.00	0.47	0.47	-0.20	-0.23	-0.22	0.01	0.02	-9.29	0.93	-8.85	0.89

Table J-3. Biology Multiple-Choice Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
1	735307	335292	0	12144	0.58	0.16	0.58	0.14	0.11	0.00	0.00	0.55	-0.23	0.55	-0.30	-0.24	-0.20	0.02	-9.90	0.89	-9.90	0.81
2	869844	469829	0	12144	0.63	0.10	0.63	0.14	0.13	0.00	0.00	0.43	-0.22	0.43	-0.25	-0.16	-0.08	0.02	-8.13	0.94	-6.88	0.91
3	978654	578639	0	12144	0.53	0.09	0.15	0.23	0.53	0.00	0.00	0.52	-0.23	-0.29	-0.22	0.52	0.15	0.02	-9.90	0.89	-9.90	0.85
4	965797	565782	0	12144	0.50	0.50	0.13	0.11	0.25	0.00	0.00	0.43	0.43	-0.19	-0.24	-0.18	0.45	0.02	-3.88	0.97	-4.29	0.95
5	975021	575006	0	12144	0.64	0.17	0.11	0.64	0.08	0.00	0.00	0.32	-0.14	-0.23	0.32	-0.11	0.20	0.02	6.97	1.05	5.15	1.06
6	974730	574715	0	12144	0.48	0.13	0.16	0.48	0.24	0.00	0.00	0.33	-0.17	-0.14	0.33	-0.13	0.64	0.02	9.90	1.09	9.90	1.11
7	740931	340916	0	12144	0.79	0.79	0.07	0.06	0.07	0.00	0.00	0.45	0.45	-0.23	-0.26	-0.21	-0.98	0.02	-9.90	0.82	-9.90	0.68
8	880321	480306	0	12144	0.73	0.06	0.13	0.73	0.08	0.00	0.00	0.45	-0.20	-0.25	0.45	-0.25	-0.61	0.02	-9.90	0.87	-9.90	0.77
9	678880	278865	0	12144	0.58	0.08	0.58	0.18	0.16	0.00	0.00	0.38	-0.17	0.38	-0.20	-0.16	0.29	0.02	2.23	1.02	0.37	1.00
10	681247	281232	0	12144	0.61	0.61	0.14	0.07	0.18	0.00	0.00	0.56	0.56	-0.27	-0.27	-0.28	-0.06	0.02	-9.90	0.83	-9.90	0.75
11	679673	279658	0	12144	0.62	0.62	0.10	0.17	0.11	0.00	0.00	0.51	0.51	-0.21	-0.28	-0.25	-0.17	0.02	-9.90	0.87	-9.90	0.80
12	679671	279656	0	12144	0.46	0.20	0.46	0.20	0.15	0.00	0.00	0.36	-0.15	0.36	-0.18	-0.12	0.67	0.02	5.13	1.04	4.82	1.05
13	868405	468390	0	12144	0.46	0.12	0.46	0.20	0.22	0.01	0.00	0.27	-0.24	0.27	-0.07	-0.05	1.01	0.02	9.90	1.23	9.90	1.32
14	809157	409142	0	12144	0.50	0.20	0.13	0.50	0.16	0.01	0.00	0.46	-0.24	-0.22	0.46	-0.15	0.08	0.02	-2.21	0.98	-4.00	0.95
15	868419	468404	0	12144	0.56	0.56	0.15	0.17	0.12	0.01	0.00	0.47	0.47	-0.25	-0.18	-0.21	0.39	0.02	-9.11	0.93	-9.18	0.90
16	868432	468417	0	12144	0.47	0.17	0.28	0.08	0.47	0.01	0.00	0.47	-0.13	-0.28	-0.21	0.47	0.46	0.02	-9.62	0.93	-8.77	0.91
17	880319	480304	0	12144	0.36	0.16	0.36	0.28	0.20	0.01	0.00	0.36	-0.06	0.36	-0.19	-0.14	0.95	0.02	-0.44	1.00	2.98	1.04
18	809460	409445	0	12144	0.63	0.11	0.18	0.63	0.08	0.00	0.00	0.44	-0.18	-0.29	0.44	-0.17	-0.21	0.02	-7.58	0.94	-8.98	0.88
19	868424	468409	0	12144	0.58	0.19	0.58	0.14	0.08	0.01	0.00	0.27	0.03	0.27	-0.26	-0.18	0.11	0.02	9.90	1.12	9.90	1.16
20	810034	410019	0	12144	0.58	0.16	0.14	0.58	0.12	0.01	0.00	0.44	-0.23	-0.21	0.44	-0.17	-0.30	0.02	3.26	1.03	-0.19	1.00
21	812547	412532	0	12144	0.53	0.05	0.12	0.30	0.53	0.01	0.00	0.34	-0.23	-0.14	-0.16	0.34	0.29	0.02	7.91	1.06	6.30	1.07
22	975022	575007	0	12144	0.52	0.15	0.14	0.52	0.18	0.01	0.00	0.33	-0.21	-0.11	0.33	-0.11	0.20	0.02	9.90	1.08	7.71	1.09
23	702077	302062	0	12144	0.60	0.60	0.10	0.17	0.13	0.01	0.00	0.48	0.48	-0.23	-0.24	-0.21	-0.30	0.02	-5.73	0.95	-4.07	0.94
24	974727	574712	0	12144	0.47	0.17	0.14	0.21	0.47	0.01	0.00	0.35	-0.11	-0.22	-0.13	0.35	0.45	0.02	6.70	1.05	7.66	1.09
25	978622	578607	0	12144	0.49	0.49	0.14	0.18	0.19	0.00	0.00	0.30	0.30	-0.12	-0.21	-0.07	0.94	0.02	9.90	1.21	9.90	1.30
26	742312	342297	0	12144	0.36	0.32	0.36	0.13	0.19	0.00	0.00	0.21	-0.02	0.21	-0.19	-0.07	1.17	0.02	9.90	1.18	9.90	1.35
27	981067	581052	0	12144	0.51	0.21	0.51	0.18	0.10	0.00	0.00	0.40	-0.18	0.40	-0.19	-0.18	0.21	0.02	2.49	1.02	-0.45	0.99
28	880346	480331	0	12144	0.42	0.14	0.30	0.42	0.14	0.00	0.00	0.23	-0.10	0.03	0.23	-0.26	1.00	0.02	9.90	1.20	9.90	1.31
29	965895	565880	0	12144	0.55	0.12	0.10	0.22	0.55	0.00	0.00	0.47	-0.26	-0.20	-0.21	0.47	0.38	0.02	-9.05	0.93	-8.51	0.91
30	969393	569378	0	12144	0.49	0.49	0.20	0.20	0.11	0.00	0.00	0.27	0.27	-0.13	-0.09	-0.16	0.74	0.02	9.90	1.18	9.90	1.22
31	810639	410624	0	12144	0.50	0.07	0.24	0.50	0.18	0.00	0.00	0.32	-0.16	-0.20	0.32	-0.08	0.21	0.02	9.90	1.10	9.90	1.13
32	975133	575118	0	12144	0.57	0.18	0.09	0.16	0.57	0.00	0.00	0.44	-0.13	-0.25	-0.25	0.44	0.19	0.02	-6.18	0.95	-5.47	0.94
33	741703	341688	0	12144	0.42	0.19	0.23	0.15	0.42	0.00	0.00	0.36	-0.18	-0.09	-0.18	0.36	0.94	0.02	7.47	1.07	8.15	1.10

Table J-3 (continued). Biology Multiple-Choice Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
34	678985	278970	0	12144	0.74	0.08	0.10	0.74	0.08	0.00	0.00	0.45	-0.27	-0.27	0.45	-0.16	-0.64	0.02	-9.90	0.85	-9.90	0.77
35	701417	301402	0	12144	0.51	0.51	0.08	0.20	0.21	0.00	0.00	0.35	0.35	-0.22	-0.13	-0.15	0.51	0.02	7.67	1.06	6.00	1.07
36	701416	301401	0	12144	0.57	0.14	0.22	0.07	0.57	0.00	0.00	0.39	-0.20	-0.13	-0.27	0.39	0.03	0.02	1.02	1.01	2.15	1.03
37	809199	409184	0	12144	0.67	0.08	0.05	0.67	0.20	0.00	0.00	0.22	-0.16	-0.21	0.22	-0.03	-0.48	0.02	9.90	1.15	9.90	1.38
38	808878	408863	0	12144	0.55	0.08	0.15	0.23	0.55	0.00	0.00	0.49	-0.23	-0.32	-0.15	0.49	-0.14	0.02	-4.14	0.97	-2.98	0.96
39	975121	575106	0	12144	0.45	0.15	0.25	0.45	0.15	0.00	0.00	0.34	-0.05	-0.17	0.34	-0.21	0.65	0.02	7.50	1.06	8.37	1.09
40	813414	413399	0	12144	0.62	0.16	0.62	0.13	0.08	0.00	0.00	0.41	-0.17	0.41	-0.25	-0.19	-0.25	0.02	-2.04	0.98	-3.21	0.95
41	742314	342299	0	12144	0.63	0.10	0.63	0.07	0.19	0.00	0.00	0.42	-0.20	0.42	-0.26	-0.18	-0.16	0.02	-6.83	0.95	-4.39	0.94
42	978008	577993	0	12144	0.69	0.69	0.06	0.06	0.18	0.00	0.00	0.43	0.43	-0.26	-0.28	-0.17	-0.49	0.02	-9.90	0.91	-7.04	0.89
43	975131	575116	0	12144	0.47	0.47	0.11	0.33	0.09	0.00	0.00	0.44	0.44	-0.27	-0.16	-0.21	0.29	0.02	-3.62	0.97	-4.83	0.95
44	813413	413398	0	12144	0.62	0.11	0.62	0.19	0.08	0.00	0.00	0.49	-0.21	0.49	-0.25	-0.24	0.32	0.02	-9.90	0.91	-9.90	0.88
45	702727	302712	0	12144	0.68	0.68	0.08	0.14	0.10	0.00	0.00	0.41	0.41	-0.25	-0.12	-0.26	-0.46	0.02	-7.10	0.94	3.44	1.06
46	965891	565876	0	12144	0.46	0.19	0.17	0.46	0.17	0.00	0.00	0.20	0.01	-0.20	0.20	-0.06	0.58	0.02	9.90	1.21	9.90	1.29
47	877359	477344	0	12144	0.45	0.13	0.25	0.17	0.45	0.00	0.00	0.43	-0.16	-0.18	-0.21	0.43	0.57	0.02	-4.22	0.97	-3.26	0.97
48	674108	274093	0	12144	0.60	0.15	0.12	0.60	0.14	0.00	0.00	0.39	-0.20	-0.20	0.39	-0.15	-0.35	0.02	7.63	1.07	4.85	1.08

Table J-4. Literature Multiple-Choice Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
1	740137	340122	0	11722	0.31	0.45	0.18	0.06	0.31	0.00	0.00	0.26	-0.01	-0.21	-0.15	0.26	1.84	0.02	6.52	1.06	9.90	1.26
2	740140	340125	0	11722	0.55	0.23	0.55	0.13	0.09	0.00	0.00	0.45	-0.31	0.45	-0.17	-0.12	1.00	0.02	-5.12	0.96	-1.78	0.98
3	740135	340120	0	11722	0.68	0.68	0.13	0.11	0.08	0.00	0.00	0.39	0.39	-0.06	-0.34	-0.20	0.26	0.02	-1.74	0.98	2.37	1.04
4	740142	340127	0	11722	0.93	0.03	0.93	0.02	0.02	0.00	0.00	0.37	-0.22	0.37	-0.19	-0.22	-1.88	0.04	-9.90	0.71	-9.90	0.49
5	740141	340126	0	11722	0.77	0.77	0.06	0.11	0.07	0.00	0.00	0.35	0.35	-0.18	-0.16	-0.21	-0.41	0.02	1.82	1.02	3.63	1.09
6	740144	340129	0	11722	0.71	0.07	0.16	0.06	0.71	0.00	0.00	0.47	-0.19	-0.34	-0.17	0.47	-0.10	0.02	-7.32	0.92	-6.46	0.88
7	740143	340128	0	11722	0.84	0.08	0.84	0.03	0.03	0.00	0.00	0.46	-0.31	0.46	-0.22	-0.21	-0.74	0.03	-9.90	0.76	-9.90	0.61
8	740139	340124	0	11722	0.48	0.07	0.03	0.42	0.48	0.00	0.00	0.37	-0.29	-0.19	-0.15	0.37	1.33	0.02	6.73	1.06	8.66	1.12
9	740136	340121	0	11722	0.71	0.02	0.06	0.71	0.20	0.00	0.00	0.38	-0.19	-0.21	0.38	-0.22	-0.17	0.02	4.11	1.05	6.84	1.15
10	993139	593124	0	11722	0.92	0.04	0.02	0.92	0.02	0.00	0.00	0.37	-0.20	-0.23	0.37	-0.20	-1.72	0.03	-9.63	0.79	-4.67	0.79
11	993144	593129	0	11722	0.71	0.08	0.07	0.71	0.14	0.00	0.00	0.36	-0.17	-0.24	0.36	-0.16	0.00	0.02	2.75	1.03	2.69	1.05
12	993133	593118	0	11722	0.73	0.73	0.12	0.05	0.09	0.00	0.00	0.43	0.43	-0.31	-0.17	-0.17	-0.50	0.02	6.30	1.08	2.24	1.06
13	993140	593125	0	11722	0.62	0.06	0.28	0.04	0.62	0.00	0.00	0.30	-0.22	-0.09	-0.26	0.30	0.45	0.02	9.90	1.13	9.90	1.20
14	993143	593128	0	11722	0.49	0.07	0.10	0.33	0.49	0.00	0.00	0.30	-0.17	-0.21	-0.08	0.30	1.11	0.02	9.90	1.13	9.90	1.22
15	993138	593123	0	11722	0.64	0.19	0.64	0.04	0.13	0.00	0.00	0.45	-0.24	0.45	-0.17	-0.25	0.30	0.02	-3.70	0.97	-2.83	0.96
16	993142	593127	0	11722	0.83	0.06	0.07	0.83	0.05	0.00	0.00	0.44	-0.20	-0.30	0.44	-0.20	-0.55	0.02	-9.90	0.79	-9.90	0.68
17	993135	593120	0	11722	0.48	0.34	0.11	0.48	0.07	0.00	0.00	0.34	-0.11	-0.20	0.34	-0.21	0.63	0.02	9.90	1.16	9.90	1.25
18	981155	581140	0	11722	0.85	0.07	0.04	0.85	0.03	0.00	0.00	0.46	-0.28	-0.26	0.46	-0.21	-0.96	0.03	-9.90	0.79	-9.90	0.60
19	981160	581145	0	11722	0.71	0.71	0.03	0.11	0.15	0.00	0.00	0.42	0.42	-0.25	-0.17	-0.26	0.09	0.02	-6.04	0.94	-6.06	0.89
20	981156	581141	0	11722	0.85	0.10	0.03	0.03	0.85	0.00	0.00	0.50	-0.36	-0.22	-0.22	0.50	-0.86	0.03	-9.90	0.75	-9.90	0.57
21	981153	581138	0	11722	0.70	0.70	0.10	0.02	0.18	0.00	0.00	0.26	0.26	-0.22	-0.24	-0.04	-0.02	0.02	9.90	1.16	9.90	1.42
22	981152	581137	0	11722	0.71	0.71	0.14	0.09	0.06	0.00	0.00	0.41	0.41	-0.25	-0.23	-0.13	-0.06	0.02	-2.37	0.97	-0.01	1.00
23	981151	581136	0	11722	0.69	0.13	0.69	0.12	0.06	0.00	0.00	0.37	-0.19	0.37	-0.20	-0.17	0.24	0.02	0.04	1.00	1.18	1.02
24	981148	581133	0	11722	0.67	0.10	0.67	0.16	0.08	0.00	0.00	0.40	-0.15	0.40	-0.26	-0.17	-0.05	0.02	7.42	1.08	6.12	1.12
25	981157	581142	0	11722	0.76	0.14	0.76	0.07	0.03	0.00	0.00	0.41	-0.16	0.41	-0.30	-0.23	-0.20	0.02	-7.76	0.91	-2.47	0.95
26	981149	581134	0	11722	0.55	0.19	0.23	0.55	0.02	0.00	0.00	0.43	-0.13	-0.31	0.43	-0.19	0.72	0.02	-1.19	0.99	0.31	1.00
27	703322	303307	0	11722	0.69	0.06	0.69	0.18	0.07	0.00	0.00	0.40	-0.28	0.40	-0.25	-0.09	0.17	0.02	-2.65	0.97	-3.11	0.95
28	703328	303313	0	11722	0.67	0.08	0.13	0.11	0.67	0.00	0.00	0.38	-0.20	-0.20	-0.17	0.38	0.01	0.02	5.90	1.06	7.17	1.14
29	703334	303319	0	11722	0.75	0.75	0.05	0.15	0.05	0.00	0.00	0.38	0.38	-0.29	-0.11	-0.28	-0.42	0.02	3.90	1.05	6.59	1.17
30	703327	303312	0	11722	0.73	0.08	0.12	0.73	0.07	0.00	0.00	0.49	-0.26	-0.23	0.49	-0.27	-0.15	0.02	-9.90	0.88	-9.12	0.82
31	703326	303311	0	11722	0.58	0.14	0.58	0.23	0.05	0.00	0.00	0.36	-0.09	0.36	-0.27	-0.13	0.23	0.02	9.90	1.17	9.90	1.23
32	703325	303310	0	11722	0.48	0.48	0.19	0.21	0.13	0.00	0.00	0.27	0.27	-0.17	-0.12	-0.05	0.95	0.02	9.90	1.18	9.90	1.32
33	703329	303314	0	11722	0.46	0.46	0.11	0.12	0.30	0.00	0.00	0.14	0.14	-0.16	-0.04	-0.01	1.07	0.02	9.90	1.31	9.90	1.47

Table J-4 (continued). Literature Multiple-Choice Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
34	703324	303309	0	11722	0.61	0.19	0.08	0.12	0.61	0.00	0.00	0.41	-0.09	-0.29	-0.26	0.41	0.67	0.02	-1.26	0.99	1.20	1.02

Table J-5. Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
1	700844	300829	0	88831	0.64	0.19	0.64	0.12	0.05	0.01	0.00	0.44	-0.18	0.44	-0.27	-0.22					-0.25	0.01	-9.90	0.95	-9.90	0.92
2	702476	302461	0	88831	0.39	0.14	0.22	0.25	0.39	0.01	0.00	0.38	-0.19	-0.11	-0.16	0.38					0.95	0.01	9.90	1.05	9.90	1.08
3	702569	302554	0	88831	0.59	0.06	0.31	0.59	0.03	0.00	0.00	0.45	-0.22	-0.31	0.45	-0.14					0.02	0.01	-9.90	0.94	-9.90	0.94
4	703347	303332	0	88831	0.32	0.45	0.10	0.12	0.32	0.00	0.00	0.46	-0.14	-0.27	-0.19	0.46					1.21	0.01	-9.90	0.91	-9.57	0.95
5	712560	312545	0	88831	0.49	0.17	0.49	0.11	0.23	0.00	0.00	0.39	-0.16	0.39	-0.22	-0.15					0.16	0.01	9.90	1.07	9.90	1.09
6	713788	313773	0	88831	0.56	0.26	0.56	0.08	0.10	0.00	0.00	0.33	-0.19	0.33	-0.20	-0.08					0.21	0.01	9.90	1.10	9.90	1.14
7	714012	313997	0	88831	0.55	0.21	0.55	0.10	0.13	0.01	0.00	0.29	-0.04	0.29	-0.22	-0.16					0.12	0.01	9.90	1.15	9.90	1.21
8	724180	324165	0	88831	0.65	0.65	0.18	0.09	0.08	0.00	0.00	0.50	0.50	-0.26	-0.26	-0.23					-0.51	0.01	-9.90	0.92	-9.90	0.87
9	817158	417143	0	88831	0.61	0.10	0.21	0.61	0.08	0.01	0.00	0.39	-0.19	-0.16	0.39	-0.23					-0.21	0.01	8.17	1.03	2.41	1.01
10	817164	417149	0	88831	0.33	0.33	0.19	0.33	0.15	0.00	0.00	0.41	0.41	-0.18	-0.21	-0.04					1.46	0.01	9.90	1.05	9.90	1.19
11	817706	417691	0	88831	0.62	0.10	0.20	0.62	0.07	0.01	0.00	0.50	-0.21	-0.28	0.50	-0.24					-0.39	0.01	-9.90	0.94	-9.90	0.89
12	817736	417721	0	88831	0.45	0.19	0.20	0.15	0.45	0.00	0.00	0.42	-0.10	-0.22	-0.21	0.42					0.60	0.01	4.94	1.02	5.69	1.03
13	818265	418250	0	88831	0.44	0.44	0.16	0.29	0.10	0.00	0.00	0.35	0.35	-0.22	-0.09	-0.17					0.58	0.01	9.90	1.08	9.90	1.12
14	818294	418279	0	88831	0.48	0.19	0.18	0.48	0.15	0.00	0.00	0.30	-0.20	-0.07	0.30	-0.11					0.93	0.01	9.90	1.23	9.90	1.33
15	818914	418899	0	88831	0.56	0.14	0.17	0.56	0.12	0.00	0.00	0.50	-0.25	-0.24	0.50	-0.20					-0.18	0.01	-9.90	0.96	-9.90	0.90
16	819224	419209	0	88831	0.52	0.16	0.15	0.16	0.52	0.00	0.00	0.46	-0.18	-0.20	-0.25	0.46					0.23	0.01	-9.90	0.96	-9.90	0.94
17	819638	419623	0	88831	0.50	0.07	0.21	0.50	0.22	0.00	0.00	0.36	-0.21	-0.24	0.36	-0.08					0.72	0.01	9.90	1.12	9.90	1.17
18	819876	419861	0	88831	0.75	0.06	0.10	0.75	0.08	0.00	0.00	0.46	-0.24	-0.23	0.46	-0.25					-0.53	0.01	-9.90	0.80	-9.90	0.70
19	871096	471081	0	88831	0.62	0.62	0.08	0.13	0.17	0.00	0.00	0.42	0.42	-0.21	-0.30	-0.13					0.00	0.01	-9.90	0.96	-9.90	0.95
20	892858	492843	0	88831	0.53	0.32	0.53	0.09	0.06	0.01	0.00	0.27	-0.06	0.27	-0.23	-0.17					0.25	0.01	9.90	1.17	9.90	1.24
21	895973	495958	0	88831	0.49	0.49	0.18	0.20	0.12	0.01	0.00	0.47	0.47	-0.28	-0.12	-0.23					0.18	0.01	-9.74	0.97	-2.47	0.99
22	896215	496200	0	88831	0.60	0.09	0.09	0.60	0.23	0.00	0.00	0.44	-0.24	-0.18	0.44	-0.23					-0.05	0.01	-9.90	0.96	-9.90	0.93
23	896400	496385	0	88831	0.70	0.08	0.70	0.06	0.16	0.00	0.00	0.35	-0.20	0.35	-0.24	-0.14					-0.57	0.01	-2.60	0.99	4.72	1.03
24	896402	496387	0	88831	0.75	0.13	0.75	0.08	0.04	0.00	0.00	0.45	-0.27	0.45	-0.25	-0.18					-0.65	0.01	-9.90	0.82	-9.90	0.73
25	896409	496394	0	88831	0.70	0.09	0.04	0.16	0.70	0.00	0.00	0.46	-0.18	-0.21	-0.31	0.46					-0.55	0.01	-9.90	0.87	-9.90	0.81
26	896428	496413	0	88831	0.34	0.11	0.34	0.34	0.21	0.00	0.00	0.39	-0.19	0.39	-0.05	-0.24					1.17	0.01	3.62	1.01	9.90	1.10
27	897702	497687	0	88831	0.60	0.04	0.11	0.25	0.60	0.00	0.00	0.41	-0.16	-0.23	-0.22	0.41					-0.18	0.01	0.20	1.00	2.36	1.01
28	901561	501546	0	88831	0.55	0.55	0.12	0.25	0.07	0.00	0.00	0.31	0.31	-0.21	-0.10	-0.17					0.21	0.01	9.90	1.12	9.90	1.15
29	905149	505134	0	88831	0.71	0.12	0.71	0.14	0.02	0.00	0.00	0.37	-0.21	0.37	-0.21	-0.15					-0.56	0.01	-9.90	0.95	-3.91	0.97
30	969293	569278	0	88831	0.62	0.62	0.17	0.13	0.08	0.00	0.00	0.46	0.46	-0.24	-0.21	-0.22					-0.38	0.01	-8.31	0.97	-5.43	0.97
31	969329	569314	0	88831	0.56	0.56	0.25	0.10	0.08	0.01	0.00	0.40	0.40	-0.18	-0.21	-0.20					0.02	0.01	7.55	1.02	4.36	1.02
32	969915	569900	0	88831	0.45	0.45	0.23	0.14	0.17	0.00	0.00	0.31	0.31	-0.11	-0.18	-0.12					0.65	0.01	9.90	1.14	9.90	1.18
33	975155	575140	0	88831	0.74	0.13	0.74	0.08	0.05	0.00	0.00	0.40	-0.19	0.40	-0.24	-0.20					-1.07	0.01	4.66	1.02	6.85	1.06

Table J-5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
34	980188	580173	0	88831	0.44	0.09	0.18	0.28	0.44	0.01	0.00	0.53	-0.15	-0.17	-0.33	0.53					0.24	0.01	-9.90	0.91	-9.90	0.90
35	985789	585774	0	88831	0.38	0.11	0.12	0.38	0.38	0.01	0.00	0.25	-0.12	-0.10	-0.09	0.25					1.03	0.01	9.90	1.20	9.90	1.29
36	985793	585778	0	88831	0.55	0.04	0.11	0.29	0.55	0.00	0.00	0.35	-0.20	-0.24	-0.12	0.35					0.33	0.01	9.90	1.09	9.90	1.11
37	1018132	618117	1	5679	0.43	0.43	0.38	0.11	0.08	0.00	0.00	0.48	0.48	-0.25	-0.21	-0.18	A+	A+	A+	A-	0.48	0.03	-3.59	0.95	-2.58	0.95
38	1018755	618740	1	5679	0.49	0.09	0.49	0.23	0.18	0.00	0.00	0.39	-0.21	0.39	-0.29	-0.03	A-	A-	A-	A+	0.13	0.03	3.94	1.05	4.44	1.09
39	1023243	623228	1	5679	0.51	0.26	0.12	0.51	0.10	0.00	0.00	0.36	-0.05	-0.29	0.36	-0.21	A-	A+	A-	A+	0.04	0.03	6.17	1.08	4.92	1.10
40	1024293	624278	1	5679	0.68	0.08	0.13	0.68	0.10	0.00	0.00	0.47	-0.22	-0.26	0.47	-0.21	A-	A-	A-	A+	-0.85	0.03	-9.73	0.88	-5.52	0.85
41	1024303	624288	1	5679	0.60	0.04	0.15	0.20	0.60	0.00	0.00	0.54	-0.18	-0.32	-0.28	0.54	A+	A+	A-	A-	-0.39	0.03	-9.90	0.84	-9.90	0.77
42	1029892	629877	1	5679	0.72	0.05	0.72	0.15	0.07	0.00	0.00	0.39	-0.18	0.39	-0.19	-0.25	A-	A-	A+	A+	-1.08	0.03	-3.73	0.95	-2.10	0.93
43	1034343	634328	1	5679	0.28	0.14	0.30	0.27	0.28	0.01	0.00	0.25	-0.10	-0.10	-0.07	0.25	A+	A-	A-	A+	1.29	0.03	9.90	1.21	9.90	1.37
44	799610	399595	1	5679	0.39	0.25	0.24	0.39	0.12	0.00	0.00	0.19	-0.03	-0.08	0.19	-0.13	A+	A-	A-	A+	0.67	0.03	9.90	1.30	9.90	1.44
45	896213	496198	1	5679	0.33	0.25	0.22	0.20	0.33	0.00	0.00	0.37	-0.12	-0.01	-0.28	0.37	A-	A+	A-	A-	1.01	0.03	5.06	1.08	5.69	1.14
46	983368	583353	1	5679	0.27	0.27	0.44	0.15	0.14	0.00	0.00	0.31	0.31	-0.24	-0.03	-0.01	A-	A+	A+	A-	1.40	0.03	6.39	1.12	9.28	1.29
47	1018139	618124	2	4386	0.59	0.17	0.14	0.10	0.59	0.00	0.00	0.32	-0.15	-0.09	-0.23	0.32	A+	A-	A+	A-	-0.04	0.03	6.79	1.09	5.16	1.13
48	1018761	618746	2	4386	0.45	0.23	0.16	0.16	0.45	0.00	0.00	0.49	-0.13	-0.27	-0.24	0.49	B+	A+	A+	A-	0.65	0.03	-5.59	0.93	-4.21	0.92
49	1018975	618960	2	4386	0.30	0.30	0.46	0.16	0.08	0.00	0.00	0.11	0.11	-0.02	-0.05	-0.07	A+	A-	A-	A-	1.48	0.04	9.90	1.32	9.90	1.60
50	1021499	621484	2	4386	0.34	0.29	0.32	0.34	0.05	0.00	0.00	0.23	-0.13	-0.04	0.23	-0.12	A+	A+	A+	A+	1.25	0.04	9.90	1.20	9.90	1.38
51	1023241	623226	2	4386	0.63	0.03	0.63	0.05	0.29	0.00	0.00	0.36	-0.16	0.36	-0.14	-0.25	A+	A+	A+	A+	-0.28	0.03	2.01	1.03	1.46	1.04
52	1023488	623473	2	4386	0.36	0.29	0.36	0.17	0.17	0.00	0.00	0.18	-0.06	0.18	-0.04	-0.10	A+	A-	A-	A+	1.10	0.04	9.90	1.26	9.90	1.43
53	1033829	633814	2	4386	0.72	0.17	0.72	0.06	0.04	0.00	0.00	0.28	-0.17	0.28	-0.22	-0.03	A-	A-	A+	A+	-0.78	0.04	3.95	1.07	6.12	1.23
54	895974	495959	2	4386	0.47	0.10	0.10	0.32	0.47	0.00	0.00	0.32	-0.21	-0.22	-0.06	0.32	A+	A+	A+	A+	0.54	0.03	8.06	1.11	7.67	1.16
55	982669	582654	2	4386	0.48	0.22	0.18	0.48	0.10	0.00	0.00	0.32	-0.20	-0.10	0.32	-0.13	A+	A-	A+	A+	0.48	0.03	8.02	1.11	5.98	1.12
56	983363	583348	2	4386	0.74	0.74	0.07	0.13	0.05	0.00	0.00	0.44	0.44	-0.25	-0.22	-0.23	A+	A+	A-	A-	-0.86	0.04	-6.56	0.89	-4.07	0.86
57	1013785	613770	3	4378	0.51	0.07	0.26	0.51	0.16	0.00	0.00	0.40	-0.15	-0.36	0.40	0.00	A-	A-	A-	A+	0.38	0.03	3.06	1.04	2.89	1.06
58	1018979	618964	3	4378	0.51	0.51	0.23	0.14	0.12	0.00	0.00	0.46	0.46	-0.16	-0.29	-0.18	A-	A-	A+	A-	0.38	0.03	-1.79	0.98	-1.50	0.97
59	1030406	630391	3	4378	0.68	0.15	0.06	0.68	0.11	0.00	0.00	0.44	-0.20	-0.28	0.44	-0.22	A+	A+	A+	A-	-0.51	0.04	-3.67	0.94	-2.64	0.92
60	1033242	633227	3	4378	0.66	0.13	0.66	0.08	0.13	0.00	0.00	0.47	-0.21	0.47	-0.23	-0.27	B-	A-	A-	A+	-0.41	0.04	-5.49	0.92	-4.06	0.89
61	1033826	633811	3	4378	0.10	0.31	0.37	0.21	0.10	0.00	0.00	-0.11	-0.14	0.37	-0.19	-0.11	A-	A-	A-	A-	3.10	0.05	9.21	1.37	9.90	3.25
62	1034348	634333	3	4378	0.53	0.06	0.06	0.35	0.53	0.00	0.00	0.24	-0.16	-0.24	-0.05	0.24	A-	A+	A+	A-	0.31	0.03	9.90	1.22	9.90	1.43
63	969325	569310	3	4378	0.75	0.09	0.75	0.07	0.09	0.00	0.00	0.31	-0.19	0.31	-0.18	-0.11	A+	A-	A-	A+	-0.91	0.04	2.80	1.05	4.24	1.17
64	984038	584023	3	4378	0.44	0.12	0.23	0.22	0.44	0.00	0.00	0.45	-0.23	-0.18	-0.18	0.45	A+	A-	A-	A+	0.77	0.03	-1.28	0.98	-0.13	1.00
65	984285	584270	3	4378	0.46	0.09	0.46	0.14	0.30	0.00	0.00	0.27	-0.17	0.27	-0.26	0.02	A+	A+	A+	A-	0.63	0.03	9.90	1.20	9.90	1.26
66	993297	593282	3	4378	0.68	0.68	0.13	0.13	0.06	0.00	0.00	0.46	0.46	-0.22	-0.25	-0.23	A+	A-	A-	A+	-0.50	0.04	-5.03	0.92	-3.89	0.89

Table J–5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
67	1018757	618742	4	4436	0.30	0.30	0.47	0.08	0.14	0.00	0.00	0.30	0.30	0.06	-0.25	-0.27	A-	A+	A-	A+	1.48	0.04	6.00	1.11	8.47	1.25
68	1021940	621925	4	4436	0.35	0.40	0.15	0.35	0.10	0.00	0.00	0.20	0.14	-0.27	0.20	-0.23	A-	A-	A-	A+	1.20	0.04	9.90	1.23	9.90	1.40
69	1023237	623222	4	4436	0.44	0.44	0.32	0.16	0.08	0.00	0.00	0.39	0.39	-0.30	-0.01	-0.16	A-	A-	A-	A+	0.71	0.03	2.98	1.04	2.55	1.05
70	1023247	623232	4	4436	0.37	0.37	0.26	0.20	0.16	0.00	0.00	0.23	0.23	-0.06	-0.13	-0.08	A-	A+	A-	A-	1.06	0.03	9.90	1.22	9.90	1.30
71	1024133	624118	4	4436	0.19	0.30	0.38	0.19	0.12	0.00	0.00	0.19	0.20	-0.19	0.19	-0.22	A-	A-	A+	A+	2.18	0.04	6.44	1.16	9.90	1.65
72	1029886	629871	4	4436	0.34	0.22	0.31	0.14	0.34	0.00	0.00	0.41	-0.25	-0.08	-0.16	0.41	A-	A-	A-	A+	1.25	0.04	-0.09	1.00	2.00	1.05
73	1034341	634326	4	4436	0.42	0.42	0.20	0.21	0.16	0.00	0.00	0.41	0.41	-0.14	-0.22	-0.15	A+	A-	A-	A+	0.81	0.03	0.55	1.01	1.54	1.03
74	969291	569276	4	4436	0.74	0.11	0.09	0.74	0.07	0.00	0.00	0.42	-0.23	-0.22	0.42	-0.19	A+	A+	A-	A+	-0.86	0.04	-4.80	0.92	-4.03	0.86
75	979850	579835	4	4436	0.49	0.14	0.49	0.17	0.19	0.00	0.00	0.32	-0.17	0.32	-0.26	0.00	A+	A-	A-	A+	0.43	0.03	8.39	1.11	7.79	1.16
76	984416	584401	4	4436	0.41	0.39	0.14	0.41	0.06	0.00	0.00	0.33	-0.15	-0.18	0.33	-0.10	A+	A-	A-	A-	0.89	0.03	6.20	1.09	7.28	1.16
77	1018128	618113	5	4355	0.55	0.18	0.09	0.55	0.17	0.00	0.00	0.25	-0.06	-0.17	0.25	-0.14	A+	A+	A+	A-	0.13	0.03	9.90	1.20	9.90	1.25
78	1018759	618744	5	4355	0.12	0.25	0.27	0.35	0.12	0.00	0.00	0.11	-0.09	0.09	-0.07	0.11	A+	A-	A-	A-	2.85	0.05	5.50	1.19	9.90	2.08
79	1018762	618747	5	4355	0.55	0.12	0.27	0.55	0.06	0.00	0.00	0.43	-0.28	-0.20	0.43	-0.14	A+	A+	A+	A+	0.15	0.03	-0.89	0.99	1.62	1.04
80	1022291	622276	5	4355	0.69	0.06	0.16	0.09	0.69	0.00	0.00	0.59	-0.23	-0.40	-0.25	0.59	A+	A+	A-	A+	-0.56	0.04	-9.90	0.76	-9.90	0.65
81	1023513	623498	5	4355	0.81	0.81	0.06	0.04	0.08	0.00	0.00	0.40	0.40	-0.21	-0.18	-0.25	A+	A-	A+	A+	-1.35	0.04	-4.16	0.91	-4.68	0.78
82	1024132	624117	5	4355	0.29	0.29	0.30	0.21	0.19	0.00	0.00	0.22	0.22	0.01	-0.10	-0.15	A+	A+	A+	A+	1.53	0.04	9.90	1.21	9.90	1.42
83	1033820	633805	5	4355	0.22	0.60	0.12	0.22	0.06	0.00	0.00	0.11	0.12	-0.22	0.11	-0.12	B-	A+	A+	A+	2.02	0.04	9.90	1.27	9.90	1.89
84	1035948	635933	5	4355	0.48	0.07	0.48	0.18	0.26	0.00	0.00	0.37	-0.10	0.37	-0.20	-0.19	A-	A+	A+	A+	0.51	0.03	4.96	1.07	3.82	1.08
85	977175	577160	5	4355	0.32	0.32	0.11	0.23	0.33	0.00	0.00	0.45	0.45	-0.12	-0.16	-0.22	A+	A+	A-	A-	1.37	0.04	-2.95	0.95	0.02	1.00
86	979980	579965	5	4355	0.64	0.08	0.64	0.19	0.08	0.00	0.00	0.39	-0.23	0.39	-0.17	-0.20	A+	A+	A-	A+	-0.30	0.04	1.28	1.02	-0.04	1.00
87	1014805	614790	6	4393	0.61	0.23	0.61	0.08	0.07	0.00	0.00	0.45	-0.21	0.45	-0.22	-0.27	A-	A+	A-	A+	-0.16	0.03	-3.58	0.95	-3.07	0.93
88	1014806	614791	6	4393	0.35	0.28	0.14	0.35	0.22	0.00	0.00	0.38	-0.19	-0.20	0.38	-0.06	A+	A+	A-	A+	1.20	0.04	2.43	1.04	5.99	1.15
89	1017845	617830	6	4393	0.42	0.15	0.26	0.42	0.16	0.00	0.00	0.33	-0.18	-0.15	0.33	-0.09	A+	A+	A+	A-	0.82	0.03	7.38	1.11	7.76	1.17
90	1018760	618745	6	4393	0.17	0.22	0.17	0.44	0.17	0.00	0.00	0.23	0.11	-0.07	-0.21	0.23	A-	A-	A-	A+	2.42	0.04	4.30	1.12	8.74	1.49
91	1022288	622273	6	4393	0.55	0.05	0.21	0.19	0.55	0.00	0.00	0.55	-0.17	-0.37	-0.22	0.55	A-	A-	A-	A-	0.16	0.03	-9.90	0.85	-9.87	0.80
92	1023242	623227	6	4393	0.42	0.18	0.17	0.23	0.42	0.00	0.00	0.35	-0.22	-0.11	-0.10	0.35	A+	A+	A+	A-	0.82	0.03	6.44	1.10	6.12	1.13
93	1033078	633063	6	4393	0.45	0.45	0.48	0.04	0.03	0.00	0.00	0.45	0.45	-0.29	-0.23	-0.20	A-	B-	A-	A+	0.69	0.03	-1.69	0.98	-1.17	0.98
94	1033245	633230	6	4393	0.72	0.72	0.08	0.15	0.04	0.00	0.00	0.40	0.40	-0.25	-0.17	-0.24	A+	A-	A+	A+	-0.78	0.04	-3.10	0.95	0.02	1.00
95	983876	583861	6	4393	0.59	0.04	0.19	0.18	0.59	0.00	0.00	0.29	-0.16	-0.29	0.01	0.29	A+	A-	A+	A+	-0.03	0.03	9.42	1.13	9.01	1.23
96	984409	584394	6	4393	0.43	0.04	0.43	0.28	0.25	0.00	0.00	0.48	-0.11	0.48	-0.09	-0.41	B-	A-	A-	A+	0.78	0.03	-4.16	0.94	-3.21	0.94
97	1018124	618109	7	4376	0.48	0.38	0.48	0.09	0.06	0.00	0.00	0.38	-0.17	0.38	-0.24	-0.17	A+	A+	A+	A+	0.52	0.03	3.41	1.05	3.84	1.08
98	1018970	618955	7	4376	0.27	0.27	0.26	0.29	0.17	0.01	0.00	0.44	0.44	-0.10	-0.14	-0.23	A-	A-	A+	A-	1.65	0.04	-3.57	0.93	1.95	1.06
99	1023238	623223	7	4376	0.49	0.09	0.15	0.27	0.49	0.00	0.00	0.47	-0.19	-0.28	-0.18	0.47	A-	A-	A-	A-	0.47	0.03	-3.87	0.95	-3.36	0.93

Table J–5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
100	1024139	624124	7	4376	0.35	0.12	0.35	0.28	0.25	0.00	0.00	0.17	-0.08	0.17	-0.10	-0.01	A-	A-	A+	A-	1.18	0.04	9.90	1.28	9.90	1.42
101	1033821	633806	7	4376	0.33	0.32	0.08	0.27	0.33	0.00	0.00	0.17	-0.05	-0.16	-0.03	0.17	A+	A+	A+	A-	1.31	0.04	9.90	1.28	9.90	1.48
102	1033823	633808	7	4376	0.48	0.23	0.13	0.48	0.17	0.00	0.00	0.22	-0.13	-0.19	0.22	0.02	A-	A-	A-	A+	0.53	0.03	9.90	1.23	9.90	1.32
103	1034346	634331	7	4376	0.13	0.48	0.29	0.13	0.09	0.00	0.00	0.14	-0.07	-0.10	0.14	0.11	A-	A+	A-	A-	2.73	0.05	4.91	1.16	9.90	1.91
104	819073	419058	7	4376	0.42	0.17	0.20	0.42	0.21	0.00	0.00	0.30	-0.11	-0.16	0.30	-0.11	A-	A-	A+	A+	0.82	0.03	8.93	1.13	8.34	1.18
105	974896	574881	7	4376	0.37	0.15	0.37	0.28	0.19	0.00	0.00	0.46	-0.14	0.46	-0.13	-0.29	B-	A-	A-	A-	1.06	0.04	-3.32	0.95	-0.28	0.99
106	984150	584135	7	4376	0.72	0.72	0.09	0.13	0.05	0.00	0.00	0.45	0.45	-0.19	-0.27	-0.22	A+	A+	A+	A-	-0.78	0.04	-7.03	0.89	-3.44	0.88
107	1021941	621926	8	4377	0.41	0.41	0.17	0.16	0.25	0.00	0.00	0.48	0.48	-0.16	-0.05	-0.36	A-	A+	A+	A+	0.85	0.03	-3.89	0.94	-2.70	0.94
108	1024135	624120	8	4377	0.43	0.22	0.43	0.23	0.12	0.00	0.00	0.31	-0.19	0.31	-0.02	-0.19	A+	A-	A+	A+	0.78	0.03	9.57	1.14	8.84	1.20
109	1024137	624122	8	4377	0.69	0.17	0.08	0.06	0.69	0.00	0.00	0.41	-0.21	-0.23	-0.20	0.41	A-	A-	A-	A+	-0.59	0.04	-2.37	0.96	-1.56	0.95
110	1024140	624125	8	4377	0.32	0.49	0.32	0.12	0.07	0.00	0.00	0.20	0.14	0.20	-0.32	-0.22	A-	A+	A+	A-	1.37	0.04	9.90	1.25	9.90	1.44
111	1033822	633807	8	4377	0.73	0.05	0.12	0.73	0.09	0.00	0.00	0.48	-0.24	-0.24	0.48	-0.27	A+	A+	A+	A+	-0.84	0.04	-7.99	0.87	-7.29	0.75
112	1033831	633816	8	4377	0.60	0.60	0.19	0.11	0.09	0.00	0.00	0.51	0.51	-0.28	-0.24	-0.21	A-	A-	A+	A+	-0.09	0.03	-8.04	0.89	-6.88	0.84
113	969324	569309	8	4377	0.35	0.35	0.27	0.22	0.16	0.00	0.00	0.23	0.23	-0.07	-0.11	-0.09	A-	A+	A+	A+	1.19	0.04	9.90	1.22	9.90	1.35
114	975161	575146	8	4377	0.62	0.62	0.10	0.21	0.06	0.00	0.00	0.46	0.46	-0.24	-0.25	-0.20	A+	A-	A+	A+	-0.20	0.03	-4.47	0.94	-3.87	0.90
115	982463	582448	8	4377	0.59	0.15	0.14	0.59	0.11	0.00	0.00	0.28	-0.19	-0.15	0.28	-0.05	A+	A+	A-	A+	-0.05	0.03	9.90	1.15	7.24	1.19
116	984083	584068	8	4377	0.21	0.24	0.12	0.43	0.21	0.00	0.00	0.25	-0.17	-0.14	0.03	0.25	A+	A+	A+	A+	2.07	0.04	3.54	1.08	9.90	1.63
117	1018129	618114	9	4380	0.39	0.33	0.12	0.15	0.39	0.00	0.00	0.47	-0.18	-0.17	-0.24	0.47	C-	A-	A-	A+	0.99	0.04	-2.12	0.97	-0.56	0.99
118	1018142	618127	9	4380	0.51	0.51	0.10	0.07	0.32	0.00	0.00	0.44	0.44	-0.21	-0.22	-0.21	A-	B-	A-	A-	0.36	0.03	0.57	1.01	0.26	1.01
119	1022290	622275	9	4380	0.60	0.08	0.60	0.14	0.17	0.00	0.00	0.44	-0.23	0.44	-0.25	-0.17	A-	A+	A+	A+	-0.12	0.03	-1.38	0.98	-2.16	0.95
120	1022293	622278	9	4380	0.39	0.08	0.35	0.17	0.39	0.00	0.00	0.33	-0.19	-0.04	-0.22	0.33	A-	A+	A+	A+	0.98	0.04	9.49	1.15	8.29	1.20
121	1022299	622284	9	4380	0.56	0.16	0.56	0.15	0.14	0.00	0.00	0.50	-0.25	0.50	-0.27	-0.18	A+	A+	A+	A-	0.13	0.03	-5.56	0.92	-5.12	0.89
122	1023510	623495	9	4380	0.63	0.13	0.11	0.63	0.14	0.00	0.00	0.51	-0.22	-0.31	0.51	-0.22	A-	A-	A-	A+	-0.24	0.04	-7.31	0.90	-7.15	0.82
123	1024295	624280	9	4380	0.59	0.10	0.12	0.59	0.18	0.01	0.00	0.49	-0.19	-0.28	0.49	-0.24	A+	A-	A-	A-	-0.08	0.03	-5.03	0.93	-5.32	0.87
124	1033247	633232	9	4380	0.36	0.26	0.36	0.14	0.24	0.00	0.00	0.49	-0.23	0.49	-0.04	-0.29	C-	A-	A-	A+	1.16	0.04	-4.50	0.93	-1.66	0.96
125	980171	580156	9	4380	0.45	0.18	0.17	0.19	0.45	0.00	0.00	0.44	-0.13	-0.23	-0.20	0.44	A+	A+	A-	A-	0.65	0.03	0.74	1.01	0.36	1.01
126	993817	593802	9	4380	0.44	0.44	0.23	0.16	0.17	0.00	0.00	0.25	0.25	0.01	-0.22	-0.12	A+	A+	A+	A+	0.74	0.03	9.90	1.24	9.90	1.34
127	1018126	618111	10	4364	0.46	0.08	0.30	0.15	0.46	0.00	0.00	0.40	-0.25	-0.13	-0.19	0.40	A-	A+	A-	A+	0.63	0.03	3.25	1.05	3.65	1.08
128	1018144	618129	10	4364	0.56	0.10	0.21	0.12	0.56	0.00	0.00	0.44	-0.28	-0.16	-0.20	0.44	A+	A-	A+	A+	0.12	0.03	-1.43	0.98	0.29	1.01
129	1018758	618743	10	4364	0.37	0.26	0.37	0.28	0.09	0.00	0.00	0.39	-0.12	0.39	-0.17	-0.20	A+	A+	A+	A+	1.10	0.04	2.52	1.04	5.45	1.14
130	1018978	618963	10	4364	0.49	0.49	0.11	0.28	0.12	0.00	0.00	0.48	0.48	-0.29	-0.15	-0.25	A-	A-	A-	A-	0.49	0.03	-3.33	0.95	-3.14	0.94
131	1021943	621928	10	4364	0.30	0.30	0.20	0.24	0.26	0.01	0.00	0.40	0.40	-0.11	-0.23	-0.08	A-	A-	A-	A+	1.50	0.04	0.71	1.01	4.58	1.14
132	1024298	624283	10	4364	0.80	0.80	0.07	0.09	0.04	0.00	0.00	0.47	0.47	-0.27	-0.26	-0.23	B+	A-	A+	A+	-1.28	0.04	-7.71	0.84	-7.85	0.66

Table J-5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
133	1024300	624285	10	4364	0.67	0.12	0.14	0.67	0.08	0.00	0.00	0.47	-0.20	-0.27	0.47	-0.23	A+	A+	A+	A+	-0.46	0.04	-4.91	0.93	-5.89	0.83
134	1030408	630393	10	4364	0.51	0.24	0.51	0.12	0.14	0.00	0.00	0.32	-0.11	0.32	-0.22	-0.13	A-	A-	A+	A-	0.39	0.03	9.21	1.13	7.21	1.16
135	1035808	635793	10	4364	0.67	0.06	0.05	0.21	0.67	0.00	0.00	0.45	-0.28	-0.24	-0.22	0.45	A-	A-	A+	A+	-0.45	0.04	-4.76	0.93	-2.78	0.92
136	985799	585784	10	4364	0.25	0.27	0.26	0.25	0.22	0.00	0.00	0.19	-0.04	-0.14	0.19	-0.01	A-	A-	A-	A-	1.85	0.04	9.79	1.21	9.90	1.64
137	1014788	614773	11	4389	0.31	0.31	0.44	0.20	0.05	0.00	0.00	0.25	0.25	-0.06	-0.15	-0.11	A-	A-	A-	A-	1.46	0.04	9.90	1.18	9.90	1.34
138	1014837	614822	11	4389	0.30	0.27	0.31	0.11	0.30	0.00	0.00	0.44	-0.23	-0.15	-0.09	0.44	A+	A+	A+	A+	1.49	0.04	-2.86	0.95	1.57	1.05
139	1021637	621622	11	4389	0.38	0.38	0.25	0.22	0.15	0.00	0.00	0.29	0.29	-0.16	-0.10	-0.07	A-	A+	A-	A-	1.07	0.04	9.83	1.16	9.90	1.26
140	1021639	621624	11	4389	0.34	0.11	0.27	0.27	0.34	0.00	0.00	0.31	-0.04	-0.21	-0.08	0.31	B-	B-	A-	A+	1.29	0.04	7.59	1.13	6.99	1.19
141	1023240	623225	11	4389	0.70	0.09	0.70	0.15	0.06	0.00	0.00	0.52	-0.20	0.52	-0.36	-0.23	A-	A+	A+	A+	-0.64	0.04	-9.90	0.83	-7.74	0.76
142	1030409	630394	11	4389	0.36	0.31	0.20	0.36	0.12	0.00	0.00	0.28	-0.14	-0.14	0.28	-0.05	A-	A+	A+	A-	1.16	0.04	9.51	1.16	9.90	1.27
143	1033075	633060	11	4389	0.59	0.17	0.59	0.09	0.14	0.00	0.00	0.38	-0.16	0.38	-0.22	-0.17	A-	A-	A+	A+	-0.03	0.03	3.74	1.05	0.87	1.02
144	1033251	633236	11	4389	0.62	0.05	0.23	0.10	0.62	0.00	0.00	0.49	-0.25	-0.28	-0.22	0.49	A+	A-	A+	A+	-0.16	0.03	-6.97	0.90	-5.88	0.86
145	1034345	634330	11	4389	0.61	0.16	0.11	0.61	0.12	0.00	0.00	0.50	-0.25	-0.28	0.50	-0.20	A+	A+	A+	A+	-0.13	0.03	-6.98	0.91	-6.42	0.85
146	985791	585776	11	4389	0.73	0.73	0.12	0.06	0.08	0.00	0.00	0.46	0.46	-0.24	-0.22	-0.26	A-	B-	A-	A+	-0.82	0.04	-6.83	0.88	-4.84	0.83
147	1013777	613762	12	4355	0.39	0.24	0.39	0.21	0.15	0.01	0.00	0.38	-0.08	0.38	-0.25	-0.12	A-	A+	A-	A+	0.96	0.04	3.23	1.05	5.57	1.13
148	1018135	618120	12	4355	0.60	0.60	0.25	0.07	0.07	0.00	0.00	0.51	0.51	-0.32	-0.18	-0.23	A-	A-	A-	A+	-0.14	0.03	-7.56	0.90	-7.25	0.83
149	1018140	618125	12	4355	0.74	0.74	0.07	0.16	0.03	0.00	0.00	0.51	0.51	-0.24	-0.36	-0.16	A-	B-	A-	A+	-0.92	0.04	-9.69	0.84	-8.99	0.70
150	1018522	618507	12	4355	0.30	0.08	0.45	0.16	0.30	0.00	0.00	0.32	-0.26	-0.08	-0.08	0.32	A-	A-	A-	A-	1.47	0.04	5.21	1.10	8.23	1.26
151	1023246	623231	12	4355	0.27	0.27	0.30	0.27	0.15	0.00	0.00	0.32	0.32	-0.02	-0.15	-0.17	C-	A+	A-	A-	1.63	0.04	4.09	1.08	8.69	1.31
152	1029888	629873	12	4355	0.48	0.17	0.48	0.17	0.17	0.00	0.00	0.26	-0.11	0.26	-0.12	-0.11	A+	A-	A-	A+	0.48	0.03	9.90	1.21	9.90	1.26
153	1033244	633229	12	4355	0.72	0.08	0.06	0.72	0.13	0.00	0.00	0.40	-0.23	-0.25	0.40	-0.16	A+	A-	A+	A+	-0.82	0.04	-2.72	0.95	0.10	1.00
154	1033249	633234	12	4355	0.52	0.52	0.18	0.15	0.15	0.00	0.00	0.39	0.39	-0.20	-0.08	-0.25	A+	A+	A+	A+	0.29	0.03	3.62	1.05	2.38	1.05
155	977998	577983	12	4355	0.52	0.12	0.20	0.52	0.17	0.00	0.00	0.33	-0.25	-0.18	0.33	-0.02	A-	A+	A-	A+	0.31	0.03	8.38	1.12	7.57	1.16
156	984412	584397	12	4355	0.41	0.05	0.25	0.41	0.29	0.00	0.00	0.45	-0.10	-0.21	0.45	-0.23	A+	A-	A+	A-	0.86	0.03	-0.95	0.99	0.43	1.01
157	1014795	614780	13	4378	0.64	0.12	0.15	0.64	0.08	0.00	0.00	0.45	-0.22	-0.24	0.45	-0.21	A-	A-	A+	A-	-0.32	0.04	-4.56	0.94	-1.85	0.95
158	1018134	618119	13	4378	0.63	0.63	0.24	0.08	0.04	0.00	0.00	0.50	0.50	-0.34	-0.24	-0.13	B-	A-	A-	A+	-0.29	0.04	-7.85	0.89	-6.88	0.83
159	1021936	621921	13	4378	0.25	0.23	0.36	0.25	0.16	0.00	0.00	0.11	0.00	-0.06	0.11	-0.05	A+	A+	A+	A-	1.82	0.04	9.90	1.32	9.90	1.79
160	1023239	623224	13	4378	0.43	0.13	0.12	0.32	0.43	0.00	0.00	0.47	-0.28	-0.26	-0.12	0.47	A-	A+	A-	A-	0.76	0.03	-3.44	0.95	-1.67	0.97
161	1023303	623288	13	4378	0.24	0.24	0.28	0.23	0.24	0.00	0.00	-0.02	0.05	0.01	-0.03	-0.02	A-	A-	A+	A-	1.83	0.04	9.90	1.46	9.90	2.01
162	1024138	624123	13	4378	0.58	0.18	0.58	0.11	0.12	0.00	0.00	0.41	-0.14	0.41	-0.29	-0.17	A-	A-	A+	A+	-0.02	0.03	0.68	1.01	0.37	1.01
163	1024296	624281	13	4378	0.51	0.08	0.17	0.51	0.24	0.00	0.00	0.44	-0.09	-0.25	0.44	-0.24	A-	A-	A-	A+	0.35	0.03	-0.70	0.99	-0.88	0.98
164	903106	503091	13	4378	0.83	0.83	0.05	0.06	0.05	0.00	0.00	0.38	0.38	-0.21	-0.22	-0.19	A+	A-	A-	A-	-1.56	0.04	-3.79	0.91	-3.72	0.81
165	969319	569304	13	4378	0.41	0.41	0.28	0.17	0.14	0.00	0.00	0.38	0.38	0.00	-0.28	-0.23	A-	A-	A-	A-	0.87	0.03	4.13	1.06	4.59	1.10

Table J-5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
166	983841	583826	13	4378	0.38	0.09	0.38	0.29	0.24	0.00	0.00	0.47	-0.22	0.47	-0.22	-0.15	C-	A-	A+	A+	1.02	0.04	-2.63	0.96	-1.89	0.96
167	1013782	613767	14	4375	0.54	0.12	0.19	0.15	0.54	0.00	0.00	0.50	-0.23	-0.21	-0.25	0.50	C-	A-	A-	A-	0.21	0.03	-7.39	0.91	-5.93	0.88
168	1022294	622279	14	4375	0.71	0.07	0.04	0.71	0.18	0.00	0.00	0.54	-0.19	-0.22	0.54	-0.39	A+	A-	A-	A-	-0.69	0.04	-9.90	0.82	-9.66	0.71
169	1023505	623490	14	4375	0.25	0.47	0.16	0.25	0.12	0.00	0.00	0.07	0.11	-0.18	0.07	-0.05	A-	A-	A+	A-	1.81	0.04	9.90	1.34	9.90	1.77
170	1024302	624287	14	4375	0.21	0.21	0.32	0.22	0.25	0.00	0.00	0.11	0.11	-0.06	-0.04	0.01	A+	A+	A-	A-	2.05	0.04	9.90	1.24	9.90	1.79
171	1024571	624556	14	4375	0.38	0.38	0.19	0.26	0.17	0.01	0.00	0.26	0.26	-0.14	-0.12	-0.04	A-	A+	A+	A-	1.00	0.03	9.90	1.17	9.90	1.29
172	1029880	629865	14	4375	0.57	0.57	0.21	0.14	0.08	0.00	0.00	0.39	0.39	-0.15	-0.25	-0.17	A-	A-	A-	A+	0.04	0.03	2.06	1.03	0.14	1.00
173	1033250	633235	14	4375	0.55	0.09	0.23	0.11	0.55	0.00	0.00	0.32	-0.19	-0.12	-0.15	0.32	A+	A-	A+	A-	0.14	0.03	6.94	1.09	7.22	1.16
174	1035812	635797	14	4375	0.40	0.14	0.40	0.26	0.19	0.00	0.00	0.20	-0.07	0.20	-0.06	-0.11	A+	A-	A+	A+	0.91	0.03	9.90	1.25	9.90	1.37
175	974505	574490	14	4375	0.74	0.11	0.74	0.07	0.07	0.00	0.00	0.47	-0.27	0.47	-0.26	-0.20	A-	A+	A+	A+	-0.85	0.04	-7.64	0.87	-6.06	0.80
176	993298	593283	14	4375	0.67	0.07	0.14	0.12	0.67	0.00	0.00	0.50	-0.21	-0.28	-0.25	0.50	A-	A+	A+	A+	-0.49	0.04	-8.70	0.87	-7.45	0.80
177	1013780	613765	15	4350	0.49	0.34	0.49	0.10	0.07	0.00	0.00	0.13	-0.05	0.13	-0.06	-0.08	A+	A-	A+	A-	0.48	0.03	9.90	1.32	9.90	1.49
178	1014804	614789	15	4350	0.46	0.14	0.18	0.22	0.46	0.00	0.00	0.47	-0.26	-0.19	-0.17	0.47	A+	A-	A+	A-	0.61	0.03	-4.18	0.94	-3.44	0.93
179	1014836	614821	15	4350	0.53	0.06	0.10	0.31	0.53	0.00	0.00	0.31	-0.14	-0.15	-0.16	0.31	B-	A-	A-	A-	0.27	0.03	8.25	1.11	9.89	1.22
180	1018756	618741	15	4350	0.26	0.45	0.26	0.12	0.18	0.00	0.00	0.26	0.17	0.26	-0.27	-0.28	A+	A+	A+	A+	1.75	0.04	6.62	1.14	9.90	1.39
181	1018971	618956	15	4350	0.49	0.49	0.27	0.19	0.05	0.00	0.00	0.41	0.41	-0.24	-0.17	-0.14	A-	A-	A-	A-	0.44	0.03	0.70	1.01	-0.24	1.00
182	1024142	624127	15	4350	0.32	0.28	0.23	0.32	0.17	0.00	0.00	0.32	-0.09	-0.16	0.32	-0.11	A+	A-	A-	A+	1.39	0.04	5.02	1.09	6.36	1.18
183	1029890	629875	15	4350	0.73	0.12	0.12	0.73	0.04	0.00	0.00	0.42	-0.24	-0.24	0.42	-0.16	A+	A+	A+	A+	-0.78	0.04	-3.85	0.94	-3.80	0.87
184	1033252	633237	15	4350	0.50	0.11	0.29	0.09	0.50	0.00	0.00	0.35	-0.26	-0.03	-0.26	0.35	B-	A-	A-	A+	0.41	0.03	5.70	1.08	5.23	1.11
185	984082	584067	15	4350	0.27	0.27	0.16	0.30	0.27	0.00	0.00	0.23	0.23	-0.07	-0.11	-0.05	A+	A-	A-	A-	1.65	0.04	8.34	1.17	9.90	1.45
186	985795	585780	15	4350	0.25	0.30	0.28	0.25	0.17	0.00	0.00	0.23	0.09	-0.19	0.23	-0.15	A-	A-	A-	A+	1.76	0.04	7.32	1.15	9.90	1.43
187	1013774	613759	16	4351	0.47	0.13	0.26	0.14	0.47	0.00	0.00	0.41	-0.19	-0.22	-0.11	0.41	B+	A+	A+	A-	0.57	0.03	1.40	1.02	2.65	1.05
188	1018136	618121	16	4351	0.48	0.48	0.21	0.22	0.09	0.00	0.00	0.37	0.37	-0.24	-0.10	-0.15	A+	A+	A+	A+	0.52	0.03	4.51	1.06	3.26	1.07
189	1021939	621924	16	4351	0.58	0.58	0.15	0.17	0.11	0.00	0.00	0.46	0.46	-0.26	-0.17	-0.23	A-	A-	A-	A+	0.02	0.03	-4.13	0.95	-2.54	0.94
190	1021945	621930	16	4351	0.49	0.14	0.23	0.49	0.14	0.00	0.00	0.35	-0.17	-0.24	0.35	-0.03	A-	A-	A-	A+	0.45	0.03	6.03	1.08	5.60	1.11
191	1030403	630388	16	4351	0.39	0.22	0.17	0.21	0.39	0.00	0.00	0.32	-0.12	-0.14	-0.13	0.32	B-	A-	A-	A-	0.98	0.04	7.55	1.12	6.88	1.16
192	1033818	633803	16	4351	0.44	0.14	0.44	0.32	0.09	0.00	0.00	0.31	-0.22	0.31	-0.04	-0.20	A+	A-	A-	A-	0.69	0.03	9.29	1.14	9.01	1.19
193	1033832	633817	16	4351	0.48	0.15	0.22	0.48	0.15	0.00	0.00	0.40	-0.10	-0.27	0.40	-0.15	A+	A-	A-	A+	0.50	0.03	1.85	1.03	1.55	1.03
194	1034347	634332	16	4351	0.33	0.11	0.33	0.53	0.03	0.00	0.00	0.37	-0.20	0.37	-0.19	-0.09	A+	A+	A+	A-	1.30	0.04	3.10	1.05	4.68	1.12
195	896398	496383	16	4351	0.59	0.24	0.59	0.13	0.04	0.00	0.00	0.43	-0.21	0.43	-0.25	-0.19	A-	A+	A+	A-	-0.05	0.03	-1.68	0.98	-0.64	0.99
196	979848	579833	16	4351	0.73	0.09	0.04	0.73	0.14	0.00	0.00	0.47	-0.27	-0.21	0.47	-0.26	C-	A-	A+	A+	-0.81	0.04	-7.57	0.88	-6.26	0.79
197	1018133	618118	17	4344	0.33	0.12	0.19	0.33	0.36	0.00	0.00	0.18	-0.11	-0.25	0.18	0.10	A-	A+	A-	A+	1.33	0.04	9.90	1.26	9.90	1.49
198	1021632	621617	17	4344	0.06	0.20	0.62	0.11	0.06	0.00	0.00	0.05	0.09	-0.02	-0.12	0.05	C-	A-	A-	A-	3.68	0.07	3.09	1.16	9.90	2.65

Table J–5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
199	1022298	622283	17	4344	0.31	0.37	0.09	0.23	0.31	0.00	0.00	0.24	-0.22	-0.22	0.14	0.24	A-	A-	A-	A+	1.41	0.04	9.90	1.20	9.90	1.36
200	1024299	624284	17	4344	0.69	0.06	0.14	0.11	0.69	0.00	0.00	0.44	-0.21	-0.22	-0.24	0.44	A+	A+	A+	A+	-0.56	0.04	-4.73	0.93	-2.84	0.91
201	1030401	630386	17	4344	0.51	0.51	0.12	0.23	0.14	0.00	0.00	0.44	0.44	-0.18	-0.17	-0.26	A+	A+	A+	A-	0.36	0.03	-0.69	0.99	-0.80	0.98
202	1033819	633804	17	4344	0.51	0.51	0.17	0.22	0.10	0.00	0.00	0.46	0.46	-0.22	-0.21	-0.21	A-	A+	A-	A+	0.36	0.03	-3.26	0.96	-2.77	0.94
203	1034342	634327	17	4344	0.36	0.21	0.36	0.14	0.28	0.00	0.00	0.35	-0.19	0.35	-0.12	-0.11	A+	A-	A-	A-	1.12	0.04	5.06	1.08	6.45	1.16
204	966687	566672	17	4344	0.44	0.13	0.25	0.44	0.17	0.00	0.00	0.25	-0.24	-0.04	0.25	-0.06	A-	A-	A-	A+	0.70	0.03	9.90	1.21	9.90	1.34
205	979842	579827	17	4344	0.57	0.09	0.57	0.25	0.10	0.00	0.00	0.43	-0.18	0.43	-0.23	-0.19	A-	A-	A+	A+	0.07	0.03	-0.77	0.99	-1.38	0.97
206	987141	587126	17	4344	0.58	0.58	0.20	0.13	0.09	0.00	0.00	0.47	0.47	-0.17	-0.33	-0.18	A-	A-	A-	B-	0.00	0.03	-5.18	0.93	-4.04	0.91
207	1015121	615106	18	4405	0.32	0.26	0.19	0.32	0.23	0.00	0.00	0.12	-0.05	-0.13	0.12	0.04	A+	A+	A+	A-	1.42	0.04	9.90	1.33	9.90	1.59
208	1018754	618739	18	4405	0.54	0.09	0.20	0.54	0.18	0.00	0.00	0.29	-0.09	-0.13	0.29	-0.18	A+	A+	A+	A+	0.24	0.03	9.90	1.15	8.50	1.20
209	1033074	633059	18	4405	0.43	0.43	0.08	0.36	0.12	0.00	0.00	0.46	0.46	-0.23	-0.18	-0.25	A-	A+	A-	A-	0.77	0.03	-2.89	0.96	-1.68	0.97
210	1033076	633061	18	4405	0.39	0.12	0.20	0.39	0.29	0.00	0.00	0.23	-0.05	-0.22	0.23	-0.02	A+	A+	A+	A+	1.01	0.03	9.90	1.22	9.90	1.36
211	1033827	633812	18	4405	0.71	0.06	0.09	0.14	0.71	0.00	0.00	0.52	-0.25	-0.27	-0.28	0.52	A+	A+	A+	A-	-0.68	0.04	-9.90	0.83	-8.79	0.73
212	1034383	634368	18	4405	0.64	0.17	0.64	0.10	0.09	0.00	0.00	0.41	-0.24	0.41	-0.10	-0.26	B-	A-	A-	A+	-0.27	0.03	-2.12	0.97	0.91	1.02
213	1035815	635800	18	4405	0.37	0.26	0.15	0.21	0.37	0.00	0.00	0.35	-0.11	-0.14	-0.17	0.35	A-	A+	A+	A+	1.09	0.04	5.86	1.09	6.23	1.15
214	1035949	635934	18	4405	0.29	0.29	0.22	0.17	0.32	0.00	0.00	0.35	0.35	-0.10	-0.17	-0.11	A-	A+	A-	A-	1.60	0.04	2.73	1.05	5.72	1.18
215	1035950	635935	18	4405	0.36	0.28	0.28	0.36	0.08	0.00	0.00	0.41	-0.22	-0.17	0.41	-0.09	A+	A-	A+	A+	1.15	0.04	0.59	1.01	2.24	1.05
216	800167	400152	18	4405	0.52	0.23	0.52	0.17	0.07	0.00	0.00	0.41	-0.16	0.41	-0.26	-0.16	A-	A-	A-	A+	0.33	0.03	0.77	1.01	0.85	1.02
217	1013784	613769	19	4404	0.43	0.14	0.18	0.24	0.43	0.00	0.00	0.32	-0.18	-0.16	-0.07	0.32	A+	A+	A+	A+	0.75	0.03	7.30	1.10	6.82	1.14
218	1014800	614785	19	4404	0.49	0.08	0.25	0.19	0.49	0.00	0.00	0.26	-0.14	0.00	-0.24	0.26	A+	A-	A+	A+	0.48	0.03	9.90	1.16	9.86	1.20
219	1015122	615107	19	4404	0.18	0.18	0.61	0.16	0.05	0.00	0.00	0.21	0.21	0.03	-0.18	-0.13	A-	A-	A-	A-	2.29	0.04	4.09	1.11	9.90	1.57
220	1018753	618738	19	4404	0.44	0.29	0.10	0.16	0.44	0.00	0.00	0.30	-0.12	-0.11	-0.17	0.30	A-	A-	A-	A+	0.71	0.03	8.50	1.12	7.95	1.16
221	1023236	623221	19	4404	0.55	0.09	0.55	0.18	0.18	0.00	0.00	0.31	-0.14	0.31	-0.13	-0.17	A-	A-	A+	A-	0.18	0.03	7.37	1.10	6.31	1.13
222	1023248	623233	19	4404	0.37	0.40	0.37	0.13	0.09	0.00	0.00	0.16	0.04	0.16	-0.17	-0.13	A-	A+	A+	A+	1.08	0.03	9.90	1.27	9.90	1.40
223	1030404	630389	19	4404	0.16	0.11	0.16	0.32	0.41	0.01	0.00	-0.07	-0.14	-0.07	0.11	0.04	A-	A-	A-	A+	2.44	0.04	9.90	1.42	9.90	2.20
224	1034349	634334	19	4404	0.48	0.48	0.21	0.17	0.14	0.00	0.00	0.32	0.32	-0.15	-0.17	-0.10	A-	A+	A+	A+	0.49	0.03	7.58	1.10	5.67	1.11
225	983361	583346	19	4404	0.69	0.12	0.09	0.69	0.10	0.00	0.00	0.47	-0.22	-0.25	0.47	-0.24	B+	A-	A+	A+	-0.56	0.04	-7.03	0.90	-7.11	0.80
226	983833	583818	19	4404	0.66	0.05	0.07	0.66	0.22	0.00	0.00	0.39	-0.23	-0.25	0.39	-0.17	A+	A+	A-	A+	-0.42	0.04	-1.32	0.98	-0.14	1.00
227	1018131	618116	20	4356	0.68	0.11	0.08	0.68	0.13	0.00	0.00	0.38	-0.17	-0.20	0.38	-0.20	A+	A-	A+	A+	-0.50	0.04	0.20	1.00	-1.01	0.97
228	1018141	618126	20	4356	0.31	0.39	0.31	0.14	0.16	0.00	0.00	0.14	0.09	0.14	-0.15	-0.15	A+	A-	A-	A-	1.43	0.04	9.90	1.28	9.90	1.55
229	1019204	619189	20	4356	0.46	0.46	0.17	0.22	0.15	0.00	0.00	0.38	0.38	-0.18	-0.14	-0.17	B+	A+	A-	A-	0.65	0.03	3.15	1.04	3.24	1.07
230	1021638	621623	20	4356	0.29	0.23	0.29	0.17	0.30	0.00	0.00	0.21	-0.08	0.21	-0.10	-0.04	A+	A+	A+	A+	1.55	0.04	9.90	1.20	9.90	1.45
231	1024136	624121	20	4356	0.50	0.50	0.14	0.23	0.13	0.00	0.00	0.46	0.46	-0.27	-0.11	-0.25	A+	A+	A-	A-	0.41	0.03	-3.07	0.96	-2.59	0.95

Table J-5 (continued). Algebra I Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
232	1029889	629874	20	4356	0.57	0.17	0.57	0.10	0.15	0.00	0.00	0.33	-0.20	0.33	-0.14	-0.11	A+	A-	A-	A+	0.05	0.03	6.80	1.09	5.03	1.12
233	1033071	633056	20	4356	0.38	0.14	0.12	0.35	0.38	0.00	0.00	0.31	-0.19	-0.17	-0.05	0.31	A-	A+	A+	A+	1.02	0.03	7.93	1.12	8.76	1.21
234	905143	505128	20	4356	0.62	0.14	0.11	0.62	0.13	0.00	0.00	0.39	-0.21	-0.19	0.39	-0.16	A-	A-	A-	A+	-0.17	0.03	0.76	1.01	0.95	1.02
235	974507	574492	20	4356	0.69	0.07	0.12	0.12	0.69	0.00	0.00	0.51	-0.26	-0.28	-0.24	0.51	A+	A+	A+	A+	-0.56	0.04	-9.90	0.84	-8.16	0.77
236	983365	583350	20	4356	0.20	0.30	0.26	0.25	0.20	0.00	0.00	0.18	-0.02	0.01	-0.15	0.18	A-	A-	A-	A-	2.20	0.04	6.54	1.17	9.90	1.67

Table J-6. Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
1	677863	277848	0	83785	0.43	0.20	0.16	0.22	0.43	0.00	0.00	0.37	-0.12	-0.14	-0.20	0.37					0.84	0.01	8.58	1.03	8.60	1.04
2	678874	278859	0	83785	0.69	0.11	0.69	0.08	0.12	0.00	0.00	0.47	-0.26	0.47	-0.26	-0.20					-0.49	0.01	-9.90	0.88	-9.90	0.79
3	678945	278930	0	83785	0.45	0.45	0.22	0.18	0.14	0.01	0.00	0.50	0.50	-0.16	-0.25	-0.23					0.89	0.01	-9.90	0.92	-9.90	0.92
4	678977	278962	0	83785	0.53	0.31	0.53	0.07	0.10	0.00	0.00	0.37	-0.13	0.37	-0.21	-0.23					0.34	0.01	8.18	1.02	6.47	1.03
5	702204	302189	0	83785	0.47	0.14	0.47	0.23	0.15	0.01	0.00	0.49	-0.15	0.49	-0.24	-0.24					0.18	0.01	-9.90	0.95	-9.90	0.92
6	702726	302711	0	83785	0.59	0.09	0.59	0.09	0.22	0.00	0.00	0.41	-0.21	0.41	-0.25	-0.16					0.16	0.01	-9.90	0.96	-9.90	0.94
7	702735	302720	0	83785	0.61	0.06	0.61	0.16	0.16	0.00	0.00	0.27	-0.18	0.27	-0.09	-0.14					0.08	0.01	9.90	1.09	9.90	1.15
8	703244	303229	0	83785	0.72	0.08	0.06	0.72	0.14	0.00	0.00	0.38	-0.24	-0.24	0.38	-0.14					-0.68	0.01	-9.90	0.95	-1.88	0.99
9	703495	303480	0	83785	0.72	0.09	0.09	0.09	0.72	0.01	0.00	0.43	-0.26	-0.25	-0.15	0.43					-0.97	0.01	8.81	1.04	5.94	1.05
10	737663	337648	0	83785	0.55	0.55	0.16	0.14	0.15	0.00	0.00	0.45	0.45	-0.25	-0.20	-0.17					0.22	0.01	-9.90	0.94	-9.90	0.92
11	737667	337652	0	83785	0.67	0.04	0.08	0.20	0.67	0.00	0.00	0.44	-0.21	-0.25	-0.24	0.44					0.04	0.01	-9.90	0.88	-9.90	0.85
12	808347	408332	0	83785	0.45	0.45	0.08	0.26	0.21	0.00	0.00	0.36	0.36	-0.15	-0.25	-0.07					0.53	0.01	9.90	1.03	8.02	1.03
13	808546	408531	0	83785	0.58	0.14	0.20	0.58	0.07	0.00	0.00	0.34	-0.18	-0.14	0.34	-0.17					-0.24	0.01	9.90	1.11	9.90	1.14
14	809059	409044	0	83785	0.66	0.08	0.66	0.09	0.17	0.00	0.00	0.51	-0.24	0.51	-0.30	-0.23					-0.52	0.01	-9.90	0.93	-9.90	0.84
15	809463	409448	0	83785	0.28	0.17	0.33	0.28	0.22	0.00	0.00	0.19	-0.12	-0.02	0.19	-0.07					1.59	0.01	9.90	1.15	9.90	1.36
16	809873	409858	0	83785	0.71	0.12	0.08	0.71	0.09	0.00	0.00	0.41	-0.14	-0.24	0.41	-0.26					-0.77	0.01	-2.74	0.99	0.33	1.00
17	810617	410602	0	83785	0.40	0.18	0.30	0.40	0.12	0.01	0.00	0.38	-0.20	-0.14	0.38	-0.12					1.32	0.01	9.90	1.10	9.90	1.20
18	813704	413689	0	83785	0.59	0.59	0.19	0.13	0.08	0.00	0.00	0.33	0.33	-0.15	-0.17	-0.17					-0.09	0.01	9.90	1.06	9.90	1.09
19	813705	413690	0	83785	0.40	0.23	0.18	0.19	0.40	0.00	0.00	0.40	-0.07	-0.22	-0.21	0.40					0.54	0.01	-4.29	0.99	-3.95	0.98
20	819086	419071	0	83785	0.61	0.09	0.15	0.15	0.61	0.00	0.00	0.48	-0.21	-0.14	-0.34	0.48					-0.33	0.01	-9.90	0.96	-9.90	0.92
21	868430	468415	0	83785	0.27	0.27	0.23	0.31	0.19	0.00	0.00	0.38	0.38	-0.16	-0.13	-0.09					1.59	0.01	-9.90	0.95	6.77	1.04
22	869043	469028	0	83785	0.42	0.42	0.30	0.13	0.15	0.00	0.00	0.32	0.32	-0.04	-0.20	-0.19					0.65	0.01	9.90	1.07	9.90	1.09
23	869826	469811	0	83785	0.49	0.18	0.49	0.21	0.12	0.00	0.00	0.43	-0.22	0.43	-0.23	-0.10					0.37	0.01	-9.90	0.97	-9.90	0.95
24	877357	477342	0	83785	0.56	0.18	0.56	0.11	0.15	0.00	0.00	0.35	-0.14	0.35	-0.19	-0.15					0.30	0.01	9.90	1.04	9.90	1.04
25	880289	480274	0	83785	0.43	0.22	0.11	0.43	0.24	0.00	0.00	0.31	-0.06	-0.21	0.31	-0.15					0.94	0.01	9.90	1.11	9.90	1.15
26	880293	480278	0	83785	0.64	0.07	0.64	0.20	0.08	0.00	0.00	0.54	-0.22	0.54	-0.31	-0.27					-0.22	0.01	-9.90	0.83	-9.90	0.75
27	880299	480284	0	83785	0.64	0.20	0.64	0.10	0.05	0.00	0.00	0.44	-0.26	0.44	-0.20	-0.20					-0.40	0.01	-9.24	0.97	-9.03	0.95
28	880325	480310	0	83785	0.60	0.13	0.60	0.13	0.13	0.00	0.00	0.47	-0.21	0.47	-0.23	-0.23					0.06	0.01	-9.90	0.90	-9.90	0.87
29	880334	480319	0	83785	0.58	0.08	0.10	0.58	0.23	0.00	0.00	0.46	-0.21	-0.24	0.46	-0.22					-0.26	0.01	-1.14	1.00	-7.47	0.96
30	880337	480322	0	83785	0.70	0.70	0.08	0.07	0.14	0.00	0.00	0.43	0.43	-0.23	-0.25	-0.20					-0.72	0.01	-5.73	0.98	-9.90	0.90
31	892744	492729	0	83785	0.47	0.14	0.47	0.22	0.16	0.00	0.00	0.38	-0.12	0.38	-0.18	-0.19					0.56	0.01	3.97	1.01	9.90	1.04
32	896412	496397	0	83785	0.49	0.14	0.22	0.49	0.15	0.00	0.00	0.30	-0.20	-0.14	0.30	-0.05					0.97	0.01	9.90	1.19	9.90	1.26
33	966516	566501	0	83785	0.60	0.08	0.07	0.25	0.60	0.00	0.00	0.42	-0.22	-0.18	-0.23	0.42					0.07	0.01	-9.90	0.95	-9.90	0.93

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
34	966709	566694	0	83785	0.64	0.21	0.09	0.64	0.07	0.00	0.00	0.40	-0.19	-0.21	0.40	-0.21					0.01	0.01	-9.90	0.95	-9.90	0.91
35	969386	569371	0	83785	0.56	0.19	0.17	0.09	0.56	0.00	0.00	0.50	-0.15	-0.29	-0.28	0.50					0.21	0.01	-9.90	0.89	-9.90	0.86
36	969400	569385	0	83785	0.50	0.11	0.50	0.25	0.15	0.00	0.00	0.35	-0.20	0.35	-0.11	-0.18					0.71	0.01	9.90	1.07	9.90	1.08
37	974719	574704	0	83785	0.56	0.56	0.11	0.14	0.18	0.00	0.00	0.40	0.40	-0.13	-0.16	-0.24					0.04	0.01	1.57	1.00	2.39	1.01
38	974722	574707	0	83785	0.67	0.10	0.12	0.67	0.11	0.00	0.00	0.44	-0.25	-0.16	0.44	-0.26					-0.06	0.01	-9.90	0.88	-9.90	0.83
39	974733	574718	0	83785	0.38	0.38	0.15	0.37	0.10	0.00	0.00	0.41	0.41	-0.21	-0.16	-0.16					0.90	0.01	-9.90	0.95	-7.50	0.97
40	974791	574776	0	83785	0.58	0.58	0.14	0.20	0.07	0.00	0.00	0.39	0.39	-0.18	-0.20	-0.18					-0.03	0.01	3.65	1.01	-2.74	0.99
41	975023	575008	0	83785	0.53	0.53	0.27	0.14	0.05	0.00	0.00	0.47	0.47	-0.17	-0.31	-0.20					0.31	0.01	-9.90	0.92	-9.90	0.89
42	975138	575123	0	83785	0.51	0.12	0.30	0.51	0.06	0.00	0.00	0.23	0.00	-0.14	0.23	-0.20					0.47	0.01	9.90	1.17	9.90	1.24
43	977601	577586	0	83785	0.35	0.35	0.25	0.22	0.18	0.00	0.00	0.23	0.23	0.00	-0.17	-0.10					1.42	0.01	9.90	1.21	9.90	1.35
44	978623	578608	0	83785	0.71	0.08	0.03	0.18	0.71	0.00	0.00	0.49	-0.23	-0.19	-0.33	0.49					-0.65	0.01	-9.90	0.87	-9.90	0.76
45	978650	578635	0	83785	0.45	0.24	0.16	0.45	0.14	0.00	0.00	0.25	-0.10	-0.12	0.25	-0.10					0.82	0.01	9.90	1.16	9.90	1.23
46	978652	578637	0	83785	0.59	0.16	0.14	0.12	0.59	0.00	0.00	0.39	-0.21	-0.19	-0.14	0.39					0.11	0.01	-5.26	0.99	-5.25	0.98
47	981069	581054	0	83785	0.51	0.14	0.19	0.16	0.51	0.00	0.00	0.39	-0.07	-0.22	-0.23	0.39					0.53	0.01	2.00	1.01	-0.49	1.00
48	981424	581409	0	83785	0.55	0.08	0.55	0.25	0.12	0.00	0.00	0.33	-0.22	0.33	-0.14	-0.13					0.35	0.01	9.90	1.06	9.69	1.04
49	1018109	618094	1	5434	0.48	0.48	0.15	0.28	0.10	0.00	0.00	0.53	0.53	-0.27	-0.23	-0.22	A-	A+	A-	A-	0.32	0.03	-9.90	0.87	-9.90	0.84
50	1020500	620485	1	5434	0.46	0.46	0.13	0.24	0.17	0.00	0.00	0.49	0.49	-0.23	-0.22	-0.20	A-	A-	A-	A-	0.39	0.03	-7.86	0.91	-6.86	0.90
51	1024198	624183	1	5434	0.34	0.34	0.14	0.17	0.34	0.00	0.00	0.37	0.37	-0.24	-0.23	-0.01	A-	A+	A-	A-	1.04	0.03	1.41	1.02	3.48	1.07
52	1024202	624187	1	5434	0.13	0.10	0.15	0.61	0.13	0.00	0.00	-0.07	-0.26	-0.20	0.36	-0.07	A+	A+	A-	A-	2.55	0.04	9.90	1.32	9.90	2.36
53	1024384	624369	1	5434	0.46	0.17	0.18	0.19	0.46	0.00	0.00	0.34	-0.15	-0.26	-0.03	0.34	A+	A+	A+	A-	0.42	0.03	5.18	1.06	5.42	1.09
54	1030597	630582	1	5434	0.49	0.20	0.49	0.15	0.16	0.00	0.00	0.40	-0.15	0.40	-0.22	-0.16	A+	A-	A+	A+	0.25	0.03	0.19	1.00	-1.01	0.98
55	1030629	630614	1	5434	0.53	0.24	0.11	0.53	0.11	0.00	0.00	0.40	-0.17	-0.27	0.40	-0.12	A+	A-	A-	A+	0.04	0.03	-0.82	0.99	-0.31	0.99
56	1030634	630619	1	5434	0.45	0.07	0.45	0.25	0.23	0.00	0.00	0.36	-0.22	0.36	-0.21	-0.08	A-	A-	A+	A+	0.44	0.03	4.13	1.05	3.77	1.06
57	1030641	630626	1	5434	0.31	0.17	0.31	0.32	0.21	0.00	0.00	0.31	-0.16	0.31	-0.16	-0.02	A-	A-	A-	A+	1.21	0.03	4.45	1.07	7.80	1.18
58	1033602	633587	1	5434	0.60	0.11	0.60	0.10	0.19	0.00	0.00	0.39	-0.17	0.39	-0.28	-0.12	A+	A+	A-	A+	-0.28	0.03	-2.26	0.98	0.48	1.01
59	1034422	634407	1	5434	0.45	0.07	0.12	0.45	0.36	0.00	0.00	0.21	-0.20	-0.22	0.21	0.04	A-	A-	A-	A+	0.45	0.03	9.90	1.20	9.90	1.24
60	1034424	634409	1	5434	0.31	0.16	0.44	0.31	0.08	0.00	0.00	0.21	-0.24	0.10	0.21	-0.21	A+	A-	A-	A+	1.17	0.03	9.90	1.19	9.90	1.31
61	1034425	634410	1	5434	0.30	0.15	0.43	0.30	0.12	0.00	0.00	0.27	-0.08	-0.01	0.27	-0.28	A-	A-	A-	A-	1.26	0.03	6.99	1.12	9.09	1.22
62	1035475	635460	1	5434	0.33	0.10	0.30	0.25	0.33	0.00	0.00	0.29	-0.20	0.04	-0.22	0.29	A-	A-	A-	A-	1.05	0.03	6.93	1.10	7.91	1.17
63	1035665	635650	1	5434	0.35	0.16	0.35	0.12	0.36	0.00	0.00	0.30	-0.25	0.30	-0.19	0.04	A-	A+	A+	A+	0.94	0.03	7.03	1.10	9.30	1.19
64	1035666	635651	1	5434	0.58	0.12	0.12	0.17	0.58	0.00	0.00	0.47	-0.23	-0.27	-0.17	0.47	A-	A+	A+	A-	-0.17	0.03	-9.20	0.90	-6.17	0.89
65	1018347	618332	2	4124	0.47	0.15	0.27	0.47	0.10	0.00	0.00	0.41	-0.13	-0.18	0.41	-0.24	A+	A+	A+	A-	0.61	0.03	-1.72	0.98	-1.30	0.98
66	1021871	621856	2	4124	0.35	0.41	0.09	0.35	0.14	0.00	0.00	0.34	-0.19	-0.21	0.34	-0.02	B-	A-	A-	A-	1.22	0.04	1.99	1.03	4.60	1.10

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
67	1021873	621858	2	4124	0.47	0.47	0.14	0.22	0.16	0.00	0.00	0.34	0.34	-0.24	-0.09	-0.13	A+	A-	A-	A+	0.62	0.03	4.11	1.05	3.79	1.07
68	1023641	623626	2	4124	0.64	0.19	0.64	0.11	0.06	0.00	0.00	0.39	-0.15	0.39	-0.23	-0.24	A+	A-	A-	A+	-0.20	0.04	-1.81	0.98	-2.25	0.95
69	1024192	624177	2	4124	0.63	0.13	0.63	0.10	0.14	0.00	0.00	0.45	-0.19	0.45	-0.27	-0.20	A-	A+	A+	A-	-0.13	0.04	-5.22	0.93	-5.63	0.88
70	1024201	624186	2	4124	0.38	0.33	0.21	0.09	0.38	0.00	0.00	0.27	0.04	-0.20	-0.23	0.27	A+	A+	A+	A-	1.09	0.04	6.99	1.11	7.41	1.16
71	1024376	624361	2	4124	0.71	0.08	0.71	0.11	0.10	0.00	0.00	0.47	-0.27	0.47	-0.25	-0.20	A+	A+	A-	A+	-0.59	0.04	-7.15	0.89	-6.87	0.81
72	1025843	625828	2	4124	0.30	0.06	0.22	0.42	0.30	0.00	0.00	0.01	-0.19	-0.08	0.15	0.01	A+	A-	A-	A-	1.52	0.04	9.90	1.35	9.90	1.59
73	1025848	625833	2	4124	0.77	0.77	0.13	0.03	0.07	0.00	0.00	0.25	0.25	-0.18	-0.17	-0.07	A-	A-	A-	A+	-0.97	0.04	1.57	1.03	4.55	1.18
74	1030649	630634	2	4124	0.44	0.44	0.22	0.19	0.15	0.00	0.00	0.24	0.24	-0.13	0.08	-0.26	A-	A-	A+	A+	0.78	0.03	9.90	1.16	9.90	1.21
75	1032285	632270	2	4124	0.51	0.16	0.12	0.20	0.51	0.00	0.00	0.43	-0.13	-0.15	-0.29	0.43	A-	A+	A+	A-	0.41	0.03	-3.27	0.96	-3.03	0.95
76	1034416	634401	2	4124	0.47	0.06	0.34	0.47	0.13	0.00	0.00	0.41	-0.19	-0.26	0.41	-0.10	A-	A-	A-	A-	0.62	0.03	-1.48	0.98	-1.05	0.98
77	1034418	634403	2	4124	0.59	0.59	0.16	0.04	0.21	0.00	0.00	0.38	0.38	-0.23	-0.14	-0.18	A+	A+	A+	A-	0.07	0.03	0.47	1.01	-1.25	0.98
78	1034419	634404	2	4124	0.62	0.02	0.10	0.62	0.27	0.00	0.00	0.39	-0.13	-0.16	0.39	-0.29	A+	A-	A-	A+	-0.09	0.03	-0.98	0.99	-2.00	0.96
79	1035661	635646	2	4124	0.66	0.18	0.09	0.66	0.07	0.00	0.00	0.39	-0.15	-0.21	0.39	-0.27	A-	A-	A-	A-	-0.33	0.04	-2.64	0.96	-0.11	1.00
80	1035664	635649	2	4124	0.51	0.06	0.27	0.16	0.51	0.00	0.00	0.19	-0.24	0.02	-0.12	0.19	A+	A-	A+	A-	0.44	0.03	9.90	1.20	9.90	1.29
81	1020420	620405	3	4093	0.60	0.12	0.24	0.60	0.05	0.00	0.00	0.41	-0.22	-0.22	0.41	-0.18	B+	A+	A-	A+	0.02	0.04	-1.52	0.98	-2.17	0.95
82	1021587	621572	3	4093	0.51	0.06	0.11	0.32	0.51	0.00	0.00	0.41	-0.24	-0.28	-0.14	0.41	A-	A-	A-	A+	0.46	0.03	-0.81	0.99	-2.10	0.96
83	1021589	621574	3	4093	0.54	0.54	0.10	0.24	0.12	0.00	0.00	0.37	0.37	-0.19	-0.21	-0.12	A-	A-	A+	A-	0.29	0.03	2.07	1.03	0.65	1.01
84	1022989	622974	3	4093	0.63	0.08	0.21	0.08	0.63	0.00	0.00	0.48	-0.21	-0.24	-0.29	0.48	A+	A-	A+	A-	-0.13	0.04	-7.14	0.91	-6.73	0.84
85	1023025	623010	3	4093	0.52	0.52	0.14	0.23	0.12	0.00	0.00	0.24	0.24	-0.10	-0.05	-0.20	A-	A-	A-	A+	0.41	0.03	9.90	1.16	9.90	1.25
86	1023310	623295	3	4093	0.35	0.13	0.45	0.35	0.07	0.00	0.00	0.28	-0.22	-0.04	0.28	-0.16	A+	A+	A+	A+	1.28	0.04	6.52	1.11	8.41	1.21
87	1024175	624160	3	4093	0.76	0.76	0.06	0.11	0.07	0.00	0.00	0.52	0.52	-0.25	-0.30	-0.26	A+	A+	A-	A-	-0.89	0.04	-9.90	0.82	-9.35	0.68
88	1026344	626329	3	4093	0.56	0.14	0.56	0.19	0.11	0.00	0.00	0.22	-0.18	0.22	-0.10	-0.02	A+	A+	A-	A-	0.18	0.03	9.90	1.17	9.90	1.39
89	1030589	630574	3	4093	0.46	0.14	0.15	0.25	0.46	0.00	0.00	0.39	-0.14	-0.17	-0.19	0.39	A-	A-	A+	A-	0.71	0.03	1.07	1.01	1.32	1.02
90	1030650	630635	3	4093	0.40	0.22	0.40	0.23	0.15	0.00	0.00	0.23	-0.14	0.23	-0.13	0.00	A+	A-	A-	A+	0.99	0.04	9.90	1.17	9.90	1.24
91	1032287	632272	3	4093	0.55	0.14	0.18	0.55	0.13	0.00	0.00	0.40	-0.14	-0.19	0.40	-0.23	A+	A-	A+	A-	0.25	0.03	0.15	1.00	-0.07	1.00
92	1034409	634394	3	4093	0.40	0.18	0.40	0.12	0.30	0.00	0.00	0.26	-0.13	0.26	-0.07	-0.12	A+	A+	A-	A-	1.00	0.04	8.74	1.13	9.90	1.22
93	1034410	634395	3	4093	0.21	0.21	0.16	0.28	0.35	0.00	0.00	0.49	0.49	-0.08	-0.11	-0.26	A+	A-	A+	A-	2.10	0.04	-7.78	0.83	-3.74	0.86
94	1035662	635647	3	4093	0.78	0.78	0.08	0.07	0.07	0.00	0.00	0.43	0.43	-0.24	-0.26	-0.18	A+	A+	A+	A-	-1.01	0.04	-5.66	0.89	-4.43	0.83
95	1035663	635648	3	4093	0.75	0.06	0.14	0.75	0.04	0.00	0.00	0.40	-0.16	-0.26	0.40	-0.21	A+	A+	A+	A-	-0.83	0.04	-3.62	0.93	-2.99	0.89
96	1035901	635886	3	4093	0.69	0.11	0.69	0.07	0.12	0.00	0.00	0.49	-0.23	0.49	-0.22	-0.30	A-	A+	A+	A+	-0.48	0.04	-8.75	0.87	-6.77	0.81
97	1018348	618333	4	4114	0.44	0.22	0.44	0.23	0.11	0.00	0.00	0.28	-0.08	0.28	-0.18	-0.10	A-	A+	A-	A-	0.79	0.03	7.23	1.10	6.66	1.12
98	1021585	621570	4	4114	0.70	0.70	0.08	0.08	0.13	0.00	0.00	0.39	0.39	-0.19	-0.17	-0.23	A+	A+	A+	A+	-0.54	0.04	-3.21	0.95	-1.73	0.95
99	1021586	621571	4	4114	0.46	0.15	0.46	0.16	0.22	0.00	0.00	0.41	-0.19	0.41	-0.22	-0.14	A-	A+	A+	A-	0.65	0.03	-2.03	0.97	-1.44	0.98

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
100	1021874	621859	4	4114	0.60	0.60	0.05	0.14	0.20	0.00	0.00	0.41	0.41	-0.23	-0.29	-0.11	A-	A-	A-	A-	-0.01	0.03	-2.89	0.96	-1.28	0.97
101	1023312	623297	4	4114	0.62	0.09	0.62	0.13	0.16	0.00	0.00	0.44	-0.26	0.44	-0.28	-0.12	A+	A+	A+	A+	-0.12	0.04	-5.25	0.93	-5.84	0.87
102	1024193	624178	4	4114	0.48	0.18	0.17	0.48	0.17	0.00	0.00	0.41	-0.35	-0.07	0.41	-0.11	A-	A+	A-	A-	0.58	0.03	-1.90	0.98	-2.17	0.96
103	1024195	624180	4	4114	0.56	0.07	0.19	0.18	0.56	0.00	0.00	0.43	-0.18	-0.10	-0.33	0.43	A-	A-	A-	A+	0.20	0.03	-3.50	0.96	-3.69	0.93
104	1024377	624362	4	4114	0.43	0.23	0.13	0.20	0.43	0.00	0.00	0.29	0.01	-0.24	-0.17	0.29	A-	B-	A-	A-	0.80	0.03	6.81	1.09	6.32	1.12
105	1026337	626322	4	4114	0.48	0.48	0.10	0.05	0.37	0.00	0.00	0.36	0.36	-0.10	-0.13	-0.25	A-	A+	A-	A-	0.55	0.03	2.16	1.03	1.60	1.03
106	1030627	630612	4	4114	0.39	0.13	0.27	0.39	0.21	0.00	0.00	0.27	-0.09	-0.08	0.27	-0.16	A-	A-	A-	A+	1.03	0.04	7.16	1.11	8.48	1.18
107	1032282	632267	4	4114	0.42	0.28	0.42	0.15	0.14	0.00	0.00	0.21	0.04	0.21	-0.17	-0.18	A-	A-	A-	A+	0.84	0.03	9.90	1.18	9.90	1.23
108	1032290	632275	4	4114	0.26	0.26	0.29	0.24	0.21	0.00	0.00	0.04	0.04	-0.19	0.03	0.15	A+	A+	A+	A+	1.73	0.04	9.90	1.33	9.90	1.53
109	1033594	633579	4	4114	0.40	0.15	0.21	0.40	0.25	0.00	0.00	0.40	-0.16	-0.06	0.40	-0.26	A-	A+	A-	A-	0.99	0.04	-1.42	0.98	-0.09	1.00
110	1033600	633585	4	4114	0.66	0.66	0.04	0.16	0.14	0.00	0.00	0.31	0.31	-0.21	-0.19	-0.10	A+	A-	A+	A+	-0.30	0.04	2.68	1.04	3.85	1.10
111	1034574	634559	4	4114	0.59	0.09	0.22	0.10	0.59	0.00	0.00	0.44	-0.19	-0.27	-0.15	0.44	B-	A-	A-	A+	0.06	0.03	-5.34	0.93	-3.00	0.94
112	1034724	634709	4	4114	0.61	0.08	0.24	0.07	0.61	0.00	0.00	0.36	-0.19	-0.19	-0.17	0.36	A+	A-	A+	A+	-0.07	0.03	0.80	1.01	-1.06	0.98
113	1017722	617707	5	4106	0.40	0.10	0.24	0.40	0.25	0.00	0.00	0.24	-0.24	0.02	0.24	-0.12	A+	A+	A+	A-	0.94	0.04	9.90	1.16	9.90	1.24
114	1021584	621569	5	4106	0.65	0.07	0.65	0.06	0.22	0.00	0.00	0.39	-0.23	0.39	-0.26	-0.17	A-	A-	A-	A-	-0.29	0.04	-1.45	0.98	-1.62	0.96
115	1021590	621575	5	4106	0.73	0.73	0.10	0.09	0.08	0.00	0.00	0.46	0.46	-0.23	-0.25	-0.23	A-	A-	A-	A+	-0.73	0.04	-6.42	0.89	-5.95	0.81
116	1023311	623296	5	4106	0.49	0.11	0.49	0.28	0.12	0.00	0.00	0.40	0.02	0.40	-0.22	-0.31	A+	A-	A+	A-	0.53	0.03	0.33	1.00	-0.35	0.99
117	1023639	623624	5	4106	0.43	0.23	0.43	0.18	0.16	0.00	0.00	0.31	-0.10	0.31	-0.02	-0.28	A+	A+	A+	A-	0.82	0.03	6.67	1.09	5.96	1.12
118	1024180	624165	5	4106	0.62	0.07	0.16	0.62	0.14	0.00	0.00	0.41	-0.17	-0.19	0.41	-0.23	A-	A+	A+	A+	-0.15	0.04	-1.49	0.98	-2.55	0.94
119	1024197	624182	5	4106	0.50	0.34	0.50	0.08	0.08	0.00	0.00	0.40	-0.19	0.40	-0.18	-0.22	A-	A-	A+	A-	0.47	0.03	-0.09	1.00	-0.52	0.99
120	1024199	624184	5	4106	0.62	0.12	0.18	0.62	0.07	0.00	0.00	0.28	-0.22	-0.04	0.28	-0.18	A+	A-	A+	A+	-0.14	0.04	7.22	1.10	6.73	1.17
121	1025002	624987	5	4106	0.58	0.11	0.18	0.58	0.13	0.00	0.00	0.49	-0.27	-0.21	0.49	-0.23	A+	A+	A+	A+	0.06	0.03	-7.70	0.90	-7.79	0.84
122	1025006	624991	5	4106	0.57	0.19	0.11	0.57	0.13	0.00	0.00	0.44	-0.21	-0.27	0.44	-0.14	A+	A-	A-	A+	0.14	0.03	-3.25	0.96	-2.86	0.94
123	1026336	626321	5	4106	0.67	0.20	0.09	0.04	0.67	0.00	0.00	0.39	-0.17	-0.26	-0.21	0.39	A-	A-	A-	A+	-0.40	0.04	-2.29	0.97	-0.12	1.00
124	1030596	630581	5	4106	0.39	0.25	0.39	0.20	0.16	0.00	0.00	0.32	-0.03	0.32	-0.21	-0.16	A-	A+	A+	A+	1.01	0.04	4.94	1.07	5.94	1.13
125	1033598	633583	5	4106	0.76	0.06	0.10	0.08	0.76	0.00	0.00	0.49	-0.24	-0.26	-0.27	0.49	A+	A-	A+	A+	-0.88	0.04	-8.74	0.84	-6.79	0.77
126	1034408	634393	5	4106	0.45	0.16	0.22	0.17	0.45	0.00	0.00	0.42	-0.18	-0.18	-0.17	0.42	A-	A+	A+	A+	0.73	0.03	-1.73	0.98	-1.49	0.97
127	1034725	634710	5	4106	0.44	0.44	0.21	0.20	0.15	0.00	0.00	0.37	0.37	-0.17	-0.18	-0.11	A+	A+	A+	A-	0.78	0.03	2.17	1.03	2.10	1.04
128	1035345	635330	5	4106	0.53	0.12	0.17	0.18	0.53	0.00	0.00	0.43	-0.16	-0.25	-0.18	0.43	A+	A-	A-	A-	0.33	0.03	-2.85	0.96	-2.80	0.95
129	1020421	620406	6	4119	0.55	0.13	0.55	0.15	0.17	0.00	0.00	0.43	-0.13	0.43	-0.25	-0.21	A+	A+	A+	A+	0.23	0.03	-3.47	0.96	-3.73	0.93
130	1020497	620482	6	4119	0.78	0.05	0.09	0.08	0.78	0.00	0.00	0.51	-0.26	-0.30	-0.25	0.51	A+	A+	A+	A-	-1.00	0.04	-9.90	0.81	-9.79	0.67
131	1020501	620486	6	4119	0.25	0.25	0.09	0.30	0.36	0.00	0.00	0.36	0.36	-0.24	0.08	-0.26	A+	A+	A+	A-	1.79	0.04	-0.87	0.98	2.16	1.07
132	1021685	621670	6	4119	0.42	0.42	0.20	0.25	0.12	0.00	0.00	0.31	0.31	-0.01	-0.12	-0.29	A-	A-	A+	A-	0.84	0.03	5.87	1.08	6.48	1.12

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
133	1023306	623291	6	4119	0.37	0.35	0.19	0.08	0.37	0.00	0.00	0.27	-0.07	-0.12	-0.17	0.27	A-	A-	A-	A+	1.09	0.04	6.77	1.10	7.59	1.16
134	1024191	624176	6	4119	0.36	0.13	0.28	0.36	0.22	0.00	0.00	0.24	-0.16	-0.11	0.24	-0.02	A-	A+	A-	A+	1.15	0.04	8.82	1.14	9.61	1.22
135	1024196	624181	6	4119	0.64	0.07	0.64	0.22	0.07	0.00	0.00	0.45	-0.22	0.45	-0.26	-0.20	A-	A-	A+	A-	-0.23	0.04	-5.77	0.92	-5.78	0.87
136	1024380	624365	6	4119	0.24	0.24	0.09	0.47	0.20	0.00	0.00	-0.01	-0.01	-0.12	0.26	-0.23	B-	A+	A-	A-	1.84	0.04	9.90	1.35	9.90	1.71
137	1025840	625825	6	4119	0.72	0.12	0.08	0.72	0.08	0.00	0.00	0.47	-0.28	-0.26	0.47	-0.18	A+	A+	A+	A+	-0.64	0.04	-7.63	0.88	-7.44	0.79
138	1030606	630591	6	4119	0.55	0.08	0.22	0.16	0.55	0.00	0.00	0.27	-0.16	-0.06	-0.19	0.27	A+	A-	A+	A+	0.22	0.03	8.17	1.10	6.41	1.12
139	1034412	634397	6	4119	0.65	0.65	0.11	0.18	0.06	0.00	0.00	0.41	0.41	-0.19	-0.25	-0.16	A+	A-	A-	A-	-0.29	0.04	-2.80	0.96	-3.64	0.91
140	1034562	634547	6	4119	0.51	0.25	0.15	0.51	0.09	0.00	0.00	0.39	-0.17	-0.14	0.39	-0.24	A-	A-	A-	A-	0.43	0.03	-0.08	1.00	-0.51	0.99
141	1035467	635452	6	4119	0.76	0.76	0.09	0.12	0.03	0.00	0.00	0.28	0.28	-0.12	-0.18	-0.15	A-	A-	A-	A+	-0.92	0.04	0.65	1.01	5.13	1.20
142	1035476	635461	6	4119	0.30	0.18	0.38	0.30	0.14	0.00	0.00	0.15	-0.06	0.00	0.15	-0.12	A+	A+	A-	A-	1.49	0.04	9.90	1.21	9.90	1.40
143	1035905	635890	6	4119	0.43	0.21	0.21	0.43	0.14	0.00	0.00	0.25	-0.01	-0.16	0.25	-0.16	A-	A-	A+	A+	0.77	0.03	9.51	1.13	9.90	1.19
144	1035907	635892	6	4119	0.51	0.06	0.19	0.51	0.24	0.00	0.00	0.29	-0.17	-0.09	0.29	-0.15	A-	A+	A+	A+	0.41	0.03	7.21	1.09	5.56	1.10
145	1018114	618099	7	4081	0.70	0.09	0.70	0.04	0.17	0.00	0.00	0.41	-0.28	0.41	-0.19	-0.19	A+	A+	A-	A-	-0.50	0.04	-3.59	0.94	-3.87	0.89
146	1023308	623293	7	4081	0.49	0.07	0.08	0.37	0.49	0.00	0.00	0.33	-0.25	-0.19	-0.10	0.33	A-	A-	A-	A-	0.58	0.03	5.03	1.07	3.98	1.07
147	1023643	623628	7	4081	0.42	0.21	0.14	0.23	0.42	0.00	0.00	0.44	-0.24	-0.15	-0.15	0.44	A-	A+	A+	A+	0.89	0.04	-2.88	0.96	-1.39	0.97
148	1025005	624990	7	4081	0.58	0.14	0.19	0.58	0.09	0.00	0.00	0.51	-0.22	-0.27	0.51	-0.24	A+	A-	A-	A+	0.09	0.03	-9.29	0.88	-8.73	0.83
149	1025069	625054	7	4081	0.25	0.12	0.25	0.25	0.37	0.00	0.00	0.30	-0.12	0.30	-0.14	-0.06	A-	A-	A-	A-	1.82	0.04	1.89	1.04	5.93	1.20
150	1025077	625062	7	4081	0.14	0.54	0.14	0.17	0.14	0.00	0.00	-0.03	0.30	-0.20	-0.19	-0.03	A+	A-	A-	A+	2.65	0.05	7.46	1.25	9.90	2.21
151	1025847	625832	7	4081	0.59	0.59	0.19	0.07	0.16	0.00	0.00	0.49	0.49	-0.28	-0.21	-0.22	A+	A-	A+	A+	0.09	0.03	-8.41	0.89	-7.51	0.85
152	1030593	630578	7	4081	0.46	0.12	0.46	0.27	0.15	0.00	0.00	0.43	-0.20	0.43	-0.28	-0.07	A-	A-	A+	A+	0.70	0.03	-2.50	0.97	-2.33	0.96
153	1030646	630631	7	4081	0.76	0.09	0.09	0.76	0.06	0.00	0.00	0.40	-0.25	-0.19	0.40	-0.18	A+	A+	A-	A+	-0.88	0.04	-3.83	0.93	-2.41	0.91
154	1032280	632265	7	4081	0.78	0.06	0.08	0.08	0.78	0.00	0.00	0.47	-0.22	-0.28	-0.25	0.47	B-	A-	A-	A+	-1.01	0.04	-7.69	0.85	-7.65	0.73
155	1035260	635245	7	4081	0.35	0.14	0.35	0.22	0.29	0.00	0.00	0.13	-0.18	0.13	-0.10	0.09	A+	A-	A-	A-	1.25	0.04	9.90	1.26	9.90	1.39
156	1035468	635453	7	4081	0.49	0.13	0.18	0.20	0.49	0.00	0.00	0.27	-0.17	-0.23	0.03	0.27	A+	A+	A+	A+	0.57	0.03	9.55	1.13	9.35	1.17
157	1035471	635456	7	4081	0.33	0.33	0.33	0.16	0.18	0.00	0.00	0.24	0.24	-0.11	-0.25	0.08	A-	A+	A-	A-	1.35	0.04	8.59	1.15	9.14	1.23
158	1035477	635462	7	4081	0.47	0.47	0.08	0.17	0.28	0.00	0.00	0.42	0.42	-0.24	-0.15	-0.20	A-	A+	A-	A-	0.66	0.03	-2.11	0.97	-2.06	0.96
159	1035902	635887	7	4081	0.58	0.13	0.58	0.17	0.12	0.00	0.00	0.49	-0.23	0.49	-0.23	-0.23	A-	A+	A-	A+	0.12	0.03	-8.17	0.90	-7.79	0.85
160	1035904	635889	7	4081	0.41	0.41	0.15	0.26	0.19	0.00	0.00	0.32	0.32	-0.17	-0.16	-0.07	A-	A+	A+	A-	0.97	0.04	4.59	1.07	5.23	1.10
161	1018349	618334	8	4089	0.51	0.18	0.51	0.21	0.09	0.00	0.00	0.28	-0.10	0.28	-0.07	-0.24	A+	A+	A+	A+	0.41	0.03	8.58	1.11	7.43	1.15
162	1024999	624984	8	4089	0.37	0.13	0.25	0.37	0.24	0.00	0.00	0.27	-0.18	-0.08	0.27	-0.07	A+	A+	A+	A-	1.13	0.04	7.49	1.12	8.37	1.19
163	1025003	624988	8	4089	0.44	0.24	0.05	0.44	0.27	0.00	0.00	0.34	-0.09	-0.23	0.34	-0.18	A-	A-	A-	A+	0.79	0.03	4.30	1.06	3.46	1.07
164	1025078	625063	8	4089	0.45	0.06	0.36	0.13	0.45	0.00	0.00	0.28	-0.25	0.05	-0.29	0.28	A+	A+	A-	A+	0.72	0.03	9.11	1.13	7.58	1.15
165	1025079	625064	8	4089	0.58	0.58	0.22	0.11	0.09	0.00	0.00	0.42	0.42	-0.13	-0.23	-0.28	A+	A-	A-	A-	0.11	0.03	-3.22	0.96	-1.68	0.96

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
166	1025844	625829	8	4089	0.65	0.15	0.11	0.65	0.09	0.00	0.00	0.40	-0.27	-0.12	0.40	-0.20	A+	A-	A+	A+	-0.24	0.04	-2.15	0.97	-2.78	0.93
167	1030639	630624	8	4089	0.39	0.08	0.29	0.39	0.23	0.00	0.00	0.31	-0.24	-0.10	0.31	-0.09	A-	A-	A+	A-	1.05	0.04	5.95	1.09	6.21	1.13
168	1032289	632274	8	4089	0.61	0.19	0.13	0.08	0.61	0.00	0.00	0.32	-0.08	-0.18	-0.23	0.32	A-	A+	A+	A-	-0.04	0.04	3.50	1.05	6.06	1.15
169	1032370	632355	8	4089	0.40	0.10	0.07	0.40	0.43	0.00	0.00	0.33	-0.22	-0.18	0.33	-0.10	A-	A-	A-	A+	0.98	0.04	4.49	1.07	4.65	1.10
170	1032371	632356	8	4089	0.69	0.11	0.11	0.09	0.69	0.00	0.00	0.52	-0.25	-0.29	-0.25	0.52	A+	A-	B-	A-	-0.49	0.04	-9.90	0.84	-9.54	0.74
171	1035248	635233	8	4089	0.57	0.08	0.12	0.23	0.57	0.00	0.00	0.43	-0.23	-0.26	-0.16	0.43	A+	A+	A+	A-	0.16	0.03	-3.80	0.95	-2.83	0.94
172	1035258	635243	8	4089	0.49	0.49	0.10	0.12	0.28	0.00	0.00	0.47	0.47	-0.26	-0.25	-0.16	A+	A-	A+	A-	0.51	0.03	-5.82	0.93	-5.62	0.90
173	1035473	635458	8	4089	0.45	0.15	0.45	0.21	0.19	0.00	0.00	0.31	-0.15	0.31	-0.19	-0.06	A+	B-	A-	A+	0.74	0.03	5.84	1.08	4.80	1.09
174	1035857	635842	8	4089	0.63	0.12	0.10	0.15	0.63	0.00	0.00	0.38	-0.16	-0.19	-0.20	0.38	A+	A+	A-	A+	-0.15	0.04	-0.01	1.00	-2.03	0.95
175	1035906	635891	8	4089	0.55	0.19	0.16	0.55	0.10	0.00	0.00	0.51	-0.23	-0.30	0.51	-0.18	A-	A+	A-	A-	0.25	0.03	-9.79	0.88	-8.54	0.84
176	1035908	635893	8	4089	0.42	0.11	0.42	0.12	0.35	0.00	0.00	0.38	-0.10	0.38	-0.17	-0.21	A+	A+	A+	A-	0.88	0.04	0.46	1.01	1.66	1.03
177	1017599	617584	9	4111	0.23	0.17	0.23	0.19	0.41	0.00	0.00	0.22	-0.11	0.22	-0.13	0.00	A+	A+	A+	A+	1.95	0.04	3.87	1.09	9.90	1.40
178	1020422	620407	9	4111	0.60	0.60	0.12	0.17	0.12	0.00	0.00	0.45	0.45	-0.18	-0.22	-0.25	A+	A-	A-	A+	0.01	0.03	-5.95	0.93	-5.46	0.89
179	1020498	620483	9	4111	0.55	0.55	0.22	0.16	0.06	0.00	0.00	0.42	0.42	-0.20	-0.19	-0.23	A-	A-	A-	A-	0.24	0.03	-3.42	0.96	-3.33	0.94
180	1021688	621673	9	4111	0.66	0.06	0.08	0.66	0.21	0.00	0.00	0.38	-0.22	-0.24	0.38	-0.16	A+	A+	A+	A+	-0.29	0.04	-1.34	0.98	-2.28	0.94
181	1024173	624158	9	4111	0.77	0.11	0.06	0.05	0.77	0.00	0.00	0.41	-0.16	-0.27	-0.25	0.41	B+	A+	A+	A-	-0.94	0.04	-5.17	0.90	-4.07	0.85
182	1024280	624265	9	4111	0.65	0.65	0.17	0.11	0.08	0.00	0.00	0.47	0.47	-0.22	-0.25	-0.26	A+	A+	A+	A-	-0.26	0.04	-8.02	0.89	-7.15	0.83
183	1024375	624360	9	4111	0.30	0.13	0.43	0.15	0.30	0.00	0.00	0.13	-0.16	0.14	-0.20	0.13	A-	A-	A-	A-	1.53	0.04	9.90	1.23	9.90	1.41
184	1025071	625056	9	4111	0.26	0.22	0.17	0.35	0.26	0.00	0.00	0.06	-0.11	-0.14	0.15	0.06	A+	A-	A-	A+	1.73	0.04	9.90	1.27	9.90	1.55
185	1025074	625059	9	4111	0.35	0.23	0.35	0.20	0.22	0.00	0.00	0.25	-0.08	0.25	-0.12	-0.08	A+	A+	A-	A-	1.25	0.04	7.51	1.12	8.82	1.21
186	1026340	626325	9	4111	0.50	0.21	0.19	0.50	0.10	0.00	0.00	0.35	-0.10	-0.17	0.35	-0.21	A-	A-	A-	A+	0.51	0.03	2.79	1.04	3.44	1.06
187	1030599	630584	9	4111	0.33	0.42	0.09	0.33	0.17	0.00	0.00	0.13	0.15	-0.26	0.13	-0.16	A-	A+	A-	A-	1.36	0.04	9.90	1.23	9.90	1.37
188	1030604	630589	9	4111	0.51	0.17	0.17	0.51	0.14	0.00	0.00	0.38	-0.18	-0.22	0.38	-0.11	A-	A+	A-	A-	0.42	0.03	0.21	1.00	0.31	1.01
189	1030612	630597	9	4111	0.46	0.18	0.25	0.11	0.46	0.00	0.00	0.16	0.03	-0.10	-0.15	0.16	A-	A-	A-	A-	0.67	0.03	9.90	1.23	9.90	1.28
190	1033593	633578	9	4111	0.55	0.14	0.55	0.09	0.22	0.00	0.00	0.44	-0.09	0.44	-0.18	-0.32	A-	A-	A+	A-	0.26	0.03	-4.37	0.95	-4.33	0.92
191	1035470	635455	9	4111	0.92	0.02	0.92	0.04	0.02	0.00	0.00	0.36	-0.18	0.36	-0.23	-0.20	B+	A-	A-	B+	-2.25	0.06	-2.88	0.88	-7.56	0.52
192	1035856	635841	9	4111	0.36	0.38	0.14	0.11	0.36	0.00	0.00	0.32	-0.11	-0.09	-0.23	0.32	A+	A-	A-	A+	1.17	0.04	2.66	1.04	4.79	1.10
193	1023307	623292	10	4107	0.50	0.28	0.11	0.50	0.11	0.00	0.00	0.26	-0.09	-0.23	0.26	-0.06	A+	A+	A-	A-	0.51	0.03	7.18	1.09	5.81	1.09
194	1023634	623619	10	4107	0.43	0.14	0.43	0.22	0.21	0.00	0.00	0.35	-0.18	0.35	-0.18	-0.08	A+	A+	A+	A-	0.82	0.03	0.34	1.00	0.49	1.01
195	1024279	624264	10	4107	0.30	0.05	0.30	0.51	0.14	0.00	0.00	0.09	-0.21	0.09	0.16	-0.21	A-	A+	A+	A+	1.50	0.04	9.90	1.22	9.90	1.38
196	1024282	624267	10	4107	0.45	0.11	0.45	0.31	0.13	0.00	0.00	0.29	-0.09	0.29	-0.13	-0.15	A+	A+	A+	A-	0.73	0.03	5.11	1.06	4.50	1.07
197	1024381	624366	10	4107	0.49	0.12	0.49	0.09	0.29	0.00	0.00	0.27	-0.26	0.27	-0.24	0.04	A+	A-	A+	A+	0.51	0.03	6.52	1.08	5.73	1.09
198	1025007	624992	10	4107	0.31	0.26	0.31	0.31	0.12	0.00	0.00	0.04	0.20	-0.08	0.04	-0.20	A-	A-	A-	A+	1.45	0.04	9.90	1.28	9.90	1.45

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
199	1025833	625818	10	4107	0.55	0.33	0.05	0.55	0.06	0.00	0.00	0.23	-0.03	-0.25	0.23	-0.18	A+	A+	A+	A-	0.23	0.03	9.14	1.11	8.27	1.14
200	1025850	625835	10	4107	0.58	0.19	0.16	0.07	0.58	0.00	0.00	0.29	-0.20	-0.03	-0.22	0.29	A+	A+	A+	A+	0.10	0.03	4.18	1.05	4.54	1.08
201	1026341	626326	10	4107	0.26	0.39	0.05	0.26	0.30	0.00	0.00	0.16	0.15	-0.11	0.16	-0.26	B-	A-	A+	A+	1.69	0.04	6.99	1.14	9.90	1.30
202	1028205	628190	10	4107	0.36	0.08	0.36	0.13	0.42	0.00	0.00	0.26	-0.22	0.26	-0.19	0.01	B-	A-	A-	A-	1.16	0.04	5.30	1.08	6.05	1.12
203	1028210	628195	10	4107	0.28	0.28	0.43	0.16	0.13	0.00	0.00	-0.05	-0.05	0.32	-0.24	-0.14	A-	A+	A-	A-	1.59	0.04	9.90	1.35	9.90	1.56
204	1030615	630600	10	4107	0.21	0.21	0.25	0.29	0.24	0.00	0.00	0.07	0.07	-0.06	0.01	0.00	A-	A+	A+	A-	2.06	0.04	7.60	1.18	9.90	1.49
205	1030623	630608	10	4107	0.20	0.12	0.20	0.54	0.13	0.00	0.00	-0.01	-0.23	-0.01	0.18	-0.02	A-	A-	A-	A-	2.11	0.04	9.90	1.26	9.90	1.63
206	1034423	634408	10	4107	0.49	0.08	0.18	0.24	0.49	0.00	0.00	0.49	-0.18	-0.30	-0.19	0.49	A+	A+	A-	A-	0.52	0.03	-9.90	0.88	-9.83	0.86
207	1035341	635326	10	4107	0.29	0.13	0.29	0.10	0.48	0.00	0.00	0.21	-0.10	0.21	-0.20	0.01	A-	A-	A+	A-	1.55	0.04	5.38	1.10	8.88	1.24
208	1035858	635843	10	4107	0.35	0.35	0.15	0.38	0.12	0.00	0.00	0.29	0.29	-0.25	0.02	-0.17	A+	A+	A-	A-	1.24	0.04	3.15	1.05	4.98	1.11
209	1023644	623629	11	4107	0.52	0.11	0.19	0.52	0.18	0.00	0.00	0.45	-0.19	-0.26	0.45	-0.16	A+	A-	A+	A+	0.39	0.03	-5.15	0.94	-5.53	0.91
210	1023646	623631	11	4107	0.50	0.13	0.25	0.50	0.12	0.00	0.00	0.36	-0.20	-0.13	0.36	-0.15	A-	A+	A+	A+	0.49	0.03	2.19	1.03	1.85	1.03
211	1024181	624166	11	4107	0.64	0.11	0.64	0.11	0.14	0.00	0.00	0.47	-0.25	0.47	-0.30	-0.15	A-	A+	A+	A-	-0.20	0.04	-8.35	0.89	-6.04	0.87
212	1024281	624266	11	4107	0.62	0.62	0.13	0.11	0.14	0.00	0.00	0.37	0.37	-0.16	-0.23	-0.15	A+	A+	A-	A+	-0.09	0.03	-0.81	0.99	1.47	1.03
213	1024283	624268	11	4107	0.51	0.16	0.19	0.14	0.51	0.00	0.00	0.27	-0.09	-0.18	-0.09	0.27	A-	A+	A-	A-	0.43	0.03	8.62	1.11	7.25	1.13
214	1028206	628191	11	4107	0.29	0.29	0.18	0.41	0.12	0.00	0.00	0.28	0.28	-0.15	0.02	-0.25	A+	A-	A+	A-	1.58	0.04	2.83	1.05	6.92	1.20
215	1028208	628193	11	4107	0.43	0.36	0.11	0.43	0.10	0.00	0.00	0.07	0.13	-0.22	0.07	-0.10	A+	A-	A-	A-	0.81	0.03	9.90	1.32	9.90	1.40
216	1030608	630593	11	4107	0.45	0.16	0.14	0.45	0.25	0.00	0.00	0.41	-0.03	-0.24	0.41	-0.26	A+	A-	A-	A-	0.73	0.03	-2.51	0.97	-1.07	0.98
217	1030624	630609	11	4107	0.66	0.12	0.12	0.66	0.09	0.00	0.00	0.37	-0.13	-0.24	0.37	-0.17	A+	A+	A+	A+	-0.32	0.04	-1.48	0.98	-0.49	0.99
218	1030630	630615	11	4107	0.61	0.11	0.61	0.18	0.11	0.00	0.00	0.44	-0.22	0.44	-0.22	-0.20	A-	A-	A-	A-	-0.06	0.03	-5.52	0.93	-5.31	0.89
219	1030653	630638	11	4107	0.16	0.46	0.14	0.16	0.23	0.00	0.00	0.01	0.29	-0.25	0.01	-0.14	A+	A-	A+	A-	2.43	0.05	8.86	1.27	9.90	1.85
220	1032288	632273	11	4107	0.56	0.07	0.07	0.56	0.29	0.00	0.00	0.15	-0.25	-0.21	0.15	0.10	A+	A+	A-	A+	0.16	0.03	9.90	1.21	9.90	1.39
221	1033603	633588	11	4107	0.35	0.35	0.27	0.15	0.22	0.00	0.00	0.30	0.30	-0.06	-0.21	-0.09	A-	A-	A+	A-	1.25	0.04	4.25	1.07	6.25	1.14
222	1034573	634558	11	4107	0.63	0.09	0.63	0.19	0.10	0.00	0.00	0.47	-0.14	0.47	-0.30	-0.22	A-	A-	A+	A+	-0.15	0.04	-7.66	0.90	-5.80	0.88
223	1034722	634707	11	4107	0.47	0.11	0.47	0.27	0.14	0.00	0.00	0.38	-0.18	0.38	-0.16	-0.17	A+	A-	A+	A+	0.60	0.03	0.06	1.00	-0.40	0.99
224	1035342	635327	11	4107	0.21	0.36	0.21	0.20	0.22	0.00	0.00	0.19	0.10	0.19	-0.17	-0.13	A-	A+	A-	A+	2.04	0.04	3.78	1.09	9.90	1.49
225	1018112	618097	12	4089	0.34	0.30	0.23	0.13	0.34	0.00	0.00	0.41	-0.13	-0.13	-0.23	0.41	A-	A-	A-	A+	1.30	0.04	-3.25	0.95	0.06	1.00
226	1020416	620401	12	4089	0.66	0.15	0.11	0.07	0.66	0.00	0.00	0.47	-0.14	-0.29	-0.29	0.47	A+	A-	A+	A-	-0.29	0.04	-8.04	0.89	-6.43	0.85
227	1021690	621675	12	4089	0.39	0.39	0.20	0.32	0.08	0.00	0.00	0.21	0.21	-0.19	-0.02	-0.05	A-	A-	A-	A-	1.01	0.04	9.90	1.17	9.90	1.25
228	1023313	623298	12	4089	0.37	0.11	0.37	0.24	0.27	0.00	0.00	0.33	-0.18	0.33	-0.16	-0.07	A+	A+	A-	A+	1.11	0.04	2.10	1.03	4.36	1.09
229	1023633	623618	12	4089	0.41	0.41	0.06	0.15	0.38	0.00	0.00	0.15	0.15	-0.11	-0.16	0.02	A+	A+	A+	A+	0.95	0.03	9.90	1.23	9.90	1.29
230	1026338	626323	12	4089	0.55	0.37	0.04	0.04	0.55	0.00	0.00	0.29	-0.18	-0.17	-0.13	0.29	A-	A+	A+	A+	0.26	0.03	6.48	1.08	4.81	1.09
231	1028207	628192	12	4089	0.52	0.09	0.08	0.52	0.30	0.00	0.00	0.29	-0.21	-0.13	0.29	-0.11	A+	A-	A-	A+	0.39	0.03	6.73	1.08	6.18	1.11

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
232	1028209	628194	12	4089	0.33	0.14	0.33	0.29	0.23	0.00	0.00	-0.03	-0.11	-0.03	0.13	-0.02	A+	A+	A+	A-	1.32	0.04	9.90	1.39	9.90	1.56
233	1030591	630576	12	4089	0.33	0.35	0.17	0.14	0.33	0.00	0.00	0.32	0.11	-0.29	-0.26	0.32	A+	A-	A+	A+	1.34	0.04	2.80	1.05	4.89	1.12
234	1030605	630590	12	4089	0.32	0.18	0.23	0.27	0.32	0.00	0.00	0.31	-0.13	-0.14	-0.08	0.31	A-	A-	A+	A-	1.38	0.04	2.64	1.04	4.64	1.11
235	1030611	630596	12	4089	0.46	0.09	0.21	0.46	0.23	0.00	0.00	0.42	-0.18	-0.12	0.42	-0.25	A+	A+	A+	A+	0.67	0.03	-2.93	0.96	-2.79	0.95
236	1032372	632357	12	4089	0.86	0.04	0.86	0.03	0.07	0.00	0.00	0.41	-0.23	0.41	-0.22	-0.23	A-	A-	A-	A+	-1.59	0.05	-4.68	0.87	-6.59	0.68
237	1034715	634700	12	4089	0.32	0.23	0.18	0.32	0.27	0.00	0.00	0.17	0.02	-0.05	0.17	-0.16	A-	A-	A-	A-	1.37	0.04	9.90	1.18	9.90	1.31
238	1034726	634711	12	4089	0.52	0.21	0.15	0.11	0.52	0.00	0.00	0.42	-0.14	-0.25	-0.21	0.42	A+	A-	A+	A-	0.40	0.03	-3.13	0.96	-3.09	0.95
239	1035257	635242	12	4089	0.59	0.14	0.14	0.59	0.13	0.00	0.00	0.39	-0.20	-0.16	0.39	-0.21	A+	A-	A-	A+	0.07	0.03	-1.23	0.98	-2.68	0.95
240	1035860	635845	12	4089	0.51	0.51	0.24	0.15	0.11	0.00	0.00	0.42	0.42	-0.14	-0.24	-0.20	A+	A-	A+	A-	0.46	0.03	-2.86	0.97	-3.35	0.94
241	1023315	623300	13	4173	0.41	0.14	0.23	0.41	0.22	0.00	0.00	0.36	-0.17	-0.08	0.36	-0.20	A-	A+	A-	A+	0.91	0.03	1.84	1.03	2.88	1.06
242	1023636	623621	13	4173	0.41	0.26	0.15	0.41	0.18	0.00	0.00	0.41	-0.12	-0.24	0.41	-0.17	A+	A+	A+	A-	0.93	0.03	-1.99	0.97	-0.91	0.98
243	1023645	623630	13	4173	0.39	0.06	0.25	0.29	0.39	0.00	0.00	0.31	-0.19	-0.21	-0.03	0.31	A+	A-	A+	A+	0.99	0.04	5.00	1.07	5.37	1.11
244	1024374	624359	13	4173	0.60	0.12	0.60	0.19	0.10	0.00	0.00	0.42	-0.18	0.42	-0.23	-0.20	A-	A+	A+	A+	0.00	0.03	-3.44	0.96	-3.33	0.93
245	1025000	624985	13	4173	0.22	0.32	0.18	0.27	0.22	0.00	0.00	0.18	0.13	-0.19	-0.14	0.18	A+	A-	A-	A-	1.97	0.04	6.41	1.15	9.90	1.44
246	1026342	626327	13	4173	0.66	0.66	0.10	0.16	0.08	0.00	0.00	0.40	0.40	-0.18	-0.27	-0.13	A-	A-	A-	A-	-0.32	0.04	-3.37	0.95	-2.27	0.94
247	1030607	630592	13	4173	0.42	0.22	0.42	0.12	0.25	0.00	0.00	0.43	-0.22	0.43	-0.22	-0.10	A+	A+	A+	A-	0.88	0.03	-2.68	0.96	-2.30	0.96
248	1030648	630633	13	4173	0.50	0.20	0.50	0.15	0.16	0.00	0.00	0.45	-0.16	0.45	-0.19	-0.26	B-	A-	A-	A-	0.48	0.03	-4.92	0.94	-4.41	0.92
249	1030654	630639	13	4173	0.39	0.15	0.23	0.39	0.22	0.00	0.00	0.34	-0.15	-0.21	0.34	-0.05	A+	A+	A+	A-	1.03	0.04	2.76	1.04	4.23	1.09
250	1034597	634582	13	4173	0.29	0.18	0.25	0.28	0.29	0.00	0.00	0.15	-0.14	-0.07	0.03	0.15	A+	A-	A-	A+	1.53	0.04	9.90	1.24	9.90	1.39
251	1034599	634584	13	4173	0.40	0.40	0.33	0.08	0.19	0.00	0.00	0.37	0.37	-0.22	-0.25	-0.01	A-	A-	A-	A-	0.95	0.03	2.17	1.03	2.75	1.05
252	1034719	634704	13	4173	0.36	0.16	0.17	0.31	0.36	0.00	0.00	0.29	-0.19	-0.26	0.06	0.29	A+	A+	A+	A-	1.19	0.04	5.61	1.09	7.01	1.16
253	1034727	634712	13	4173	0.51	0.18	0.16	0.51	0.15	0.00	0.00	0.38	-0.09	-0.21	0.38	-0.21	A+	A+	A+	A-	0.43	0.03	1.20	1.02	0.31	1.01
254	1035264	635249	13	4173	0.48	0.21	0.48	0.17	0.13	0.00	0.00	0.29	-0.16	0.29	-0.03	-0.20	A+	A+	A+	A+	0.57	0.03	7.96	1.10	6.66	1.13
255	1035346	635331	13	4173	0.64	0.17	0.12	0.64	0.07	0.00	0.00	0.33	-0.22	-0.16	0.33	-0.08	A+	A+	A+	A+	-0.20	0.04	2.58	1.03	2.07	1.05
256	1035859	635844	13	4173	0.60	0.18	0.60	0.10	0.12	0.00	0.00	0.39	-0.18	0.39	-0.27	-0.12	A+	A+	A-	A+	-0.02	0.03	-1.79	0.98	0.37	1.01
257	1020417	620402	14	4171	0.49	0.49	0.15	0.25	0.10	0.00	0.00	0.46	0.46	-0.22	-0.18	-0.25	A+	A-	A-	A-	0.49	0.03	-5.84	0.93	-5.55	0.91
258	1021682	621667	14	4171	0.56	0.56	0.09	0.18	0.17	0.00	0.00	0.54	0.54	-0.24	-0.31	-0.21	A-	A+	A-	A+	0.15	0.03	-9.90	0.85	-9.90	0.80
259	1023018	623003	14	4171	0.27	0.11	0.27	0.48	0.13	0.00	0.00	0.06	-0.16	0.06	0.17	-0.18	A-	A-	A-	A+	1.65	0.04	9.90	1.31	9.90	1.54
260	1023020	623005	14	4171	0.41	0.32	0.20	0.41	0.07	0.00	0.00	0.26	-0.06	-0.14	0.26	-0.18	A+	A+	A-	A+	0.93	0.03	8.49	1.12	8.78	1.18
261	1024176	624161	14	4171	0.61	0.18	0.11	0.10	0.61	0.00	0.00	0.38	-0.18	-0.18	-0.20	0.38	A-	A-	A-	A+	-0.09	0.03	-0.71	0.99	0.75	1.02
262	1024177	624162	14	4171	0.71	0.08	0.13	0.08	0.71	0.00	0.00	0.50	-0.24	-0.30	-0.22	0.50	A+	A-	A+	A-	-0.61	0.04	-9.34	0.86	-9.42	0.74
263	1025067	625052	14	4171	0.24	0.37	0.19	0.20	0.24	0.00	0.00	0.17	-0.01	-0.03	-0.13	0.17	A+	A+	A+	A+	1.86	0.04	6.98	1.15	9.90	1.44
264	1025070	625055	14	4171	0.23	0.23	0.18	0.34	0.26	0.00	0.00	0.18	0.18	-0.11	-0.11	0.04	A-	A+	A+	A+	1.95	0.04	7.00	1.16	9.90	1.38

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
265	1025835	625820	14	4171	0.41	0.41	0.07	0.14	0.38	0.00	0.00	0.26	0.26	0.03	-0.13	-0.18	A-	A+	A+	A+	0.92	0.03	8.75	1.13	8.88	1.18
266	1030651	630636	14	4171	0.33	0.37	0.33	0.20	0.09	0.00	0.00	0.17	0.07	0.17	-0.21	-0.09	A-	A+	A+	A-	1.33	0.04	9.90	1.18	9.90	1.36
267	1030658	630643	14	4171	0.77	0.77	0.10	0.05	0.07	0.00	0.00	0.34	0.34	-0.21	-0.17	-0.16	A+	A+	A+	A+	-0.96	0.04	-1.50	0.97	-1.03	0.96
268	1034572	634557	14	4171	0.45	0.20	0.45	0.21	0.14	0.00	0.00	0.36	-0.23	0.36	-0.09	-0.14	A-	A-	A+	A-	0.70	0.03	2.47	1.03	2.05	1.04
269	1034602	634587	14	4171	0.59	0.12	0.59	0.21	0.08	0.00	0.00	0.22	-0.19	0.22	-0.03	-0.11	A-	A-	A+	A-	0.02	0.03	9.90	1.15	9.15	1.20
270	1034603	634588	14	4171	0.84	0.84	0.07	0.05	0.03	0.00	0.00	0.40	0.40	-0.26	-0.20	-0.18	A+	B-	A-	A+	-1.45	0.04	-4.54	0.88	-5.15	0.76
271	1035253	635238	14	4171	0.74	0.04	0.09	0.74	0.12	0.00	0.00	0.42	-0.24	-0.24	0.42	-0.21	A-	A-	A-	A-	-0.79	0.04	-5.13	0.91	-4.83	0.85
272	1035903	635888	14	4171	0.49	0.27	0.49	0.18	0.06	0.00	0.00	0.41	-0.20	0.41	-0.14	-0.26	A+	A+	A-	A+	0.52	0.03	-1.44	0.98	-1.21	0.98
273	1021683	621668	15	4156	0.62	0.62	0.16	0.10	0.11	0.00	0.00	0.39	0.39	-0.18	-0.22	-0.19	A+	A+	A+	A-	-0.10	0.03	-1.21	0.98	-1.59	0.96
274	1023635	623620	15	4156	0.41	0.20	0.41	0.22	0.16	0.00	0.00	0.14	-0.01	0.14	-0.17	0.02	A+	A+	A+	A+	0.92	0.03	9.90	1.26	9.90	1.35
275	1024174	624159	15	4156	0.56	0.09	0.24	0.56	0.11	0.00	0.00	0.36	-0.24	-0.13	0.36	-0.17	A+	A+	A-	A-	0.18	0.03	2.41	1.03	1.30	1.03
276	1025068	625053	15	4156	0.46	0.17	0.11	0.26	0.46	0.00	0.00	0.34	-0.14	-0.20	-0.12	0.34	A+	A-	A-	A+	0.71	0.03	4.00	1.05	3.75	1.07
277	1025076	625061	15	4156	0.28	0.27	0.11	0.28	0.34	0.00	0.00	0.20	-0.01	-0.24	0.20	-0.01	A-	A+	A-	A-	1.62	0.04	8.00	1.15	9.90	1.37
278	1030598	630583	15	4156	0.51	0.18	0.51	0.16	0.15	0.00	0.00	0.45	-0.14	0.45	-0.27	-0.20	A+	A-	A+	A-	0.45	0.03	-4.27	0.95	-4.45	0.92
279	1030609	630594	15	4156	0.39	0.39	0.22	0.27	0.12	0.00	0.00	0.36	0.36	-0.23	-0.04	-0.20	A+	A-	A+	A+	1.04	0.04	2.41	1.04	4.28	1.09
280	1030622	630607	15	4156	0.69	0.69	0.21	0.07	0.03	0.00	0.00	0.28	0.28	-0.17	-0.13	-0.16	A+	A+	A+	A-	-0.44	0.04	4.80	1.07	2.59	1.07
281	1034411	634396	15	4156	0.71	0.71	0.10	0.05	0.13	0.00	0.00	0.46	0.46	-0.28	-0.25	-0.20	A-	A+	A+	A+	-0.60	0.04	-6.98	0.89	-6.52	0.81
282	1034564	634549	15	4156	0.45	0.16	0.13	0.45	0.26	0.00	0.00	0.40	-0.25	-0.21	0.40	-0.09	A+	A+	A-	A+	0.75	0.03	-0.71	0.99	0.59	1.01
283	1034601	634586	15	4156	0.64	0.64	0.10	0.15	0.11	0.00	0.00	0.46	0.46	-0.21	-0.24	-0.24	A+	A-	A-	A-	-0.20	0.04	-6.71	0.91	-5.54	0.87
284	1034604	634589	15	4156	0.74	0.14	0.74	0.06	0.06	0.00	0.00	0.34	-0.14	0.34	-0.22	-0.20	A-	A-	A-	A-	-0.77	0.04	-1.27	0.98	1.64	1.06
285	1034716	634701	15	4156	0.52	0.10	0.10	0.52	0.28	0.00	0.00	0.42	-0.21	-0.24	0.42	-0.16	A-	A+	A-	A-	0.41	0.03	-1.63	0.98	-2.27	0.96
286	1035343	635328	15	4156	0.30	0.28	0.30	0.11	0.30	0.00	0.00	0.46	-0.24	-0.15	-0.11	0.46	A-	A-	A-	A-	1.52	0.04	-6.13	0.90	-2.03	0.95
287	1035472	635457	15	4156	0.26	0.26	0.15	0.40	0.19	0.00	0.00	0.31	0.31	-0.18	-0.05	-0.12	A-	A-	A+	A+	1.78	0.04	1.24	1.02	4.88	1.16
288	1035855	635840	15	4156	0.38	0.36	0.12	0.14	0.38	0.00	0.00	0.25	0.09	-0.22	-0.27	0.25	A+	A+	A+	A-	1.08	0.04	9.33	1.14	8.85	1.19
289	1020423	620408	16	4190	0.44	0.44	0.25	0.18	0.13	0.00	0.00	0.37	0.37	-0.03	-0.26	-0.21	A+	A-	A-	A-	0.71	0.03	1.57	1.02	1.50	1.03
290	1021686	621671	16	4190	0.65	0.06	0.09	0.65	0.19	0.00	0.00	0.41	-0.29	-0.23	0.41	-0.15	B-	A-	A-	A+	-0.32	0.04	-3.35	0.95	-2.77	0.93
291	1025001	624986	16	4190	0.47	0.47	0.17	0.27	0.09	0.00	0.00	0.34	0.34	-0.24	-0.06	-0.18	A+	A+	A+	A-	0.60	0.03	4.36	1.06	4.28	1.08
292	1025837	625822	16	4190	0.38	0.30	0.25	0.07	0.38	0.00	0.00	0.27	-0.24	0.09	-0.23	0.27	A+	A+	A-	A-	1.04	0.04	8.11	1.12	8.68	1.19
293	1025842	625827	16	4190	0.45	0.15	0.18	0.21	0.45	0.00	0.00	0.39	-0.18	-0.15	-0.17	0.39	A-	A-	A+	A-	0.68	0.03	0.26	1.00	0.32	1.01
294	1026339	626324	16	4190	0.23	0.30	0.18	0.23	0.30	0.00	0.00	0.06	0.03	-0.26	0.06	0.14	A-	A+	A-	A+	1.91	0.04	9.90	1.27	9.90	1.63
295	1030600	630585	16	4190	0.49	0.18	0.10	0.49	0.23	0.00	0.00	0.32	-0.15	-0.26	0.32	-0.05	A-	A-	A-	A+	0.49	0.03	5.65	1.07	6.31	1.11
296	1034415	634400	16	4190	0.59	0.13	0.15	0.59	0.13	0.00	0.00	0.46	-0.20	-0.24	0.46	-0.21	B+	A-	A-	A+	0.01	0.03	-5.97	0.93	-6.41	0.88
297	1034420	634405	16	4190	0.64	0.21	0.11	0.64	0.04	0.00	0.00	0.43	-0.20	-0.28	0.43	-0.18	A+	A+	A-	A-	-0.23	0.04	-4.55	0.94	-3.49	0.92

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
298	1034561	634546	16	4190	0.43	0.07	0.26	0.23	0.43	0.00	0.00	0.24	-0.14	-0.06	-0.13	0.24	A-	A-	A+	A+	0.77	0.03	9.90	1.16	9.90	1.19
299	1034570	634555	16	4190	0.73	0.13	0.09	0.73	0.06	0.00	0.00	0.48	-0.27	-0.24	0.48	-0.22	A-	A-	A-	A+	-0.72	0.04	-8.05	0.87	-7.43	0.78
300	1034721	634706	16	4190	0.44	0.14	0.27	0.15	0.44	0.00	0.00	0.36	-0.16	-0.12	-0.18	0.36	A+	A-	A-	A-	0.75	0.03	2.75	1.04	1.82	1.03
301	1035533	635518	16	4190	0.71	0.08	0.71	0.15	0.06	0.00	0.00	0.46	-0.24	0.46	-0.27	-0.20	A-	A-	A-	A-	-0.65	0.04	-7.17	0.89	-6.47	0.82
302	1035535	635520	16	4190	0.69	0.69	0.09	0.13	0.08	0.00	0.00	0.49	0.49	-0.24	-0.24	-0.26	A+	A+	A+	A-	-0.52	0.04	-8.82	0.87	-7.65	0.80
303	1036075	636060	16	4190	0.22	0.11	0.22	0.35	0.32	0.00	0.00	0.03	-0.19	0.03	-0.10	0.20	A-	A-	A-	A+	2.00	0.04	9.90	1.28	9.90	1.74
304	1036079	636064	16	4190	0.48	0.19	0.23	0.48	0.10	0.00	0.00	0.36	-0.10	-0.17	0.36	-0.24	A+	A-	A-	A+	0.51	0.03	2.62	1.03	2.21	1.04
305	1021689	621674	17	4132	0.47	0.19	0.26	0.47	0.08	0.00	0.00	0.42	-0.19	-0.15	0.42	-0.23	A+	A+	A+	A+	0.64	0.03	-1.61	0.98	-1.53	0.97
306	1021875	621860	17	4132	0.77	0.04	0.10	0.77	0.09	0.00	0.00	0.34	-0.17	-0.17	0.34	-0.19	A+	A-	A-	A+	-0.96	0.04	-1.52	0.97	-0.56	0.98
307	1023019	623004	17	4132	0.64	0.11	0.17	0.64	0.08	0.00	0.00	0.53	-0.24	-0.31	0.53	-0.22	A-	A-	A+	A-	-0.20	0.04	-9.90	0.85	-9.90	0.78
308	1023024	623009	17	4132	0.59	0.19	0.05	0.18	0.59	0.00	0.00	0.49	-0.21	-0.24	-0.28	0.49	A-	A-	A+	A+	0.07	0.03	-8.24	0.90	-7.81	0.85
309	1024383	624368	17	4132	0.16	0.18	0.16	0.17	0.49	0.00	0.00	-0.06	-0.18	-0.06	-0.17	0.31	A-	A+	A+	A-	2.51	0.05	9.38	1.29	9.90	2.29
310	1030601	630586	17	4132	0.46	0.29	0.17	0.46	0.09	0.00	0.00	0.44	-0.07	-0.29	0.44	-0.28	A+	A+	A-	A+	0.70	0.03	-3.44	0.96	-2.55	0.95
311	1030602	630587	17	4132	0.39	0.22	0.39	0.25	0.15	0.00	0.00	0.40	-0.09	0.40	-0.23	-0.15	B-	A-	A+	A-	1.05	0.04	-0.95	0.99	1.14	1.02
312	1030613	630598	17	4132	0.51	0.17	0.23	0.09	0.51	0.00	0.00	0.16	0.02	-0.08	-0.19	0.16	A+	A-	A+	A-	0.43	0.03	9.90	1.24	9.90	1.40
313	1030625	630610	17	4132	0.41	0.23	0.19	0.41	0.16	0.00	0.00	0.26	-0.06	-0.12	0.26	-0.14	A+	A+	A+	A+	0.92	0.04	9.18	1.13	9.37	1.19
314	1030647	630632	17	4132	0.67	0.13	0.67	0.09	0.11	0.00	0.00	0.43	-0.21	0.43	-0.20	-0.22	A-	A+	A+	A-	-0.37	0.04	-4.59	0.93	-2.16	0.94
315	1030657	630642	17	4132	0.33	0.21	0.27	0.18	0.33	0.00	0.00	0.21	-0.12	-0.05	-0.06	0.21	A+	A+	A+	A+	1.33	0.04	9.66	1.17	9.90	1.31
316	1035479	635464	17	4132	0.45	0.45	0.18	0.22	0.14	0.00	0.00	0.41	0.41	-0.16	-0.20	-0.17	A-	A-	A+	A-	0.71	0.03	-0.75	0.99	-0.63	0.99
317	1035534	635519	17	4132	0.28	0.28	0.08	0.13	0.51	0.00	0.00	0.13	0.13	-0.27	-0.24	0.20	A-	A-	A-	A-	1.64	0.04	9.90	1.25	9.90	1.46
318	1035536	635521	17	4132	0.79	0.09	0.06	0.06	0.79	0.00	0.00	0.41	-0.25	-0.25	-0.15	0.41	A-	A-	A-	A+	-1.10	0.04	-4.92	0.90	-3.20	0.87
319	1036076	636061	17	4132	0.85	0.04	0.08	0.85	0.04	0.00	0.00	0.35	-0.23	-0.18	0.35	-0.18	A-	A-	A-	A-	-1.50	0.05	-2.48	0.93	-3.25	0.84
320	1036078	636063	17	4132	0.70	0.08	0.70	0.12	0.11	0.00	0.00	0.49	-0.25	0.49	-0.29	-0.20	A+	B-	A+	A-	-0.51	0.04	-8.42	0.87	-7.85	0.79
321	1023314	623299	18	4141	0.59	0.59	0.19	0.13	0.09	0.00	0.00	0.52	0.52	-0.21	-0.30	-0.25	A+	A+	A+	A-	0.03	0.03	-9.90	0.88	-9.09	0.81
322	1024179	624164	18	4141	0.53	0.13	0.13	0.22	0.53	0.00	0.00	0.41	-0.12	-0.21	-0.22	0.41	A+	A-	A-	A-	0.35	0.03	-0.83	0.99	0.73	1.01
323	1025834	625819	18	4141	0.41	0.07	0.16	0.36	0.41	0.00	0.00	0.41	-0.16	-0.29	-0.11	0.41	A-	A-	A-	A-	0.93	0.04	-0.38	0.99	0.21	1.00
324	1025836	625821	18	4141	0.56	0.13	0.11	0.20	0.56	0.00	0.00	0.50	-0.26	-0.26	-0.20	0.50	A+	A+	A+	A+	0.16	0.03	-8.41	0.90	-8.01	0.84
325	1025841	625826	18	4141	0.61	0.10	0.15	0.13	0.61	0.00	0.00	0.46	-0.21	-0.18	-0.27	0.46	A+	A-	A+	A-	-0.09	0.04	-5.52	0.93	-5.21	0.88
326	1030594	630579	18	4141	0.69	0.10	0.14	0.69	0.08	0.00	0.00	0.51	-0.28	-0.25	0.51	-0.25	A-	A-	A+	A+	-0.48	0.04	-9.00	0.87	-9.08	0.75
327	1030645	630630	18	4141	0.65	0.08	0.08	0.65	0.20	0.00	0.00	0.40	-0.19	-0.25	0.40	-0.17	A+	A-	A+	A-	-0.27	0.04	-1.51	0.98	-0.69	0.98
328	1032291	632276	18	4141	0.59	0.07	0.20	0.59	0.12	0.00	0.00	0.39	-0.22	-0.16	0.39	-0.21	A-	A-	A-	A+	0.01	0.03	0.36	1.00	0.17	1.00
329	1034563	634548	18	4141	0.80	0.80	0.04	0.11	0.05	0.00	0.00	0.32	0.32	-0.19	-0.19	-0.15	A+	A-	A+	A+	-1.21	0.04	-1.04	0.98	0.37	1.02
330	1034565	634550	18	4141	0.61	0.61	0.25	0.05	0.08	0.00	0.00	0.41	0.41	-0.24	-0.25	-0.14	A-	A-	A-	A+	-0.10	0.04	-1.94	0.97	-2.04	0.95

Table J-6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
331	1035252	635237	18	4141	0.41	0.41	0.12	0.20	0.27	0.00	0.00	0.45	0.45	-0.22	-0.20	-0.14	A-	A+	A+	A-	0.95	0.04	-4.09	0.94	-1.57	0.97
332	1035259	635244	18	4141	0.35	0.15	0.35	0.34	0.16	0.00	0.00	0.20	-0.14	0.20	-0.07	-0.03	A+	A+	A-	A+	1.24	0.04	9.90	1.21	9.90	1.31
333	1035262	635247	18	4141	0.27	0.27	0.37	0.14	0.21	0.00	0.00	0.21	0.21	0.09	-0.23	-0.14	A+	A-	A-	A-	1.69	0.04	7.52	1.15	9.90	1.40
334	1035268	635253	18	4141	0.62	0.12	0.10	0.62	0.15	0.00	0.00	0.53	-0.21	-0.27	0.53	-0.30	A-	A+	A-	A-	-0.15	0.04	-9.90	0.86	-9.53	0.78
335	1035960	635945	18	4141	0.57	0.09	0.22	0.13	0.57	0.00	0.00	0.33	-0.20	-0.14	-0.14	0.33	A+	A-	A+	A-	0.15	0.03	4.69	1.06	5.17	1.11
336	1035961	635946	18	4141	0.56	0.24	0.56	0.10	0.10	0.00	0.00	0.40	-0.15	0.40	-0.20	-0.25	A-	A+	A+	A+	0.19	0.03	-0.01	1.00	-1.69	0.97
337	1021684	621669	19	4132	0.44	0.32	0.44	0.11	0.13	0.00	0.00	0.38	-0.09	0.38	-0.26	-0.19	A+	A+	A+	A-	0.77	0.03	1.12	1.02	1.27	1.02
338	1030618	630603	19	4132	0.61	0.61	0.09	0.20	0.11	0.00	0.00	0.42	0.42	-0.26	-0.16	-0.21	A-	A-	A+	A-	-0.05	0.03	-3.34	0.96	-1.83	0.96
339	1030652	630637	19	4132	0.62	0.18	0.14	0.62	0.06	0.00	0.00	0.51	-0.29	-0.25	0.51	-0.21	A+	A+	A-	A-	-0.12	0.04	-9.69	0.87	-9.39	0.80
340	1033592	633577	19	4132	0.25	0.25	0.25	0.25	0.25	0.00	0.00	0.04	-0.06	0.05	-0.02	0.04	A-	A-	A+	A+	1.80	0.04	9.90	1.29	9.90	1.69
341	1033596	633581	19	4132	0.55	0.55	0.15	0.21	0.09	0.00	0.00	0.45	0.45	-0.20	-0.19	-0.25	A+	A-	A-	A-	0.22	0.03	-4.27	0.95	-3.89	0.93
342	1033599	633584	19	4132	0.29	0.29	0.15	0.19	0.37	0.00	0.00	0.28	0.28	-0.21	-0.25	0.11	A+	A-	A-	A-	1.55	0.04	4.28	1.08	7.98	1.23
343	1033604	633589	19	4132	0.29	0.08	0.27	0.29	0.36	0.00	0.00	0.12	-0.21	-0.20	0.12	0.20	A+	A-	A+	A+	1.58	0.04	9.90	1.26	9.90	1.47
344	1034413	634398	19	4132	0.46	0.34	0.46	0.09	0.10	0.00	0.00	0.31	-0.02	0.31	-0.26	-0.24	A+	A+	A-	A-	0.65	0.03	6.33	1.09	5.51	1.10
345	1034414	634399	19	4132	0.25	0.19	0.29	0.27	0.25	0.00	0.00	0.23	0.01	-0.14	-0.09	0.23	A-	A+	A+	A-	1.82	0.04	5.07	1.11	8.46	1.30
346	1035249	635234	19	4132	0.54	0.13	0.18	0.14	0.54	0.00	0.00	0.54	-0.19	-0.26	-0.30	0.54	A+	A-	A-	A+	0.26	0.03	-9.90	0.85	-9.90	0.81
347	1035255	635240	19	4132	0.63	0.63	0.12	0.20	0.04	0.00	0.00	0.38	0.38	-0.25	-0.15	-0.21	A+	A+	A+	A-	-0.20	0.04	-0.69	0.99	-0.69	0.98
348	1035266	635251	19	4132	0.35	0.16	0.11	0.38	0.35	0.00	0.00	0.13	-0.18	-0.16	0.12	0.13	A-	A-	A+	A-	1.25	0.04	9.90	1.27	9.90	1.41
349	1035347	635332	19	4132	0.48	0.15	0.25	0.13	0.48	0.00	0.00	0.39	-0.32	-0.01	-0.22	0.39	A-	A-	A-	A+	0.58	0.03	0.94	1.01	2.45	1.04
350	1035478	635463	19	4132	0.41	0.25	0.15	0.41	0.19	0.00	0.00	0.39	-0.18	-0.21	0.39	-0.08	A-	A-	A+	A-	0.93	0.04	-0.15	1.00	1.96	1.04
351	1035962	635947	19	4132	0.64	0.04	0.18	0.14	0.64	0.00	0.00	0.50	-0.21	-0.31	-0.23	0.50	A-	A-	A-	A+	-0.22	0.04	-8.90	0.88	-8.34	0.81
352	1035963	635948	19	4132	0.39	0.39	0.15	0.28	0.17	0.00	0.00	0.40	0.40	-0.09	-0.21	-0.17	A-	A-	A+	A-	1.01	0.04	-0.12	1.00	1.11	1.02
353	1021872	621857	20	4116	0.74	0.07	0.09	0.09	0.74	0.00	0.00	0.41	-0.17	-0.23	-0.23	0.41	A+	A-	A+	A+	-0.76	0.04	-3.88	0.93	-3.28	0.89
354	1023638	623623	20	4116	0.57	0.22	0.12	0.57	0.08	0.00	0.00	0.39	-0.04	-0.30	0.39	-0.29	A-	A+	A-	A+	0.13	0.03	0.30	1.00	1.58	1.03
355	1024382	624367	20	4116	0.20	0.18	0.20	0.46	0.16	0.00	0.00	-0.08	-0.12	-0.08	0.26	-0.14	A-	A-	A+	A-	2.18	0.04	9.90	1.40	9.90	2.10
356	1030588	630573	20	4116	0.75	0.10	0.08	0.08	0.75	0.00	0.00	0.45	-0.25	-0.23	-0.21	0.45	A+	A-	A+	A-	-0.82	0.04	-5.95	0.89	-4.88	0.83
357	1030592	630577	20	4116	0.34	0.07	0.08	0.51	0.34	0.00	0.00	0.09	-0.25	-0.22	0.16	0.09	A+	A-	A-	A+	1.32	0.04	9.90	1.33	9.90	1.49
358	1030616	630601	20	4116	0.28	0.19	0.28	0.21	0.33	0.00	0.00	0.09	-0.16	0.09	-0.05	0.10	A+	A-	A-	A-	1.64	0.04	9.90	1.26	9.90	1.60
359	1030628	630613	20	4116	0.56	0.08	0.15	0.56	0.20	0.00	0.00	0.32	-0.22	-0.19	0.32	-0.06	A+	A+	A-	A+	0.17	0.03	5.80	1.08	5.28	1.11
360	1032278	632263	20	4116	0.61	0.06	0.61	0.14	0.19	0.00	0.00	0.36	-0.23	0.36	-0.25	-0.08	A-	A-	A+	A+	-0.04	0.04	2.46	1.03	2.44	1.06
361	1032421	632406	20	4116	0.55	0.55	0.29	0.11	0.05	0.00	0.00	0.41	0.41	-0.16	-0.26	-0.22	A-	A-	A-	A-	0.26	0.03	-0.39	0.99	-1.50	0.97
362	1033601	633586	20	4116	0.72	0.72	0.11	0.08	0.08	0.00	0.00	0.46	0.46	-0.24	-0.24	-0.24	A+	A-	A-	A+	-0.66	0.04	-6.03	0.90	-6.77	0.79
363	1035254	635239	20	4116	0.72	0.05	0.14	0.72	0.09	0.00	0.00	0.41	-0.20	-0.20	0.41	-0.25	A+	A-	A-	A-	-0.65	0.04	-3.02	0.95	-2.84	0.91

Table J–6 (continued). Biology Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
364	1035265	635250	20	4116	0.44	0.12	0.44	0.20	0.23	0.00	0.00	0.43	-0.09	0.43	-0.28	-0.16	A+	A-	A-	A-	0.77	0.03	-1.70	0.98	-1.37	0.97
365	1035267	635252	20	4116	0.60	0.15	0.13	0.12	0.60	0.00	0.00	0.50	-0.21	-0.27	-0.23	0.50	A+	A+	A+	A-	-0.02	0.04	-7.91	0.90	-6.88	0.85
366	1035474	635459	20	4116	0.56	0.09	0.10	0.56	0.24	0.00	0.00	0.44	-0.17	-0.19	0.44	-0.26	A-	A+	A+	A+	0.17	0.03	-3.04	0.96	-2.01	0.96
367	1035959	635944	20	4116	0.62	0.62	0.10	0.15	0.13	0.00	0.00	0.58	0.58	-0.25	-0.31	-0.28	A-	A-	A-	A-	-0.12	0.04	-9.90	0.81	-9.90	0.73
368	1035964	635949	20	4116	0.54	0.09	0.54	0.11	0.26	0.00	0.00	0.52	-0.22	0.52	-0.27	-0.25	A-	A-	A+	A-	0.31	0.03	-9.62	0.88	-8.79	0.84

Table J-7. Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
1	809472	409457	0	82141	0.50	0.21	0.13	0.50	0.15	0.00	0.00	0.32	-0.17	-0.25	0.32	-0.01					1.12	0.01	9.90	1.11	9.90	1.21
2	809473	409458	0	82141	0.71	0.11	0.71	0.13	0.04	0.00	0.00	0.46	-0.26	0.46	-0.19	-0.28					-0.08	0.01	-9.90	0.94	-8.39	0.94
3	809474	409459	0	82141	0.57	0.23	0.15	0.05	0.57	0.00	0.00	0.48	-0.21	-0.29	-0.21	0.48					0.54	0.01	-9.90	0.96	-9.90	0.94
4	809475	409460	0	82141	0.48	0.14	0.27	0.11	0.48	0.00	0.00	0.42	-0.19	-0.17	-0.21	0.42					0.86	0.01	5.72	1.02	8.25	1.04
5	809477	409462	0	82141	0.65	0.65	0.19	0.06	0.10	0.00	0.00	0.49	0.49	-0.21	-0.31	-0.25					-0.12	0.01	9.90	1.06	6.67	1.05
6	809479	409464	0	82141	0.72	0.07	0.07	0.72	0.14	0.00	0.00	0.49	-0.27	-0.25	0.49	-0.24					-0.11	0.01	-9.90	0.92	-9.90	0.84
7	809480	409465	0	82141	0.70	0.70	0.08	0.06	0.15	0.00	0.00	0.53	0.53	-0.29	-0.30	-0.24					-0.21	0.01	-9.90	0.94	-9.90	0.84
8	809481	409466	0	82141	0.73	0.73	0.09	0.12	0.05	0.00	0.00	0.51	0.51	-0.30	-0.23	-0.27					-0.42	0.01	-2.19	0.99	-9.90	0.90
9	824168	424153	0	82141	0.60	0.12	0.60	0.23	0.05	0.00	0.00	0.40	-0.28	0.40	-0.11	-0.25					0.34	0.01	9.90	1.07	9.90	1.08
10	824169	424154	0	82141	0.66	0.66	0.05	0.14	0.14	0.00	0.00	0.48	0.48	-0.24	-0.31	-0.17					0.17	0.01	-9.90	0.96	-9.90	0.90
11	824170	424155	0	82141	0.66	0.66	0.16	0.09	0.08	0.00	0.00	0.40	0.40	-0.14	-0.25	-0.23					0.19	0.01	7.35	1.03	7.43	1.05
12	824171	424156	0	82141	0.51	0.02	0.37	0.09	0.51	0.00	0.00	0.32	-0.20	-0.06	-0.34	0.32					0.91	0.01	9.90	1.11	9.90	1.16
13	824172	424157	0	82141	0.47	0.16	0.47	0.29	0.07	0.00	0.00	0.31	-0.22	0.31	-0.05	-0.17					1.27	0.01	9.90	1.10	9.90	1.22
14	824173	424158	0	82141	0.66	0.08	0.24	0.02	0.66	0.00	0.00	0.35	-0.25	-0.15	-0.24	0.35					-0.07	0.01	9.90	1.18	9.90	1.23
15	824175	424160	0	82141	0.86	0.06	0.04	0.86	0.04	0.00	0.00	0.48	-0.26	-0.27	0.48	-0.25					-1.40	0.01	2.36	1.02	-9.90	0.79
16	824177	424162	0	82141	0.78	0.04	0.14	0.78	0.04	0.00	0.00	0.45	-0.25	-0.25	0.45	-0.24					-0.63	0.01	-2.72	0.99	-9.72	0.91
17	824219	424204	0	82141	0.41	0.21	0.41	0.21	0.17	0.00	0.00	0.15	0.01	0.15	-0.12	-0.07					1.59	0.01	9.90	1.29	9.90	1.53
18	980172	580157	0	82141	0.53	0.04	0.53	0.08	0.35	0.00	0.00	0.33	-0.27	0.33	-0.22	-0.10					1.23	0.01	9.90	1.09	9.90	1.19
19	980174	580159	0	82141	0.73	0.16	0.05	0.73	0.06	0.00	0.00	0.42	-0.22	-0.26	0.42	-0.21					-0.24	0.01	-2.30	0.99	-6.16	0.95
20	980176	580161	0	82141	0.65	0.65	0.11	0.23	0.02	0.00	0.00	0.39	0.39	-0.11	-0.31	-0.16					0.26	0.01	9.90	1.04	1.81	1.01
21	980177	580162	0	82141	0.65	0.16	0.65	0.15	0.03	0.00	0.00	0.31	-0.14	0.31	-0.16	-0.20					0.30	0.01	9.90	1.11	9.90	1.14
22	980178	580163	0	82141	0.80	0.09	0.04	0.80	0.07	0.00	0.00	0.41	-0.21	-0.24	0.41	-0.22					-0.58	0.01	-9.90	0.92	-9.23	0.91
23	980180	580165	0	82141	0.66	0.66	0.11	0.15	0.09	0.00	0.00	0.36	0.36	-0.11	-0.26	-0.16					0.34	0.01	9.90	1.05	9.90	1.06
24	980181	580166	0	82141	0.84	0.06	0.04	0.84	0.06	0.00	0.00	0.42	-0.26	-0.25	0.42	-0.18					-1.10	0.01	-4.98	0.97	-9.31	0.88
25	980183	580168	0	82141	0.52	0.52	0.11	0.31	0.06	0.00	0.00	0.19	0.19	-0.15	-0.05	-0.11					1.04	0.01	9.90	1.25	9.90	1.35
26	980184	580169	0	82141	0.68	0.07	0.68	0.20	0.05	0.00	0.00	0.38	-0.18	0.38	-0.19	-0.26					-0.09	0.01	9.90	1.10	9.90	1.11
27	986943	586928	0	82141	0.54	0.34	0.04	0.54	0.07	0.00	0.00	0.38	-0.19	-0.17	0.38	-0.26					1.10	0.01	9.90	1.04	9.90	1.08
28	986944	586929	0	82141	0.62	0.09	0.03	0.27	0.62	0.00	0.00	0.31	-0.24	-0.21	-0.11	0.31					0.31	0.01	9.90	1.15	9.90	1.26
29	986945	586930	0	82141	0.83	0.83	0.02	0.11	0.04	0.00	0.00	0.33	0.33	-0.20	-0.21	-0.16					-0.84	0.01	-0.90	0.99	2.98	1.03
30	986947	586932	0	82141	0.51	0.09	0.05	0.34	0.51	0.00	0.00	0.25	-0.14	-0.17	-0.10	0.25					1.54	0.01	9.90	1.26	9.90	1.47
31	986948	586933	0	82141	0.89	0.89	0.05	0.03	0.03	0.00	0.00	0.44	0.44	-0.28	-0.25	-0.20					-1.27	0.01	-9.90	0.73	-9.90	0.58
32	986951	586936	0	82141	0.61	0.18	0.61	0.11	0.09	0.00	0.00	0.36	-0.08	0.36	-0.25	-0.22					0.79	0.01	9.90	1.05	9.90	1.09
33	986952	586937	0	82141	0.77	0.04	0.12	0.77	0.07	0.00	0.00	0.23	-0.09	-0.16	0.23	-0.11					-0.15	0.01	9.90	1.07	9.90	1.18

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
34	987009	586994	0	82141	0.67	0.11	0.12	0.67	0.10	0.00	0.00	0.44	-0.30	-0.12	0.44	-0.25					0.55	0.01	-9.90	0.93	-9.90	0.91
35	1021304	621289	1	4159	0.58	0.23	0.09	0.58	0.11	0.00	0.00	0.34	-0.09	-0.21	0.34	-0.22	A-	A-	A-	A-	0.65	0.04	6.63	1.10	6.27	1.15
36	1021307	621292	1	4159	0.49	0.11	0.49	0.18	0.22	0.00	0.00	0.28	-0.12	0.28	-0.20	-0.06	A+	A+	A+	A+	1.08	0.03	9.90	1.15	9.82	1.22
37	1021308	621293	1	4159	0.32	0.57	0.06	0.05	0.32	0.00	0.00	0.08	0.14	-0.21	-0.27	0.08	A-	A+	A+	A-	2.00	0.04	9.90	1.33	9.90	1.67
38	1021312	621297	1	4159	0.89	0.07	0.01	0.03	0.89	0.00	0.00	0.41	-0.25	-0.23	-0.23	0.41	A-	A-	A-	A-	-1.55	0.05	-3.38	0.88	-3.18	0.79
39	1021314	621299	1	4159	0.79	0.04	0.79	0.11	0.05	0.00	0.00	0.26	-0.15	0.26	-0.16	-0.10	A+	A+	A+	A-	-0.64	0.04	4.74	1.11	7.21	1.35
40	1021317	621302	1	4159	0.74	0.05	0.74	0.19	0.03	0.00	0.00	0.40	-0.22	0.40	-0.26	-0.17	C-	A-	A+	A-	-0.25	0.04	0.29	1.01	-1.47	0.95
41	1032810	632795	1	4159	0.67	0.04	0.16	0.67	0.14	0.00	0.00	0.52	-0.24	-0.31	0.52	-0.24	A+	A-	A+	A-	0.16	0.04	-6.69	0.89	-5.67	0.85
42	1032812	632797	1	4159	0.77	0.77	0.07	0.13	0.03	0.00	0.00	0.49	0.49	-0.25	-0.29	-0.25	A+	A-	A+	A+	-0.45	0.04	-4.74	0.90	-4.85	0.82
43	1032814	632799	1	4159	0.80	0.08	0.80	0.08	0.03	0.00	0.00	0.44	-0.28	0.44	-0.23	-0.21	A-	A+	A+	A+	-0.70	0.04	-2.83	0.93	-3.11	0.86
44	1032817	632802	1	4159	0.77	0.03	0.06	0.14	0.77	0.00	0.00	0.49	-0.22	-0.26	-0.30	0.49	A+	A-	A-	A-	-0.47	0.04	-4.93	0.90	-4.16	0.84
45	1032818	632803	1	4159	0.81	0.81	0.05	0.03	0.11	0.00	0.00	0.48	0.48	-0.31	-0.29	-0.22	A-	A-	A-	A+	-0.75	0.04	-4.84	0.89	-4.51	0.80
46	1032882	632867	1	4159	0.47	0.10	0.10	0.47	0.32	0.00	0.00	0.20	-0.21	-0.26	0.20	0.09	A-	A-	A-	A-	1.18	0.03	9.90	1.24	9.90	1.37
47	1021303	621288	2	4109	0.66	0.05	0.22	0.66	0.08	0.00	0.00	0.38	-0.11	-0.27	0.38	-0.17	A-	A+	A-	A-	0.22	0.04	2.71	1.05	2.34	1.07
48	1021305	621290	2	4109	0.66	0.66	0.16	0.07	0.10	0.00	0.00	0.34	0.34	-0.07	-0.18	-0.30	A-	A-	A-	A-	0.18	0.04	4.78	1.08	6.99	1.21
49	1021310	621295	2	4109	0.74	0.06	0.04	0.16	0.74	0.00	0.00	0.45	-0.24	-0.25	-0.25	0.45	A-	A-	A-	A-	-0.27	0.04	-3.02	0.94	-1.28	0.95
50	1021315	621300	2	4109	0.56	0.56	0.31	0.09	0.04	0.00	0.00	0.26	0.26	0.00	-0.29	-0.22	A-	A-	A+	A-	0.73	0.04	9.90	1.19	9.90	1.26
51	1021319	621304	2	4109	0.95	0.01	0.02	0.95	0.02	0.00	0.00	0.39	-0.20	-0.24	0.39	-0.22	A-	C-	B-	A+	-2.41	0.07	-2.54	0.86	-6.63	0.43
52	1021422	621407	2	4109	0.70	0.22	0.07	0.02	0.70	0.00	0.00	0.45	-0.26	-0.26	-0.23	0.45	A-	A-	A+	A+	-0.01	0.04	-2.49	0.96	-1.69	0.95
53	1032809	632794	2	4109	0.92	0.03	0.03	0.92	0.02	0.00	0.00	0.39	-0.28	-0.13	0.39	-0.23	A+	A+	A+	A+	-1.89	0.06	-2.48	0.90	-3.96	0.69
54	1032811	632796	2	4109	0.83	0.07	0.07	0.83	0.02	0.00	0.00	0.46	-0.32	-0.21	0.46	-0.23	A+	A+	A+	B+	-0.97	0.05	-3.94	0.90	-4.38	0.78
55	1032815	632800	2	4109	0.38	0.38	0.05	0.16	0.42	0.00	0.00	0.11	0.11	-0.26	-0.18	0.14	A-	A-	A-	A+	1.66	0.04	9.90	1.29	9.90	1.61
56	1032816	632801	2	4109	0.77	0.13	0.77	0.05	0.05	0.00	0.00	0.47	-0.22	0.47	-0.31	-0.27	A-	A-	A-	B+	-0.46	0.04	-3.86	0.92	-4.32	0.83
57	1032819	632804	2	4109	0.69	0.12	0.69	0.06	0.13	0.00	0.00	0.45	-0.28	0.45	-0.27	-0.16	A-	A-	A-	A+	0.01	0.04	-2.08	0.96	-1.36	0.96
58	1035939	635924	2	4109	0.94	0.02	0.94	0.03	0.01	0.00	0.00	0.38	-0.18	0.38	-0.26	-0.19	A-	A+	A+	A+	-2.29	0.07	-2.54	0.87	-3.86	0.65
59	1034523	634508	3	4090	0.84	0.05	0.03	0.08	0.84	0.00	0.00	0.54	-0.30	-0.29	-0.29	0.54	A+	A-	A-	A+	-0.92	0.05	-7.07	0.82	-8.98	0.60
60	1034529	634514	3	4090	0.27	0.49	0.27	0.10	0.14	0.00	0.00	0.22	0.01	0.22	-0.20	-0.12	A-	A-	A+	A+	2.32	0.04	3.82	1.07	9.90	1.53
61	1034531	634516	3	4090	0.79	0.05	0.09	0.06	0.79	0.00	0.00	0.41	-0.25	-0.16	-0.27	0.41	A-	A-	A-	A+	-0.56	0.04	-2.18	0.95	2.28	1.10
62	1034532	634517	3	4090	0.77	0.09	0.06	0.09	0.77	0.00	0.00	0.55	-0.28	-0.31	-0.28	0.55	A+	A-	A+	A-	-0.42	0.04	-7.89	0.84	-8.24	0.71
63	1034533	634518	3	4090	0.69	0.05	0.69	0.20	0.06	0.00	0.00	0.42	-0.28	0.42	-0.17	-0.28	A+	A-	A-	A+	0.08	0.04	-0.66	0.99	1.14	1.03
64	1035312	635297	3	4090	0.70	0.09	0.13	0.70	0.08	0.00	0.00	0.26	-0.19	-0.15	0.26	-0.05	A-	A+	A-	A-	0.02	0.04	8.63	1.17	8.66	1.29
65	1035317	635302	3	4090	0.78	0.12	0.04	0.78	0.05	0.00	0.00	0.37	-0.18	-0.22	0.37	-0.21	A+	A+	A-	A-	-0.49	0.04	0.78	1.02	-0.37	0.98
66	1035318	635303	3	4090	0.36	0.03	0.36	0.30	0.31	0.00	0.00	0.12	-0.13	0.12	-0.14	0.06	A+	A+	A-	A+	1.80	0.04	9.90	1.28	9.90	1.55

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
67	1035319	635304	3	4090	0.82	0.04	0.82	0.09	0.04	0.00	0.00	0.24	-0.18	0.24	-0.05	-0.19	A+	A-	A+	A-	-0.81	0.04	4.04	1.11	6.81	1.38
68	1035320	635305	3	4090	0.27	0.44	0.11	0.27	0.18	0.00	0.00	0.15	-0.07	-0.05	0.15	-0.03	A-	A-	A-	A+	2.32	0.04	9.66	1.19	9.90	1.60
69	1036717	636702	3	4090	0.89	0.05	0.89	0.03	0.03	0.00	0.00	0.40	-0.26	0.40	-0.19	-0.20	A+	A-	A-	B+	-1.42	0.05	-2.48	0.92	-3.37	0.78
70	1036820	636805	3	4090	0.58	0.28	0.10	0.58	0.04	0.00	0.00	0.29	-0.12	-0.14	0.29	-0.25	A-	B-	B-	A-	0.66	0.04	8.82	1.14	8.43	1.20
71	1034522	634507	4	4122	0.83	0.07	0.04	0.83	0.05	0.00	0.00	0.45	-0.22	-0.26	0.45	-0.25	A-	A+	A-	A-	-0.90	0.05	-3.36	0.91	-2.47	0.87
72	1034524	634509	4	4122	0.85	0.05	0.04	0.85	0.06	0.00	0.00	0.52	-0.27	-0.28	0.52	-0.29	A-	A-	A-	A+	-1.07	0.05	-5.98	0.84	-7.08	0.64
73	1034525	634510	4	4122	0.71	0.14	0.71	0.10	0.05	0.00	0.00	0.39	-0.22	0.39	-0.23	-0.13	A-	A-	A-	A+	-0.06	0.04	2.36	1.05	-0.01	1.00
74	1034528	634513	4	4122	0.51	0.12	0.13	0.24	0.51	0.00	0.00	0.35	-0.21	-0.22	-0.08	0.35	A-	A-	A+	A-	1.02	0.04	4.84	1.07	5.54	1.12
75	1034530	634515	4	4122	0.34	0.13	0.34	0.34	0.18	0.00	0.00	0.25	-0.09	-0.25	0.25	0.09	A-	A-	A+	A+	1.90	0.04	8.57	1.14	9.90	1.42
76	1034534	634519	4	4122	0.79	0.79	0.08	0.05	0.07	0.00	0.00	0.53	0.53	-0.26	-0.27	-0.31	A+	A-	A-	A-	-0.59	0.04	-6.92	0.85	-6.45	0.74
77	1035313	635298	4	4122	0.38	0.38	0.08	0.11	0.42	0.00	0.00	0.19	0.19	-0.13	-0.07	-0.06	A-	A-	A+	A+	1.71	0.04	9.90	1.25	9.90	1.47
78	1035314	635299	4	4122	0.61	0.61	0.05	0.25	0.08	0.00	0.00	0.25	0.25	-0.29	-0.04	-0.15	A-	A-	A+	A-	0.48	0.04	9.90	1.22	9.90	1.34
79	1035315	635300	4	4122	0.81	0.07	0.10	0.81	0.02	0.00	0.00	0.32	-0.26	-0.15	0.32	-0.09	A+	A+	A-	A+	-0.75	0.04	2.52	1.06	3.55	1.18
80	1035316	635301	4	4122	0.85	0.08	0.04	0.02	0.85	0.00	0.00	0.44	-0.26	-0.26	-0.21	0.44	A+	A+	A-	A-	-1.13	0.05	-3.60	0.90	-2.35	0.86
81	1035323	635308	4	4122	0.83	0.06	0.06	0.04	0.83	0.00	0.00	0.42	-0.18	-0.19	-0.32	0.42	A+	A-	A-	A+	-0.93	0.05	-2.34	0.94	-0.10	0.99
82	1035324	635309	4	4122	0.88	0.02	0.88	0.06	0.03	0.00	0.00	0.36	-0.22	0.36	-0.16	-0.23	A+	A-	A-	A-	-1.45	0.05	-1.28	0.96	2.01	1.15
83	1035625	635610	5	4111	0.74	0.10	0.03	0.74	0.12	0.00	0.00	0.46	-0.30	-0.25	0.46	-0.20	A-	A-	A+	A-	-0.30	0.04	-2.99	0.94	-3.02	0.89
84	1035626	635611	5	4111	0.61	0.11	0.18	0.11	0.61	0.00	0.00	0.34	-0.17	-0.17	-0.15	0.34	A+	A+	A+	A-	0.49	0.04	6.80	1.11	6.23	1.16
85	1035629	635614	5	4111	0.82	0.82	0.04	0.08	0.05	0.00	0.00	0.45	0.45	-0.22	-0.28	-0.21	A-	A-	A-	A-	-0.88	0.04	-3.08	0.92	-4.45	0.78
86	1035630	635615	5	4111	0.88	0.06	0.88	0.04	0.02	0.00	0.00	0.38	-0.23	0.38	-0.25	-0.13	A-	A+	A+	A+	-1.45	0.05	-2.20	0.93	-1.74	0.88
87	1035634	635619	5	4111	0.77	0.10	0.77	0.04	0.08	0.00	0.00	0.44	-0.22	0.44	-0.23	-0.26	A-	A+	A-	A-	-0.48	0.04	-2.46	0.95	-3.07	0.87
88	1036551	636536	5	4111	0.47	0.06	0.09	0.38	0.47	0.00	0.00	0.18	-0.24	-0.29	0.10	0.18	A-	A-	A-	A+	1.22	0.04	9.90	1.29	9.90	1.45
89	1036552	636537	5	4111	0.52	0.18	0.22	0.52	0.08	0.00	0.00	0.30	-0.05	-0.15	0.30	-0.25	A+	A-	A-	A-	0.95	0.04	9.90	1.16	8.72	1.20
90	1036553	636538	5	4111	0.77	0.10	0.77	0.08	0.05	0.00	0.00	0.47	-0.21	0.47	-0.31	-0.23	A+	A-	A-	A+	-0.47	0.04	-3.92	0.92	-2.56	0.89
91	1036554	636539	5	4111	0.63	0.10	0.63	0.06	0.21	0.00	0.00	0.46	-0.23	0.46	-0.24	-0.24	A-	A+	A-	A+	0.38	0.04	-2.46	0.96	-4.39	0.89
92	1036558	636543	5	4111	0.64	0.64	0.08	0.14	0.13	0.00	0.00	0.43	0.43	-0.32	-0.15	-0.20	B-	A-	A-	A+	0.30	0.04	-0.11	1.00	-0.28	0.99
93	1037049	637034	5	4111	0.59	0.06	0.21	0.59	0.14	0.00	0.00	0.33	-0.10	-0.20	0.33	-0.17	A-	A-	A-	A+	0.58	0.04	7.53	1.12	7.62	1.19
94	1037172	637157	5	4111	0.77	0.09	0.77	0.08	0.05	0.00	0.00	0.34	-0.17	0.34	-0.14	-0.25	A-	A-	A-	A+	-0.49	0.04	2.63	1.06	3.15	1.14
95	1035622	635607	6	4109	0.73	0.11	0.12	0.04	0.73	0.00	0.00	0.43	-0.28	-0.17	-0.23	0.43	A-	A-	A-	A+	-0.21	0.04	-1.15	0.98	-0.09	1.00
96	1035623	635608	6	4109	0.73	0.12	0.10	0.73	0.06	0.00	0.00	0.43	-0.20	-0.27	0.43	-0.19	B+	A+	A+	A+	-0.22	0.04	-0.52	0.99	-2.04	0.93
97	1035624	635609	6	4109	0.69	0.16	0.04	0.69	0.10	0.00	0.00	0.43	-0.21	-0.24	0.43	-0.23	A+	A-	A-	A-	-0.01	0.04	-0.31	0.99	-1.25	0.96
98	1035627	635612	6	4109	0.92	0.92	0.03	0.02	0.02	0.00	0.00	0.41	0.41	-0.23	-0.25	-0.22	A+	A+	A+	A-	-2.02	0.06	-3.61	0.84	-4.06	0.67
99	1035631	635616	6	4109	0.72	0.06	0.05	0.72	0.18	0.00	0.00	0.24	-0.21	-0.13	0.24	-0.07	A+	A+	A+	A-	-0.15	0.04	9.90	1.21	9.07	1.34

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
100	1035632	635617	6	4109	0.49	0.23	0.49	0.10	0.18	0.00	0.00	0.23	-0.04	0.23	-0.20	-0.10	A+	A-	A-	A-	1.08	0.04	9.90	1.23	9.90	1.34
101	1036547	636532	6	4109	0.69	0.10	0.15	0.06	0.69	0.00	0.00	0.48	-0.27	-0.20	-0.28	0.48	B-	A-	A-	A-	0.01	0.04	-3.89	0.93	-2.90	0.91
102	1036548	636533	6	4109	0.60	0.27	0.08	0.60	0.05	0.00	0.00	0.43	-0.17	-0.30	0.43	-0.23	A-	A-	A-	A-	0.51	0.04	-0.20	1.00	-0.36	0.99
103	1036549	636534	6	4109	0.55	0.08	0.55	0.23	0.14	0.00	0.00	0.35	-0.22	0.35	-0.17	-0.10	A-	A-	A+	A+	0.80	0.04	6.46	1.10	6.90	1.16
104	1036550	636535	6	4109	0.77	0.05	0.10	0.08	0.77	0.00	0.00	0.51	-0.23	-0.27	-0.30	0.51	A-	B-	B-	A+	-0.50	0.04	-5.54	0.88	-6.27	0.76
105	1036556	636541	6	4109	0.89	0.01	0.07	0.89	0.03	0.00	0.00	0.30	-0.17	-0.16	0.30	-0.18	A-	A+	A-	A+	-1.49	0.05	0.00	1.00	2.02	1.15
106	1036559	636544	6	4109	0.63	0.26	0.63	0.09	0.03	0.00	0.00	0.29	-0.08	0.29	-0.27	-0.19	A-	B-	A-	A+	0.37	0.04	9.45	1.16	9.49	1.27
107	1032798	632783	7	4075	0.57	0.09	0.11	0.57	0.23	0.00	0.00	0.31	-0.17	-0.16	0.31	-0.13	A-	A-	A-	A+	0.66	0.04	7.30	1.11	6.79	1.15
108	1032799	632784	7	4075	0.78	0.04	0.15	0.78	0.03	0.00	0.00	0.42	-0.25	-0.22	0.42	-0.28	A+	A+	B+	A+	-0.56	0.04	-2.65	0.94	-1.17	0.95
109	1032804	632789	7	4075	0.72	0.08	0.09	0.72	0.10	0.00	0.00	0.43	-0.27	-0.18	0.43	-0.21	A+	A-	A-	A+	-0.15	0.04	-2.29	0.96	-2.22	0.93
110	1032805	632790	7	4075	0.51	0.10	0.20	0.19	0.51	0.00	0.00	0.41	-0.17	-0.23	-0.14	0.41	A-	B-	B-	A-	0.97	0.04	-0.23	1.00	0.66	1.01
111	1032806	632791	7	4075	0.37	0.37	0.34	0.21	0.08	0.00	0.00	-0.04	-0.04	0.08	0.09	-0.21	A-	A-	A+	A-	1.68	0.04	9.90	1.50	9.90	1.78
112	1035328	635313	7	4075	0.71	0.71	0.07	0.15	0.06	0.00	0.00	0.29	0.29	-0.14	-0.20	-0.09	A-	A-	A-	A-	-0.10	0.04	6.13	1.12	4.69	1.16
113	1035331	635316	7	4075	0.69	0.69	0.18	0.03	0.10	0.00	0.00	0.41	0.41	-0.14	-0.23	-0.31	A+	A-	A-	A-	0.03	0.04	-0.36	0.99	0.68	1.02
114	1035332	635317	7	4075	0.45	0.07	0.19	0.29	0.45	0.00	0.00	0.35	-0.22	-0.12	-0.15	0.35	A+	A-	A-	A-	1.30	0.04	3.48	1.05	4.42	1.09
115	1035334	635319	7	4075	0.45	0.09	0.45	0.41	0.05	0.00	0.00	0.25	-0.20	0.25	-0.01	-0.26	A-	A-	A+	A+	1.27	0.04	9.90	1.16	9.90	1.25
116	1035335	635320	7	4075	0.74	0.03	0.74	0.12	0.11	0.00	0.00	0.31	-0.21	0.31	-0.08	-0.23	A-	A-	A+	A-	-0.30	0.04	4.15	1.09	2.69	1.10
117	1035337	635322	7	4075	0.63	0.28	0.63	0.04	0.05	0.00	0.00	0.25	-0.10	0.25	-0.24	-0.14	A-	B+	A-	A+	0.37	0.04	9.90	1.17	9.90	1.27
118	1035943	635928	7	4075	0.70	0.18	0.08	0.70	0.04	0.00	0.00	0.41	-0.22	-0.27	0.41	-0.15	B-	A-	A-	A+	-0.03	0.04	-1.01	0.98	0.18	1.01
119	1032795	632780	8	4095	0.35	0.04	0.47	0.14	0.35	0.00	0.00	0.15	-0.25	0.04	-0.14	0.15	A+	A-	A+	A+	1.83	0.04	9.90	1.27	9.90	1.54
120	1032796	632781	8	4095	0.81	0.07	0.07	0.05	0.81	0.00	0.00	0.49	-0.24	-0.28	-0.28	0.49	A-	A-	A-	A+	-0.74	0.04	-5.17	0.88	-6.10	0.73
121	1032800	632785	8	4095	0.82	0.82	0.08	0.07	0.03	0.00	0.00	0.45	0.45	-0.26	-0.22	-0.27	A-	A-	A-	A+	-0.85	0.04	-3.86	0.90	-2.25	0.89
122	1032801	632786	8	4095	0.81	0.11	0.03	0.81	0.05	0.00	0.00	0.46	-0.27	-0.23	0.46	-0.23	C-	A-	B-	A+	-0.72	0.04	-3.31	0.92	-4.24	0.81
123	1032803	632788	8	4095	0.68	0.08	0.68	0.20	0.04	0.00	0.00	0.24	-0.17	0.24	-0.06	-0.22	A+	A-	A-	A+	0.10	0.04	9.90	1.20	9.90	1.37
124	1032807	632792	8	4095	0.34	0.34	0.32	0.03	0.31	0.00	0.00	0.20	0.20	0.01	-0.28	-0.11	B-	A-	A-	A+	1.92	0.04	9.90	1.17	9.90	1.51
125	1035325	635310	8	4095	0.82	0.08	0.82	0.05	0.05	0.00	0.00	0.46	-0.26	0.46	-0.26	-0.23	A+	A+	A+	A-	-0.81	0.04	-3.72	0.91	-4.41	0.79
126	1035327	635312	8	4095	0.76	0.12	0.07	0.76	0.05	0.00	0.00	0.31	-0.17	-0.14	0.31	-0.19	A-	A-	A+	A+	-0.37	0.04	3.69	1.08	6.73	1.29
127	1035329	635314	8	4095	0.83	0.83	0.03	0.12	0.01	0.00	0.00	0.27	0.27	-0.20	-0.16	-0.12	A-	A-	A-	A+	-0.95	0.05	2.64	1.07	7.42	1.47
128	1035330	635315	8	4095	0.64	0.10	0.12	0.64	0.14	0.00	0.00	0.43	-0.21	-0.17	0.43	-0.24	A-	A+	A-	A-	0.31	0.04	-0.13	1.00	-1.99	0.95
129	1035333	635318	8	4095	0.71	0.03	0.06	0.20	0.71	0.00	0.00	0.37	-0.26	-0.16	-0.22	0.37	A+	A+	A+	A-	-0.08	0.04	2.65	1.05	1.91	1.07
130	1035336	635321	8	4095	0.78	0.05	0.06	0.78	0.11	0.00	0.00	0.48	-0.24	-0.21	0.48	-0.31	A+	A-	A-	A-	-0.50	0.04	-4.70	0.90	-3.92	0.84
131	1025655	625640	9	4129	0.87	0.87	0.09	0.03	0.01	0.00	0.00	0.41	0.41	-0.23	-0.27	-0.21	A-	A+	A-	A+	-1.28	0.05	-2.70	0.92	-2.88	0.82
132	1025660	625645	9	4129	0.68	0.06	0.07	0.18	0.68	0.00	0.00	0.40	-0.16	-0.18	-0.25	0.40	A+	A+	A+	A+	0.08	0.04	1.35	1.02	2.24	1.07

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
133	1025662	625647	9	4129	0.49	0.07	0.15	0.49	0.29	0.00	0.00	0.28	-0.11	-0.32	0.28	0.01	A-	A+	A+	A-	1.10	0.04	9.90	1.16	9.90	1.25
134	1025666	625651	9	4129	0.69	0.07	0.06	0.69	0.19	0.00	0.00	0.43	-0.24	-0.18	0.43	-0.24	A+	A-	A+	A-	0.05	0.04	-0.38	0.99	0.07	1.00
135	1025669	625654	9	4129	0.92	0.92	0.03	0.02	0.04	0.00	0.00	0.43	0.43	-0.30	-0.23	-0.20	B+	A-	A+	A-	-1.86	0.06	-3.81	0.84	-5.72	0.58
136	1025670	625655	9	4129	0.48	0.25	0.15	0.12	0.48	0.00	0.00	0.24	0.10	-0.20	-0.27	0.24	A-	A-	A+	A-	1.15	0.04	9.90	1.21	9.90	1.34
137	1034510	634495	9	4129	0.91	0.02	0.91	0.03	0.03	0.00	0.00	0.43	-0.19	0.43	-0.26	-0.26	A-	A-	A-	A+	-1.81	0.06	-3.57	0.86	-5.49	0.60
138	1034511	634496	9	4129	0.63	0.15	0.07	0.63	0.15	0.00	0.00	0.39	-0.31	-0.27	0.39	-0.03	A-	A+	A-	A-	0.37	0.04	2.44	1.04	2.75	1.07
139	1034512	634497	9	4129	0.65	0.03	0.65	0.25	0.07	0.00	0.00	0.38	-0.19	0.38	-0.22	-0.18	A-	A-	A-	A+	0.28	0.04	3.27	1.05	2.44	1.07
140	1034515	634500	9	4129	0.65	0.05	0.20	0.10	0.65	0.00	0.00	0.41	-0.20	-0.19	-0.25	0.41	A-	A-	A-	A+	0.28	0.04	0.83	1.01	1.39	1.04
141	1034519	634504	9	4129	0.83	0.83	0.03	0.11	0.03	0.00	0.00	0.49	0.49	-0.24	-0.30	-0.28	B-	A+	A-	A-	-0.91	0.05	-5.25	0.87	-5.66	0.73
142	1034520	634505	9	4129	0.59	0.27	0.08	0.59	0.06	0.00	0.00	0.45	-0.25	-0.23	0.45	-0.20	A-	A-	A-	A+	0.58	0.04	-2.65	0.96	-2.39	0.95
143	1025654	625639	10	4107	0.33	0.20	0.27	0.20	0.33	0.00	0.00	0.29	-0.25	0.02	-0.11	0.29	A+	C-	A-	A-	1.98	0.04	6.30	1.10	7.32	1.23
144	1025656	625641	10	4107	0.69	0.02	0.12	0.69	0.15	0.00	0.00	0.42	-0.26	-0.24	0.42	-0.20	A-	A-	A+	B-	0.04	0.04	0.39	1.01	-1.58	0.95
145	1025658	625643	10	4107	0.48	0.34	0.48	0.06	0.13	0.00	0.00	0.13	0.17	0.13	-0.27	-0.24	A-	A+	A-	A+	1.19	0.04	9.90	1.36	9.90	1.51
146	1025664	625649	10	4107	0.78	0.78	0.07	0.04	0.12	0.00	0.00	0.37	0.37	-0.30	-0.20	-0.12	A-	A+	A-	A+	-0.48	0.04	0.96	1.02	2.11	1.09
147	1025668	625653	10	4107	0.81	0.05	0.81	0.05	0.09	0.00	0.00	0.44	-0.23	0.44	-0.24	-0.24	A-	A+	A+	A-	-0.75	0.04	-2.45	0.94	-3.60	0.83
148	1034509	634494	10	4107	0.54	0.18	0.21	0.07	0.54	0.00	0.00	0.41	-0.21	-0.19	-0.19	0.41	A-	A+	A-	A-	0.89	0.04	1.09	1.02	2.76	1.06
149	1034513	634498	10	4107	0.55	0.55	0.33	0.04	0.08	0.00	0.00	0.29	0.29	-0.04	-0.28	-0.26	A+	A+	A+	A+	0.82	0.04	9.90	1.16	9.90	1.23
150	1034514	634499	10	4107	0.76	0.10	0.08	0.76	0.05	0.00	0.00	0.45	-0.22	-0.31	0.45	-0.17	A-	A-	A-	A+	-0.40	0.04	-2.96	0.94	-1.79	0.93
151	1034517	634502	10	4107	0.54	0.13	0.54	0.17	0.16	0.00	0.00	0.32	-0.11	0.32	-0.17	-0.16	A-	A-	A-	A+	0.87	0.04	7.92	1.12	6.95	1.15
152	1034518	634503	10	4107	0.57	0.09	0.10	0.23	0.57	0.00	0.00	0.44	-0.22	-0.18	-0.24	0.44	A-	A-	B-	A-	0.71	0.04	-1.44	0.98	-0.57	0.99
153	1035969	635954	10	4107	0.90	0.04	0.90	0.02	0.05	0.00	0.00	0.37	-0.16	0.37	-0.18	-0.27	B-	A-	C-	B+	-1.56	0.05	-1.52	0.94	-3.33	0.77
154	1036196	636181	10	4107	0.95	0.03	0.95	0.02	0.01	0.00	0.00	0.35	-0.22	0.35	-0.22	-0.16	A+	A-	A-	A+	-2.36	0.07	-2.06	0.89	-2.98	0.70
155	1025641	625626	11	4108	0.66	0.03	0.18	0.13	0.66	0.00	0.00	0.22	-0.25	-0.21	0.05	0.22	A+	A+	A-	A+	0.21	0.04	9.90	1.22	9.90	1.44
156	1025643	625628	11	4108	0.68	0.18	0.68	0.06	0.08	0.00	0.00	0.39	-0.14	0.39	-0.22	-0.27	A+	A-	A-	A+	0.09	0.04	1.93	1.03	1.86	1.05
157	1025645	625630	11	4108	0.50	0.16	0.31	0.03	0.50	0.00	0.00	0.39	-0.18	-0.20	-0.18	0.39	A-	A+	A+	A-	1.06	0.04	1.99	1.03	3.09	1.07
158	1025647	625632	11	4108	0.80	0.09	0.05	0.80	0.06	0.00	0.00	0.33	-0.18	-0.14	0.33	-0.22	A-	A-	A-	A+	-0.70	0.04	1.62	1.04	2.87	1.14
159	1025651	625636	11	4108	0.63	0.63	0.23	0.09	0.05	0.00	0.00	0.41	0.41	-0.15	-0.26	-0.26	A-	A+	A-	A+	0.34	0.04	0.71	1.01	1.55	1.04
160	1034536	634521	11	4108	0.66	0.05	0.06	0.23	0.66	0.00	0.00	0.51	-0.27	-0.19	-0.32	0.51	A+	A-	A-	A+	0.21	0.04	-6.09	0.90	-6.15	0.84
161	1034539	634524	11	4108	0.40	0.03	0.13	0.44	0.40	0.00	0.00	0.22	-0.24	-0.20	0.01	0.22	A+	A+	A-	A-	1.55	0.04	9.90	1.16	9.90	1.38
162	1034540	634525	11	4108	0.30	0.14	0.30	0.39	0.16	0.00	0.00	-0.01	-0.06	-0.01	0.19	-0.17	A+	A-	A+	A-	2.07	0.04	9.90	1.41	9.90	1.97
163	1034545	634530	11	4108	0.86	0.86	0.06	0.04	0.04	0.00	0.00	0.50	0.50	-0.24	-0.31	-0.27	A+	A-	A-	A+	-1.19	0.05	-5.65	0.84	-6.43	0.66
164	1034560	634545	11	4108	0.72	0.10	0.12	0.72	0.06	0.00	0.00	0.36	-0.24	-0.14	0.36	-0.17	B-	A-	A-	A+	-0.16	0.04	2.55	1.05	4.13	1.14
165	1036020	636005	11	4108	0.91	0.04	0.04	0.91	0.02	0.00	0.00	0.37	-0.23	-0.22	0.37	-0.14	A-	B-	A-	A+	-1.77	0.06	-2.26	0.91	-2.58	0.80

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
166	1036096	636081	11	4108	0.75	0.07	0.12	0.06	0.75	0.00	0.00	0.48	-0.33	-0.20	-0.24	0.48	A-	A+	A-	A+	-0.34	0.04	-4.32	0.91	-4.35	0.85
167	1025644	625629	12	4148	0.76	0.10	0.06	0.76	0.08	0.00	0.00	0.42	-0.21	-0.25	0.42	-0.20	A+	A-	A+	A-	-0.39	0.04	-1.25	0.97	-1.27	0.95
168	1025646	625631	12	4148	0.74	0.04	0.74	0.10	0.12	0.00	0.00	0.46	-0.27	0.46	-0.26	-0.20	A-	A+	A-	A-	-0.28	0.04	-3.17	0.94	-4.15	0.85
169	1025648	625633	12	4148	0.78	0.12	0.06	0.04	0.78	0.00	0.00	0.53	-0.30	-0.26	-0.28	0.53	A+	A-	A-	A-	-0.52	0.04	-7.30	0.85	-7.58	0.71
170	1025650	625635	12	4148	0.73	0.08	0.73	0.14	0.05	0.00	0.00	0.32	-0.22	0.32	-0.08	-0.25	A+	A+	A+	A+	-0.23	0.04	4.01	1.08	6.50	1.25
171	1025652	625637	12	4148	0.67	0.14	0.67	0.04	0.15	0.00	0.00	0.45	-0.31	0.45	-0.21	-0.18	C-	A-	A-	A-	0.12	0.04	-2.28	0.96	-3.24	0.91
172	1025653	625638	12	4148	0.77	0.07	0.10	0.77	0.06	0.00	0.00	0.50	-0.24	-0.29	0.50	-0.27	A-	A+	A+	A-	-0.49	0.04	-5.70	0.88	-6.27	0.76
173	1034538	634523	12	4148	0.64	0.06	0.64	0.15	0.15	0.00	0.00	0.39	-0.22	0.39	-0.20	-0.17	B-	A+	A-	A+	0.31	0.04	2.35	1.04	2.26	1.06
174	1034541	634526	12	4148	0.59	0.15	0.14	0.12	0.59	0.00	0.00	0.42	-0.11	-0.17	-0.33	0.42	A-	A-	A-	A-	0.60	0.04	0.33	1.00	2.26	1.05
175	1034542	634527	12	4148	0.59	0.21	0.16	0.59	0.03	0.00	0.00	0.35	-0.12	-0.23	0.35	-0.21	A-	A+	A-	A+	0.56	0.04	5.42	1.08	3.74	1.09
176	1034543	634528	12	4148	0.53	0.04	0.53	0.04	0.39	0.00	0.00	0.14	-0.25	0.14	-0.19	0.03	A-	A-	A-	A+	0.86	0.04	9.90	1.34	9.90	1.50
177	1034544	634529	12	4148	0.64	0.16	0.10	0.64	0.10	0.00	0.00	0.36	-0.11	-0.25	0.36	-0.18	A-	A+	A-	A+	0.31	0.04	3.94	1.06	2.78	1.08
178	1034546	634531	12	4148	0.67	0.67	0.10	0.19	0.04	0.00	0.00	0.33	0.33	-0.17	-0.14	-0.24	B-	A-	A-	A+	0.13	0.04	5.82	1.10	3.79	1.11
179	1035636	635621	13	4089	0.44	0.20	0.05	0.32	0.44	0.00	0.00	0.24	-0.22	-0.13	-0.01	0.24	A+	A+	A-	A+	1.38	0.04	9.90	1.21	9.90	1.32
180	1035637	635622	13	4089	0.70	0.21	0.70	0.07	0.02	0.00	0.00	0.40	-0.25	0.40	-0.22	-0.19	A-	A-	A-	A+	0.01	0.04	0.84	1.02	-0.17	0.99
181	1035638	635623	13	4089	0.81	0.81	0.12	0.03	0.04	0.00	0.00	0.47	0.47	-0.35	-0.19	-0.18	A-	A-	A+	A+	-0.71	0.04	-4.09	0.90	-5.25	0.77
182	1035640	635625	13	4089	0.48	0.23	0.19	0.48	0.10	0.00	0.00	0.45	-0.25	-0.15	0.45	-0.19	A-	A+	A+	A-	1.16	0.04	-3.15	0.96	0.30	1.01
183	1035646	635631	13	4089	0.80	0.04	0.09	0.80	0.07	0.00	0.00	0.40	-0.26	-0.28	0.40	-0.11	A+	A+	A-	A-	-0.68	0.04	-1.30	0.97	-1.44	0.93
184	1035647	635632	13	4089	0.66	0.11	0.14	0.09	0.66	0.00	0.00	0.41	-0.20	-0.24	-0.18	0.41	A+	A-	A-	A-	0.23	0.04	0.52	1.01	0.18	1.00
185	1036562	636547	13	4089	0.50	0.15	0.12	0.24	0.50	0.00	0.00	0.46	-0.25	-0.27	-0.12	0.46	B-	A+	A+	A-	1.08	0.04	-4.48	0.94	-2.46	0.95
186	1036567	636552	13	4089	0.45	0.07	0.31	0.45	0.17	0.00	0.00	0.28	-0.20	-0.17	0.28	-0.03	A-	A+	A+	A+	1.29	0.04	8.33	1.12	9.90	1.27
187	1036568	636553	13	4089	0.73	0.12	0.07	0.73	0.08	0.00	0.00	0.37	-0.14	-0.20	0.37	-0.24	A-	A-	A+	A+	-0.18	0.04	2.21	1.04	2.06	1.07
188	1036569	636554	13	4089	0.71	0.04	0.71	0.10	0.15	0.00	0.00	0.40	-0.27	0.40	-0.19	-0.20	A-	A-	A-	A-	-0.05	0.04	0.64	1.01	-0.25	0.99
189	1036571	636556	13	4089	0.62	0.17	0.09	0.12	0.62	0.00	0.00	0.40	-0.21	-0.19	-0.18	0.40	A+	A-	A-	A+	0.44	0.04	1.96	1.03	2.41	1.06
190	1037156	637141	13	4089	0.81	0.01	0.81	0.14	0.03	0.00	0.00	0.36	-0.20	0.36	-0.23	-0.19	A-	A+	A-	A-	-0.78	0.04	0.70	1.02	-0.05	1.00
191	1035635	635620	14	4117	0.52	0.52	0.22	0.15	0.10	0.00	0.00	0.30	0.30	-0.11	-0.24	-0.08	A-	A+	A+	A-	0.95	0.04	8.75	1.13	7.75	1.17
192	1035642	635627	14	4117	0.56	0.16	0.16	0.12	0.56	0.00	0.00	0.39	-0.13	-0.17	-0.25	0.39	A+	A-	A-	A+	0.75	0.04	2.21	1.03	1.72	1.04
193	1035643	635628	14	4117	0.49	0.05	0.49	0.26	0.20	0.00	0.00	0.15	-0.25	0.15	0.11	-0.17	A+	A+	A-	A+	1.11	0.04	9.90	1.31	9.90	1.47
194	1035644	635629	14	4117	0.37	0.09	0.49	0.37	0.05	0.00	0.00	0.06	-0.23	0.13	0.06	-0.13	A-	A-	A+	A-	1.75	0.04	9.90	1.37	9.90	1.67
195	1035645	635630	14	4117	0.42	0.38	0.06	0.42	0.13	0.00	0.00	0.25	0.00	-0.23	0.25	-0.20	A+	A+	A+	A+	1.45	0.04	9.90	1.16	9.90	1.33
196	1036564	636549	14	4117	0.52	0.14	0.15	0.19	0.52	0.00	0.00	0.50	-0.25	-0.27	-0.17	0.50	A-	A+	A+	A+	0.97	0.04	-7.70	0.90	-4.61	0.91
197	1036565	636550	14	4117	0.61	0.20	0.11	0.08	0.61	0.00	0.00	0.39	-0.13	-0.21	-0.28	0.39	A-	A-	A-	A-	0.52	0.04	2.19	1.03	3.08	1.07
198	1036566	636551	14	4117	0.63	0.23	0.10	0.63	0.03	0.00	0.00	0.34	-0.15	-0.23	0.34	-0.17	A-	A-	A-	A-	0.39	0.04	5.73	1.09	5.83	1.15

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
199	1036570	636555	14	4117	0.81	0.81	0.06	0.09	0.04	0.00	0.00	0.32	0.32	-0.26	-0.07	-0.22	A-	A-	A-	A+	-0.73	0.04	0.85	1.02	5.47	1.28
200	1036572	636557	14	4117	0.83	0.83	0.07	0.04	0.06	0.00	0.00	0.47	0.47	-0.23	-0.29	-0.25	A+	A-	A+	A+	-0.93	0.05	-5.20	0.87	-4.10	0.80
201	1036975	636960	14	4117	0.63	0.11	0.24	0.02	0.63	0.00	0.00	0.42	-0.20	-0.26	-0.20	0.42	B-	A-	A+	A-	0.41	0.04	0.22	1.00	-1.57	0.96
202	1037158	637143	14	4117	0.61	0.61	0.22	0.08	0.09	0.00	0.00	0.45	0.45	-0.18	-0.27	-0.25	A-	A-	A-	A+	0.50	0.04	-2.21	0.97	-3.15	0.93
203	1032782	632767	15	4142	0.59	0.05	0.12	0.23	0.59	0.00	0.00	0.36	-0.28	-0.15	-0.14	0.36	A-	A+	A+	A-	0.60	0.04	4.73	1.07	3.70	1.09
204	1032787	632772	15	4142	0.74	0.74	0.04	0.09	0.13	0.00	0.00	0.46	0.46	-0.27	-0.28	-0.21	A-	A+	A+	A+	-0.23	0.04	-3.29	0.94	-1.77	0.94
205	1032788	632773	15	4142	0.29	0.29	0.15	0.23	0.32	0.00	0.00	-0.02	-0.02	0.03	0.05	-0.04	A+	A-	A-	A+	2.16	0.04	9.90	1.41	9.90	2.02
206	1032791	632776	15	4142	0.79	0.06	0.79	0.07	0.08	0.00	0.00	0.53	-0.18	0.53	-0.37	-0.28	A+	A+	A-	A+	-0.56	0.04	-6.72	0.85	-7.31	0.72
207	1032792	632777	15	4142	0.71	0.10	0.71	0.11	0.08	0.00	0.00	0.44	-0.35	0.44	-0.13	-0.20	A-	A+	A-	A+	-0.05	0.04	-1.85	0.97	-1.26	0.96
208	1035648	635633	15	4142	0.57	0.06	0.12	0.57	0.24	0.00	0.00	0.28	-0.23	-0.10	0.28	-0.11	A+	A+	A-	A-	0.70	0.04	9.90	1.16	9.79	1.23
209	1035653	635638	15	4142	0.59	0.22	0.59	0.16	0.02	0.00	0.00	0.25	-0.10	0.25	-0.12	-0.22	A+	A-	A-	A+	0.59	0.04	9.90	1.20	9.90	1.34
210	1035654	635639	15	4142	0.81	0.81	0.03	0.03	0.13	0.00	0.00	0.48	0.48	-0.30	-0.28	-0.28	A+	B-	B-	A-	-0.75	0.04	-4.53	0.89	-6.31	0.73
211	1035656	635641	15	4142	0.79	0.05	0.06	0.79	0.10	0.00	0.00	0.43	-0.30	-0.27	0.43	-0.15	A+	A-	A-	A-	-0.58	0.04	-2.28	0.95	-2.58	0.89
212	1035658	635643	15	4142	0.83	0.83	0.08	0.05	0.04	0.00	0.00	0.50	0.50	-0.29	-0.27	-0.26	A+	A+	A-	A-	-0.88	0.05	-5.77	0.85	-5.77	0.73
213	1035711	635696	15	4142	0.75	0.17	0.75	0.02	0.06	0.00	0.00	0.44	-0.32	0.44	-0.26	-0.14	A+	A-	A+	A-	-0.30	0.04	-2.29	0.95	-1.06	0.96
214	1035764	635749	15	4142	0.70	0.11	0.70	0.06	0.13	0.00	0.00	0.38	-0.24	0.38	-0.19	-0.16	A+	A+	A+	A+	0.01	0.04	2.10	1.04	1.13	1.03
215	1032784	632769	16	4063	0.84	0.03	0.05	0.84	0.08	0.00	0.00	0.42	-0.26	-0.25	0.42	-0.20	A-	A-	A+	A+	-0.94	0.05	-2.30	0.94	-2.90	0.85
216	1032785	632770	16	4063	0.69	0.09	0.08	0.69	0.13	0.00	0.00	0.42	-0.15	-0.24	0.42	-0.24	A-	A+	A-	A+	0.07	0.04	0.03	1.00	-0.66	0.98
217	1032789	632774	16	4063	0.78	0.06	0.11	0.05	0.78	0.00	0.00	0.48	-0.25	-0.26	-0.26	0.48	A+	A-	A-	A+	-0.53	0.04	-4.04	0.91	-4.90	0.80
218	1032790	632775	16	4063	0.27	0.67	0.27	0.03	0.02	0.00	0.00	0.22	-0.03	0.22	-0.24	-0.27	A+	A+	A-	A-	2.32	0.04	6.61	1.13	9.90	1.45
219	1032793	632778	16	4063	0.79	0.06	0.13	0.79	0.02	0.00	0.00	0.42	-0.24	-0.26	0.42	-0.18	A+	A-	A-	A-	-0.60	0.04	-1.46	0.97	-1.19	0.95
220	1032794	632779	16	4063	0.54	0.54	0.15	0.11	0.21	0.00	0.00	0.45	0.45	-0.30	-0.29	-0.07	A-	A-	A-	A-	0.90	0.04	-2.74	0.96	-2.15	0.95
221	1035649	635634	16	4063	0.64	0.08	0.09	0.64	0.20	0.00	0.00	0.41	-0.18	-0.25	0.41	-0.19	A+	A-	A-	A+	0.38	0.04	1.27	1.02	0.45	1.01
222	1035650	635635	16	4063	0.82	0.82	0.08	0.05	0.05	0.00	0.00	0.43	0.43	-0.27	-0.23	-0.19	A+	A-	B-	A+	-0.81	0.04	-2.38	0.94	-3.64	0.82
223	1035652	635637	16	4063	0.65	0.17	0.65	0.05	0.13	0.00	0.00	0.25	-0.09	0.25	-0.23	-0.10	A-	A+	A+	A+	0.28	0.04	9.90	1.20	8.86	1.26
224	1035655	635640	16	4063	0.33	0.29	0.18	0.19	0.33	0.00	0.00	0.13	-0.03	-0.06	-0.06	0.13	A-	A+	A+	A+	1.95	0.04	9.90	1.30	9.90	1.56
225	1035657	635642	16	4063	0.83	0.04	0.05	0.08	0.83	0.00	0.00	0.39	-0.19	-0.31	-0.14	0.39	A-	A-	A-	A-	-0.91	0.05	-1.52	0.96	1.87	1.10
226	1035659	635644	16	4063	0.89	0.06	0.04	0.89	0.01	0.00	0.00	0.36	-0.20	-0.26	0.36	-0.17	A-	A-	A-	A+	-1.53	0.05	-1.62	0.94	-0.15	0.99
227	1034862	634847	17	4104	0.64	0.14	0.06	0.15	0.64	0.00	0.00	0.41	-0.23	-0.29	-0.13	0.41	A+	A-	A-	A+	0.34	0.04	0.18	1.00	-1.59	0.96
228	1034863	634848	17	4104	0.76	0.14	0.76	0.07	0.04	0.00	0.00	0.50	-0.26	0.50	-0.28	-0.29	A-	A+	A+	A+	-0.39	0.04	-5.64	0.89	-5.77	0.80
229	1034864	634849	17	4104	0.37	0.37	0.15	0.12	0.36	0.00	0.00	0.29	0.29	-0.33	-0.05	-0.01	A-	A-	A+	A-	1.71	0.04	4.81	1.07	8.36	1.22
230	1034868	634853	17	4104	0.73	0.10	0.06	0.11	0.73	0.00	0.00	0.41	-0.22	-0.25	-0.17	0.41	B-	A+	A+	A+	-0.20	0.04	-0.51	0.99	0.72	1.02
231	1034869	634854	17	4104	0.45	0.14	0.16	0.45	0.25	0.00	0.00	0.12	-0.25	-0.10	0.12	0.14	A+	B-	A-	A-	1.30	0.04	9.90	1.32	9.90	1.46

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
232	1035918	635903	17	4104	0.39	0.23	0.11	0.26	0.39	0.00	0.00	0.12	-0.06	-0.10	0.00	0.12	A+	A+	A+	A+	1.60	0.04	9.90	1.29	9.90	1.53
233	1035919	635904	17	4104	0.78	0.78	0.03	0.02	0.17	0.00	0.00	0.36	0.36	-0.19	-0.24	-0.22	A+	A-	A-	A+	-0.51	0.04	0.28	1.01	4.01	1.17
234	1035923	635908	17	4104	0.86	0.05	0.04	0.05	0.86	0.00	0.00	0.47	-0.26	-0.25	-0.26	0.47	A+	A+	A-	A-	-1.15	0.05	-4.86	0.86	-6.13	0.69
235	1035924	635909	17	4104	0.38	0.55	0.38	0.03	0.04	0.00	0.00	-0.05	0.24	-0.05	-0.26	-0.24	A-	A-	A+	A-	1.67	0.04	9.90	1.48	9.90	1.83
236	1035927	635912	17	4104	0.83	0.07	0.07	0.83	0.04	0.00	0.00	0.44	-0.21	-0.27	0.44	-0.25	A-	A-	B-	A+	-0.88	0.04	-3.07	0.92	-4.93	0.78
237	1036745	636730	17	4104	0.77	0.07	0.10	0.77	0.07	0.00	0.00	0.49	-0.23	-0.31	0.49	-0.23	A-	A+	A-	A+	-0.43	0.04	-4.72	0.90	-6.69	0.77
238	1037191	637176	17	4104	0.62	0.62	0.09	0.10	0.19	0.00	0.00	0.22	0.22	-0.10	-0.14	-0.09	A-	A+	A-	A-	0.45	0.04	9.90	1.22	9.81	1.24
239	1034865	634850	18	4098	0.58	0.58	0.07	0.10	0.25	0.00	0.00	0.26	0.26	-0.28	-0.25	0.05	A-	A-	A-	A-	0.65	0.04	9.90	1.18	9.90	1.30
240	1034866	634851	18	4098	0.61	0.61	0.11	0.21	0.07	0.00	0.00	0.28	0.28	0.00	-0.19	-0.21	A-	A+	A+	A+	0.49	0.04	9.90	1.16	9.90	1.27
241	1034870	634855	18	4098	0.62	0.16	0.09	0.12	0.62	0.00	0.00	0.36	-0.05	-0.29	-0.21	0.36	A-	A-	A+	A+	0.45	0.04	3.84	1.06	5.67	1.14
242	1034871	634856	18	4098	0.50	0.14	0.50	0.15	0.22	0.00	0.00	0.08	-0.04	0.08	0.02	-0.07	A-	A+	A+	A-	1.08	0.04	9.90	1.39	9.90	1.56
243	1034872	634857	18	4098	0.54	0.54	0.14	0.17	0.15	0.00	0.00	0.48	0.48	-0.24	-0.24	-0.18	C-	A-	A-	A+	0.87	0.04	-6.27	0.91	-5.38	0.89
244	1034874	634859	18	4098	0.73	0.07	0.15	0.73	0.05	0.00	0.00	0.49	-0.23	-0.29	0.49	-0.25	B-	A+	A-	A+	-0.20	0.04	-4.70	0.91	-5.97	0.81
245	1035916	635901	18	4098	0.86	0.08	0.03	0.86	0.03	0.00	0.00	0.31	-0.16	-0.20	0.31	-0.19	A+	A+	A+	A-	-1.20	0.05	-0.02	1.00	2.44	1.15
246	1035917	635902	18	4098	0.58	0.10	0.19	0.13	0.58	0.00	0.00	0.41	-0.21	-0.21	-0.16	0.41	A+	A-	A-	A+	0.66	0.04	1.15	1.02	1.34	1.03
247	1035920	635905	18	4098	0.51	0.51	0.10	0.32	0.07	0.00	0.00	0.24	0.24	-0.03	-0.18	-0.09	A-	A-	A-	A-	1.00	0.04	9.90	1.20	9.90	1.29
248	1035921	635906	18	4098	0.82	0.04	0.82	0.07	0.07	0.00	0.00	0.43	-0.21	0.43	-0.22	-0.26	A+	A-	A-	A+	-0.82	0.04	-2.85	0.93	-3.72	0.83
249	1035925	635910	18	4098	0.47	0.29	0.47	0.11	0.13	0.00	0.00	0.44	-0.19	0.44	-0.13	-0.25	A-	A-	A-	A+	1.23	0.04	-3.22	0.96	-0.19	1.00
250	1035926	635911	18	4098	0.95	0.02	0.95	0.02	0.02	0.00	0.00	0.30	-0.20	0.30	-0.19	-0.11	A-	A-	A-	B+	-2.40	0.07	-1.54	0.91	-1.53	0.84
251	1018843	618828	19	4097	0.75	0.05	0.06	0.14	0.75	0.00	0.00	0.52	-0.31	-0.25	-0.27	0.52	A-	A-	A-	A-	-0.35	0.04	-6.84	0.87	-6.35	0.78
252	1018845	618830	19	4097	0.74	0.08	0.08	0.09	0.74	0.00	0.00	0.40	-0.16	-0.19	-0.27	0.40	A+	A+	A-	A+	-0.29	0.04	-0.50	0.99	0.19	1.01
253	1018846	618831	19	4097	0.83	0.04	0.83	0.09	0.04	0.00	0.00	0.45	-0.25	0.45	-0.23	-0.27	A+	A-	A-	A+	-0.93	0.05	-4.23	0.89	-2.01	0.90
254	1018850	618835	19	4097	0.58	0.58	0.13	0.20	0.10	0.00	0.00	0.31	0.31	-0.23	-0.07	-0.17	A-	A-	A-	A-	0.65	0.04	7.59	1.11	8.68	1.20
255	1018852	618837	19	4097	0.84	0.07	0.03	0.05	0.84	0.00	0.00	0.38	-0.17	-0.18	-0.27	0.38	A-	A-	A-	A+	-0.99	0.05	-1.78	0.95	-0.39	0.98
256	1018854	618839	19	4097	0.54	0.20	0.54	0.15	0.11	0.00	0.00	0.35	-0.23	0.35	-0.03	-0.23	B-	A+	A-	A+	0.84	0.04	4.56	1.06	4.15	1.09
257	1035747	635732	19	4097	0.58	0.05	0.58	0.20	0.17	0.00	0.00	0.22	-0.19	0.22	-0.12	-0.06	A-	A-	A-	A+	0.62	0.04	9.90	1.22	9.90	1.31
258	1035749	635734	19	4097	0.51	0.20	0.51	0.18	0.11	0.00	0.00	0.13	-0.01	0.13	-0.07	-0.10	A+	A+	A-	A+	0.99	0.04	9.90	1.33	9.90	1.46
259	1035750	635735	19	4097	0.82	0.07	0.08	0.82	0.03	0.00	0.00	0.41	-0.19	-0.29	0.41	-0.17	A-	A-	A-	A-	-0.80	0.04	-2.39	0.94	-0.93	0.95
260	1035753	635738	19	4097	0.79	0.12	0.79	0.06	0.03	0.00	0.00	0.42	-0.21	0.42	-0.30	-0.20	A+	A-	A+	A+	-0.62	0.04	-2.66	0.94	-1.45	0.94
261	1035756	635741	19	4097	0.59	0.06	0.25	0.10	0.59	0.00	0.00	0.36	-0.19	-0.18	-0.18	0.36	A+	A-	A-	A-	0.59	0.04	3.72	1.06	3.79	1.09
262	1037069	637054	19	4097	0.50	0.50	0.21	0.13	0.16	0.00	0.00	0.30	0.30	-0.19	-0.06	-0.13	A+	A-	A-	A-	1.06	0.04	7.98	1.11	8.40	1.18
263	1018844	618829	20	4069	0.45	0.12	0.18	0.45	0.24	0.00	0.00	0.40	-0.16	-0.21	0.40	-0.15	A-	A+	A-	A-	1.34	0.04	-0.85	0.99	3.63	1.08
264	1018847	618832	20	4069	0.62	0.62	0.29	0.05	0.03	0.00	0.00	0.28	0.28	-0.07	-0.29	-0.19	A+	A-	A-	A+	0.44	0.04	9.62	1.16	9.33	1.24

Table J-7 (continued). Literature Multiple-Choice Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	M/F	W/B	W/H	O/P	Meas	SEM	z	MS	z	MS
265	1018848	618833	20	4069	0.83	0.08	0.83	0.05	0.04	0.00	0.00	0.48	-0.31	0.48	-0.22	-0.25	A+	A-	A-	A-	-0.89	0.05	-5.14	0.87	-6.21	0.71
266	1018849	618834	20	4069	0.64	0.10	0.18	0.08	0.64	0.00	0.00	0.37	-0.14	-0.11	-0.33	0.37	A-	A+	A-	A+	0.34	0.04	3.05	1.05	3.73	1.10
267	1018853	618838	20	4069	0.54	0.54	0.04	0.07	0.36	0.00	0.00	0.22	0.22	-0.24	-0.32	0.04	A-	A-	A-	A-	0.88	0.04	9.90	1.23	9.90	1.36
268	1018952	618937	20	4069	0.46	0.40	0.07	0.46	0.07	0.00	0.00	0.27	-0.03	-0.27	0.27	-0.18	A+	A+	A+	A-	1.30	0.04	9.90	1.14	9.90	1.26
269	1035746	635731	20	4069	0.38	0.38	0.14	0.30	0.17	0.00	0.00	0.17	0.17	-0.19	0.12	-0.19	A-	A+	A+	A-	1.67	0.04	9.90	1.27	9.90	1.41
270	1035751	635736	20	4069	0.74	0.21	0.03	0.74	0.02	0.00	0.00	0.29	-0.16	-0.22	0.29	-0.15	A+	A+	A-	A-	-0.22	0.04	5.07	1.10	6.66	1.25
271	1035752	635737	20	4069	0.54	0.07	0.35	0.04	0.54	0.00	0.00	0.43	-0.29	-0.18	-0.28	0.43	A-	A-	A-	A-	0.87	0.04	-1.14	0.98	-0.48	0.99
272	1035755	635740	20	4069	0.37	0.09	0.18	0.36	0.37	0.00	0.00	0.25	-0.16	-0.02	-0.14	0.25	A-	A-	A-	A+	1.71	0.04	9.22	1.14	9.90	1.31
273	1035757	635742	20	4069	0.47	0.47	0.03	0.28	0.21	0.00	0.00	0.29	0.29	-0.23	-0.07	-0.18	A+	A-	A+	A+	1.23	0.04	7.92	1.11	9.58	1.21
274	1035758	635743	20	4069	0.75	0.09	0.75	0.04	0.12	0.00	0.00	0.43	-0.22	0.43	-0.24	-0.22	A+	A-	A-	A-	-0.30	0.04	-1.75	0.96	-2.56	0.91

Table J–8. Algebra I Multiple-Choice Item Statistics: Summer

Ref	ID	PubID	Form	<i>N</i>	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
1																						
2																						
3																						

Table J–9. Biology Multiple-Choice Item Statistics: Summer

Ref	ID	PubID	Form	<i>N</i>	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
1																						
2																						
3																						

Table J–10. Literature Multiple-Choice Item Statistics: Summer

Ref	ID	PubID	Form	<i>N</i>	Pval	P(A)	P(B)	P(C)	P(D)	P(-)	P(*)	ITCorr	Corr(A)	Corr(B)	Corr(C)	Corr(D)	Meas	SEM	z	MS	z	MS
1																						
2																						
3																						

CONSTRUCTED-RESPONSE ITEMS

Table J–11. Algebra I Constructed-Response Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(4)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Corr(4)	Meas	SEM	z	MS	z	MS
1	673351	273336	0	13200	0.38	0.21	0.29	0.23	0.16	0.07	0.03	0.66	-0.38	-0.21	0.09	0.34	0.42	0.83	0.02	-5.22	0.94	-5.85	0.93
2	734697	334682	0	13200	0.38	0.13	0.34	0.28	0.19	0.01	0.03	0.65	-0.36	-0.27	0.20	0.46	0.18	1.40	0.03	-8.02	0.91	-7.22	0.92
3	818209	418194	0	13200	0.23	0.40	0.33	0.17	0.05	0.02	0.02	0.70	-0.52	0.05	0.39	0.34	0.26	1.53	0.02	-9.90	0.79	-9.90	0.75
4	818616	418601	0	13200	0.26	0.42	0.27	0.14	0.14	0.02	0.01	0.66	-0.53	0.05	0.23	0.42	0.24	1.93	0.02	1.70	1.02	-0.97	0.98
5	818665	418650	0	13200	0.16	0.54	0.28	0.08	0.06	0.01	0.02	0.67	-0.53	0.22	0.32	0.38	0.20	2.62	0.02	-7.74	0.89	-9.90	0.80
6	877382	477367	0	13200	0.17	0.49	0.35	0.09	0.04	0.01	0.02	0.72	-0.58	0.26	0.39	0.35	0.18	2.62	0.02	-9.90	0.74	-9.90	0.67

Table J–12. Biology Constructed-Response Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Meas	SEM	z	MS	z	MS
1	702742	302727	0	12144	0.31	0.34	0.38	0.21	0.05	0.02	0.59	-0.42	0.05	0.38	0.29	1.56	0.02	-5.77	0.93	-4.75	0.94
2	741576	341561	0	12144	0.28	0.48	0.20	0.15	0.12	0.03	0.71	-0.56	0.14	0.35	0.47	1.33	0.02	-9.90	0.79	-9.90	0.69
3	819535	419520	0	12144	0.53	0.18	0.25	0.29	0.25	0.02	0.54	-0.36	-0.14	0.18	0.37	0.51	0.02	9.90	1.16	9.90	1.20
4	877366	477351	0	12144	0.44	0.11	0.42	0.41	0.02	0.02	0.55	-0.25	-0.28	0.45	0.22	0.93	0.03	-9.90	0.79	-9.90	0.79
5	966775	566760	0	12144	0.45	0.29	0.23	0.20	0.24	0.02	0.70	-0.49	-0.11	0.20	0.55	0.21	0.03	-2.62	0.97	-0.58	0.99
6	978211	578196	0	12144	0.46	0.18	0.33	0.31	0.14	0.02	0.64	-0.42	-0.17	0.31	0.40	0.59	0.02	-9.90	0.84	-9.90	0.84

Table J–13. Literature Constructed-Response Item Statistics: Winter

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Meas	SEM	z	MS	z	MS
1	704766	304751	0	11722	0.49	0.11	0.34	0.38	0.13	0.03	0.70	-0.43	-0.22	0.38	0.37	1.16	0.03	-9.90	0.75	-9.90	0.75
2	704767	304752	0	11722	0.52	0.08	0.28	0.45	0.13	0.04	0.70	-0.35	-0.34	0.37	0.40	0.75	0.03	-9.90	0.69	-9.90	0.69
3	742085	342070	0	11722	0.53	0.04	0.40	0.43	0.12	0.01	0.67	-0.35	-0.39	0.37	0.36	0.91	0.03	-9.90	0.60	-9.90	0.62
4	986358	586343	0	11722	0.56	0.03	0.34	0.47	0.13	0.02	0.67	-0.26	-0.45	0.34	0.39	0.80	0.04	-9.90	0.80	-9.90	0.80
5	994603	594588	0	11722	0.55	0.06	0.37	0.38	0.17	0.02	0.70	-0.37	-0.38	0.32	0.42	0.72	0.04	-9.90	0.75	-9.90	0.74
6	994606	594591	0	11722	0.51	0.06	0.37	0.39	0.13	0.05	0.70	-0.37	-0.34	0.39	0.39	0.80	0.03	-9.90	0.73	-9.90	0.73

Table J–14. Algebra I Constructed-Response Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(4)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Corr(4)	Meas	SEM	z	MS	z	MS
1	817339	417324	0	88831	0.34	0.27	0.27	0.26	0.18	0.01	0.01	0.71	-0.52	-0.14	0.24	0.51	0.15	2.11	0.01	-9.90	0.86	-9.90	0.85
2	820072	420057	0	88831	0.20	0.49	0.29	0.12	0.06	0.02	0.02	0.65	-0.51	0.12	0.34	0.35	0.24	2.26	0.01	-9.90	0.91	-9.90	0.86
3	821550	421535	0	88831	0.43	0.11	0.27	0.38	0.16	0.05	0.03	0.73	-0.37	-0.37	0.11	0.45	0.37	0.87	0.01	-9.90	0.70	-9.90	0.71
4	821569	421554	0	88831	0.29	0.25	0.43	0.20	0.09	0.02	0.01	0.74	-0.55	-0.09	0.39	0.40	0.23	2.00	0.01	-9.90	0.87	-9.90	0.83
5	905404	505389	0	88831	0.16	0.45	0.42	0.07	0.03	0.00	0.03	0.64	-0.51	0.30	0.35	0.26	0.11	2.51	0.01	-9.90	0.79	-9.90	0.78
6	969452	569437	0	88831	0.29	0.35	0.28	0.17	0.11	0.06	0.03	0.73	-0.55	-0.03	0.27	0.39	0.37	1.31	0.01	-9.90	0.83	-9.90	0.78

Table J–15. Biology Constructed-Response Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Meas	SEM	z	MS	z	MS
1	741444	341429	0	83785	0.39	0.31	0.30	0.23	0.14	0.02	0.44	-0.25	-0.13	0.22	0.32	0.84	0.01	9.90	1.33	9.90	1.35
2	741581	341566	0	83785	0.41	0.27	0.25	0.25	0.17	0.04	0.62	-0.40	-0.09	0.28	0.43	0.92	0.01	3.83	1.02	-2.13	0.99
3	966416	566401	0	83785	0.47	0.26	0.25	0.26	0.21	0.02	0.71	-0.51	-0.16	0.28	0.50	0.60	0.01	-9.90	0.79	-9.90	0.78
4	978209	578194	0	83785	0.44	0.20	0.34	0.28	0.14	0.02	0.58	-0.29	-0.19	0.20	0.46	1.01	0.01	5.57	1.03	9.90	1.09
5	978647	578632	0	83785	0.49	0.23	0.27	0.27	0.22	0.01	0.67	-0.47	-0.18	0.19	0.52	0.49	0.01	-9.90	0.81	-9.90	0.81
6	983824	583809	0	83785	0.42	0.27	0.32	0.25	0.15	0.01	0.68	-0.49	-0.10	0.29	0.46	0.67	0.01	-9.90	0.79	-9.90	0.80

Table J–16. Literature Constructed-Response Item Statistics: Spring

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Meas	SEM	z	MS	z	MS
1	824935	424920	0	82141	0.53	0.09	0.31	0.46	0.12	0.02	0.68	-0.43	-0.29	0.38	0.34	0.76	0.01	-9.90	0.81	-9.90	0.81
2	824936	424921	0	82141	0.53	0.06	0.30	0.46	0.13	0.04	0.69	-0.39	-0.35	0.37	0.37	0.95	0.01	-9.90	0.77	-9.90	0.76
3	826264	426249	0	82141	0.53	0.09	0.27	0.51	0.09	0.02	0.73	-0.47	-0.31	0.46	0.32	0.93	0.01	-9.90	0.63	-9.90	0.63
4	826283	426268	0	82141	0.54	0.09	0.26	0.45	0.16	0.04	0.72	-0.45	-0.29	0.36	0.39	0.81	0.01	-9.90	0.92	-9.90	0.92
5	984231	584216	0	82141	0.50	0.05	0.40	0.45	0.06	0.02	0.65	-0.33	-0.35	0.46	0.27	1.12	0.01	-9.90	0.75	-9.90	0.75
6	994432	594417	0	82141	0.62	0.03	0.23	0.52	0.19	0.01	0.64	-0.29	-0.41	0.22	0.39	0.44	0.02	-9.90	0.81	-9.90	0.80

Table J–17. Algebra I Constructed-Response Item Statistics: Summer

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(4)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Corr(4)	Meas	SEM	z	MS	z	MS		
1																									
2																									
3																									
4																									
5																									
6																									

Table J–18. Biology Constructed-Response Item Statistics: Summer

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Meas	SEM	z	MS	z	MS					
1																										
2																										
3																										
4																										
5																										
6																										

Table J–19. Literature Constructed-Response Item Statistics: Summer

Ref	ID	PubID	Form	N	Pval	P(0)	P(1)	P(2)	P(3)	P(B)	ITCorr	Corr(0)	Corr(1)	Corr(2)	Corr(3)	Meas	SEM	z	MS	z	MS						
1																											
2																											
3																											
4																											
5																											
6																											

APPENDIX K: RAW-TO-SCALE SCORE CONVERSION TABLES

Appendix K provides the raw-to-scaled score conversion tables for each administration and content area. The scaled score conversions are administration specific and therefore cannot be used for highest total scaled score to date.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

Table K-1. Raw-to-Scaled Score Conversion Tables

Column Heading	Definition
Raw	Raw score
SS	Scaled score
CSEM	Conditional standard error of measurement
LCI	Lower confidence interval
UCI	Upper confidence interval

WINTER

Table K-2. Algebra I Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0	1218	92	1200	1310
1	1279	51	1228	1330
2	1315	36	1279	1351
3	1337	30	1307	1367
4	1353	26	1327	1379
5	1365	24	1341	1389
6	1376	22	1354	1398
7	1385	21	1364	1406
8	1393	20	1373	1413
9	1401	19	1382	1420
10	1408	18	1390	1426
11	1414	18	1396	1432
12	1420	17	1403	1437
13	1426	17	1409	1443
14	1431	16	1415	1447
15	1436	16	1420	1452
16	1441	16	1425	1457
17	1446	15	1431	1461
18	1450	15	1435	1465
19	1455	15	1440	1470
20	1459	15	1444	1474
21	1464	15	1449	1479
22	1468	15	1453	1483
23	1472	14	1458	1486
24	1476	14	1462	1490
25	1481	14	1467	1495
26	1485	14	1471	1499
27	1489	14	1475	1503
28	1493	14	1479	1507
29	1497	14	1483	1511
30	1501	14	1487	1515
31	1505	14	1491	1519
32	1509	14	1495	1523
33	1513	14	1499	1527
34	1517	14	1503	1531
35	1521	14	1507	1535
36	1525	15	1510	1540
37	1530	15	1515	1545
38	1534	15	1519	1549
39	1539	15	1524	1554
40	1543	15	1528	1558

Table K-2 (continued). Algebra I Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
41	1548	16	1532	1564
42	1553	16	1537	1569
43	1558	16	1542	1574
44	1563	17	1546	1580
45	1569	17	1552	1586
46	1575	17	1558	1592
47	1581	18	1563	1599
48	1588	19	1569	1607
49	1595	19	1576	1614
50	1603	20	1583	1623
51	1611	21	1590	1632
52	1621	23	1598	1644
53	1632	24	1608	1656
54	1645	26	1619	1671
55	1660	29	1631	1689
56	1678	32	1646	1710
57	1700	35	1665	1735
58	1730	41	1689	1771
59	1773	54	1719	1800
60	1800	94	1706	1800

Table K-3. Biology Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0	1222	92	1200	1314
1	1283	51	1232	1334
2	1318	36	1282	1354
3	1340	30	1310	1370
4	1355	26	1329	1381
5	1367	24	1343	1391
6	1378	22	1356	1400
7	1386	20	1366	1406
8	1394	19	1375	1413
9	1401	18	1383	1419
10	1408	17	1391	1425
11	1413	17	1396	1430
12	1419	16	1403	1435
13	1424	16	1408	1440
14	1429	15	1414	1444
15	1434	15	1419	1449
16	1438	15	1423	1453
17	1442	14	1428	1456
18	1446	14	1432	1460
19	1450	14	1436	1464
20	1454	14	1440	1468
21	1458	14	1444	1472
22	1461	13	1448	1474
23	1465	13	1452	1478
24	1468	13	1455	1481
25	1472	13	1459	1485
26	1475	13	1462	1488
27	1478	13	1465	1491
28	1482	13	1469	1495
29	1485	13	1472	1498
30	1488	13	1475	1501
31	1491	13	1478	1504
32	1494	13	1481	1507
33	1498	13	1485	1511
34	1501	13	1488	1514
35	1504	13	1491	1517
36	1507	13	1494	1520
37	1510	13	1497	1523
38	1514	13	1501	1527
39	1517	13	1504	1530
40	1520	13	1507	1533
41	1524	13	1511	1537
42	1527	13	1514	1540

Table K-3 (continued). Biology Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
43	1530	13	1517	1543
44	1534	13	1521	1547
45	1538	14	1524	1552
46	1541	14	1527	1555
47	1545	14	1531	1559
48	1549	14	1535	1563
49	1553	14	1539	1567
50	1558	15	1543	1573
51	1562	15	1547	1577
52	1567	15	1552	1582
53	1572	16	1556	1588
54	1577	16	1561	1593
55	1582	17	1565	1599
56	1588	18	1570	1606
57	1595	18	1577	1613
58	1602	19	1583	1621
59	1610	21	1589	1631
60	1619	22	1597	1641
61	1630	24	1606	1654
62	1642	26	1616	1668
63	1658	30	1628	1688
64	1680	37	1643	1717
65	1717	51	1666	1768
66	1778	92	1686	1800

Table K-4. Literature Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0	1200	92	1200	1292
1	1258	51	1207	1309
2	1295	37	1258	1332
3	1317	31	1286	1348
4	1334	27	1307	1361
5	1347	24	1323	1371
6	1358	23	1335	1381
7	1367	21	1346	1388
8	1376	20	1356	1396
9	1384	19	1365	1403
10	1391	19	1372	1410
11	1398	18	1380	1416
12	1404	18	1386	1422
13	1410	17	1393	1427
14	1416	17	1399	1433
15	1422	17	1405	1439
16	1427	16	1411	1443
17	1432	16	1416	1448
18	1437	16	1421	1453
19	1442	16	1426	1458
20	1447	16	1431	1463
21	1452	15	1437	1467
22	1457	15	1442	1472
23	1461	15	1446	1476
24	1466	15	1451	1481
25	1471	15	1456	1486
26	1476	15	1461	1491
27	1480	15	1465	1495
28	1485	15	1470	1500
29	1490	16	1474	1506
30	1495	16	1479	1511
31	1500	16	1484	1516
32	1505	16	1489	1521
33	1510	16	1494	1526
34	1515	16	1499	1531
35	1520	17	1503	1537
36	1526	17	1509	1543
37	1532	17	1515	1549
38	1538	18	1520	1556
39	1544	18	1526	1562
40	1551	19	1532	1570
41	1558	19	1539	1577
42	1565	20	1545	1585

Table K-4 (continued). Literature Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
43	1574	21	1553	1595
44	1582	21	1561	1603
45	1592	23	1569	1615
46	1603	24	1579	1627
47	1615	26	1589	1641
48	1629	28	1601	1657
49	1647	32	1615	1679
50	1671	38	1633	1709
51	1709	52	1657	1761
52	1772	92	1680	1800

SPRING

Table K-5. Algebra I Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0	1215	92	1200	1307
1	1276	51	1225	1327
2	1312	36	1276	1348
3	1334	30	1304	1364
4	1350	26	1324	1376
5	1362	24	1338	1386
6	1373	22	1351	1395
7	1382	21	1361	1403
8	1391	20	1371	1411
9	1398	19	1379	1417
10	1405	18	1387	1423
11	1411	18	1393	1429
12	1417	17	1400	1434
13	1423	17	1406	1440
14	1428	16	1412	1444
15	1434	16	1418	1450
16	1439	16	1423	1455
17	1443	15	1428	1458
18	1448	15	1433	1463
19	1453	15	1438	1468
20	1457	15	1442	1472
21	1462	15	1447	1477
22	1466	15	1451	1481
23	1470	15	1455	1485
24	1475	15	1460	1490
25	1479	15	1464	1494
26	1483	14	1469	1497
27	1487	14	1473	1501
28	1491	14	1477	1505
29	1496	15	1481	1511
30	1500	15	1485	1515
31	1504	15	1489	1519
32	1508	15	1493	1523
33	1513	15	1498	1528
34	1517	15	1502	1532
35	1521	15	1506	1536
36	1526	15	1511	1541
37	1530	15	1515	1545
38	1535	15	1520	1550
39	1540	16	1524	1556
40	1545	16	1529	1561
41	1550	16	1534	1566

Table K-5 (continued). Algebra I Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
42	1555	16	1539	1571
43	1560	17	1543	1577
44	1566	17	1549	1583
45	1572	17	1555	1589
46	1578	18	1560	1596
47	1584	18	1566	1602
48	1591	19	1572	1610
49	1598	19	1579	1617
50	1606	20	1586	1626
51	1614	21	1593	1635
52	1623	22	1601	1645
53	1634	23	1611	1657
54	1645	25	1620	1670
55	1659	28	1631	1687
56	1677	31	1646	1708
57	1699	36	1663	1735
58	1729	43	1686	1772
59	1777	57	1720	1800
60	1800	96	1704	1800

Table K-6. Biology Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0	1224	92	1200	1316
1	1285	51	1234	1336
2	1321	36	1285	1357
3	1342	30	1312	1372
4	1358	26	1332	1384
5	1370	24	1346	1394
6	1380	22	1358	1402
7	1389	20	1369	1409
8	1397	19	1378	1416
9	1404	18	1386	1422
10	1410	18	1392	1428
11	1416	17	1399	1433
12	1422	16	1406	1438
13	1427	16	1411	1443
14	1432	15	1417	1447
15	1436	15	1421	1451
16	1441	15	1426	1456
17	1445	14	1431	1459
18	1449	14	1435	1463
19	1453	14	1439	1467
20	1457	14	1443	1471
21	1460	13	1447	1473
22	1464	13	1451	1477
23	1467	13	1454	1480
24	1471	13	1458	1484
25	1474	13	1461	1487
26	1477	13	1464	1490
27	1481	13	1468	1494
28	1484	13	1471	1497
29	1487	13	1474	1500
30	1490	12	1478	1502
31	1493	12	1481	1505
32	1496	12	1484	1508
33	1499	12	1487	1511
34	1502	12	1490	1514
35	1505	12	1493	1517
36	1509	12	1497	1521
37	1512	12	1500	1524
38	1515	12	1503	1527
39	1518	13	1505	1531
40	1521	13	1508	1534
41	1524	13	1511	1537
42	1528	13	1515	1541

Table K-6 (continued). Biology Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
43	1531	13	1518	1544
44	1534	13	1521	1547
45	1538	13	1525	1551
46	1541	13	1528	1554
47	1545	14	1531	1559
48	1549	14	1535	1563
49	1553	14	1539	1567
50	1557	14	1543	1571
51	1561	15	1546	1576
52	1565	15	1550	1580
53	1570	16	1554	1586
54	1575	16	1559	1591
55	1580	17	1563	1597
56	1586	17	1569	1603
57	1592	18	1574	1610
58	1599	19	1580	1618
59	1607	20	1587	1627
60	1615	22	1593	1637
61	1625	23	1602	1648
62	1637	26	1611	1663
63	1653	30	1623	1683
64	1674	36	1638	1710
65	1710	50	1660	1760
66	1770	92	1678	1800

Table K-7. Literature Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0	1201	92	1200	1293
1	1263	51	1212	1314
2	1299	37	1262	1336
3	1321	30	1291	1351
4	1337	27	1310	1364
5	1350	24	1326	1374
6	1361	23	1338	1384
7	1371	21	1350	1392
8	1379	20	1359	1399
9	1387	19	1368	1406
10	1394	19	1375	1413
11	1401	18	1383	1419
12	1407	18	1389	1425
13	1413	17	1396	1430
14	1419	17	1402	1436
15	1424	16	1408	1440
16	1430	16	1414	1446
17	1435	16	1419	1451
18	1440	16	1424	1456
19	1445	16	1429	1461
20	1450	16	1434	1466
21	1455	15	1440	1470
22	1459	15	1444	1474
23	1464	15	1449	1479
24	1469	15	1454	1484
25	1474	15	1459	1489
26	1478	15	1463	1493
27	1483	15	1468	1498
28	1488	16	1472	1504
29	1493	16	1477	1509
30	1498	16	1482	1514
31	1503	16	1487	1519
32	1508	16	1492	1524
33	1513	16	1497	1529
34	1519	17	1502	1536
35	1524	17	1507	1541
36	1530	17	1513	1547
37	1536	18	1518	1554
38	1542	18	1524	1560
39	1549	18	1531	1567
40	1556	19	1537	1575
41	1563	20	1543	1583
42	1571	20	1551	1591

Table K-7 (continued). Literature Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
43	1580	21	1559	1601
44	1589	22	1567	1611
45	1599	23	1576	1622
46	1610	24	1586	1634
47	1623	26	1597	1649
48	1638	29	1609	1667
49	1656	32	1624	1688
50	1680	38	1642	1718
51	1719	52	1667	1771
52	1782	92	1690	1800

SUMMER

Table K-8. Algebra Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0				
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				

Table K-8 (continued). Algebra Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
42				
43				
44				
45				
46				
47				
48				
49				
50				
51				
52				
53				
54				
55				
56				
57				
58				
59				
60				

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Table K-9. Biology Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0				
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				
42				

Table K-9 (continued). Biology Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
43				
44				
45				
46				
47				
48				
49				
50				
51				
52				
53				
54				
55				
56				
57				
58				
59				
60				
61				
62				
63				
64				
65				
66				

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Table K-10. Literature Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
0				
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				
42				

Table K-10 (continued). Literature Raw-to-Scaled Score Conversion Table

Raw	SS	CSEM	LCI	UCI
43				
44				
45				
46				
47				
48				
49				
50				
51				
52				

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

APPENDIX L: PRE-EQUATING VERIFICATION RESULTS

Appendix L includes person fit summary box plots, normalized scale score difference plots, and raw to scale score table comparisons generated during the pre-equating verification process. Figure L-1 presents the person infit boxplots for the pre-equated (left) and post-equated (right) solutions by administration and content area for each of the following subgroups: race/ethnicity, gender, IEP, and EL status. Figure L-2 presents the normalized scale score difference plots that show this difference between the pre-equated and post-equated solutions. The vertical dashed lines depict the raw cut-scores based on the pre-equated solutions.

Table L-1 shows both the pre-equated and post-equated raw-to-scaled score conversion tables. Tables provide details for both pre-equated (Pre) and post-equated (Post) solutions with respect to the scaled score (SS), standard error of measurement (SEM), performance level (PL), overall proportion of students earning each raw score, and whether there were differences in the performance level classification for pre- and post-equated solutions (Same PL). All student reporting is based on the tables in Appendix K, which use the pre-equated solutions.

Results are presented for both the Winter and Spring administrations.

Appendix T contains additional information and results based on the data used for the pre-equating verification. Results are presented for the fully-anchored pre-equating solution (hereinafter “pre-equating”) and the partially anchored pre-equating solution when misfitting items were freely calibrated (hereinafter “post-equating”). The results presented in this appendix provide support for utilizing the pre-equated solution for all student reporting. A complete description of the pre-equating verification process is discussed in Chapter Fifteen.

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

WINTER

Figure L-1. Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

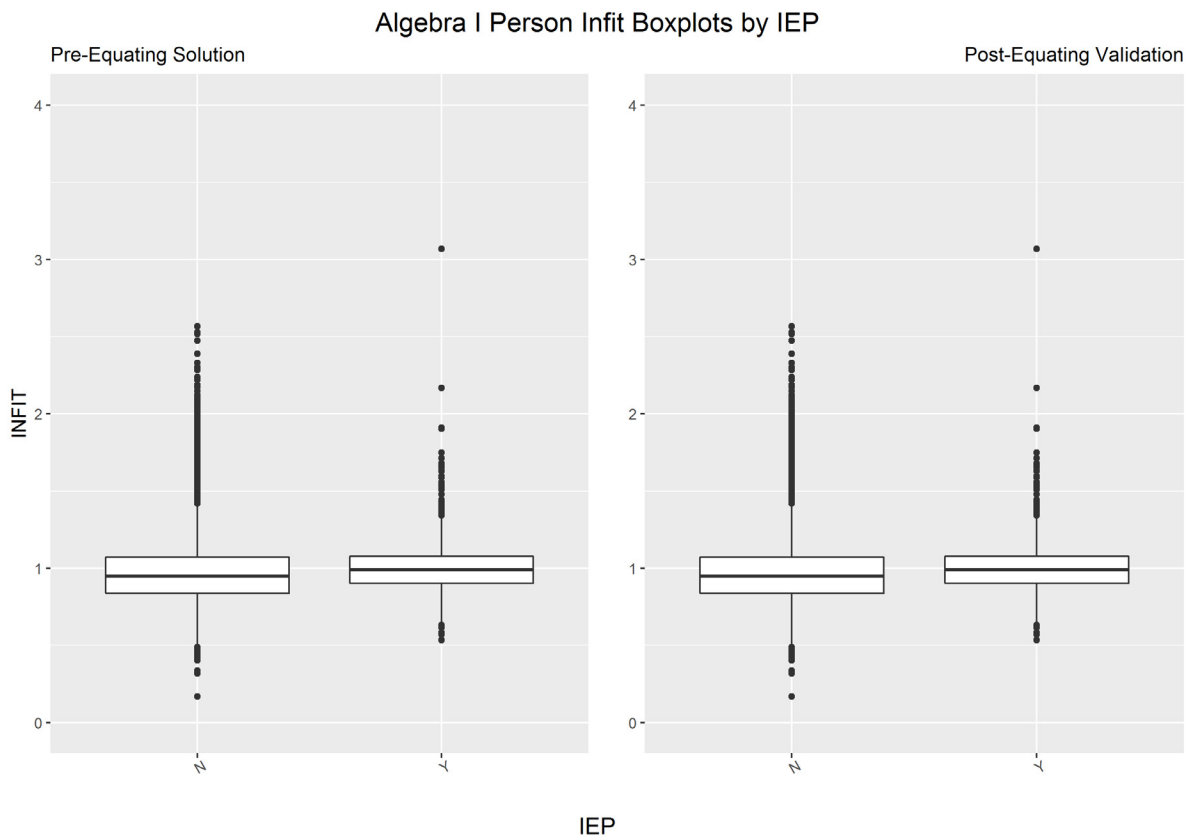
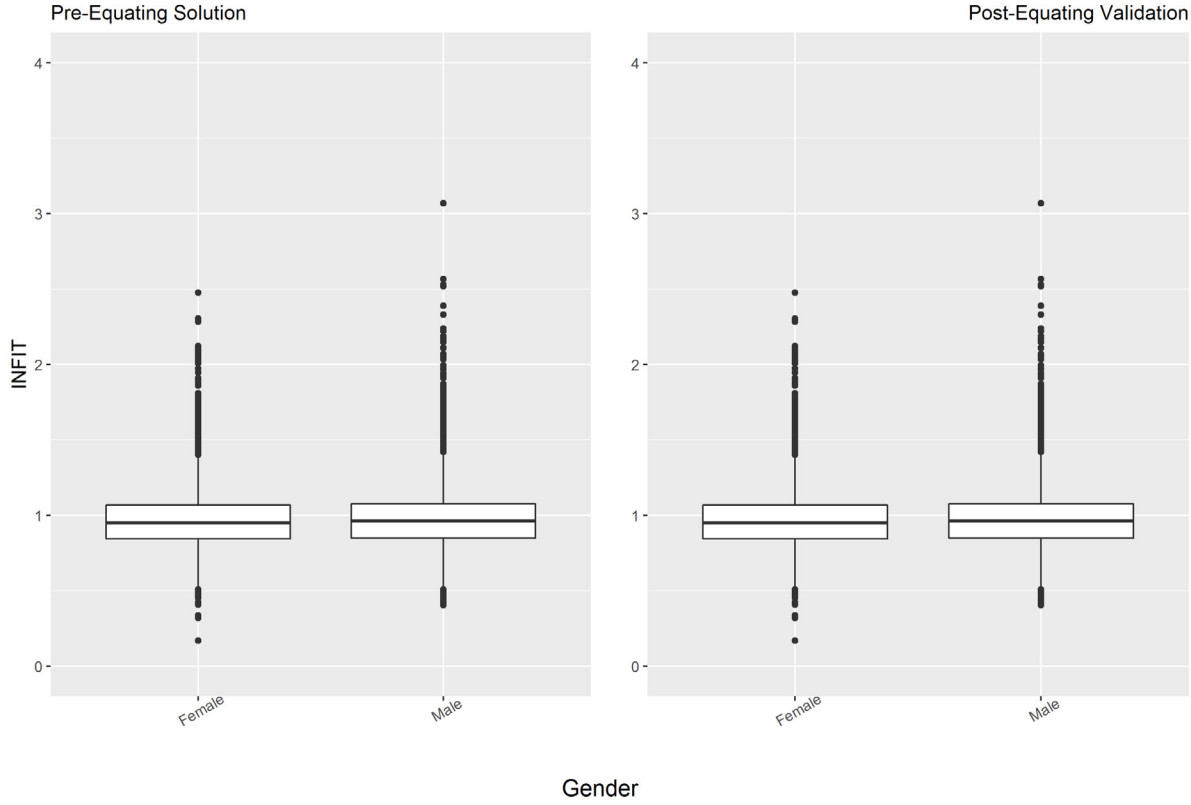


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

Algebra I Person Infit Boxplots by Gender



Algebra I Person Infit Boxplots by Ethnicity

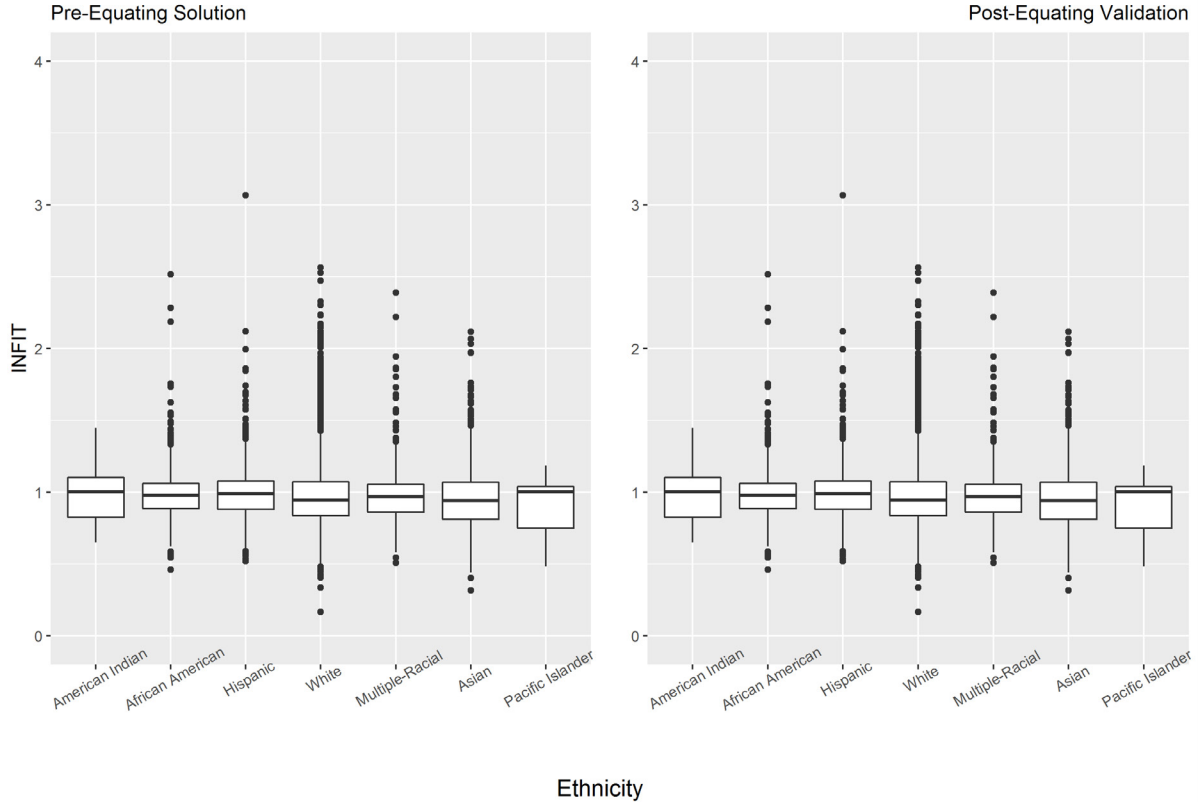


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

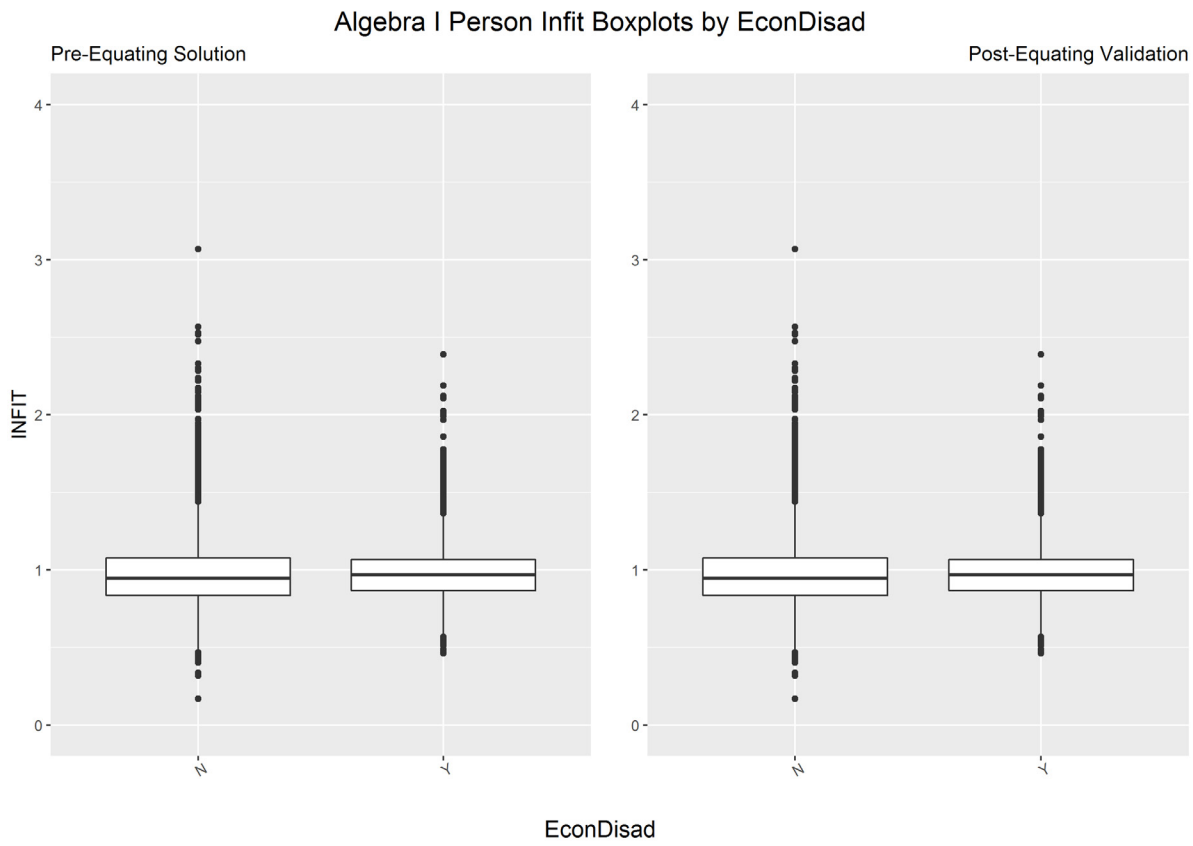
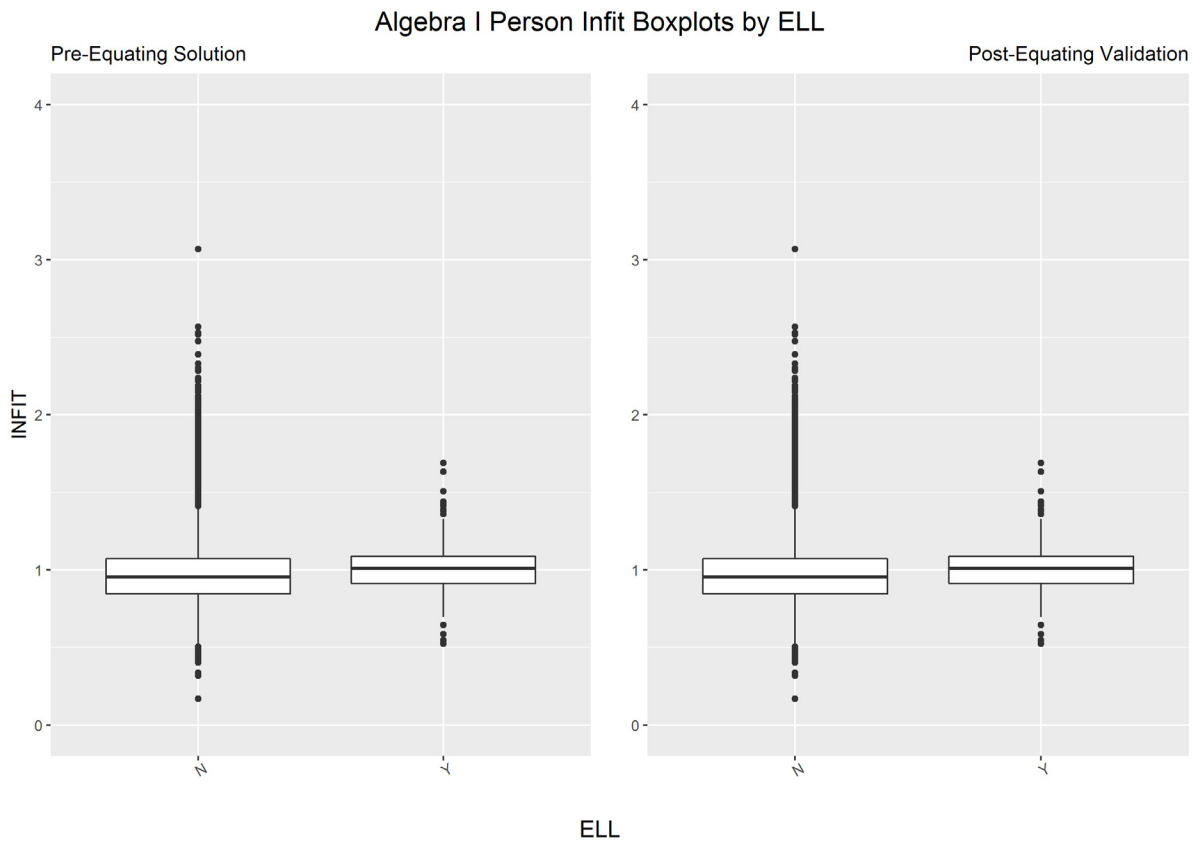


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

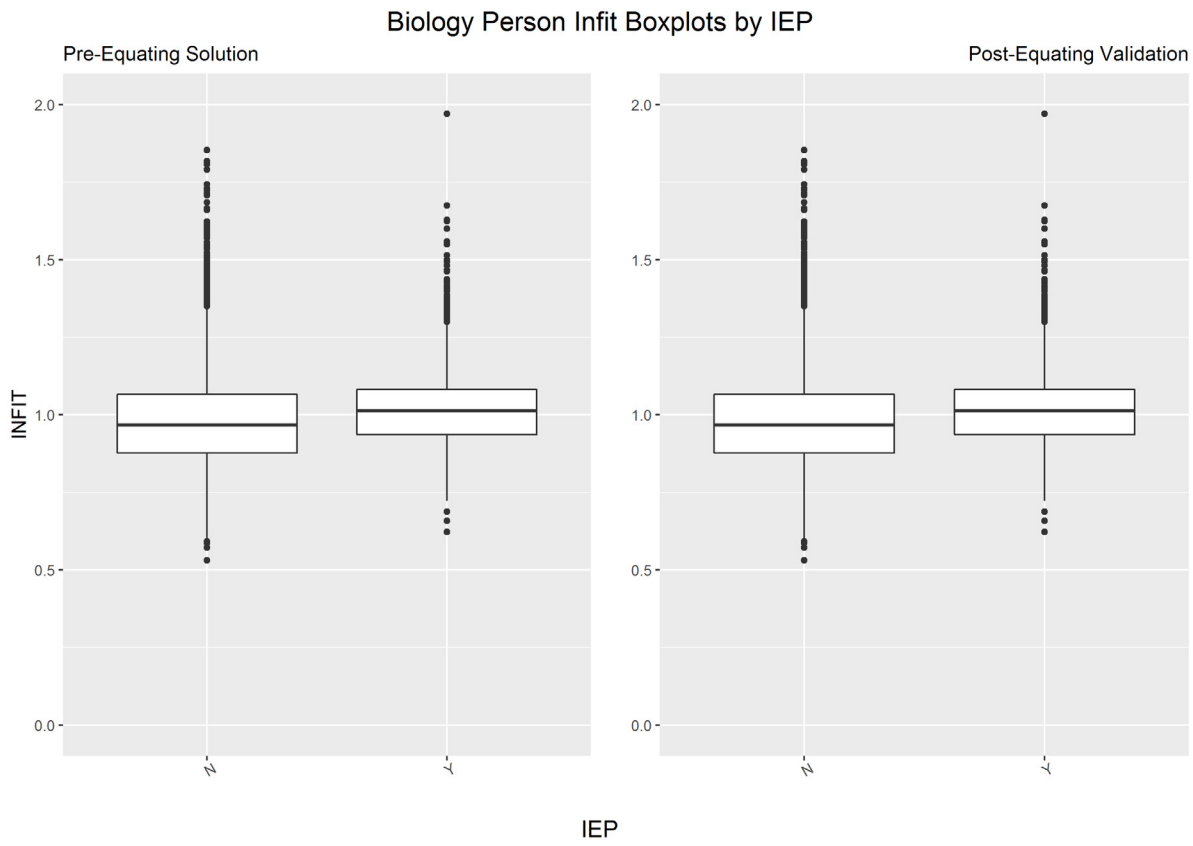


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

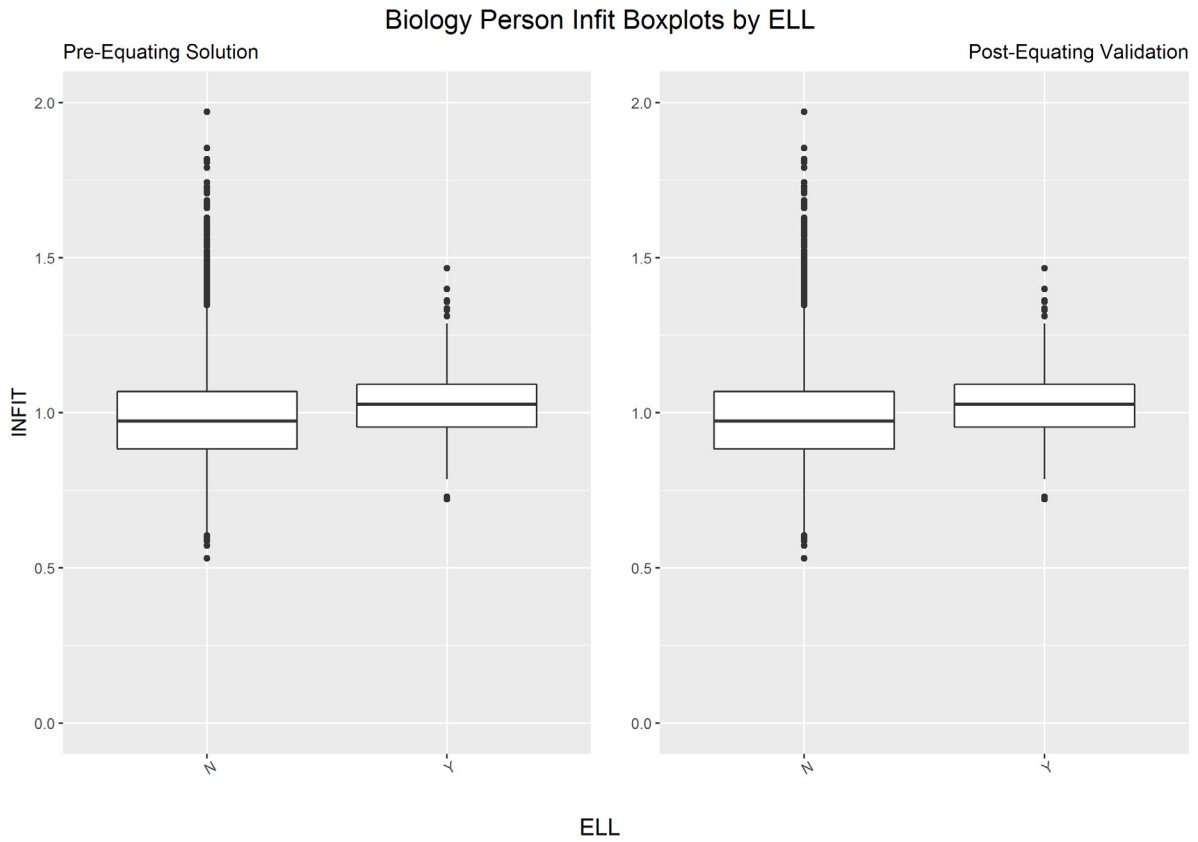
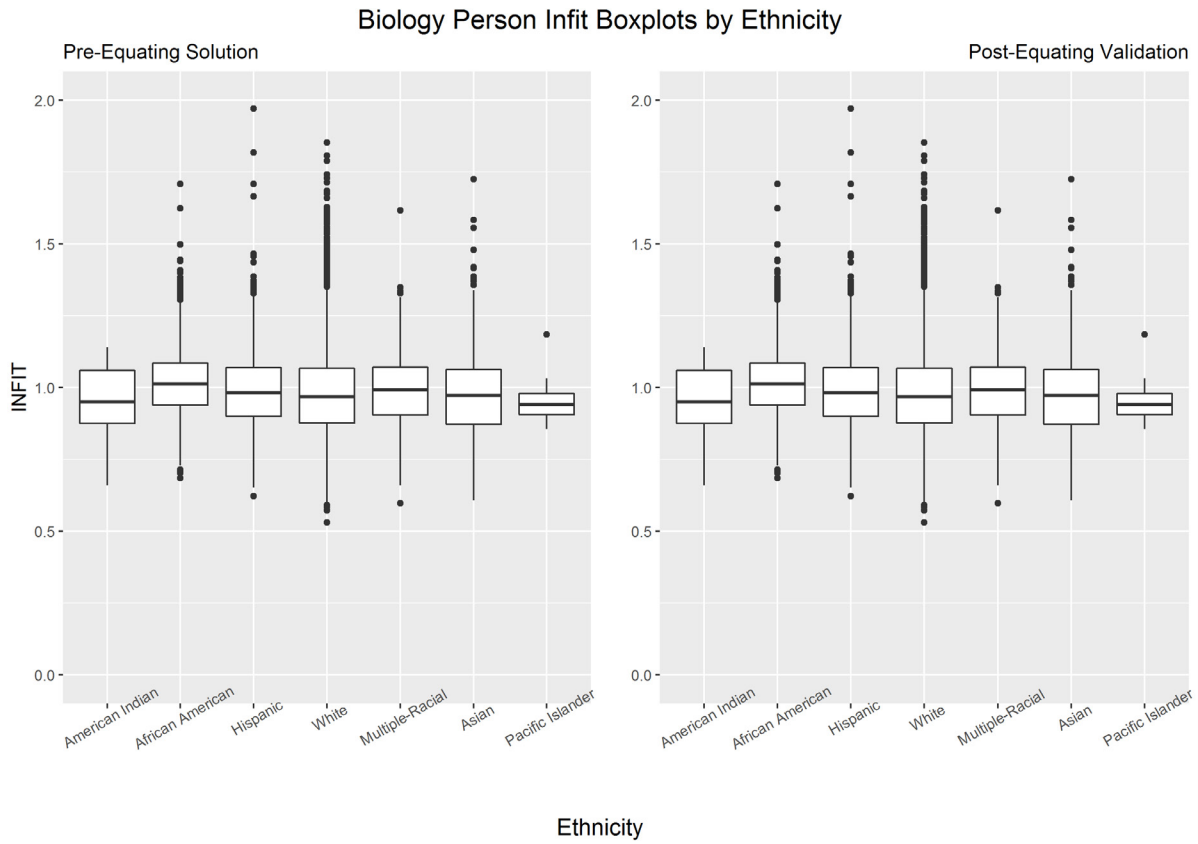


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

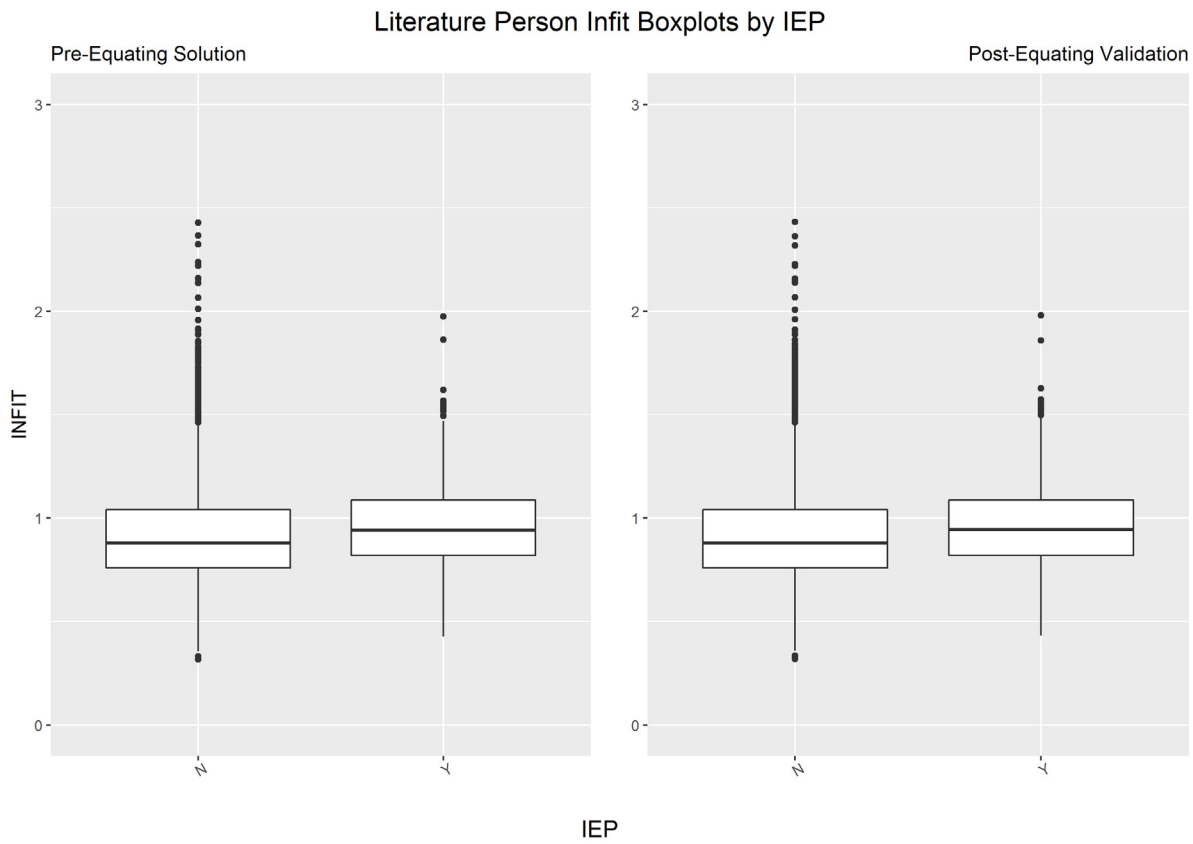
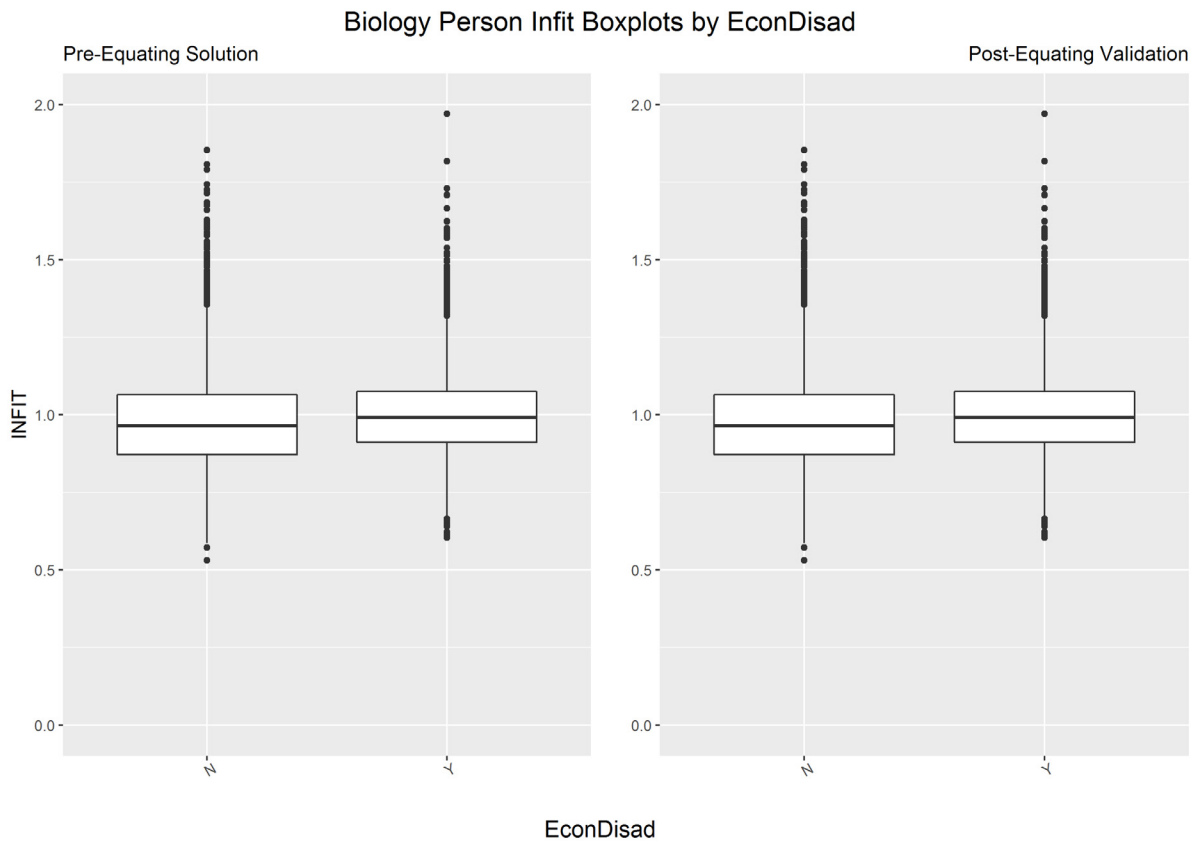


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

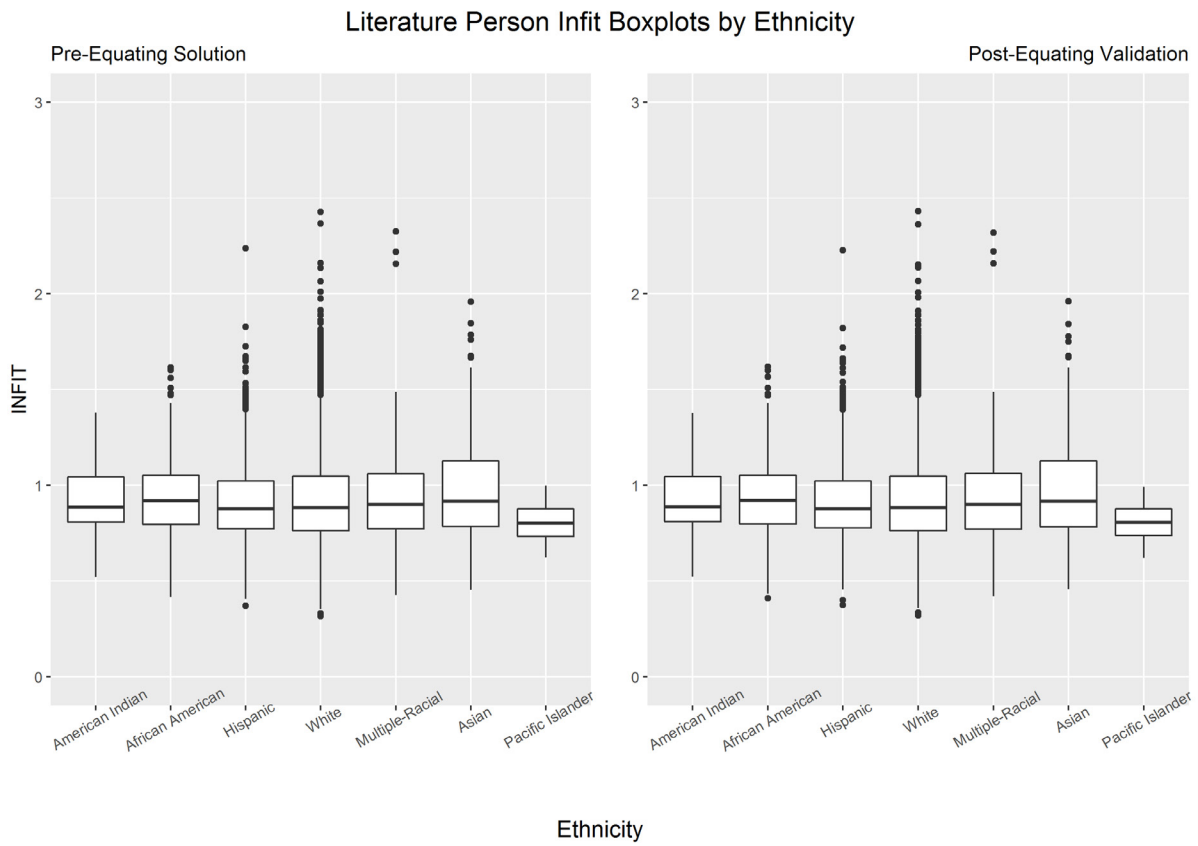
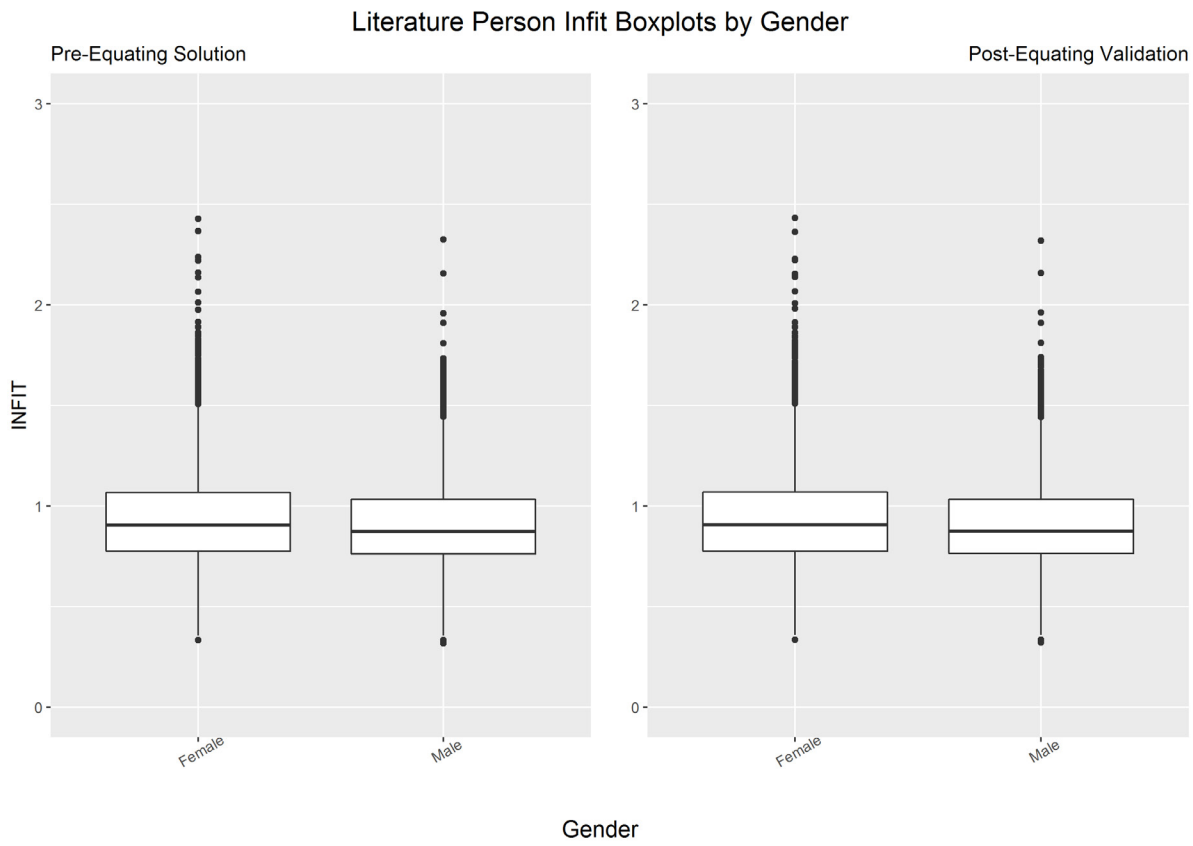


Figure L-1 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Winter

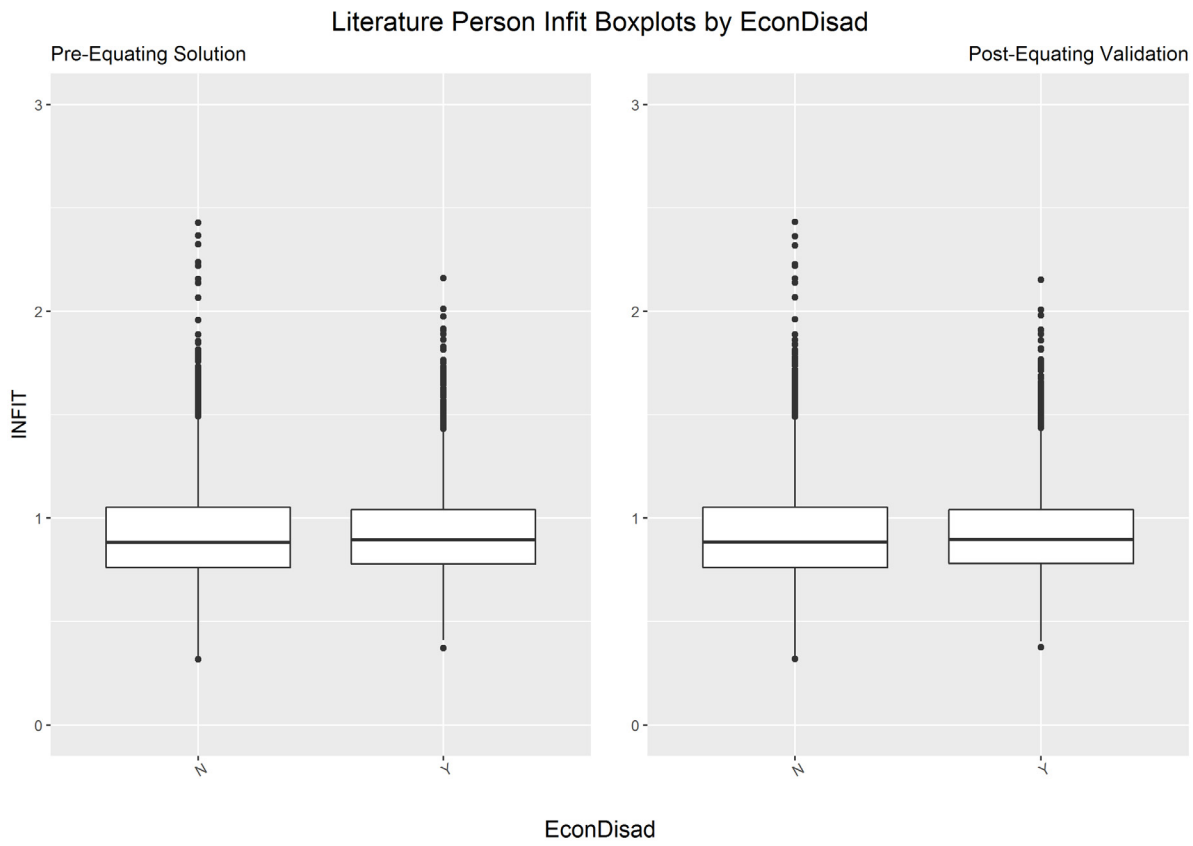
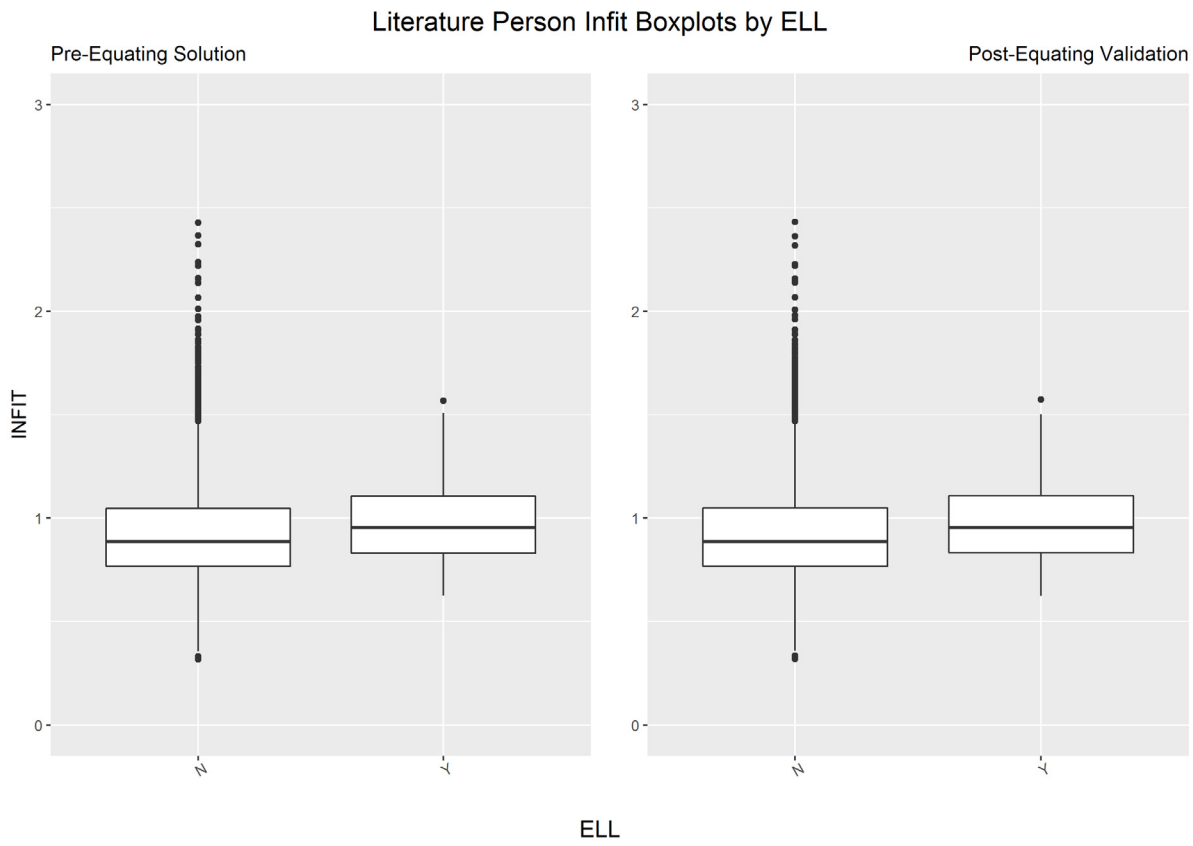


Figure L-2. Normalized Scale Score Distributions by Content Area: Winter

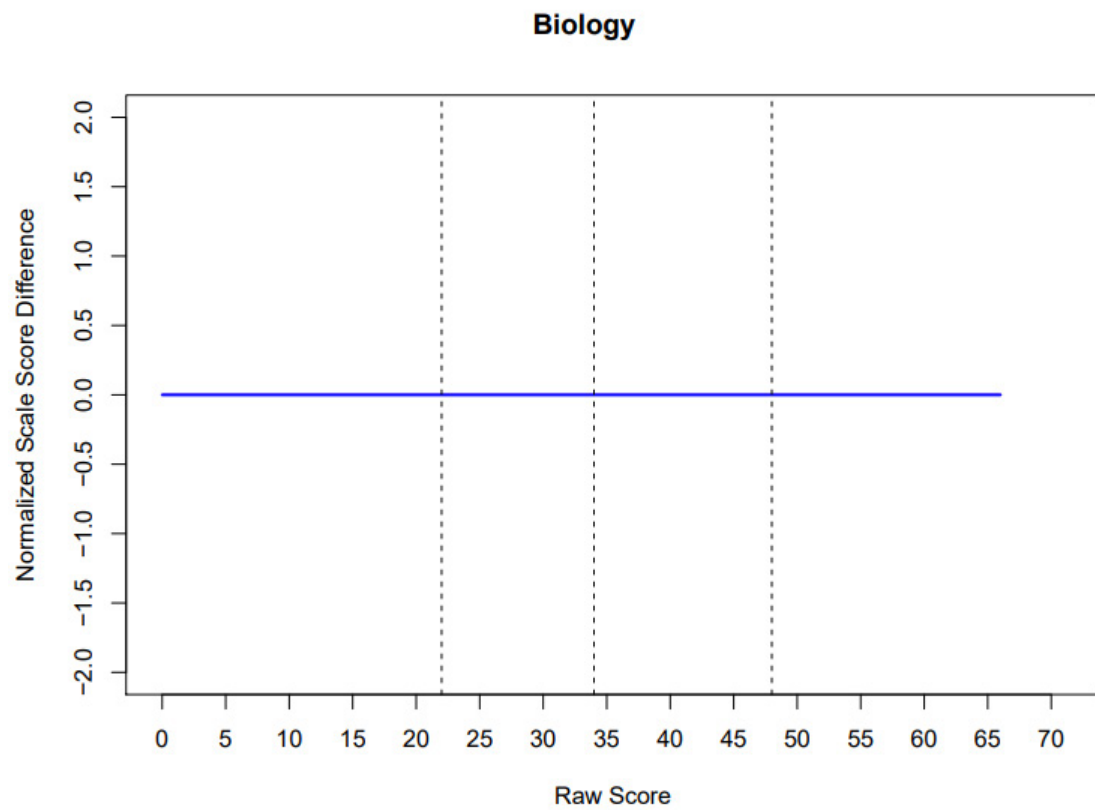
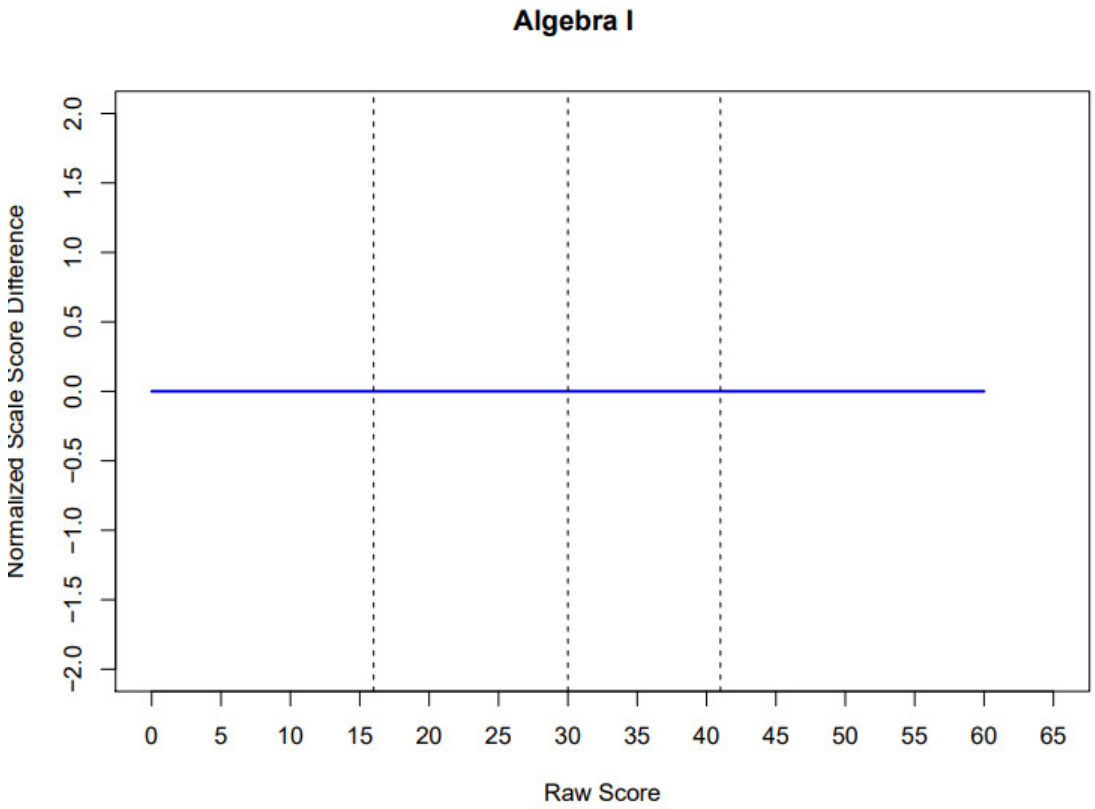


Figure L-2 (continued). Normalized Scale Score Distributions by Content Area: Winter

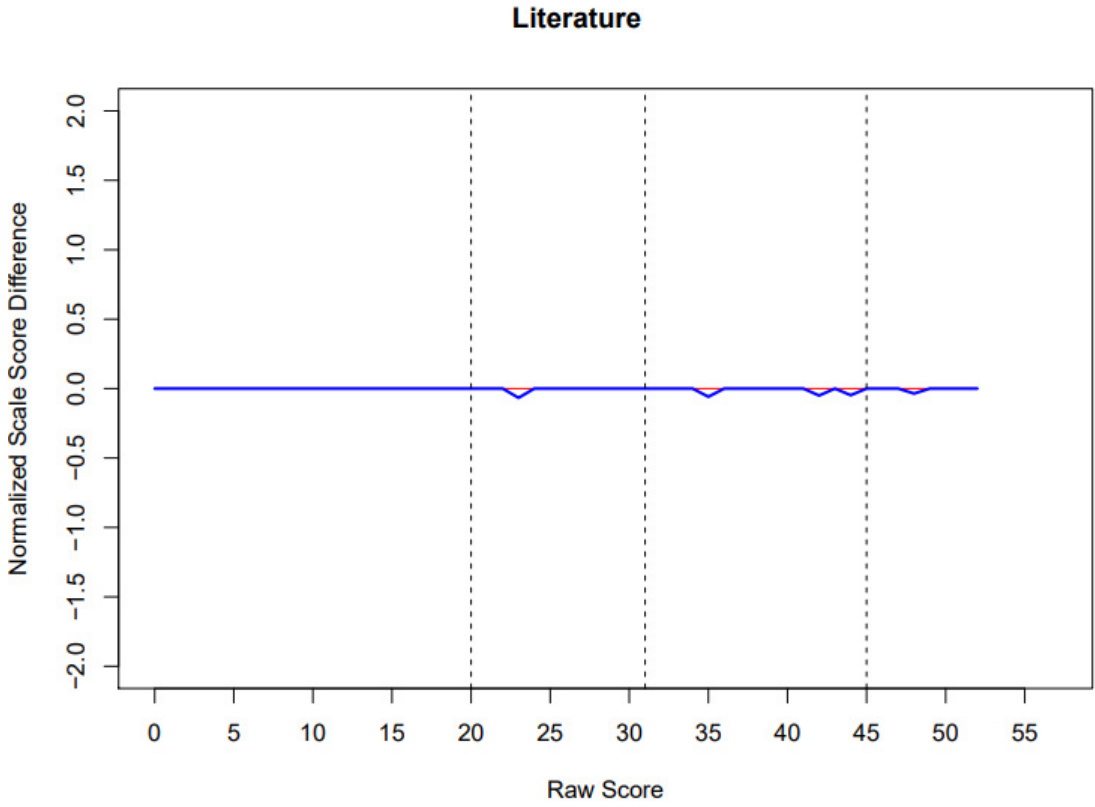


Table L-1. Algebra I Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
0	1218	1218	92	92	BB	BB	0.0	True
1	1279	1279	51	51	BB	BB	0.0	True
2	1315	1315	36	36	BB	BB	0.0	True
3	1337	1337	30	30	BB	BB	0.1	True
4	1353	1353	26	26	BB	BB	0.1	True
5	1365	1365	24	24	BB	BB	0.2	True
6	1376	1376	22	22	BB	BB	0.5	True
7	1385	1385	21	21	BB	BB	0.8	True
8	1393	1393	20	20	BB	BB	1.3	True
9	1401	1401	19	19	BB	BB	1.8	True
10	1408	1408	18	18	BB	BB	2.0	True
11	1414	1414	18	18	BB	BB	2.7	True
12	1420	1420	17	17	BB	BB	3.4	True
13	1426	1426	17	17	BB	BB	3.0	True
14	1431	1431	16	16	BB	BB	3.3	True
15	1436	1436	16	16	BB	BB	3.4	True
16	1441	1441	16	16	B	B	3.9	True
17	1446	1446	15	15	B	B	3.2	True
18	1450	1450	15	15	B	B	3.3	True
19	1455	1455	15	15	B	B	3.0	True
20	1459	1459	15	15	B	B	3.2	True
21	1464	1464	15	15	B	B	3.1	True
22	1468	1468	15	15	B	B	3.3	True
23	1472	1472	14	14	B	B	3.1	True
24	1476	1476	14	14	B	B	2.9	True
25	1481	1481	14	14	B	B	3.0	True
26	1485	1485	14	14	B	B	2.8	True
27	1489	1489	14	14	B	B	2.7	True
28	1493	1493	14	14	B	B	2.7	True
29	1497	1497	14	14	B	B	2.7	True
30	1501	1501	14	14	P	P	2.7	True
31	1505	1505	14	14	P	P	2.7	True
32	1509	1509	14	14	P	P	2.3	True
33	1513	1513	14	14	P	P	2.2	True
34	1517	1517	14	14	P	P	2.0	True
35	1521	1521	14	14	P	P	2.2	True
36	1525	1525	15	15	P	P	2.0	True
37	1530	1530	15	15	P	P	1.7	True
38	1534	1534	15	15	P	P	1.6	True
39	1539	1539	15	15	P	P	1.5	True
40	1543	1543	15	15	P	P	1.4	True
41	1548	1548	16	16	A	A	1.4	True

Table L-1 (continued). Algebra I Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
42	1553	1553	16	16	A	A	1.4	True
43	1558	1558	16	16	A	A	1.2	True
44	1563	1563	17	17	A	A	1.2	True
45	1569	1569	17	17	A	A	0.9	True
46	1575	1575	17	17	A	A	1.0	True
47	1581	1581	18	18	A	A	0.9	True
48	1588	1588	19	19	A	A	0.8	True
49	1595	1595	19	19	A	A	0.7	True
50	1603	1603	20	20	A	A	0.6	True
51	1611	1611	21	21	A	A	0.5	True
52	1621	1621	23	23	A	A	0.5	True
53	1632	1632	24	24	A	A	0.4	True
54	1645	1645	26	26	A	A	0.2	True
55	1660	1660	29	29	A	A	0.2	True
56	1678	1678	32	32	A	A	0.1	True
57	1700	1700	35	35	A	A	0.1	True
58	1730	1730	41	41	A	A	0.1	True
59	1773	1773	54	54	A	A	0.0	True
60	1800	1800	94	94	A	A	0.0	True

Table L-2. Biology Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
0	1222	1222	92	92	BB	BB	0.0	True
1	1283	1283	51	51	BB	BB	0.0	True
2	1318	1318	36	36	BB	BB	0.0	True
3	1340	1340	30	30	BB	BB	0.0	True
4	1355	1355	26	26	BB	BB	0.0	True
5	1367	1367	24	24	BB	BB	0.0	True
6	1378	1378	22	22	BB	BB	0.1	True
7	1386	1386	20	20	BB	BB	0.1	True
8	1394	1394	19	19	BB	BB	0.3	True
9	1401	1401	18	18	BB	BB	0.4	True
10	1408	1408	17	17	BB	BB	0.8	True
11	1413	1413	17	17	BB	BB	0.9	True
12	1419	1419	16	16	BB	BB	1.3	True
13	1424	1424	16	16	BB	BB	1.7	True
14	1429	1429	15	15	BB	BB	1.7	True
15	1434	1434	15	15	BB	BB	2.0	True
16	1438	1438	15	15	BB	BB	2.2	True
17	1442	1442	14	14	BB	BB	2.3	True
18	1446	1446	14	14	BB	BB	2.3	True
19	1450	1450	14	14	BB	BB	2.4	True
20	1454	1454	14	14	BB	BB	2.4	True
21	1458	1458	14	14	BB	BB	2.5	True
22	1461	1461	13	13	B	B	2.7	True
23	1465	1465	13	13	B	B	2.3	True
24	1468	1468	13	13	B	B	2.3	True
25	1472	1472	13	13	B	B	2.1	True
26	1475	1475	13	13	B	B	2.4	True
27	1478	1478	13	13	B	B	2.3	True
28	1482	1482	13	13	B	B	2.5	True
29	1485	1485	13	13	B	B	2.2	True
30	1488	1488	13	13	B	B	2.3	True
31	1491	1491	13	13	B	B	2.4	True
32	1494	1494	13	13	B	B	2.2	True
33	1498	1498	13	13	B	B	2.3	True
34	1501	1501	13	13	P	P	2.2	True
35	1504	1504	13	13	P	P	2.3	True
36	1507	1507	13	13	P	P	1.9	True
37	1510	1510	13	13	P	P	2.2	True
38	1514	1514	13	13	P	P	2.0	True
39	1517	1517	13	13	P	P	2.1	True
40	1520	1520	13	13	P	P	1.9	True
41	1524	1524	13	13	P	P	2.1	True

Table L-2 (continued). Biology Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
42	1527	1527	13	13	P	P	2.0	True
43	1530	1530	13	13	P	P	2.0	True
44	1534	1534	13	13	P	P	2.1	True
45	1538	1538	14	14	P	P	2.1	True
46	1541	1541	14	14	P	P	1.9	True
47	1545	1545	14	14	P	P	1.9	True
48	1549	1549	14	14	A	A	1.8	True
49	1553	1553	14	14	A	A	1.7	True
50	1558	1558	15	15	A	A	1.8	True
51	1562	1562	15	15	A	A	1.6	True
52	1567	1567	15	15	A	A	1.9	True
53	1572	1572	16	16	A	A	1.8	True
54	1577	1577	16	16	A	A	1.6	True
55	1582	1582	17	17	A	A	1.4	True
56	1588	1588	18	18	A	A	1.5	True
57	1595	1595	18	18	A	A	1.3	True
58	1602	1602	19	19	A	A	0.9	True
59	1610	1610	21	21	A	A	0.8	True
60	1619	1619	22	22	A	A	0.7	True
61	1630	1630	24	24	A	A	0.5	True
62	1642	1642	26	26	A	A	0.5	True
63	1658	1658	30	30	A	A	0.2	True
64	1680	1680	37	37	A	A	0.1	True
65	1717	1717	51	51	A	A	0.0	True
66	1778	1778	92	92	A	A	0.0	True

Table L-3. Literature Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
0	1200	1200	92	92	BB	BB	0.0	True
1	1258	1258	51	51	BB	BB	0.0	True
2	1295	1295	37	37	BB	BB	0.0	True
3	1317	1317	31	31	BB	BB	0.0	True
4	1334	1334	27	27	BB	BB	0.1	True
5	1347	1347	24	24	BB	BB	0.1	True
6	1358	1358	23	23	BB	BB	0.2	True
7	1367	1367	21	21	BB	BB	0.3	True
8	1376	1376	20	20	BB	BB	0.5	True
9	1384	1384	19	19	BB	BB	0.4	True
10	1391	1391	19	19	BB	BB	0.7	True
11	1398	1398	18	18	BB	BB	0.9	True
12	1404	1404	18	18	BB	BB	0.8	True
13	1410	1410	17	17	BB	BB	1.1	True
14	1416	1416	17	17	BB	BB	1.3	True
15	1422	1422	17	17	BB	BB	1.3	True
16	1427	1427	16	16	BB	BB	1.4	True
17	1432	1432	16	16	BB	BB	1.3	True
18	1437	1437	16	16	BB	BB	1.4	True
19	1442	1442	16	16	BB	BB	1.7	True
20	1447	1447	16	16	B	B	1.4	True
21	1452	1452	15	15	B	B	1.6	True
22	1457	1457	15	15	B	B	1.7	True
23	1461	1462	15	15	B	B	2.0	True
24	1466	1466	15	15	B	B	2.2	True
25	1471	1471	15	15	B	B	2.3	True
26	1476	1476	15	15	B	B	2.3	True
27	1480	1480	15	15	B	B	2.8	True
28	1485	1485	15	15	B	B	2.7	True
29	1490	1490	16	16	B	B	2.8	True
30	1495	1495	16	16	B	B	3.1	True
31	1500	1500	16	16	P	P	3.3	True
32	1505	1505	16	16	P	P	3.4	True
33	1510	1510	16	16	P	P	3.8	True
34	1515	1515	16	16	P	P	3.9	True
35	1520	1521	17	17	P	P	3.8	True
36	1526	1526	17	17	P	P	3.9	True
37	1532	1532	17	17	P	P	4.1	True
38	1538	1538	18	18	P	P	4.2	True
39	1544	1544	18	18	P	P	4.0	True
40	1551	1551	19	19	P	P	4.0	True
41	1558	1558	19	19	P	P	3.9	True

Table L-3 (continued). Literature Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
42	1565	1566	20	20	P	P	3.7	True
43	1574	1574	21	21	P	P	3.3	True
44	1582	1583	21	21	P	P	3.1	True
45	1592	1592	23	23	A	A	2.7	True
46	1603	1603	24	24	A	A	2.2	True
47	1615	1615	26	26	A	A	1.8	True
48	1629	1630	28	28	A	A	1.2	True
49	1647	1647	32	32	A	A	0.9	True
50	1671	1671	38	38	A	A	0.4	True
51	1709	1709	52	52	A	A	0.2	True
52	1772	1772	92	92	A	A	0.0	True

SPRING

Figure L-3. Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

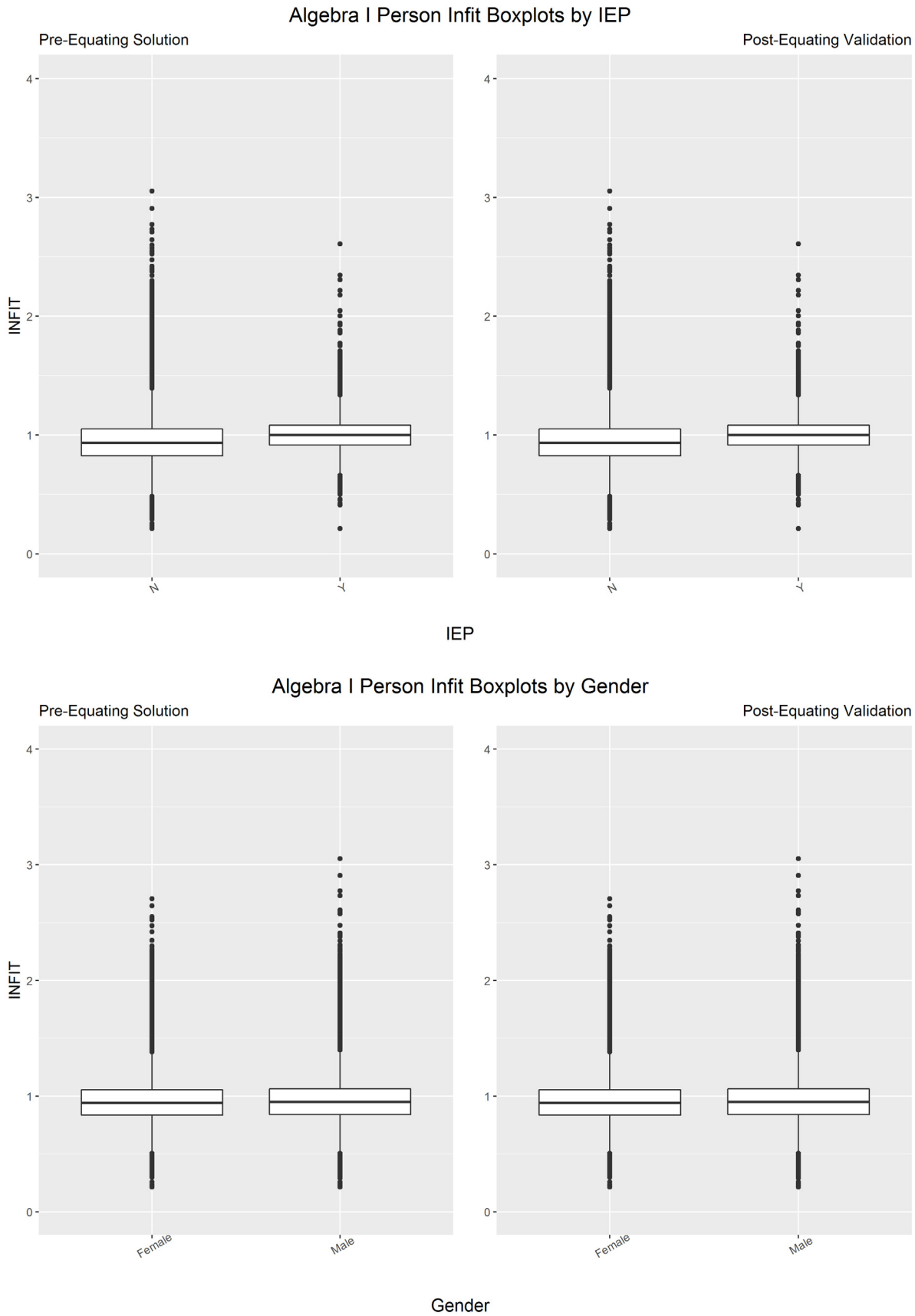


Figure L-3 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

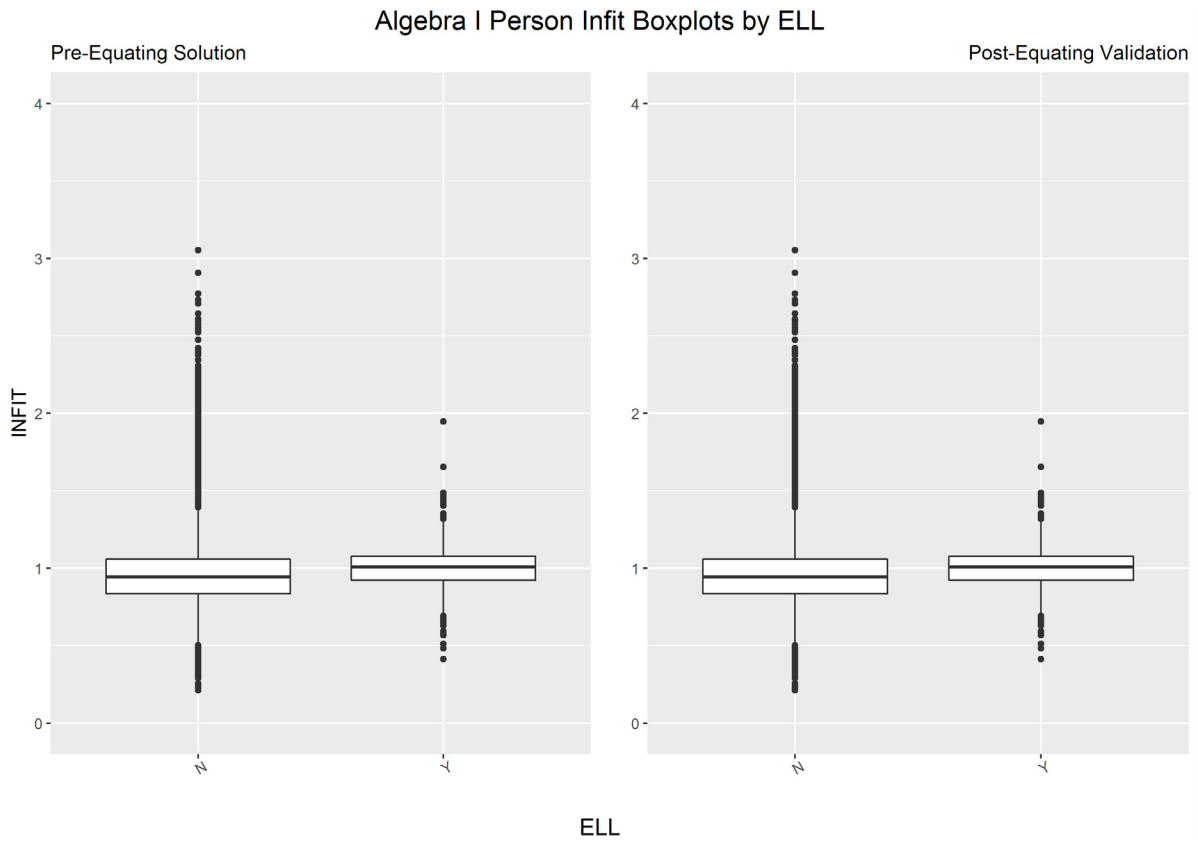
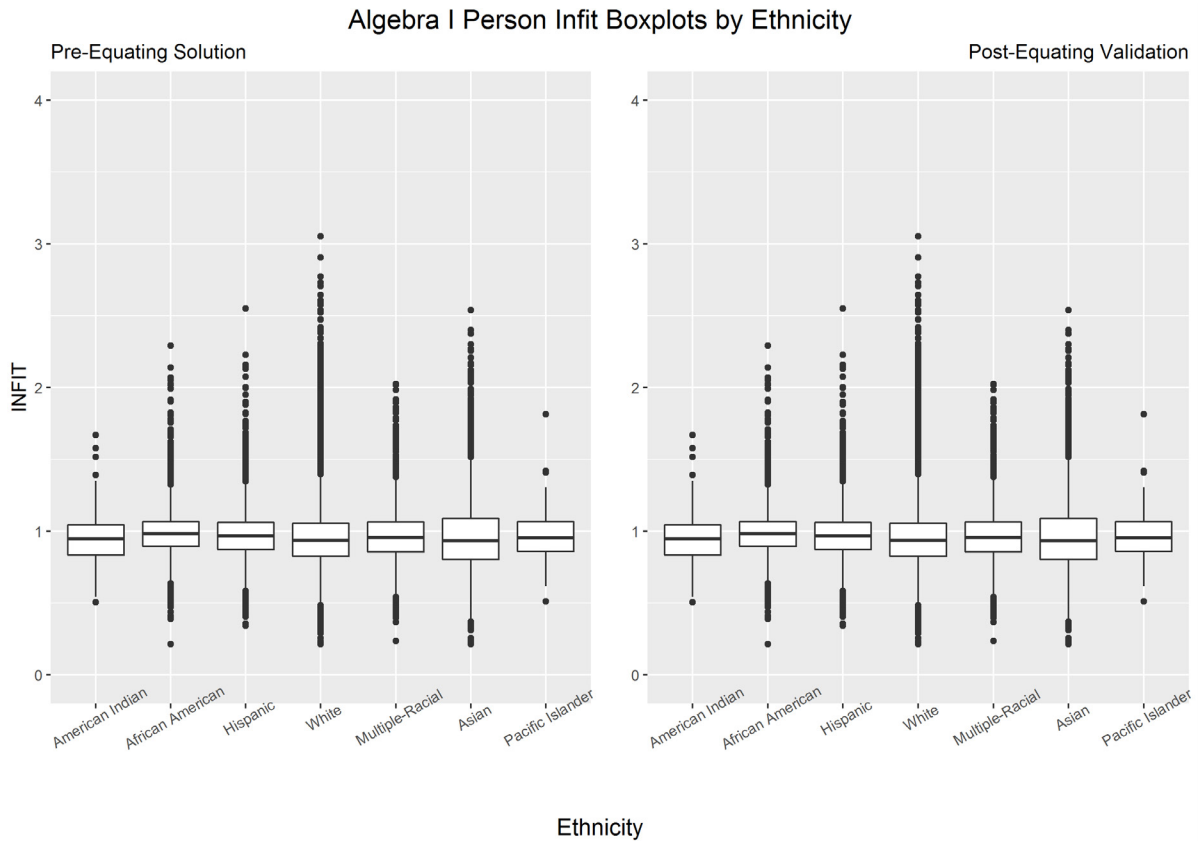


Figure L-3 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

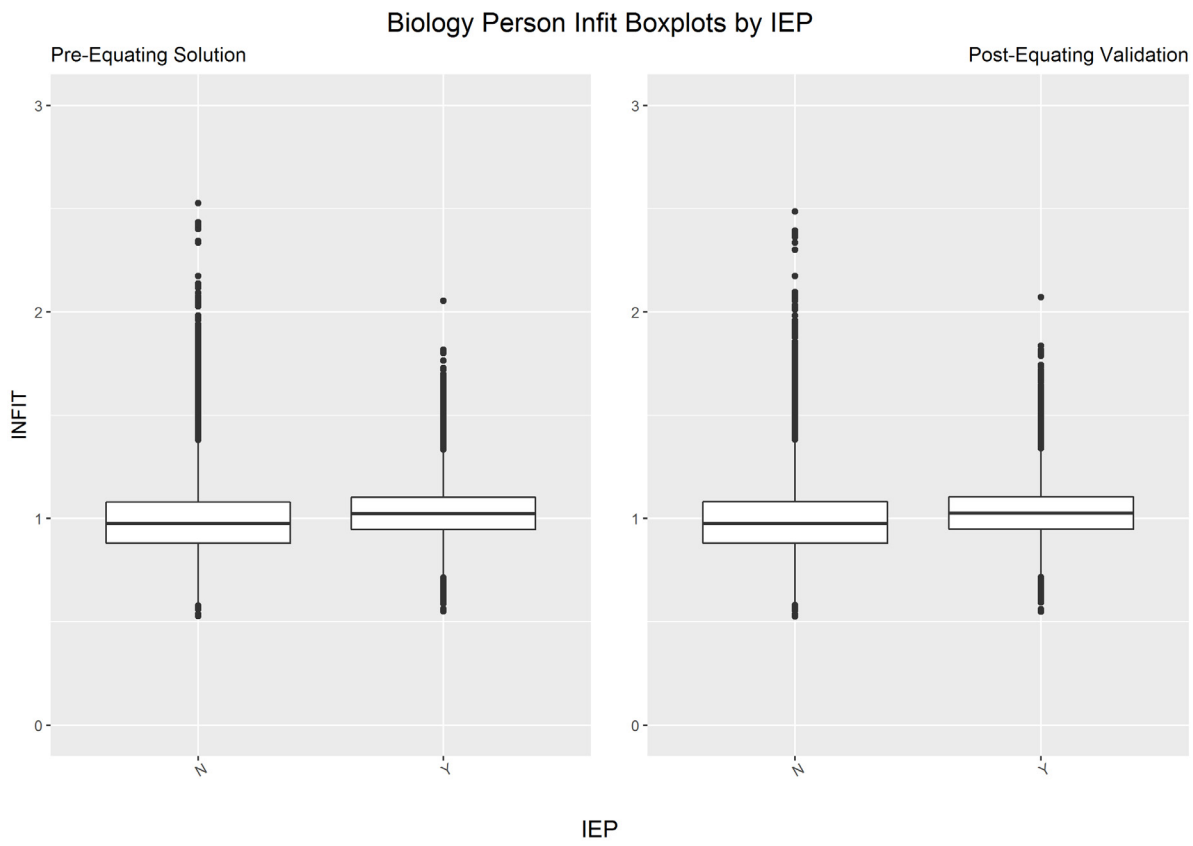
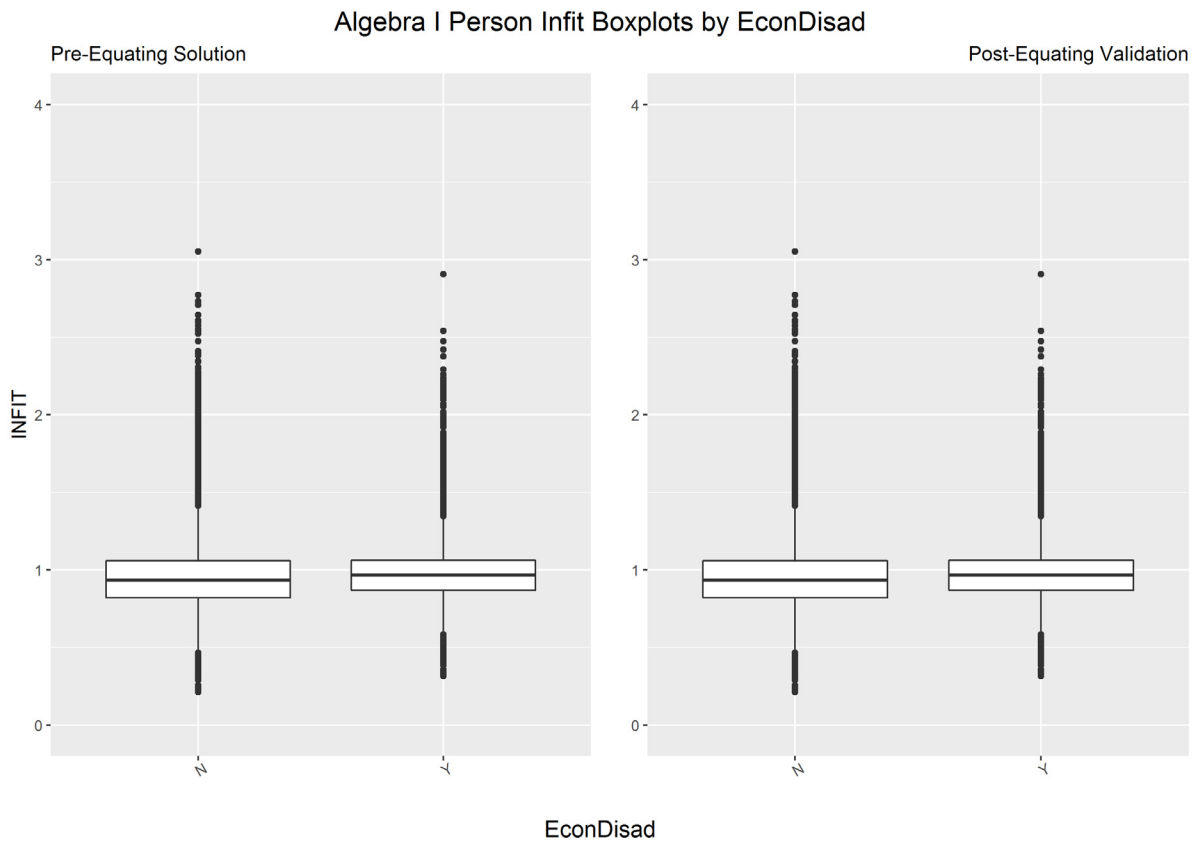


Figure L-3 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

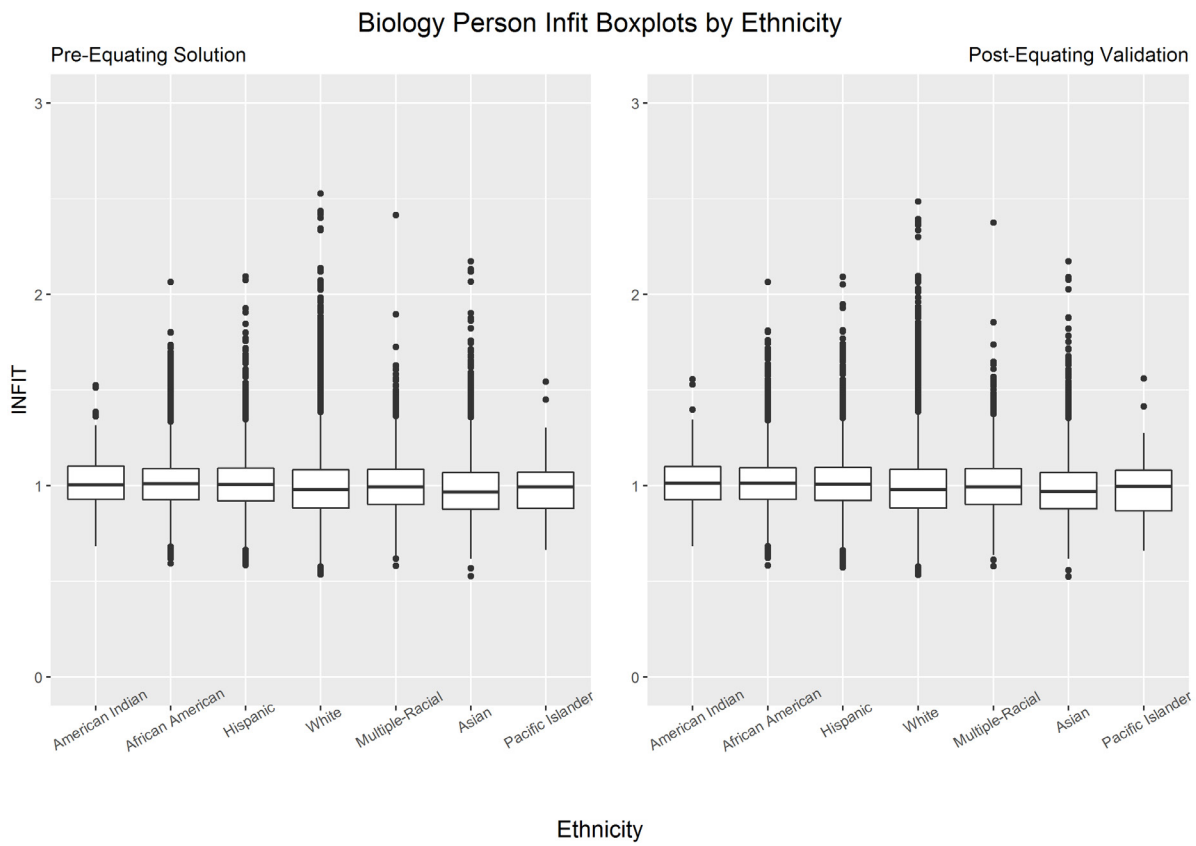


Figure L-3 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

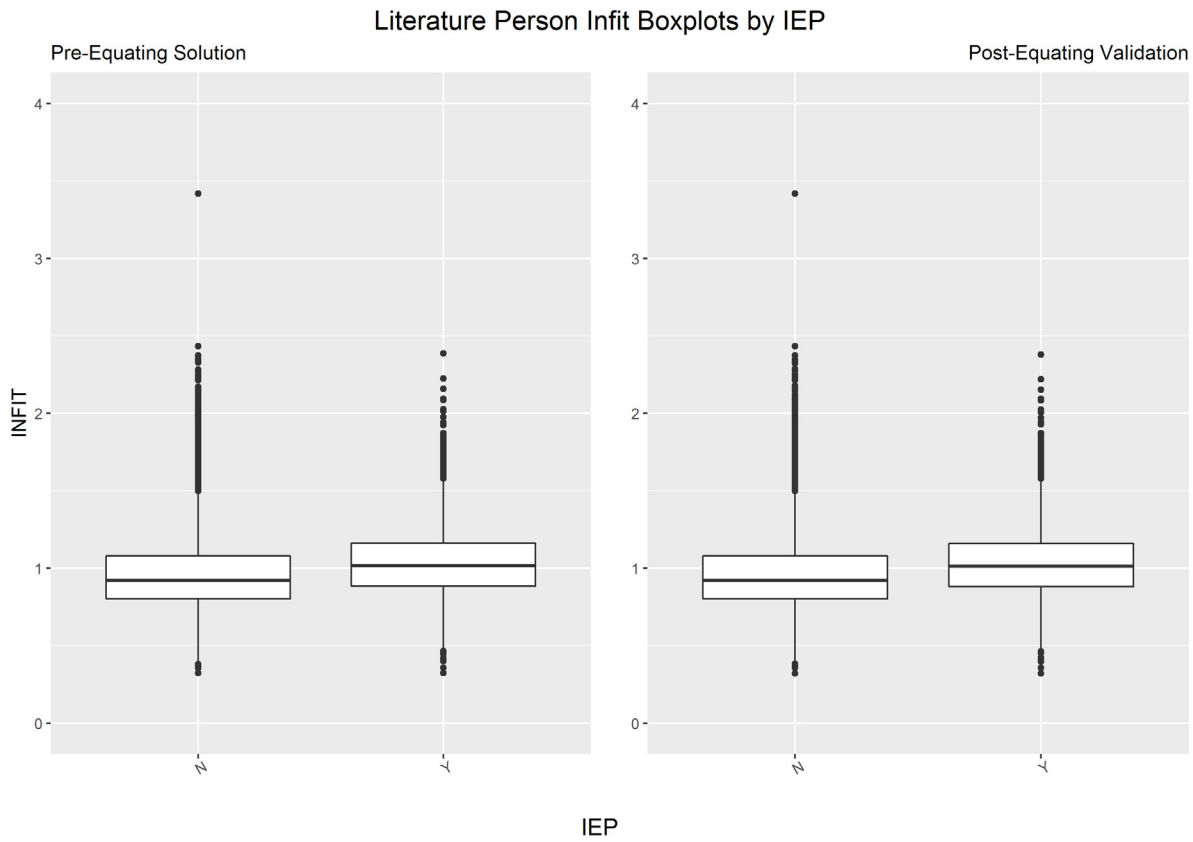
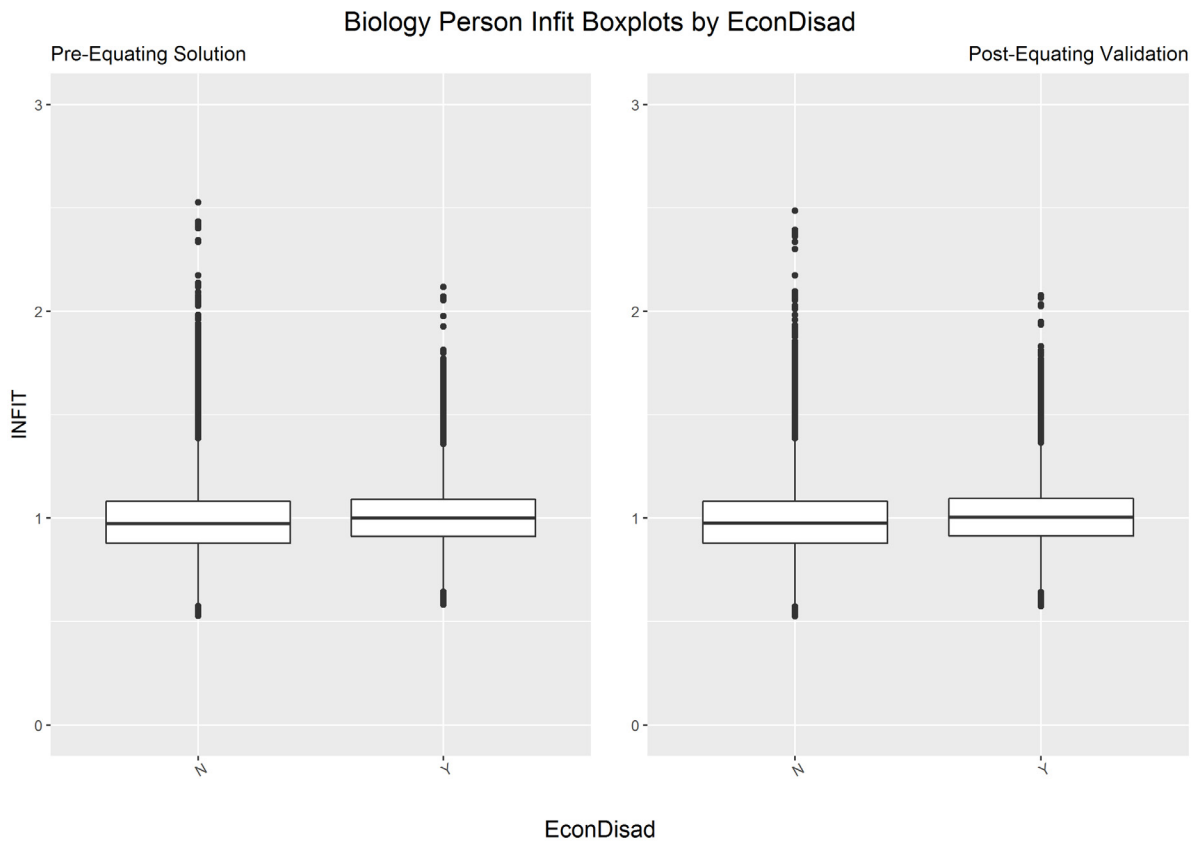


Figure L-3 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

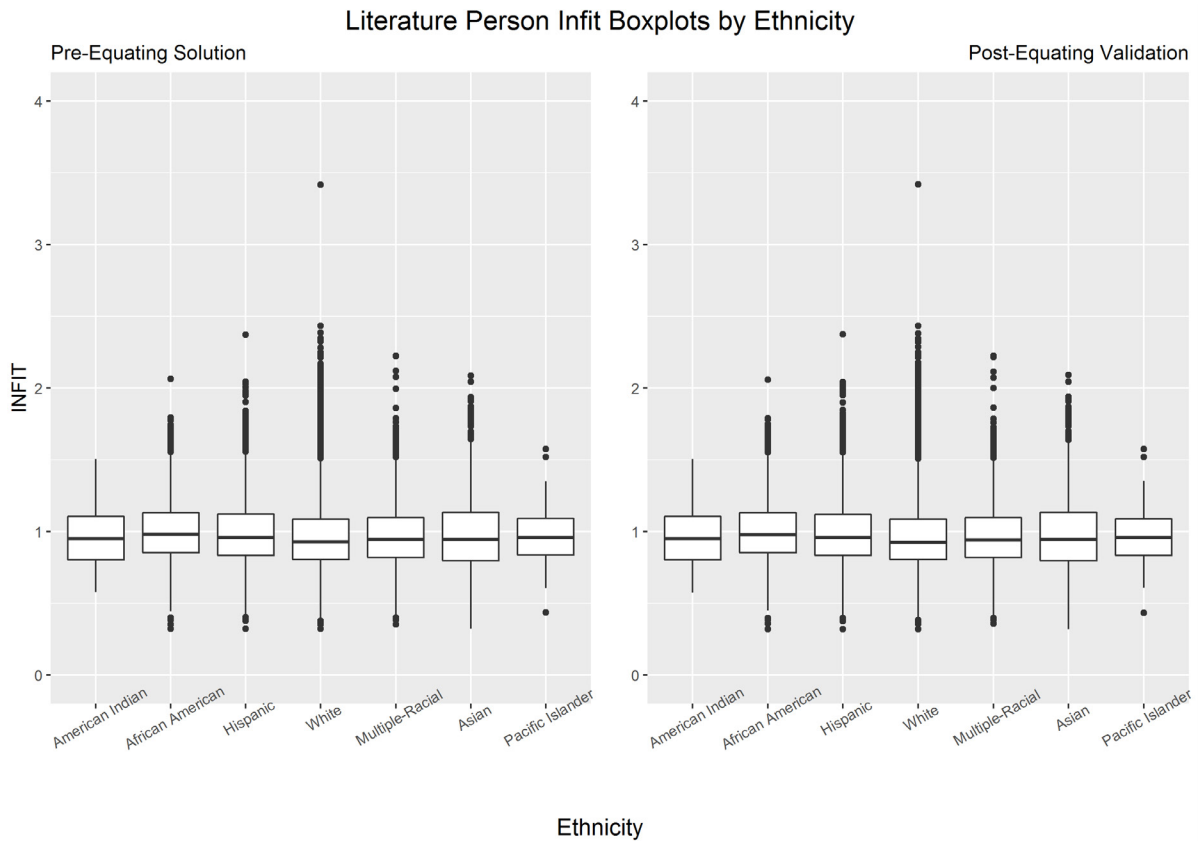
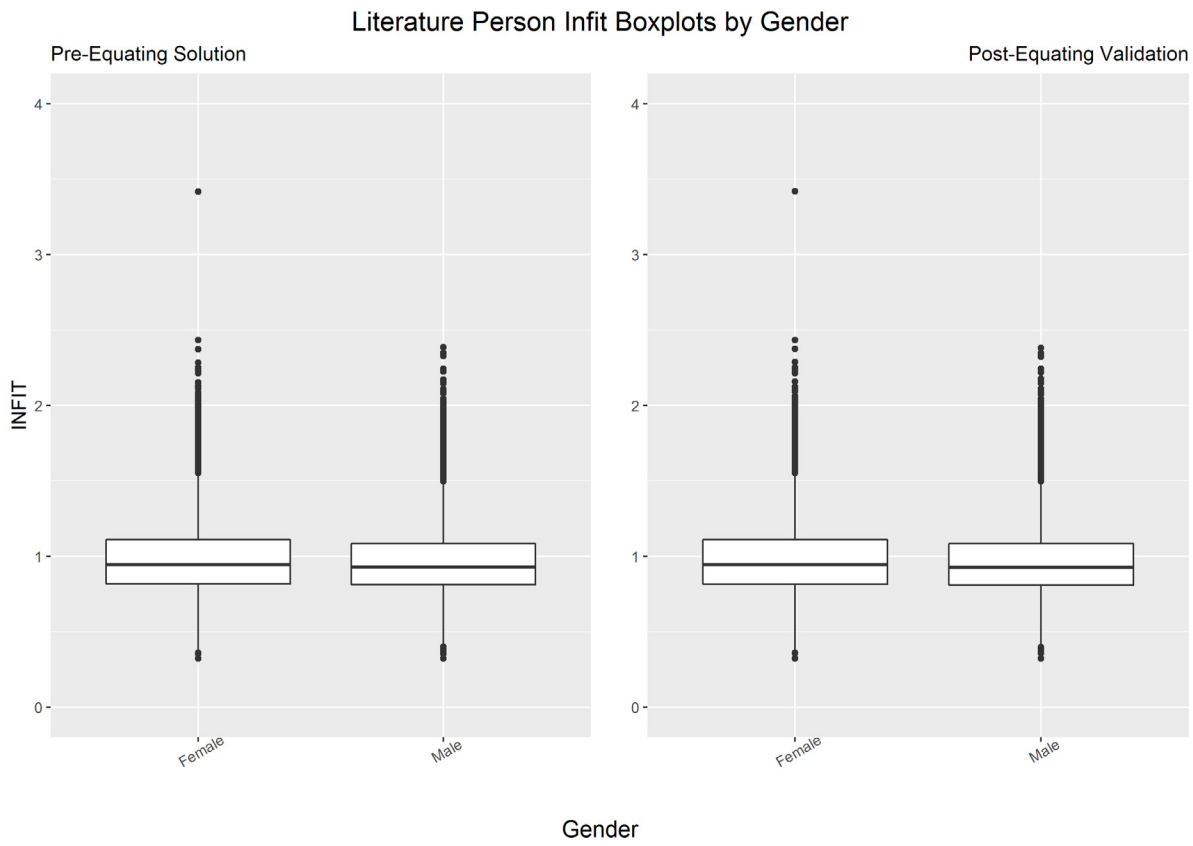


Figure L-3 (continued). Person Infit Boxplots by Content Area for Pre- and Post-Equated Solutions: Spring

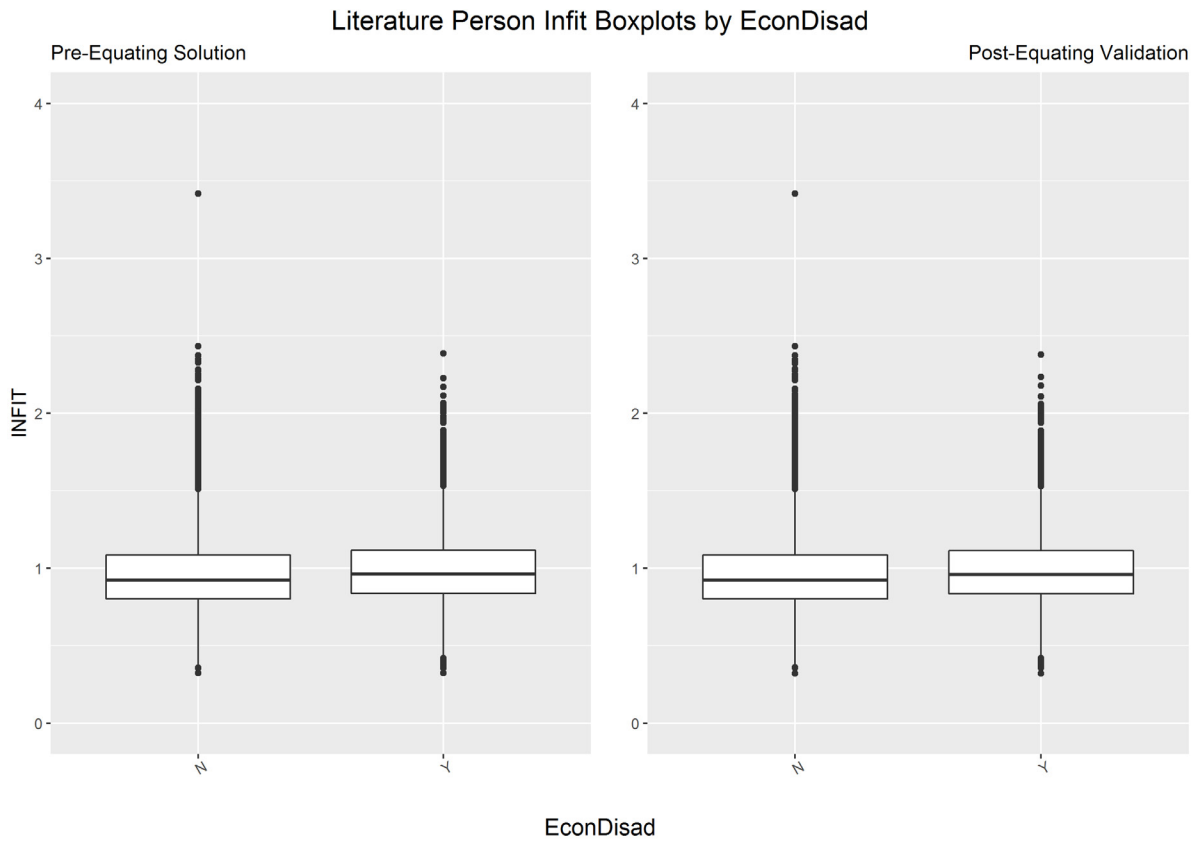
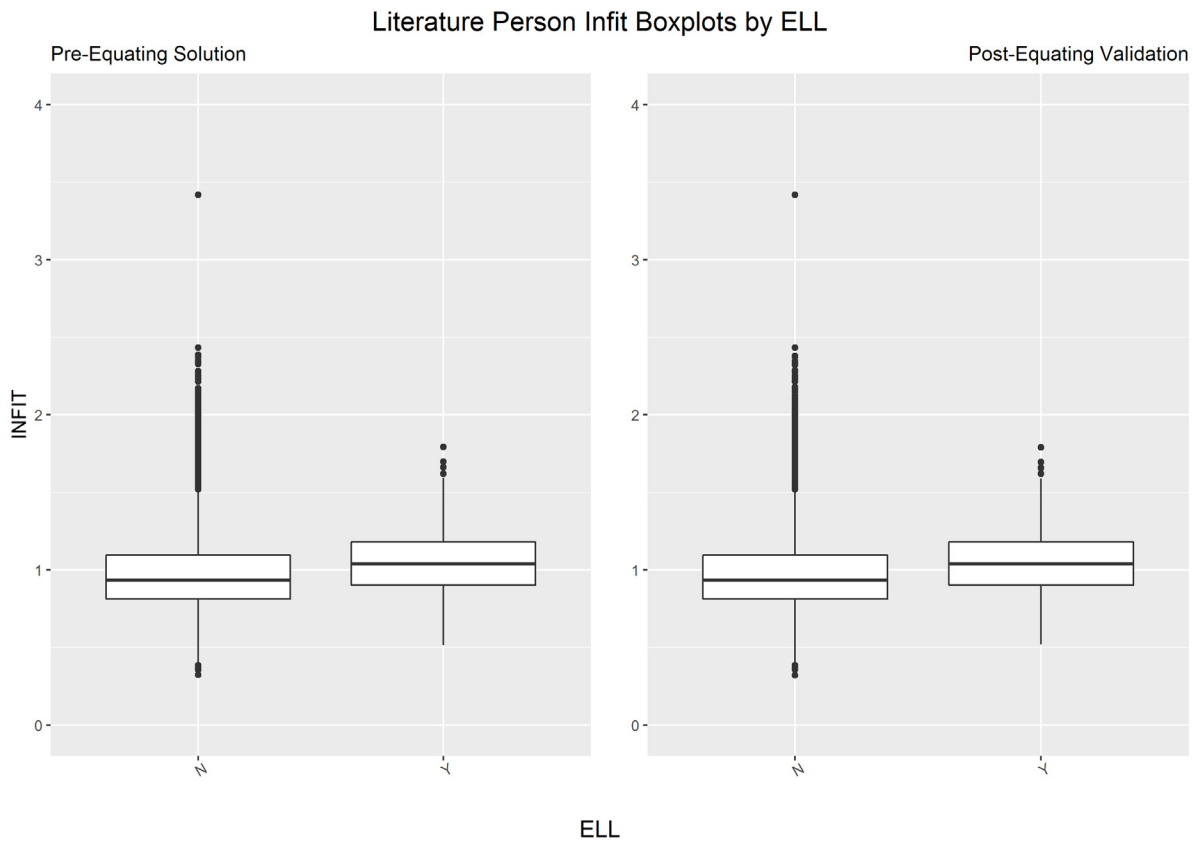


Figure L-4. Normalized Scale Score Distributions by Content Area: Spring

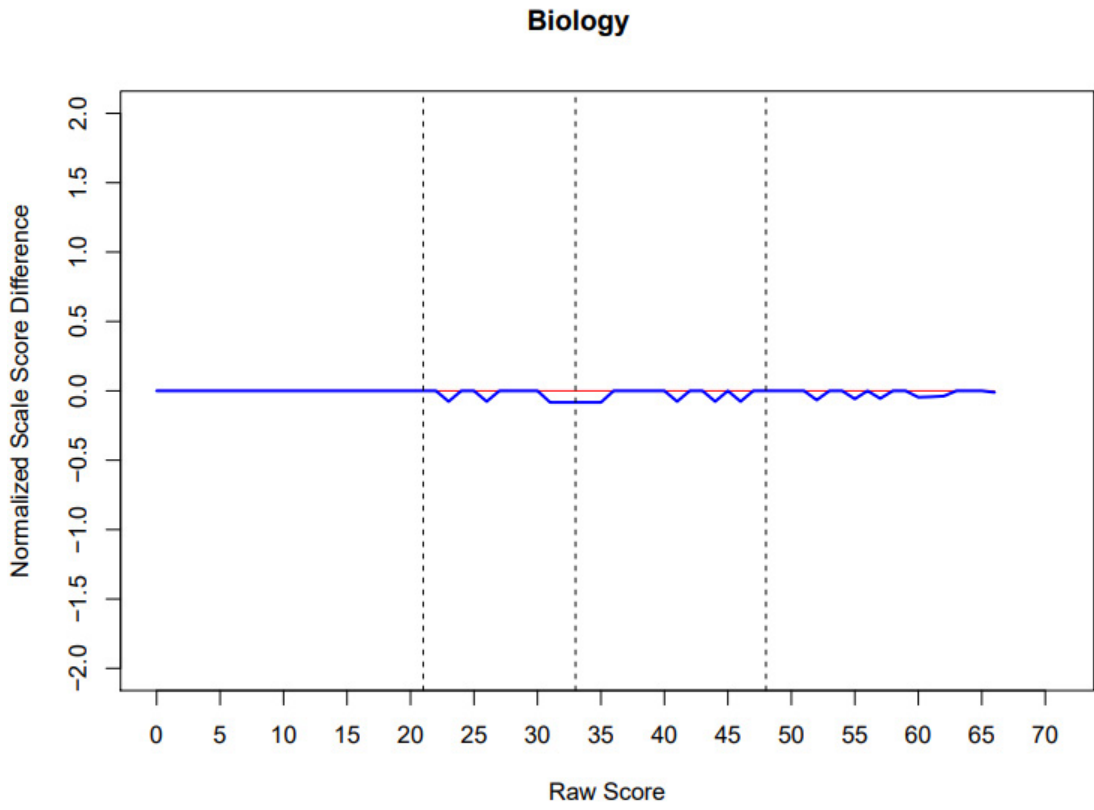
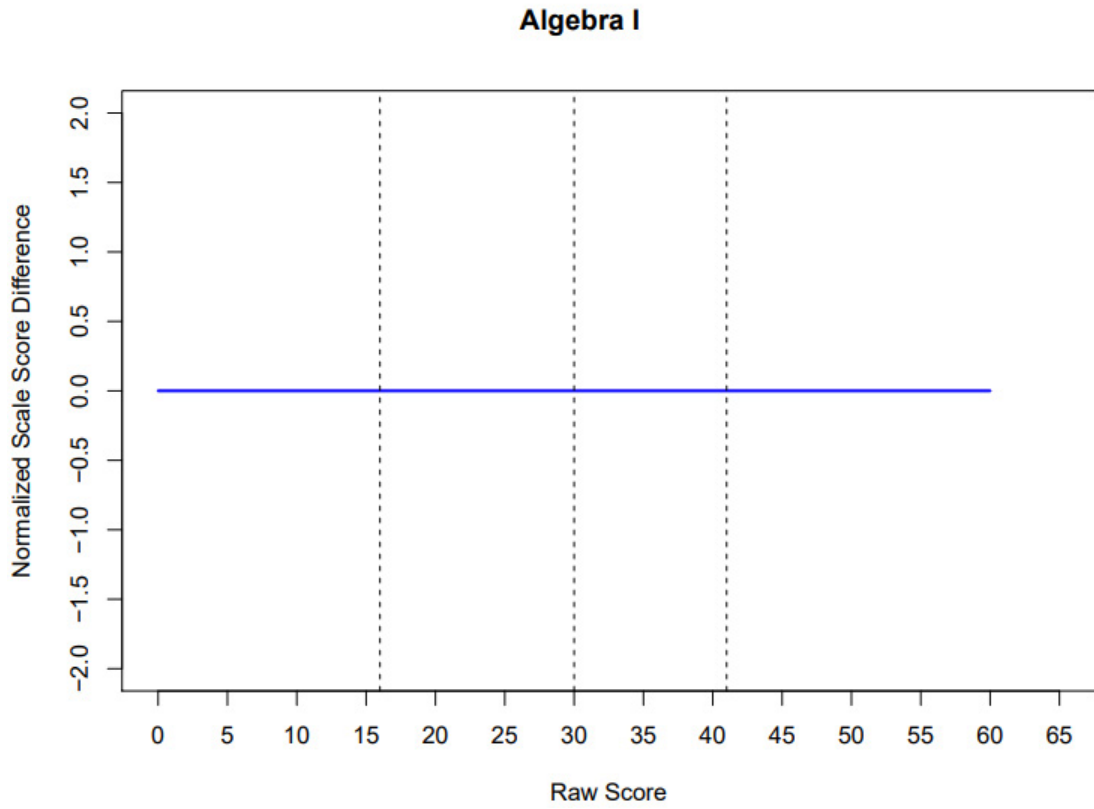


Figure L-4 (continued). Normalized Scale Score Distributions by Content Area: Spring

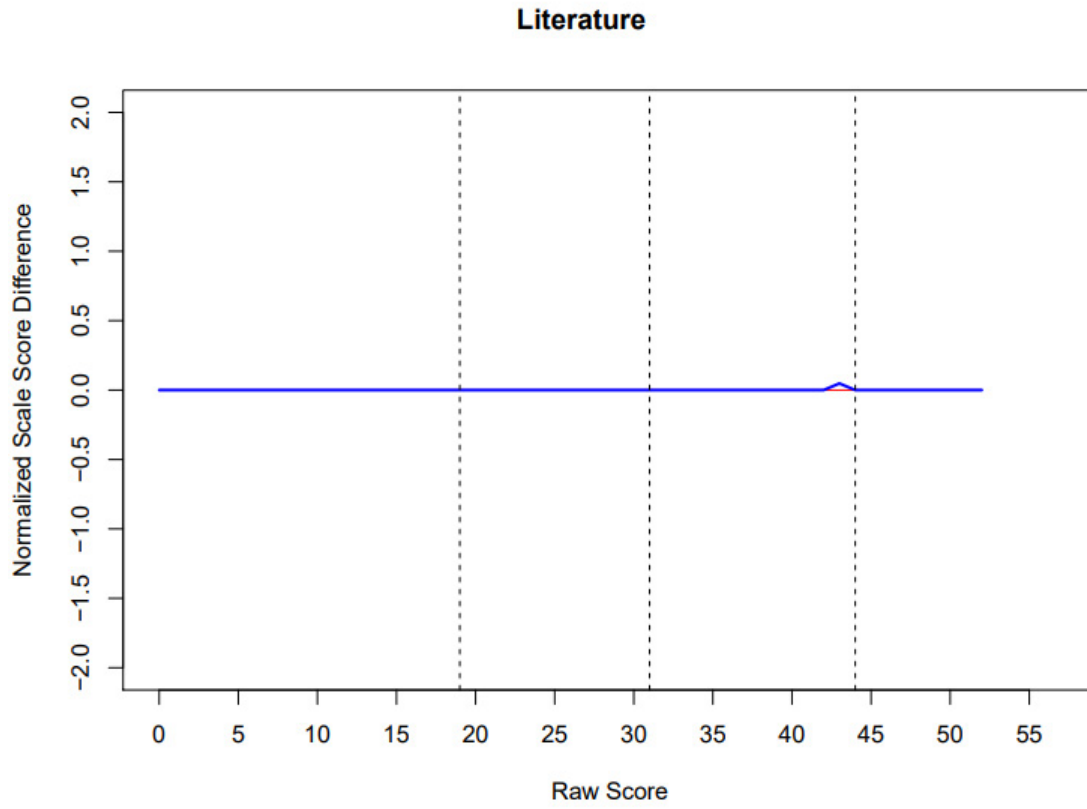


Table L-4. Algebra I Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
0	1215	1215	92	92	BB	BB	0.0	True
1	1276	1276	51	51	BB	BB	0.0	True
2	1312	1312	36	36	BB	BB	0.0	True
3	1334	1334	30	30	BB	BB	0.1	True
4	1350	1350	26	26	BB	BB	0.1	True
5	1362	1362	24	24	BB	BB	0.3	True
6	1373	1373	22	22	BB	BB	0.6	True
7	1382	1382	21	21	BB	BB	1.1	True
8	1391	1391	20	20	BB	BB	1.5	True
9	1398	1398	19	19	BB	BB	2.0	True
10	1405	1405	18	18	BB	BB	2.5	True
11	1411	1411	18	18	BB	BB	2.8	True
12	1417	1417	17	17	BB	BB	2.9	True
13	1423	1423	17	17	BB	BB	3.0	True
14	1428	1428	16	16	BB	BB	3.0	True
15	1434	1434	16	16	BB	BB	2.9	True
16	1439	1439	16	16	B	B	2.9	True
17	1443	1443	15	15	B	B	2.8	True
18	1448	1448	15	15	B	B	2.8	True
19	1453	1453	15	15	B	B	2.9	True
20	1457	1457	15	15	B	B	2.9	True
21	1462	1462	15	15	B	B	2.8	True
22	1466	1466	15	15	B	B	2.9	True
23	1470	1470	15	15	B	B	2.8	True
24	1475	1475	15	15	B	B	2.7	True
25	1479	1479	15	15	B	B	2.6	True
26	1483	1483	14	14	B	B	2.7	True
27	1487	1487	14	14	B	B	2.6	True
28	1491	1491	14	14	B	B	2.6	True
29	1496	1496	15	15	B	B	2.5	True
30	1500	1500	15	15	P	P	2.6	True
31	1504	1504	15	15	P	P	2.4	True
32	1508	1508	15	15	P	P	2.5	True
33	1513	1513	15	15	P	P	2.4	True
34	1517	1517	15	15	P	P	2.3	True
35	1521	1521	15	15	P	P	2.1	True
36	1526	1526	15	15	P	P	2.1	True
37	1530	1530	15	15	P	P	2.1	True
38	1535	1535	15	15	P	P	1.9	True
39	1540	1540	16	16	P	P	1.9	True
40	1545	1545	16	16	P	P	1.8	True
41	1550	1550	16	16	A	A	1.7	True

Table L-4 (continued). Algebra I Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
42	1555	1555	16	16	A	A	1.6	True
43	1560	1560	17	17	A	A	1.5	True
44	1566	1566	17	17	A	A	1.4	True
45	1572	1572	17	17	A	A	1.3	True
46	1578	1578	18	18	A	A	1.2	True
47	1584	1584	18	18	A	A	1.1	True
48	1591	1591	19	19	A	A	1.0	True
49	1598	1598	19	19	A	A	0.8	True
50	1606	1606	20	20	A	A	0.7	True
51	1614	1614	21	21	A	A	0.6	True
52	1623	1623	22	22	A	A	0.5	True
53	1634	1634	23	23	A	A	0.4	True
54	1645	1645	25	25	A	A	0.3	True
55	1659	1659	28	28	A	A	0.2	True
56	1677	1677	31	31	A	A	0.2	True
57	1699	1699	36	36	A	A	0.1	True
58	1729	1729	43	43	A	A	0.0	True
59	1777	1777	57	57	A	A	0.0	True
60	1800	1800	96	96	A	A	0.0	True

Table L-5. Biology Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
0	1224	1224	92	92	BB	BB	0.0	True
1	1285	1285	51	51	BB	BB	0.0	True
2	1321	1321	36	36	BB	BB	0.0	True
3	1342	1342	30	30	BB	BB	0.0	True
4	1358	1358	26	26	BB	BB	0.0	True
5	1370	1370	24	24	BB	BB	0.0	True
6	1380	1380	22	22	BB	BB	0.1	True
7	1389	1389	20	20	BB	BB	0.2	True
8	1397	1397	19	19	BB	BB	0.3	True
9	1404	1404	18	18	BB	BB	0.5	True
10	1410	1410	18	18	BB	BB	0.8	True
11	1416	1416	17	17	BB	BB	1.1	True
12	1422	1422	16	16	BB	BB	1.4	True
13	1427	1427	16	16	BB	BB	1.7	True
14	1432	1432	15	15	BB	BB	1.9	True
15	1436	1436	15	15	BB	BB	2.1	True
16	1441	1441	15	15	BB	BB	2.2	True
17	1445	1445	14	14	BB	BB	2.2	True
18	1449	1449	14	14	BB	BB	2.2	True
19	1453	1453	14	14	BB	BB	2.4	True
20	1457	1457	14	14	BB	BB	2.4	True
21	1460	1460	13	13	B	B	2.5	True
22	1464	1464	13	13	B	B	2.4	True
23	1467	1468	13	13	B	B	2.5	True
24	1471	1471	13	13	B	B	2.4	True
25	1474	1474	13	13	B	B	2.4	True
26	1477	1478	13	13	B	B	2.3	True
27	1481	1481	13	13	B	B	2.3	True
28	1484	1484	13	13	B	B	2.3	True
29	1487	1487	13	13	B	B	2.4	True
30	1490	1490	12	13	B	B	2.3	True
31	1493	1494	12	12	B	B	2.2	True
32	1496	1497	12	12	B	B	2.2	True
33	1499	1500	12	12	B	P	2.2	False
34	1502	1503	12	12	P	P	2.2	True
35	1505	1506	12	12	P	P	2.2	True
36	1509	1509	12	12	P	P	2.0	True
37	1512	1512	12	12	P	P	2.2	True
38	1515	1515	12	13	P	P	2.1	True
39	1518	1518	13	13	P	P	2.0	True
40	1521	1521	13	13	P	P	2.1	True
41	1524	1525	13	13	P	P	2.0	True

Table L-5 (continued). Biology Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
42	1528	1528	13	13	P	P	2.1	True
43	1531	1531	13	13	P	P	2.0	True
44	1534	1535	13	13	P	P	2.0	True
45	1538	1538	13	13	P	P	1.8	True
46	1541	1542	13	13	P	P	1.9	True
47	1545	1545	14	14	P	P	1.8	True
48	1549	1549	14	14	A	A	1.7	True
49	1553	1553	14	14	A	A	1.7	True
50	1557	1557	14	14	A	A	1.7	True
51	1561	1561	15	15	A	A	1.6	True
52	1565	1566	15	15	A	A	1.5	True
53	1570	1570	16	16	A	A	1.5	True
54	1575	1575	16	16	A	A	1.4	True
55	1580	1581	17	17	A	A	1.3	True
56	1586	1586	17	17	A	A	1.3	True
57	1592	1593	18	18	A	A	1.1	True
58	1599	1599	19	19	A	A	1.0	True
59	1607	1607	20	20	A	A	0.9	True
60	1615	1616	22	21	A	A	0.8	True
61	1625	1626	23	23	A	A	0.7	True
62	1637	1638	26	26	A	A	0.6	True
63	1653	1653	30	30	A	A	0.4	True
64	1674	1674	36	36	A	A	0.2	True
65	1710	1710	50	50	A	A	0.1	True
66	1770	1771	92	92	A	A	0.0	True

Table L-6. Literature Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
0	1201	1201	92	92	BB	BB	0.0	True
1	1263	1263	51	51	BB	BB	0.0	True
2	1299	1299	37	37	BB	BB	0.0	True
3	1321	1321	30	30	BB	BB	0.0	True
4	1337	1337	27	27	BB	BB	0.1	True
5	1350	1350	24	24	BB	BB	0.1	True
6	1361	1361	23	23	BB	BB	0.2	True
7	1371	1371	21	21	BB	BB	0.3	True
8	1379	1379	20	20	BB	BB	0.5	True
9	1387	1387	19	19	BB	BB	0.6	True
10	1394	1394	19	19	BB	BB	0.8	True
11	1401	1401	18	18	BB	BB	1.0	True
12	1407	1407	18	18	BB	BB	1.1	True
13	1413	1413	17	17	BB	BB	1.2	True
14	1419	1419	17	17	BB	BB	1.2	True
15	1424	1424	16	16	BB	BB	1.4	True
16	1430	1430	16	16	BB	BB	1.4	True
17	1435	1435	16	16	BB	BB	1.6	True
18	1440	1440	16	16	BB	BB	1.5	True
19	1445	1445	16	16	B	B	1.7	True
20	1450	1450	16	16	B	B	1.7	True
21	1455	1455	15	15	B	B	1.9	True
22	1459	1459	15	15	B	B	1.9	True
23	1464	1464	15	15	B	B	2.1	True
24	1469	1469	15	15	B	B	2.1	True
25	1474	1474	15	15	B	B	2.4	True
26	1478	1478	15	15	B	B	2.4	True
27	1483	1483	15	15	B	B	2.4	True
28	1488	1488	16	16	B	B	2.6	True
29	1493	1493	16	16	B	B	2.9	True
30	1498	1498	16	16	B	B	3.0	True
31	1503	1503	16	16	P	P	3.2	True
32	1508	1508	16	16	P	P	3.4	True
33	1513	1513	16	16	P	P	3.6	True
34	1519	1519	17	17	P	P	3.6	True
35	1524	1524	17	17	P	P	3.8	True
36	1530	1530	17	17	P	P	4.0	True
37	1536	1536	18	18	P	P	4.2	True
38	1542	1542	18	18	P	P	4.1	True
39	1549	1549	18	18	P	P	4.0	True
40	1556	1556	19	19	P	P	4.0	True
41	1563	1563	20	19	P	P	3.9	True

Table L-6 (continued). Literature Raw-to-Scaled Score Comparison for Pre-Equated and Post-Equated Solutions

Raw Score	Pre-SS	Post-SS	Pre-SEM	Post-SEM	Pre-PL	Post-PL	Proportion (%)	Same PL
42	1571	1571	20	20	P	P	3.6	True
43	1580	1579	21	21	P	P	3.3	True
44	1589	1589	22	22	A	A	2.9	True
45	1599	1599	23	23	A	A	2.5	True
46	1610	1610	24	24	A	A	1.9	True
47	1623	1623	26	26	A	A	1.5	True
48	1638	1638	29	29	A	A	1.0	True
49	1656	1656	32	32	A	A	0.7	True
50	1680	1680	38	38	A	A	0.4	True
51	1719	1719	52	52	A	A	0.2	True
52	1782	1782	92	92	A	A	0.0	True

APPENDIX M: RELIABILITIES

The Summer 2021 administration of the Keystone exams was cancelled due to the elongated spring testing window, which lasted from May 2021 to September 2021. Consequently, tables and graphs that usually display Summer Keystone test data will not be populated within this section of the 2021 Keystone Exams Technical Report, including any form-level or item-level information. Refer to the Preface for additional information.

Table M-1. Reliabilities

Column Heading	Definition
Level	Total test or module level
Group	Student group: all students or subgroup
Pts.	Max points possible
Len.	Test length
<i>N</i>	Number of students
Mean	Mean of raw score
SD	Standard deviation of raw score
<i>r</i>	Reliability coefficient: Cronbach's alpha
<i>SEM</i>	Standard error of measurement

Note. "DNR" in the tables below represents "Do Not Report". This happened only when the *N* count was small.

Table M–2. Winter: Algebra I Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	60	42	13200	25.43	11.44	0.91	3.46
	Module 1	All	30	21	13200	12.78	6.08	0.83	2.49
	Module 2	All	30	21	13200	12.65	5.88	0.83	2.40
Gender	Total	Female	60	42	6535	25.62	11.14	0.90	3.45
		Male	60	42	6660	25.23	11.72	0.91	3.46
	Module 1	Female	30	21	6535	12.95	5.93	0.83	2.47
		Male	30	21	6660	12.61	6.21	0.84	2.50
	Module 2	Female	30	21	6535	12.67	5.73	0.82	2.40
		Male	30	21	6660	12.63	6.02	0.84	2.39
Ethnicity	Total	African American	60	42	1065	18.42	9.27	0.88	3.23
		American Indian	60	42	26	25.42	9.84	0.87	3.49
		Asian	60	42	430	32.99	12.73	0.92	3.53
		Hispanic	60	42	1536	19.73	9.66	0.88	3.29
		Multi-racial	60	42	457	23.63	10.94	0.9	3.44
		Native Hawaiian/ Pacific Islander	60	42	12	28.58	15.95	0.95	3.67
		White	60	42	9669	26.85	11.21	0.9	3.47
	Module 1	African American	30	21	1065	9.28	5.06	0.8	2.28
		American Indian	30	21	26	12.19	5.27	0.78	2.48
		Asian	30	21	430	16.49	6.67	0.85	2.56
		Hispanic	30	21	1536	9.82	5.29	0.81	2.34
		Multi-racial	30	21	457	11.99	5.93	0.83	2.45
		Native Hawaiian/ Pacific Islander	30	21	12	13.83	8.82	0.91	2.60
		White	30	21	9669	13.51	5.95	0.82	2.50
	Module 2	African American	30	21	1065	9.14	4.80	0.78	2.27
		American Indian	30	21	26	13.23	4.80	0.73	2.49
		Asian	30	21	430	16.5	6.48	0.86	2.44
		Hispanic	30	21	1536	9.91	4.93	0.78	2.31
		Multi-racial	30	21	457	11.64	5.57	0.81	2.41
		Native Hawaiian/ Pacific Islander	30	21	12	14.75	7.33	0.87	2.61
		White		21	9669	13.34	5.79	0.83	2.41
EL	Total	All	60	42	302	16.16	8.36	0.87	3.06
	Module 1	All	30	21	302	8.06	4.79	0.80	2.16
	Module 2	All	30	21	302	8.10	4.21	0.74	2.15
IEP	Total	All	60	42	1936	16.77	8.54	0.86	3.15
	Module 1	All	30	21	1936	8.35	4.66	0.77	2.24
	Module 2	All	30	21	1936	8.42	4.5	0.76	2.21
ED	Total	All	60	42	5121	21.14	10.03	0.89	3.34
	Module 1	All	30	21	5121	10.61	5.41	0.81	2.38
	Module 2	All	30	21	5121	10.53	5.19	0.80	2.33

Table M–3. Winter: Biology Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	66	54	12144	33.80	13.83	0.93	3.75
	Module 1	All	33	27	12144	16.94	7.35	0.88	2.56
	Module 2	All	33	27	12144	16.86	7.01	0.85	2.74
Gender	Total	Female	66	54	5936	34.39	13.45	0.92	3.76
		Male	66	54	6207	33.23	14.15	0.93	3.73
	Module 1	Female	33	27	5936	17.12	7.22	0.87	2.58
		Male	33	27	6207	16.76	7.47	0.88	2.55
	Module 2	Female	33	27	5936	17.27	6.78	0.84	2.74
		Male	33	27	6207	16.47	7.20	0.86	2.72
Ethnicity	Total	African American	66	54	1108	24.15	10.93	0.88	3.71
		American Indian	66	54	15	28.33	10.91	0.88	3.72
		Asian	66	54	441	43.56	13.84	0.93	3.56
		Hispanic	66	54	1049	28.16	12.63	0.91	3.76
		Multi-racial	66	54	412	30.84	14.02	0.93	3.75
		Native Hawaiian/ Pacific Islander	66	54	10	33.1	18.74	0.96	3.69
		White	66	54	9109	35.29	13.46	0.92	3.74
	Module 1	African American	33	27	1108	12.03	5.86	0.81	2.54
		American Indian	33	27	15	13.13	6.60	0.85	2.57
		Asian	33	27	441	22.14	7.37	0.89	2.42
		Hispanic	33	27	1049	13.93	6.75	0.86	2.56
		Multi-racial	33	27	412	15.42	7.39	0.88	2.59
		Native Hawaiian/ Pacific Islander	33	27	10	17	9.83	0.94	2.41
		White	33	27	9109	17.7	7.16	0.87	2.56
	Module 2	African American	33	27	1108	12.12	5.73	0.78	2.70
		American Indian	33	27	15	15.2	5.37	0.76	2.64
		Asian	33	27	441	21.43	6.9	0.86	2.61
		Hispanic	33	27	1049	14.23	6.46	0.82	2.74
		Multi-racial	33	27	412	15.42	7.21	0.86	2.70
		Native Hawaiian/ Pacific Islander	33	27	10	16.10	9.04	0.90	2.82
		White	33	27	9109	17.59	6.84	0.84	2.72
EL	Total	All	66	54	195	19.76	9.31	0.85	3.56
	Module 1	All	33	27	195	10.19	5.22	0.77	2.48
	Module 2	All	33	27	195	9.56	4.78	0.72	2.55
IEP	Total	All	66	54	1814	21.52	10.27	0.88	3.6
	Module 1	All	33	27	1814	10.81	5.46	0.79	2.48
	Module 2	All	33	27	1814	10.71	5.45	0.77	2.61
ED	Total	All	66	54	4214	27.56	12.21	0.91	3.75
	Module 1	All	33	27	4214	13.76	6.51	0.85	2.56
	Module 2	All	33	27	4214	13.8	6.33	0.81	2.73

Table M–4. Winter: Literature Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	52	40	11722	32.33	9.99	0.91	2.97
	Module 1	All	26	20	11722	16.15	5.13	0.84	2.07
	Module 2	All	26	20	11722	16.18	5.31	0.84	2.13
Gender	Total	Female	52	40	5570	34.15	9.47	0.9	2.92
		Male	52	40	6149	30.68	10.17	0.91	2.98
	Module 1	Female	26	20	5570	17.21	4.81	0.82	2.03
		Male	26	20	6149	15.19	5.21	0.84	2.07
	Module 2	Female	26	20	5570	16.94	5.11	0.83	2.11
		Male	26	20	6149	15.49	5.39	0.84	2.13
Ethnicity	Total	African American	52	40	881	25.51	10.05	0.91	3.07
		American Indian	52	40	17	31.12	9.05	0.89	3.03
		Asian	52	40	384	37.47	9.1	0.91	2.78
		Hispanic	52	40	1069	28.7	9.84	0.91	3.03
		Multi-racial	52	40	380	30.75	9.77	0.9	3.03
		Native Hawaiian/ Pacific Islander	52	40	4	27.25	15.71	0.96	3.02
		White	52	40	8985	33.29	9.63	0.91	2.94
	Module 1	African American	26	20	881	12.84	5.13	0.83	2.12
		American Indian	26	20	17	15.47	4.72	0.80	2.13
		Asian	26	20	384	18.5	4.71	0.83	1.95
		Hispanic	26	20	1069	14.33	5.05	0.83	2.10
		Multi-racial	26	20	380	15.24	5.11	0.83	2.10
		Native Hawaiian/ Pacific Islander	26	20	4	13.75	8.62	0.94	2.14
		White	26	20	8985	16.64	4.96	0.83	2.05
	Module 2	African American	26	20	881	12.67	5.44	0.84	2.20
		American Indian	26	20	17	15.65	4.94	0.81	2.14
		Asian	26	20	384	18.97	4.79	0.83	1.99
		Hispanic	26	20	1069	14.37	5.32	0.83	2.18
		Multi-racial	26	20	380	15.51	5.19	0.82	2.17
		Native Hawaiian/ Pacific Islander	26	20	4	13.5	7.14	0.91	2.16
		White	26	20	8985	16.65	5.11	0.83	2.11
EL	Total	All	52	40	138	18.10	7.60	0.84	3.04
	Module 1	All	26	20	138	9.04	4.07	0.73	2.10
	Module 2	All	26	20	138	9.07	4.03	0.70	2.20
IEP	Total	All	52	40	1709	21.31	9.28	0.89	3.03
	Module 1	All	26	20	1709	10.79	4.87	0.81	2.10
	Module 2	All	26	20	1709	10.52	4.96	0.81	2.17
ED	Total	All	52	40	3965	28.03	10.03	0.91	3.04
	Module 1	All	26	20	3965	14.06	5.15	0.83	2.10
	Module 2	All	26	20	3965	13.97	5.37	0.83	2.19

Table M–5. Spring: Algebra Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	60	42	109710	25.25	12.12	0.92	3.38
	Module 1	All	30	21	109710	12.56	6.06	0.85	2.31
	Module 2	All	30	21	109710	12.69	6.53	0.86	2.47
Gender	Total	Female	60	42	53557	25.59	12.01	0.92	3.38
		Male	60	42	56034	24.93	12.21	0.92	3.37
	Module 1	Female	30	21	53557	12.95	6.04	0.85	2.31
		Male	30	21	56034	12.19	6.04	0.85	2.31
	Module 2	Female	30	21	53557	12.64	6.43	0.85	2.47
		Male	30	21	56034	12.74	6.63	0.86	2.46
Ethnicity	Total	African American	60	42	13269	16.63	8.75	0.87	3.13
		American Indian	60	42	186	24.39	11.46	0.91	3.37
		Asian	60	42	4985	34.15	13.31	0.93	3.46
		Hispanic	60	42	12162	19.06	10.23	0.90	3.23
		Multi-racial	60	42	3975	23.83	11.88	0.92	3.37
		Native Hawaiian/ Pacific Islander	60	42	94	24.82	12.19	0.92	3.43
		White	60	42	74921	27.27	11.65	0.92	3.39
	Module 1	African American	30	21	13269	8.52	4.53	0.76	2.20
		American Indian	30	21	186	11.95	5.79	0.84	2.32
		Asian	30	21	4985	17.05	6.68	0.88	2.33
		Hispanic	30	21	12162	9.65	5.20	0.81	2.25
		Multi-racial	30	21	3975	11.88	5.95	0.85	2.31
		Native Hawaiian/ Pacific Islander	30	21	94	12.21	5.94	0.85	2.32
		White	30	21	74921	13.49	5.84	0.84	2.30
	Module 2	African American	30	21	13269	8.11	4.79	0.79	2.22
		American Indian	30	21	186	12.45	6.11	0.84	2.45
		Asian	30	21	4985	17.10	7.05	0.87	2.55
		Hispanic	30	21	12162	9.41	5.54	0.83	2.32
		Multi-racial	30	21	3975	11.95	6.41	0.85	2.45
		Native Hawaiian/ Pacific Islander	30	21	94	12.61	6.72	0.86	2.53
		White	30	21	74921	13.78	6.31	0.85	2.48
EL	Total	All	60	42	3794	14.44	7.66	0.85	3.01
	Module 1	All	30	21	3794	7.50	4.13	0.73	2.16
	Module 2	All	30	21	3794	6.93	4.11	0.74	2.09
IEP	Total	All	60	42	16978	15.88	8.70	0.87	3.11
	Module 1	All	30	21	16978	8.04	4.46	0.76	2.19
	Module 2	All	30	21	16978	7.84	4.81	0.79	2.21
ED	Total	All	60	42	42192	20.15	10.46	0.90	3.26
	Module 1	All	30	21	42192	10.13	5.27	0.82	2.26
	Module 2	All	30	21	42192	10.01	5.70	0.83	2.35

Table M–6. Spring: Biology Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	66	54	100976	32.41	14.13	0.93	3.83
	Module 1	All	33	27	100976	16.08	7.41	0.87	2.72
	Module 2	All	33	27	100976	16.33	7.27	0.86	2.70
Gender	Total	Female	66	54	49205	32.62	13.76	0.92	3.85
		Male	66	54	51707	32.23	14.47	0.93	3.82
	Module 1	Female	33	27	49205	16.08	7.24	0.86	2.72
		Male	33	27	51707	16.10	7.58	0.87	2.72
	Module 2	Female	33	27	49205	16.54	7.10	0.85	2.71
		Male	33	27	51707	16.13	7.43	0.87	2.68
Ethnicity	Total	African American	66	54	11987	22.54	10.67	0.88	3.70
		American Indian	66	54	166	29.51	13.23	0.91	3.86
		Asian	66	54	4607	41.37	14.93	0.94	3.68
		Hispanic	66	54	10788	24.85	12.17	0.90	3.77
		Multi-racial	66	54	3443	30.20	13.75	0.92	3.83
		Native Hawaiian/ Pacific Islander	66	54	85	33.67	14.64	0.93	3.82
		White	66	54	69835	34.81	13.59	0.92	3.83
	Module 1	African American	33	27	11987	11.10	5.75	0.79	2.62
		American Indian	33	27	166	14.57	6.87	0.84	2.76
		Asian	33	27	4607	20.78	7.82	0.89	2.60
		Hispanic	33	27	10788	12.22	6.48	0.83	2.67
		Multi-racial	33	27	3443	14.93	7.23	0.86	2.71
		Native Hawaiian/ Pacific Islander	33	27	85	16.59	7.45	0.87	2.73
		White	33	27	69835	17.29	7.14	0.86	2.71
	Module 2	African American	33	27	11987	11.44	5.62	0.78	2.61
		American Indian	33	27	166	14.94	6.88	0.85	2.71
		Asian	33	27	4607	20.59	7.60	0.88	2.60
		Hispanic	33	27	10788	12.63	6.31	0.82	2.65
		Multi-racial	33	27	3443	15.27	7.10	0.86	2.70
		Native Hawaiian/ Pacific Islander	33	27	85	17.08	7.76	0.88	2.66
		White	33	27	69835	17.52	7.04	0.85	2.70
EL	Total	All	66	54	3192	18.23	8.09	0.81	3.54
	Module 1	All	33	27	3192	8.98	4.59	0.70	2.50
	Module 2	All	33	27	3192	9.25	4.30	0.66	2.50
IEP	Total	All	66	54	16202	21.69	10.79	0.88	3.67
	Module 1	All	33	27	16202	10.77	5.78	0.80	2.61
	Module 2	All	33	27	16202	10.92	5.67	0.79	2.58
ED	Total	All	66	54	38330	26.24	12.33	0.91	3.79
	Module 1	All	33	27	38330	12.97	6.54	0.83	2.69
	Module 2	All	33	27	38330	13.27	6.42	0.83	2.67

Table M-7. Spring: Literature Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	52	40	97928	31.09	10.41	0.92	3.03
	Module 1	All	26	20	97928	15.69	5.20	0.83	2.16
	Module 2	All	26	20	97928	15.40	5.67	0.86	2.12
Gender	Total	Female	52	40	48041	32.88	9.83	0.91	2.98
		Male	52	40	49797	29.37	10.67	0.92	3.04
	Module 1	Female	26	20	48041	16.60	4.90	0.81	2.12
		Male	26	20	49797	14.83	5.32	0.83	2.17
	Module 2	Female	26	20	48041	16.28	5.40	0.85	2.09
		Male	26	20	49797	14.54	5.79	0.87	2.13
Ethnicity	Total	African American	52	40	11754	24.83	9.97	0.90	3.16
		American Indian	52	40	170	28.96	10.44	0.91	3.08
		Asian	52	40	4490	36.78	9.29	0.91	2.79
		Hispanic	52	40	10420	26.06	10.43	0.91	3.14
		Multi-racial	52	40	3270	30.04	10.48	0.92	3.05
		Native Hawaiian/ Pacific Islander	52	40	81	31.31	10.97	0.93	2.99
		White	52	40	67654	32.64	9.80	0.91	2.97
	Module 1	African American	26	20	11754	12.80	5.01	0.80	2.24
		American Indian	26	20	170	14.60	5.29	0.83	2.21
		Asian	26	20	4490	18.26	4.74	0.82	2.01
		Hispanic	26	20	10420	13.24	5.32	0.82	2.23
		Multi-racial	26	20	3270	15.29	5.23	0.83	2.17
		Native Hawaiian/ Pacific Islander	26	20	81	15.77	5.61	0.86	2.08
		White	26	20	67654	16.43	4.91	0.81	2.12
	Module 2	African American	26	20	11754	12.04	5.50	0.84	2.22
		American Indian	26	20	170	14.36	5.69	0.86	2.14
		Asian	26	20	4490	18.53	4.98	0.85	1.94
		Hispanic	26	20	10420	12.82	5.62	0.85	2.21
		Multi-racial	26	20	3270	14.75	5.69	0.86	2.15
		Native Hawaiian/ Pacific Islander	26	20	81	15.54	5.71	0.86	2.15
		White	26	20	67654	16.21	5.36	0.85	2.08
EL	Total	All	52	40	2814	18.51	8.04	0.85	3.14
	Module 1	All	26	20	2814	9.29	4.21	0.72	2.23
	Module 2	All	26	20	2814	9.22	4.47	0.75	2.21
IEP	Total	All	52	40	15602	20.90	9.27	0.89	3.11
	Module 1	All	26	20	15602	10.93	4.80	0.79	2.22
	Module 2	All	26	20	15602	9.97	5.03	0.81	2.18
ED	Total	All	52	40	36966	26.76	10.34	0.91	3.13
	Module 1	All	26	20	36966	13.65	5.19	0.82	2.22
	Module 2	All	26	20	36966	13.10	5.65	0.85	2.20

Table M–8. Summer: Algebra Reliabilities

	Level	Group	Pts.	Len.	N	Mean	SD	r	SEM
Overall	Total	All	60	42					
	Module 1	All	30	21					
	Module 2	All	30	21					
Gender	Total	Female	60	42					
		Male	60	42					
	Module 1	Female	30	21					
		Male	30	21					
	Module 2	Female	30	21					
		Male	30	21					
Ethnicity	Total	African American	60	42					
		American Indian	60	42					
		Asian	60	42					
		Hispanic	60	42					
		Multi-racial	60	42					
		Native Hawaiian/ Pacific Islander	60	42					
		White	60	42					
	Module 1	African American	30	21					
		American Indian	30	21					
		Asian	30	21					
		Hispanic	30	21					
		Multi-racial	30	21					
		Native Hawaiian/ Pacific Islander	30	21					
		White	30	21					
	Module 2	African American	30	21					
		American Indian	30	21					
		Asian	30	21					
		Hispanic	30	21					
		Multi-racial	30	21					
		Native Hawaiian/ Pacific Islander	30	21					
		White	30	21					
EL	Total	All	60	42					
	Module 1	All	30	21					
	Module 2	All	30	21					
IEP	Total	All	60	42					
	Module 1	All	30	21					
	Module 2	All	30	21					
ED	Total	All	60	42					
	Module 1	All	30	21					
	Module 2	All	30	21					

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Table M–9. Summer: Biology Reliabilities

	Level	Group	Pts.	Len.	<i>N</i>	Mean	SD	<i>r</i>	SEM
Overall	Total	All	66	54					
	Module 1	All	33	27					
	Module 2	All	33	27					
Gender	Total	Female	66	54					
		Male	66	54					
	Module 1	Female	33	27					
		Male	33	27					
	Module 2	Female	33	27					
		Male	33	27					
Ethnicity	Total	African American	66	54					
		American Indian	66	54					
		Asian	66	54					
		Hispanic	66	54					
		Multi-racial	66	54					
		Native Hawaiian/ Pacific Islander	66	54					
		White	66	54					
	Module 1	African American	33	27					
		American Indian	33	27					
		Asian	33	27					
		Hispanic	33	27					
		Multi-racial	33	27					
		Native Hawaiian/ Pacific Islander	33	27					
		White	33	27					
	Module 2	African American	33	27					
		American Indian	33	27					
		Asian	33	27					
		Hispanic	33	27					
		Multi-racial	33	27					
		Native Hawaiian/ Pacific Islander	33	27					
		White	33	27					
EL	Total	All	66	54					
	Module 1	All	33	27					
	Module 2	All	33	27					
IEP	Total	All	66	54					
	Module 1	All	33	27					
	Module 2	All	33	27					
ED	Total	All	66	54					
	Module 1	All	33	27					
	Module 2	All	33	27					

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

Table M–10. Summer: Literature Reliabilities

	Level	Group	Pts.	Len.	<i>N</i>	Mean	SD	<i>r</i>	SEM
Overall	Total	All	52	40					
	Module 1	All	26	20					
	Module 2	All	26	20					
Gender	Total	Female	52	40					
		Male	52	40					
	Module 1	Female	26	20					
		Male	26	20					
	Module 2	Female	26	20					
		Male	26	20					
Ethnicity	Total	African American	52	40					
		American Indian	52	40					
		Asian	52	40					
		Hispanic	52	40					
		Multi-racial	52	40					
		Native Hawaiian/ Pacific Islander	52	40					
		White	52	40					
	Module 1	African American	26	20					
		American Indian	26	20					
		Asian	26	20					
		Hispanic	26	20					
		Multi-racial	26	20					
		Native Hawaiian/ Pacific Islander	26	20					
		White	26	20					
	Module 2	African American	26	20					
		American Indian	26	20					
		Asian	26	20					
		Hispanic	26	20					
		Multi-racial	26	20					
		Native Hawaiian/ Pacific Islander	26	20					
		White	26	20					
EL	Total	All	52	40					
	Module 1	All	26	20					
	Module 2	All	26	20					
IEP	Total	All	52	40					
	Module 1	All	26	20					
	Module 2	All	26	20					
ED	Total	All	52	40					
	Module 1	All	26	20					
	Module 2	All	26	20					

Note. There was no Summer 2021 administration due to the elongated spring testing windows from May through September 2021.

APPENDIX N: OPPORTUNITY TO LEARN SURVEY

During the Spring Keystone administration, students responded to a survey aimed to collect information regarding their opportunity to learn (OTL) during the 2020–2021 school year. The results of surveys processed during the first reporting window (before July 14) are presented in this Appendix. Data is only reported for test-takers who responded to at least one of the survey questions. The total count of responses and complete responses are shown in Table N–1.

Table N–1. Survey Response Rate

Subject	Group	Total	Complete Survey	% Complete
Algebra I	Online	28889	27859	96.4
	Paper	76828	25304	32.9
	Total	105717	53163	50.3
Biology	Online	28696	27430	95.6
	Paper	66810	24534	36.7
	Total	95506	51964	54.4
Literature	Online	25884	24869	96.1
	Paper	67156	25984	38.7
	Total	93040	50853	54.7
Total	Online	83469	80158	96.0
	Paper	210794	75822	36.0
	Total	294263	155980	53.0

SUMMARY

For each survey question, a table reports the frequencies (counts) and percentages of responses for each option by subject.

- The response rate for online administrations was higher than the response rate for paper-based assessments (96.0% versus 36.0%); however, response patterns were not substantially different between administration mode and are therefore reported in aggregate. Online response rates tended to be higher than paper-based response rates likely due to the placement of the survey at the beginning of the test and minimal teacher involvement needed for completion.
- Approximately 42–49% of the test-takers indicated that they attended school in-person, 21–26% did not attend school in a school building, and the remaining percentage attended school in a hybrid model (1–4 days in school per week). Data is also reported by instructional model.
- Approximately 90% of respondents said that they felt somewhat or very comfortable using online tools for schoolwork.
- Between 61% and 65% of respondents said they felt positive about their learning this school year.

RESULTS BY SUBJECT AND MODE

Table N–2. Survey Results for Question 1 by Subject and Mode

Question: In the past month, how many days have you attended school in a school building?

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
I usually attended school in a school building every day (5 days per week)	14207 (51.0%)	11903 (47.0%)	12552 (45.8%)	9764 (39.8%)	11161 (44.9%)	10206 (39.3%)
I usually attended school in a school building 4 days per week	4141 (14.9%)	4705 (18.6%)	3904 (14.2%)	4841 (19.7%)	3411 (13.7%)	4899 (18.9%)
I usually attended school in a school building 3 days per week	721 (2.6%)	773 (3.1%)	1017 (3.7%)	947 (3.9%)	821 (3.3%)	953 (3.7%)
I usually attended school in a school building 2 days per week	2433 (8.7%)	2423 (9.6%)	2416 (8.8%)	2754 (11.2%)	2161 (8.7%)	2994 (11.5%)
I usually attended school in a school building 1 day per week	320 (1.1%)	249 (1.0%)	341 (1.2%)	245 (1.0%)	322 (1.3%)	283 (1.1%)
I have not attended school in a school building at all	6012 (21.6%)	5095 (20.1%)	7182 (26.2%)	5859 (23.9%)	6971 (28.0%)	6506 (25.0%)
Blank	25 (0.1%)	156 (0.6%)	18 (0.1%)	124 (0.5%)	22 (0.1%)	143 (0.6%)

Table N–3. Survey Results for Question 2 by Subject and Mode

Question: In the past month, how many days have you used the Internet for schoolwork outside of a school building?

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
I have used the Internet every day (7 days per week)	6301 (22.6%)	8948 (35.4%)	7820 (28.5%)	10023 (40.9%)	7682 (30.9%)	10852 (41.8%)
I have used the Internet 4–6 days per week	11847 (42.5%)	7527 (29.7%)	11453 (41.8%)	8045 (32.8%)	10290 (41.4%)	8492 (32.7%)
I have used the Internet 2–3 days per week	5676 (20.4%)	4463 (17.6%)	4790 (17.5%)	3637 (14.8%)	4077 (16.4%)	3649 (14%)
I have used the Internet 0–1 days per week	3967 (14.2%)	2773 (11.0%)	3286 (12.0%)	2156 (8.8%)	2761 (11.1%)	2241 (8.6%)
Blank	68 (0.2%)	1593 (6.3%)	81 (0.3%)	673 (2.7%)	59 (0.2%)	750 (2.9%)

Table N–4. Survey Results for Question 3 by Subject and Mode

Question: When you are not in a school building, what device do you use most often to complete your schoolwork?

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
I use a personal computer (PC), Laptop, Chromebook, iPad or other tablet	23914 (85.8%)	20174 (79.7%)	23758 (86.6%)	20588 (83.9%)	21738 (87.4%)	21769 (83.8%)
I use a smartphone (cell phone)	2853 (10.2%)	4096 (16.2%)	2692 (9.8%)	3061 (12.5%)	2319 (9.3%)	3294 (12.7%)
I am not sure what kind of device I use	312 (1.1%)	492 (1.9%)	252 (0.9%)	364 (1.5%)	187 (0.8%)	357 (1.4%)
I do not use a device when I am not in a school building	699 (2.5%)	386 (1.5%)	636 (2.3%)	374 (1.5%)	545 (2.2%)	411 (1.6%)
Blank	81 (0.3%)	156 (0.6%)	92 (0.3%)	147 (0.6%)	80 (0.3%)	153 (0.6%)

Table N-5. Survey Results for Question 4 by Subject and Mode

Question: In the past month, where have you completed most of your schoolwork?

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
in my own home	13128 (47.1%)	12856 (50.8%)	14879 (54.2%)	14106 (57.5%)	14021 (56.4%)	15274 (58.8%)
at the home of a friend, neighbor, or relative	255 (0.9%)	1615 (6.4%)	303 (1.1%)	849 (3.5%)	244 (1.0%)	882 (3.4%)
at the place where my parent/guardian works	101 (0.4%)	415 (1.6%)	95 (0.3%)	229 (0.9%)	90 (0.4%)	276 (1.1%)
at a library, store, restaurant, or other public building	60 (0.2%)	222 (0.9%)	112 (0.4%)	159 (0.6%)	86 (0.3%)	183 (0.7%)
in a school building	14238 (51.1%)	10021 (39.6%)	11955 (43.6%)	9058 (36.9%)	10349 (41.6%)	9210 (35.4%)
Blank	77 (0.3%)	175 (0.7%)	86 (0.3%)	133 (0.5%)	79 (0.3%)	159 (0.6%)

Table N-6. Survey Results for Question 5 by Subject and Mode

Question: Which statement best describes how comfortable you are with using online tools for schoolwork (such as Zoom, Schoology, Google Meet, or Teams)?

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
I feel very comfortable using online tools for schoolwork	12046 (43.2%)	10660 (42.1%)	12016 (43.8%)	10864 (44.3%)	11530 (46.4%)	11914 (45.9%)
I feel somewhat comfortable using online tools for schoolwork	13173 (47.3%)	11507 (45.5%)	12634 (46.1%)	10968 (44.7%)	10876 (43.7%)	11237 (43.2%)
I do not feel comfortable using online tools for schoolwork	2247 (8.1%)	2618 (10.3%)	2369 (8.6%)	2262 (9.2%)	2097 (8.4%)	2362 (9.1%)
I have not used online tools for schoolwork	306 (1.1%)	334 (1.3%)	313 (1.1%)	269 (1.1%)	281 (1.1%)	299 (1.2%)
Blank	87 (0.3%)	185 (0.7%)	98 (0.4%)	171 (0.7%)	85 (0.3%)	172 (0.7%)

Table N-7. Survey Results for Question 6 by Subject and Mode

Question: In the past month, I have understood almost all my assignments.

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
Strongly agree	8132 (29.2%)	6197 (24.5%)	6771 (24.7%)	6234 (25.4%)	6518 (26.2%)	6544 (25.2%)
Somewhat agree	14384 (51.6%)	12827 (50.7%)	14266 (52.0%)	12311 (50.2%)	12663 (50.9%)	13014 (50.1%)
Somewhat disagree	3816 (13.7%)	4446 (17.6%)	4509 (16.4%)	4195 (17.1%)	4023 (16.2%)	4459 (17.2%)
Strongly disagree	1386 (5.0%)	1411 (5.6%)	1726 (6.3%)	1380 (5.6%)	1537 (6.2%)	1547 (6.0%)
Blank	141 (0.5%)	423 (1.7%)	158 (0.6%)	414 (1.7%)	128 (0.5%)	420 (1.6%)

Table N-8. Survey Results for Question 7 by Subject and Mode

Question: In the past month, I have completed almost all my assignments.

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
Strongly agree	13752 (49.4%)	9856 (39.0%)	12427 (45.3%)	11176 (45.6%)	11492 (46.2%)	11821 (45.5%)
Somewhat agree	9492 (34.1%)	10328 (40.8%)	9637 (35.1%)	8505 (34.7%)	8469 (34.1%)	8790 (33.8%)
Somewhat disagree	3214 (11.5%)	3529 (13.9%)	3729 (13.6%)	3256 (13.3%)	3310 (13.3%)	3624 (13.9%)
Strongly disagree	1273 (4.6%)	1098 (4.3%)	1499 (5.5%)	1110 (4.5%)	1471 (5.9%)	1281 (4.9%)
Blank	128 (0.5%)	493 (1.9%)	138 (0.5%)	487 (2.0%)	127 (0.5%)	468 (1.8%)

Table N-9. Survey Results for Question 8 by Subject and Mode

Question: I feel positive about my learning so far this school year.

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
Strongly agree	7649 (27.5%)	4528 (17.9%)	5905 (21.5%)	4121 (16.8%)	5214 (21.0%)	4146 (16.0%)
Somewhat agree	12824 (46.0%)	9588 (37.9%)	12722 (46.4%)	9742 (39.7%)	11379 (45.8%)	10429 (40.1%)
Somewhat disagree	4764 (17.1%)	4331 (17.1%)	5568 (20.3%)	4945 (20.2%)	5202 (20.9%)	5436 (20.9%)
Strongly disagree	2336 (8.4%)	2569 (10.2%)	2913 (10.6%)	3017 (12.3%)	2814 (11.3%)	3291 (12.7%)
Blank	286 (1.0%)	4288 (16.9%)	322 (1.2%)	2709 (11%)	260 (1.0%)	2682 (10.3%)

Table N-10. Survey Results for Question 9 by Subject and ModeQuestion: Indicate the month in which you took this assessment. (*Select the month when you started the assessment if the administration crossed over two months.*)

Option	Algebra Online	Algebra Paper	Biology Online	Biology Paper	Literature Online	Literature Paper
May	23721 (82.1%)	17478 (22.7%)	24494 (85.4%)	18182 (27.2%)	22326 (86.3%)	20253 (30.2%)
June	5146 (17.8%)	2972 (3.9%)	4193 (14.6%)	2963 (4.4%)	3548 (13.7%)	2478 (3.7%)
July	20 (0.1%)	50 (0.1%)	4 (0.0%)	58 (0.1%)	4 (0.0%)	46 (0.1%)
August		26 (0.0%)		16 (0.0%)		22 (0.0%)
September		27 (0.0%)		18 (0.0%)		21 (0.0%)
Blank	2 (0.0%)	56275 (73.2%)	5 (0.0%)	45573 (68.2%)	6 (0.0%)	44336 (66.0%)

RESULTS BY INSTRUCTIONAL MODE

Results are disaggregated and presented for each instructional mode (in-person, hybrid, and virtual) students selected in survey question 1. Students that indicated that they attended school in-person every day were categorized as in-person, students that indicated that they did not attend school in-person at all were classified as virtual, and students that indicated that they attended school 1 to 4 days per week were classified as hybrid. It is important to note that students answered this question with respect to their instructional model the month prior to test administration, so it does not give a complete picture of how test-takers attended school across instructional modes for the entire 2020–2021 school year. Moreover, students have transitioned between instructional models frequently over the past year, and the data does not capture this fluidity.

Table N–11. Survey Results for Question 2 by Subject and Instructional Mode

Question: In the past month, how many days have you used the Internet for schoolwork outside of a school building?

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
I have used the Internet every day (7 days per week)	6124 (23.5%)	3649 (23.1%)	5428 (48.9%)	6142 (27.5%)	4805 (29.2%)	6862 (52.6%)	6214 (29.1%)	4951 (31.2%)	7321 (54.3%)
I have used the Internet 4–6 days per week	8972 (34.4%)	5724 (36.3%)	4645 (41.8%)	8044 (36.0%)	6243 (37.9%)	5186 (39.8%)	7724 (36.1%)	5807 (36.7%)	5222 (38.7%)
I have used the Internet 2–3 days per week	5432 (20.8%)	4127 (26.2%)	558 (5.0%)	4229 (19.0%)	3638 (22.1%)	543 (4.2%)	3780 (17.7%)	3393 (21.4%)	536 (4.0%)
I have used the Internet 0–1 days per week	4280 (16.4%)	2020 (12.8%)	427 (3.8%)	3378 (15.1%)	1645 (10.0%)	411 (3.2%)	3121 (14.6%)	1515 (9.6%)	357 (2.6%)

Table N–12. Survey Results for Question 3 by Subject and Instructional Mode

Question: When you are not in a school building, what device do you use most often to complete your schoolwork?

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
I use a personal computer (PC), Laptop, Chromebook, iPad or other tablet	20567 (78.8%)	13067 (82.9%)	10358 (93.3%)	17970 (80.5%)	14069 (85.4%)	12242 (93.9%)	17296 (80.9%)	13389 (84.5%)	12743 (94.6%)
I use a smartphone (cell phone)	4284 (16.4%)	2104 (13.3%)	537 (4.8%)	3264 (14.6%)	1845 (11.2%)	631 (4.8%)	3078 (14.4%)	1928 (12.2%)	583 (4.3%)
I am not sure what kind of device I use	442 (1.7%)	266 (1.7%)	89 (0.8%)	327 (1.5%)	222 (1.3%)	62 (0.5%)	275 (1.3%)	208 (1.3%)	54 (0.4%)
I do not use a device when I am not in a school building	731 (2.8%)	255 (1.6%)	89 (0.8%)	659 (3.0%)	270 (1.6%)	75 (0.6%)	633 (3.0%)	254 (1.6%)	60 (0.4%)

Table N–13. Survey Results for Question 4 by Subject and Instructional Mode

Question: In the past month, where have you completed most of your schoolwork?

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
in my own home	7111 (27.2%)	8002 (50.8%)	10786 (97.1%)	6802 (30.5%)	9397 (57.1%)	12737 (97.7%)	6699 (31.4%)	9357 (59.1%)	13176 (97.8%)
at the home of a friend, neighbor, or relative	1367 (5.2%)	356 (2.3%)	143 (1.3%)	697 (3.1%)	298 (1.8%)	150 (1.2%)	663 (3.1%)	297 (1.9%)	157 (1.2%)
at the place where my parent/guardian works	363 (1.4%)	113 (0.7%)	37 (0.3%)	188 (0.8%)	104 (0.6%)	29 (0.2%)	222 (1.0%)	103 (0.7%)	36 (0.3%)
at a library, store, restaurant, or other public building	157 (0.6%)	93 (0.6%)	29 (0.3%)	127 (0.6%)	109 (0.7%)	33 (0.3%)	129 (0.6%)	113 (0.7%)	22 (0.2%)
in a school building	17029 (65.2%)	7121 (45.2%)	68 (0.6%)	14423 (64.6%)	6498 (39.5%)	61 (0.5%)	13564 (63.5%)	5900 (37.2%)	57 (0.4%)

Table N–14. Survey Results for Question 5 by Subject and Instructional Mode

Question: Which statement best describes how comfortable you are with using online tools for schoolwork (such as Zoom, Schoology, Google Meet, or Teams)?

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
I feel very comfortable using online tools for schoolwork	11079 (42.4%)	5996 (38.0%)	5573 (50.2%)	9501 (42.6%)	6518 (39.6%)	6820 (52.3%)	9327 (43.7%)	6700 (42.3%)	7367 (54.7%)
I feel somewhat comfortable using online tools for schoolwork	12120 (46.4%)	7774 (49.3%)	4722 (42.5%)	10322 (46.3%)	7957 (48.3%)	5283 (40.5%)	9589 (44.9%)	7256 (45.8%)	5222 (38.7%)
I do not feel comfortable using online tools for schoolwork	2482 (9.5%)	1744 (11.1%)	627 (5.6%)	2122 (9.5%)	1779 (10.8%)	720 (5.5%)	2101 (9.8%)	1672 (10.6%)	671 (5.0%)
I have not used online tools for schoolwork	326 (1.2%)	164 (1.0%)	146 (1.3%)	249 (1.1%)	149 (0.9%)	181 (1.4%)	260 (1.2%)	141 (0.9%)	172 (1.3%)

Table N–15. Survey Results for Question 6 by Subject and Instructional Mode

Question: In the past month, I have understood almost all my assignments.

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
Strongly agree	8050 (30.8%)	3533 (22.4%)	2717 (24.5%)	6276 (28.1%)	3559 (21.6%)	3136 (24.0%)	6138 (28.7%)	3506 (22.1%)	3389 (25.1%)
Somewhat agree	13376 (51.2%)	8017 (50.9%)	5737 (51.7%)	11511 (51.6%)	8253 (50.1%)	6750 (51.8%)	10845 (50.8%)	7805 (49.3%)	6957 (51.6%)
Somewhat disagree	3409 (13.1%)	2953 (18.7%)	1860 (16.7%)	3233 (14.5%)	3289 (20.0%)	2164 (16.6%)	3150 (14.7%)	3116 (19.7%)	2189 (16.2%)
Strongly disagree	1065 (4.1%)	1059 (6.7%)	665 (6.0%)	1050 (4.7%)	1190 (7.2%)	858 (6.6%)	1030 (4.8%)	1244 (7.9%)	794 (5.9%)

Table N–16. Survey Results for Question 7 by Subject and Instructional Mode

Question: In the past month, I have completed almost all my assignments.

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
Strongly agree	12441 (47.6%)	5982 (37.9%)	5136 (46.2%)	10786 (48.3%)	6499 (39.5%)	6270 (48.1%)	10254 (48%)	6307 (39.8%)	6704 (49.7%)
Somewhat agree	10134 (38.8%)	5996 (38%)	3615 (32.5%)	7895 (35.4%)	6062 (36.8%)	4142 (31.8%)	7488 (35.0%)	5627 (35.5%)	4086 (30.3%)
Somewhat disagree	2625 (10.1%)	2571 (16.3%)	1521 (13.7%)	2626 (11.8%)	2664 (16.2%)	1676 (12.9%)	2598 (12.2%)	2587 (16.3%)	1727 (12.8%)
Strongly disagree	686 (2.6%)	989 (6.3%)	685 (6.2%)	743 (3.3%)	1045 (6.3%)	812 (6.2%)	786 (3.7%)	1148 (7.2%)	807 (6%)

Table N–17. Survey Results for Question 8 by Subject and Instructional Mode

Question: I feel positive about my learning so far this school year.

Option Text	Algebra In-person	Algebra Hybrid	Algebra Virtual	Biology In-person	Biology Hybrid	Biology Virtual	Literature In-person	Literature Hybrid	Literature Virtual
Strongly agree	6779 (26.0%)	2686 (17.0%)	2682 (24.1%)	5044 (22.6%)	2305 (14.0%)	2649 (20.3%)	4595 (21.5%)	2016 (12.7%)	2723 (20.2%)
Somewhat agree	10853 (41.6%)	6523 (41.4%)	4957 (44.6%)	9767 (43.8%)	6742 (40.9%)	5900 (45.2%)	9270 (43.4%)	6407 (40.4%)	6073 (45.1%)
Somewhat disagree	3610 (13.8%)	3338 (21.2%)	2125 (19.1%)	3793 (17.0%)	3935 (23.9%)	2766 (21.2%)	3862 (18.1%)	3851 (24.3%)	2903 (21.5%)
Strongly disagree	1694 (6.5%)	2027 (12.9%)	1164 (10.5%)	1889 (8.5%)	2486 (15.1%)	1543 (11.8%)	1948 (9.1%)	2540 (16.0%)	1596 (11.8%)

APPENDIX O: EXAMINING STUDENT PERFORMANCE IN SPRING 2019 AND SPRING 2021 USING PROPENSITY SCORE MATCHING

SUMMARY

The purpose of the research study being discussed was to contextualize results from the Keystone Spring 2021 test-taking population. To explain observed differences in performance, we compared covariates and outcomes from two testing populations: students who were administered the Keystone end-of-course examination in spring 2019 and students who were administered the examination in spring 2021. The two groups of students were significantly different on most covariates prior to matching, including race/ethnicity, locale (e.g., city, suburb), mode of administration, English Learner (EL) status, whether the student had an Individualized Educational Plan (IEP), and whether the student was economically disadvantaged. Moreover, the spring 2021 test-takers had higher scores on prior assessments (PSSA) than test-takers in prior administrations. To develop comparable groups for analyses, propensity scores were estimated using covariates, and then the two groups of test-takers were matched using nearest neighbor algorithm. Once two comparable groups were developed, independent samples t-tests were employed to assess whether there were significant differences in mean scaled scores between the two groups. After matching, results indicated that the 2021 test-takers earned lower scores than the 2019 test-takers for all subjects. These results provide evidence that the group of spring 2021 test-takers were not representative of prior spring Keystone administrations and that the continued disruption to learning caused by Covid-19 influenced student performance.

INTRODUCTION

The Pennsylvania Department of Education (PDE) develops Keystone as an end-of-course assessment for Algebra I, Biology, and Literature. The Keystone assessment program is administered three times annually (winter, spring, and summer) with the bulk of the administrations occurring during the spring. For each administration, students earn an administration-specific score, and then each scaled score is classified into a proficiency level (i.e., Below Basic, Basic, Proficient, Advanced). In addition, students earn scores on two independent modules representing unique standards, and each module score is classified into a passing status (i.e., Pass, Not Pass). For example, the Keystone Literature examination consists of both a nonfiction module and a fiction module. Earning a proficient or advanced score for each subject is one of several pathways to graduating from a Pennsylvania high school. Therefore, a student may take the test multiple times until earning a passing score. Although a test-taker earns a scaled score for each administration, Pennsylvania policy also stipulates the calculation of a best-score-to-date, and a passing status is dependent on the best-score-to-date. Only administration-specific scaled scores were used for this study.

In any given year, the testing populations across the winter, spring, and summer administration windows differ both in volume of administration and in the test-taking population's background characteristics. For example, spring test-takers tend to be predominantly first-time test-takers who are completing the corresponding course, in contrast to winter test-takers who are predominantly re-testers. The tables in the Supplemental Information section show the demographic characteristics for two administrations prior to Covid-19 (Winter 2019 and Spring 2019) and three administrations following the outbreak of Covid-19 (Summer/Fall 2020, Winter 2020–2021, and Spring 2021).

In spring 2020, the Keystone spring administration was cancelled due to the Covid-19 pandemic and resultant school closures. In addition, the Summer 2020 administration was delayed from June 2020 to September 2020, and the Winter 2020–2021 administration was extended to a four-month testing window (December 1, 2020 through March 31, 2021) to provide students with additional opportunities to take Keystone exams. In addition, the Spring 2021 administration was also extended through the start of the 2021–2022 school year (through September 30, 2021). It is important to note that for the purpose of the analyses presented here, all students who were administered Keystone were included.

There are many factors that may have influenced the test-taking population in the 2020–2021 school year. In Fall 2020, schools and districts employed a wide variety of instructional models, including a full in-school instructional model, a full virtual instructional model, and hybrid models. Hybrid models were implemented in a variety of ways across the state and could be specific to the district or school. Moreover, prior to the launch of the Winter 2020–2021 testing window, PDE passed legislation indicating that students who took the trigger course in the 2019–2020 school year did not need to take the corresponding Keystone Exam. Rather, these students

were automatically considered “Proficient” for graduation requirements. PDE also pushed back the graduation requirement so that the class of 2023 (sophomores in the 2020–2021 school year) would be the first class the graduation requirement would impact. Lastly, PDE was granted a waiver from the United States Department of Education that allowed lower participation in statewide assessments and an elongated testing window. Therefore, there were many confounding variables that led to the group of spring test-takers, and as such this group cannot be considered a representative group and should not be compared directly to any single prior administration without the use of Propensity Score Matching.

During the Winter 2020–2021 and Spring 2021 Keystone administrations, test-takers were asked to respond to a survey as a part of their exam. Although there was a very low response rate for paper-based administrations compared to online administrations (36% versus 96%), responses did not vary significantly by administration mode or by subject. During the first reporting window of spring 2021, when school was still in session, the highest proportion of students responded that they attended school fully in-person (45%), compared to fully remote (24%) and hybrid (31%). These results and the results of other questions asked on the survey support that the spring 2021 test-takers represented a variety of modes and had different experiences during the 2020–2021 school year.

Due to these factors, the spring 2021 testing population was substantially different from that of any prior administration in both student count and demographic composition. The spring 2021 administration counts were approximately 55% of the expected total administrations and approximately 74% of the spring 2019 administrations. In prior administrations, approximately 20–25% of tests were taken by students in city settings (e.g., Pittsburgh, Philadelphia); however due to differences in instructional models during the first semester of the 2020–21 academic year, there was a lower proportion of students testing in city settings (17%). The proportion of online test administrations exceeded all prior administrations, including the two most recent administrations. As such, the group of spring 2021 test-takers was a self-selected group who had an opportunity to test based on local decisions. These differences are critical to note because they contextualize observed differences in student performance across Keystone administrations. The tables in the Supplemental Information section present a comparison between key demographic characteristics for the Spring 2021, Winter 2020–2021, Summer 2020, Winter 2019–2020, and Spring 2019 administrations.

Although it has been well-established that the Covid-19 pandemic is a wide-spread public health crisis, minority groups and vulnerable populations have been disproportionately impacted. Factors that contribute to an increased risk of contracting and dying from the virus, including occupation and housing (Center for Disease Control [CDC], 2020), may have also influenced local policies that informed school closures in the 2019–2020 academic year. Therefore, districts and communities have not been uniformly impacted by the pandemic, and we expect that some of these factors influenced whether students decided to take the Keystone Exams. Demographic and background characteristics are presented in the tables in the Supplemental Information section. Additional PSM analyses were conducted by subgroups of students to investigate the specific impact on economically disadvantaged students, students with IEPs, and English Learners (ELs). See the Supplemental Information section for additional information.

The purpose of this study was to compare the student performance on Keystone examinations between two spring administrations to understand both the influence of non-traditional instructional models and the influence that self-selection to test had on mean performance and performance-level classification. This study employed propensity score matching to compare performance for the two groups of test-takers. After identifying key covariates to balance the typical group of test-takers, we then compared the total scaled score and performance-level classifications between matched groups.

METHODS

DATA

The data used for this study included two groups of test-takers for each subject: test-takers from the previous spring administration (spring 2019, hereinafter “control”) and test-takers from the most recent administration (spring 2021, hereinafter “treatment”). Test-takers were included if they met the following criteria: (a) earned a valid total score on the examination, (b) earned scores on both sessions of the examination, (c) were administered the same form for both sessions of the examination, and (d) had complete demographic and background characteristics. Any duplicate records that could not be resolved were removed from the analysis set. The same inclusion and exclusion criteria were applied to both the control and the treatment groups. Additional variables were provided to DRC by PDE and used for analyses, including demographic characteristics (i.e., race/ethnicity, gender), whether the student was identified as an EL or economically disadvantaged student, or whether the student had an IEP.

Data was merged with National Center for Educational Statistics (NCES) data that assigns districts into location classifications, such as city, suburb, town, and rural based on the physical address of a district and its proximity to an urbanized area (NCES, 2020a). These classifications were used as covariates in propensity score matching.

In addition, prior scores on the Pennsylvania System of School Assessment (PSSA) English Language Arts (ELA) and Mathematics tests were aggregated from 2015 to 2019 for grades 5 through 8. The PSSA is a summative assessment administered once a year to students in grades 3 through 8. Because the PSSA is not vertically scaled, scaled scores were z-transformed within each grade level and administration year to allow for data aggregation. The z-scores represent a test-taker’s performance relative to their cohort and were used as a covariate in the regression model for propensity score matching. Once aggregated within each subject, the most recent score was identified and merged with the Keystone data. If a Keystone test-taker did not have a most recent PSSA ELA or math score, they were excluded from the analysis set. Table O–1 shows the final dataset counts for each subject and condition that were used for matching (after common support was established).

Table O–1. Total Counts by Subject and Condition

Subject	Control (Prior)	Treatment (New)
Algebra	141,934	100,219
Biology	122,940	93,447
Literature	114,736	91,114

ANALYSES

Table O–1 shows that the count of test-takers between the treatment and control group differ, and as such comparing mean student performance between the two groups would be inappropriate. Propensity Score Matching (PSM) was used to develop comparable groups for analyzing mean differences of the total scaled score. Propensity score matching allows researchers to model the conditional probability of assignment to a condition given a set of covariates (Rosenbaum & Rubin, 1983). For the current study, we modeled the probability of a student taking the Keystone assessment in spring 2021 on the set of covariates presented in Table O–2. Unless otherwise specified, the covariates were provided to DRC by PDE.

Table O–2. Covariate Descriptions

Variable Category	Covariate	Values	Description
Race/Ethnicity	Asian	0/1	Mutually exclusive ethnicities included in analysis are Black/African American (not Hispanic), Hispanic, White/Caucasian (not Hispanic), Multi-Racial (not Hispanic), Asian (not Hispanic). Due to small counts, other races/ethnicities were considered as the reference group (Native Hawaiian or Pacific Islander and American Indian Alaskan Native).
	Black	0/1	
	Hispanic	0/1	
	Multi-Race	0/1	
	White	0/1	
Gender	Female	0/1	Genders included Male and Female and were defined according to student registration data. Male was used as the reference group.
District Location	City	0/1	NCES defines the district classifications as four mutually exclusive categories of school location: city, suburb, town, and rural. Schools are assigned to these categories in the NCES Common Core of Data based the proximity to an urbanized area. Town was used as the reference group.
	Rural	0/1	
	Suburb	0/1	
	Town	0/1	
Form	Accommodations	0/1	Whether the student was administered an accommodated form (i.e., Spanish, Braille, large print, visual sign language, etc.).
	Retester	0/1	Identifies whether the test-taker has previously taken a prior Keystone form in the same subject.
Subgroup	Economically Disadvantaged	0/1	Defines students from various poverty data sources, such as Temporary Assistance for Needy Families cases, census poor, Medicaid, children living in institutions for the neglected or delinquent, or those supported in foster homes may be used. If such data are not available, the most recent reliable data available (e.g., free and reduced-price lunch eligibility) (NCES, 2020b).
	English Learner	0/1	Defines students as English Learners (ELs) for the current administration. The classification of ELs does not include students who exited EL and are currently monitored, or students in Limited or Interrupted Formal Education (LIFE) programs.
	Individualized Education Plan	0/1	Identifies students enrolled in special education programs.
Grade	Grade Level	5–12	Students grade level at time of testing.
Prior Performance	ELA Performance (Z)	600–1800	Student's most recent ELA score, z-standardized based on the cohort of test-takers. Linear and quadratic terms were included in the model.
	Math Performance (Z)	600–1800	Student's most recent math score, z-standardized based on the cohort of test-takers. Linear and quadratic terms were included in the model.

Propensity scores were calculated using a multinomial logistic model (see equation 1). The model adequacy was examined by evaluating the area under the curve and the match quality on each covariate. To ensure that students at each propensity score level had equal probability of being in either condition group, common support techniques were employed. Specifically, after propensity scores were calculated, descriptive statistics for each condition were computed and the dataset was refined to only include students with propensity scores greater than the maximum of the minimum of propensity scores (by condition group) and the minimum of the maximum of propensity scores by condition group.

$$Pr(y) = \frac{1}{1 + e^{-(B_1x_1 + B_2x_2 + \dots + B_nx_n + B_0)}}$$

Once common support was established, propensity scores were matched (1:1) and a matched sample was created. The match quality was assessed for the covariates using chi-square tests for categorical variables and t-tests for continuous variables. The Bonferroni correction for family-wise error was applied when appropriate. In addition, PSM results were evaluated by comparing the pseudo-R2 value from the logistic regression model on the unmatched raw data and the matched data (Staffa & Zurakowski, 2018). The matched pseudo-R2 should be considerably lower than that of the unmatched data and should be close to 0 implying that the covariates do not predict the treatment. We calculated and compared McFadden’s pseudo-R2 for the total (unmatched) dataset and the matched dataset (McFadden, 1974).

Then, additional analyses were conducted to assess differences in outcomes (assessment scaled score and performance level classifications). Specifically, independent samples t-tests were conducted to determine whether there were significant differences in the outcome variables (assessment total scaled score and module scaled scores) between the treatment conditions. Lastly, chi-square tests were conducted to determine whether there were significant differences in the percentages of students in each performance-level classification. The Bonferroni correction for family-wise error was applied when appropriate.

SAS was used for all data aggregation and management. Matching samples and conducting analyses on outcome variables were conducted using MatchIt in R (Ho, Imai, King & Stuart, 2011).

RESULTS

After computing propensity scores, we established common support by refining both datasets. After computing propensity scores and ensuring common support, the match quality was assessed (see Table O-3). The means of dichotomous variables represent proportions, and the means of continuous variables represent the average within the dataset. For dichotomous variables, two-proportion Z-tests were conducted to assess whether the proportions between the control and treatment groups were significantly different. For continuous variables, independent sample t-tests were used to test whether the means between the two groups were significantly different. Then, mean difference and Cohen's D were calculated to assess match quality. These parameters are included for each subject in Table O-3.

Table O-3. Group Comparisons on Covariates After Propensity Score Matching

Subject	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen's D
Algebra I (N=100,218)	Asian	0.043	0.202	0.041	0.199	1.683	-0.002	-0.008
	Black	0.123	0.328	0.120	0.325	2.052	-0.003	-0.009
	Hispanic	0.106	0.308	0.103	0.304	2.198	-0.003	-0.010
	Multi-race	0.036	0.185	0.036	0.186	0.301	0.000	0.001
	White	0.690	0.462	0.697	0.459	3.605*	0.007	0.016
	Female	0.490	0.500	0.489	0.500	0.438	-0.001	-0.002
	Ec. Disadvantaged	0.403	0.490	0.388	0.487	6.757*	-0.015	-0.030
	EL	0.023	0.149	0.023	0.149	0.045	0.000	0.000
	IEP	0.161	0.367	0.159	0.365	1.092	-0.002	-0.005
	Accommodation	0.010	0.097	0.011	0.103	2.611	0.001	0.012
	Retester	0.059	0.236	0.059	0.236	0.246	0.000	0.001
	City	0.172	0.377	0.163	0.369	5.345*	-0.009	-0.024
	Rural	0.179	0.384	0.178	0.382	0.892	-0.002	-0.004
	Suburb	0.537	0.499	0.550	0.498	5.955*	0.013	0.027
	Town	0.103	0.304	0.102	0.302	1.135	-0.002	-0.005
	Prior ELA Scaled Score	0.056	0.981	0.080	0.981	-5.304*	0.023	0.024
	Prior Math Scaled Score	0.042	0.980	0.070	1.003	-6.256*	0.028	0.028
Grade	8.882	0.930	8.897	0.933	-3.669*	0.015	0.016	

Table O-3 (continued). Group Comparisons on Covariates After Propensity Score Matching

Subject	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen's D
Biology (N=93,446)	Asian	0.044	0.206	0.043	0.202	1.801	-0.002	-0.008
	Black	0.122	0.327	0.117	0.321	3.432*	-0.005	-0.016
	Hispanic	0.101	0.301	0.099	0.299	1.026	-0.001	-0.005
	Multi-race	0.033	0.178	0.034	0.181	1.239	0.001	0.006
	White	0.698	0.459	0.705	0.456	3.442*	0.007	0.016
	Female	0.484	0.500	0.488	0.500	1.643	0.004	0.008
	Ec. Disadvantaged	0.392	0.488	0.379	0.485	5.575*	-0.013	-0.026
	EL	0.020	0.140	0.021	0.143	1.274	0.001	0.006
	IEP	0.166	0.372	0.163	0.369	2.103	-0.004	-0.010
	Accommodation	0.012	0.107	0.012	0.110	1.428	0.001	0.007
	Retester	0.026	0.158	0.026	0.158	0.000	0.000	0.000
	City	0.167	0.373	0.161	0.368	3.209*	-0.006	-0.015
	Rural	0.175	0.380	0.170	0.375	3.306*	-0.006	-0.015
	Suburb	0.541	0.498	0.556	0.497	6.405*	0.015	0.030
	Town	0.108	0.310	0.104	0.306	2.583	-0.004	-0.012
	Prior ELA Scaled Score	0.066	0.979	0.085	0.975	-4.247*	0.019	0.020
	Prior Math Scaled Score	0.079	1.001	0.095	0.994	-3.564*	0.016	0.016
	Grade	9.623	0.585	9.614	0.605	3.13*	-0.009	-0.015
Literature (N=91,111)	Asian	0.043	0.202	0.043	0.204	0.855	0.001	0.004
	Black	0.119	0.324	0.118	0.322	1.022	-0.002	-0.005
	Hispanic	0.096	0.294	0.099	0.299	2.788	0.004	0.013
	Multi-race	0.030	0.172	0.033	0.179	3.062*	0.003	0.014
	White	0.710	0.454	0.704	0.456	2.697	-0.006	-0.013
	Female	0.489	0.500	0.490	0.500	0.567	0.001	0.003
	Ec. Disadvantaged	0.378	0.485	0.378	0.485	0.193	0.000	-0.001
	EL	0.018	0.133	0.020	0.140	3.142*	0.002	0.015
	IEP	0.161	0.368	0.161	0.368	0.070	0.000	0.000
	Accommodation	0.001	0.025	0.001	0.025	0.094	0.000	0.000
	Retester	0.015	0.123	0.015	0.123	0.000	0.000	0.000
	City	0.166	0.373	0.164	0.371	1.236	-0.002	-0.006
	Rural	0.178	0.383	0.174	0.379	2.288	-0.004	-0.011
	Suburb	0.540	0.498	0.547	0.498	3.048*	0.007	0.014
	Town	0.106	0.308	0.105	0.307	0.709	-0.001	-0.003
	Prior ELA Scaled Score	0.098	0.984	0.091	0.982	1.412	-0.007	-0.007
	Prior Math Scaled Score	0.106	1.001	0.098	0.995	1.770	-0.008	-0.008
	Grade	9.991	0.301	10.010	0.344	-12.41*	0.019	0.058

* $p < .05$

To evaluate the matched sample, we aim for the absolute value of Cohen’s D effect size for each covariate to be less than 0.10 and the average effect size to be less than 0.05 (see Tables O–3 and O–4). Both criteria were met for all subjects and covariates where the average effect sizes were less than 0.15. These findings support the matching results from PSM.

Table O–4. Average Mean Difference and Cohen’s D to Support Covariate Balance

Subject	Avg. Mean Difference	Max. Mean Difference	Avg. Cohen’s D	Max. Cohen’s D
Algebra I (N=100,218)	0.007	0.028	0.012	0.030
Biology (N=93,446)	0.006	0.019	0.013	0.030
Literature (N=91,111)	0.004	0.019	0.010	0.058

Table O–5 reports McFadden’s pseudo-R2 for the unmatched raw data, the matched raw data for each subject (McFadden, 1974). For each subject, the total pseudo-R2 is greater than the matched sample pseudo-R2. Moreover, the pseudo-R2 is very close to 0 for each subject, however the initial values for the total unmatched population are very low, possibly indicating that the PSM did not succeed at creating two well-matched samples for comparison. On the other hand, the other match quality evidence presented in Tables O–3 and O–4 support match quality by the small mean differences and effect sizes that meet acceptable criteria.

Table O–5. McFadden’s Pseudo-R2

Subject	Total (Unmatched)	Matched
Algebra I	0.0561	0.0010
Biology	0.0413	0.0004
Literature	0.0400	0.0008

An independent samples t-test was conducted to test whether there were significant differences in the scaled score between the control and treatment groups (see Table O–6). After controlling for differences in the demographic background of the two groups, we hypothesized that the matched samples would earn similar mean scaled scores. Prior to employing PSM, the two groups performed significantly differently. For Literature, the spring 2021 test-takers earned significantly higher scores than prior administrations, whereas for Algebra I and Biology, the spring 2021 test-takers earned significantly lower scores. However, after PSM was performed, the t-test results showed that student performance in 2021 was significantly lower than 2019 across all subjects, for total scaled scores as well as module scaled scores ($p < .05$). The largest mean differences were observed for Algebra I, in which total scaled score differences were 19 scaled score points lower in 2021 compared to 2019. In terms of module scaled scores, there were larger mean differences on Biology module 1 (cells and cell processes) than Biology module 2 (continuity and unity of life) as well as larger mean differences on Literature module 2 (nonfiction) than Literature module 1 (fiction). These findings support that comparisons should not be made directly between the spring 2019 and spring 2021 test-takers.

Table O–6. T-test Results for Total Sample and Matched Sample

Subject	Group	Variable	Control N	Control Mean	Control SD	Treatment N	Treatment Mean	Treatment SD	Mean Diff.	SE	t	Cohen's D
Algebra I (N=100,218)	Matched	Mod. 1 SS	100218	1496.526	66.629	100218	1478.304	61.603	-18.222	0.287	63.569*	-0.284
	Total	Mod. 1 SS	141934	1487.543	65.533	100219	1478.303	61.604	-9.240	0.261	35.400*	-0.145
	Matched	Mod. 2 SS	100218	1500.361	67.256	100218	1479.901	64.544	-20.460	0.294	69.487*	-0.310
	Total	Mod. 2 SS	141934	1490.650	66.044	100219	1479.900	64.544	-10.750	0.269	39.979*	-0.164
	Matched	Total SS	100218	1498.570	62.863	100218	1479.439	58.677	-19.131	0.272	70.427*	-0.315
	Total	Total SS	141934	1489.430	61.678	100219	1479.439	58.677	-9.991	0.247	40.403*	-0.165
Biology (N=93,446)	Matched	Mod. 1 SS	93446	1512.802	62.065	93446	1497.140	57.790	-15.662	0.277	56.459*	-0.261
	Total	Mod. 1 SS	122940	1504.279	61.144	93447	1497.139	57.790	-7.140	0.257	27.762*	-0.120
	Matched	Mod. 2 SS	93446	1512.889	60.413	93446	1500.948	55.935	-11.941	0.269	44.334*	-0.205
	Total	Mod. 2 SS	122940	1504.047	60.410	93447	1500.948	55.935	-3.099	0.251	12.332*	-0.053
	Matched	Total SS	93446	1512.311	56.772	93446	1499.026	52.690	-13.285	0.253	52.432*	-0.243
	Total	Total SS	122940	1503.896	56.407	93447	1499.026	52.690	-4.870	0.236	20.655*	-0.089
Literature (N=91,111)	Matched	Mod. 1 SS	91111	1516.952	64.044	91111	1513.603	62.692	-3.349	0.297	11.280*	-0.053
	Total	Mod. 1 SS	114736	1507.362	66.082	91114	1513.603	62.691	6.241	0.285	-21.900*	0.097
	Matched	Mod. 2 SS	91111	1517.592	64.772	91111	1509.020	68.335	-8.572	0.312	27.480*	-0.129
	Total	Mod. 2 SS	114736	1507.762	67.117	91114	1509.020	68.334	1.258	0.301	-4.180*	0.019
	Matched	Total SS	91111	1516.322	59.559	91111	1510.485	60.788	-5.837	0.282	20.704*	-0.097
	Total	Total SS	114736	1506.864	61.772	91114	1510.485	60.787	3.621	0.272	-13.329*	0.059

* $p < .05$

After conducting significance tests for scaled score differences between the control and treatment group, we conducted chi-square tests to determine whether there were significant differences in the percentages of students in each performance level classification both overall and for each module. Table O–7 shows the proportion of students classified into each of the four ordered performance levels (Below Basic, Basic, Proficient, and Advanced) for the control and treatment groups, the associated chi-square statistic, and the p -value. In addition, we also conducted the same tests for the proportion of students who earned passing scores overall and for each module. The results showed that there were significant differences in the proportion of students in all performance levels across subjects, except the proportion of proficient students in Algebra I (22.6% in control and 22.4% in treatment). However, there were significant differences in the proportion of Algebra I test-takers earning either proficient or advanced classifications (47.2% in control and 35.8% in treatment). Larger differences in proficiency were observed for Algebra and Biology than for Literature. Moreover, in Literature there were significant differences in the proportion of students earning passing scores for module 2, informational text (10.4%), and module 1, fictional text (3.1%), indicating a noticeable difference in topic areas. For Biology, there were smaller differences in the proportion of students earning passing scores for each module (8.9% for module 1 and 11.3% for module 2).

Table O–7. Chi-Square Results for Performance Level Differences

Subject	Performance Level	Module	Control N	Control Proportion	Treatment N	Treatment Proportion	χ^2	p
Algebra I (N=100,218)	Below Basic	Total	19223	19.2%	26117	26.1%	1354.587*	0.000
	Basic	Total	33662	33.6%	38179	38.1%	442.598*	0.000
	Proficient	Total	22603	22.6%	22495	22.4%	0.335	0.563
	Advanced	Total	24730	24.7%	13427	13.4%	4135.621*	0.000
	Pass (Prof/Adv)	Total	47333	47.2%	35922	35.8%	2675.381*	0.000
	Pass	Module 1	48245	48.1%	36708	36.6%	2719.537*	0.000
	Pass	Module 2	48861	48.8%	38119	38.0%	2343.865*	0.000
Biology (N=93,446)	Below Basic	Total	19298	20.7%	22724	24.3%	360.29*	0.000
	Basic	Total	22005	23.5%	28517	30.5%	1150.226*	0.000
	Proficient	Total	26787	28.7%	25088	26.8%	77.050*	0.000
	Advanced	Total	25356	27.1%	17117	18.3%	2068.358*	0.000
	Pass (Prof/Adv)	Total	52143	55.8%	42205	45.2%	2114.217*	0.000
	Pass	Module 1	51148	54.7%	42803	45.8%	1490.681*	0.000
	Pass	Module 2	54522	58.3%	43954	47.0%	2397.484*	0.000
Literature (N=91,111)	Below Basic	Total	11580	12.7%	13061	14.3%	102.913*	0.000
	Basic	Total	20249	22.2%	25456	27.9%	791.741*	0.000
	Proficient	Total	48932	53.7%	42866	47.0%	807.901*	0.000
	Advanced	Total	10350	11.4%	9728	10.7%	21.663*	0.000
	Pass (Prof/Adv)	Total	59282	65.1%	52594	57.7%	1035.841*	0.000
	Pass	Module 1	60214	66.1%	57382	63.0%	192.389*	0.000
	Pass	Module 2	59987	65.8%	50450	55.4%	2090.873*	0.000

DISCUSSION

The results of this set of analyses provide some evidence of differences in student performance between these two populations, likely influenced by the disruption to learning during the Covid-19 pandemic. In the educational measurement community, we acknowledge that student ability will likely be impacted by the disruption to the spring semester of the 2019–2020 school year and the non-streamlined approach to the re-opening of Pennsylvania schools and instructional models employed in the 2020–2021 school year.

This research study provides evidence that the groups of spring 2019 and spring 2021 test-takers are both qualitatively and quantitatively different. The 2021 test-takers were majorly white, from suburban or rural populations, and earned higher scores on prior PSSAs in comparison to the typical group of spring test-takers. After attempting to control for these differences by developing comparable groups on key covariates, the 2021 test-takers performed significantly lower than the 2019 test-takers across all subjects, both overall and by module. Larger differences were observed for Algebra and Biology than Literature, and there were differences in student performance between modules for both Biology and Literature, suggesting that specific topic areas were more impacted than others. These findings suggest that the disruption to education in spring 2020 and throughout the 2020–2021 school year may have influenced (among other factors) performance on the Keystone Exams. Moreover, the non-uniform findings across Keystone subjects may suggest that student knowledge and skills were impacted differently for Algebra and Biology than Literature. Some high school subjects may have been more conducive to online or hybrid instructional models than others. Further research in this area is warranted.

There are several limitations to the results of this study. First, it is important to note that neither students, communities, nor districts were randomly assigned to different instructional models. Districts and communities across the United States have been differently impacted by the Covid-19 pandemic due to a variety of factors. We did not have sufficient information to match each student to a single instructional model as these have also changed over the course of the first semester of the 2020–21 school year. For example, students in large city settings may have had less opportunity for in-person instructional models because large urban areas were

impacted more heavily by Covid-19 outbreaks than students in small, rural communities. As such, students from city settings were underrepresented in this study.

Second, the initial pseudo-R2 values for the total unmatched population are very low. Although the match quality was also established via alternative means (i.e., mean differences and effect sizes), the pseudo-R2 values are low. If one was to establish PSM within a high-stakes linking study, these values should be close to .9, in comparison to .03 found in this study. Although linking and equating is not the purpose of the current research study nor will this information be used for high-stakes purposes, it is important to note that the initial values of the pseudo-R2 values do not suggest matching adequacy.

This line of research has now been repeated for the following groups of test-takers: fall 2020, winter 2020–2021, and spring 2021. Results continue to show a larger impact for Algebra and Biology than Literature. It is important to note that the Keystone Exams are offered three times annually, and there is variability among the group of test-takers across administrations. Yet the results from this and prior studies suggest a certain disruption to learning evidenced by student performance on the Keystone Exams over the past year.

SUPPLEMENTAL INFORMATION

COMPOSITION OF STUDENTS BY ADMINISTRATION (IN PERCENTAGES)

Table O–8. Algebra I

Group	Category	Spring 2021 Waves 1 & 2 (N =109,678)	Spring 2021 Wave 1 only (N =94,861)	Winter 2020 (N =13,201)	Summer 2020 (N = 6,955)	Winter 2019 (N = 46,385)	Spring 2019 (N = 155,427)
Ethnicity	AIAN	0.17	0.15	0.18	0.07	0.16	0.16
	Asian	4.53	3.99	3.22	4.10	2.54	3.96
	Black	12.00	8.12	8.06	4.51	18.37	15.43
	Hispanic	10.96	9.17	11.52	5.79	14.78	12.29
	Multi-Race	3.61	3.50	3.43	3.16	3.40	3.26
	NHPI	0.09	0.08	0.09	0.12	0.08	0.08
	White	68.09	74.65	73.02	81.64	60.63	64.65
Gender	Female	48.61	48.52	49.29	48.77	48.70	49.05
	Male	50.84	51.15	50.26	50.67	51.26	50.79
Subgroup	Ec. Disadvantaged	38.05	34.52	37.83	29.27	48.44	44.11
	EL	3.37	2.25	2.53	1.85	4.87	4.08
	IEP	15.37	15.11	14.42	10.29	18.92	16.46
Mode	Online	27.31	30.05	28.97	20.92	11.56	8.75
	Paper	72.69	69.95	71.03	79.08	88.44	91.25
Locale	City	16.92	7.87	13.22	2.98	24.28	21.74
	Rural	17.14	19.10	27.55	31.34	17.21	15.62
	Suburb	54.82	61.11	46.53	49.98	46.97	51.45
	Town	9.92	10.88	11.92	15.70	9.81	9.64
Grade	6	0.27	0.30	0.07	0.78	0.00	0.25
	7	5.75	6.49	1.27	7.94	0.09	5.44
	8	24.84	27.55	4.30	26.86	1.53	21.69
	9	46.74	44.36	34.55	38.22	17.13	39.53
	10	16.38	15.67	41.83	19.35	47.81	21.57
	11	5.68	5.37	17.10	6.47	32.68	11.23
	12	0.22	0.17	0.72	0.30	0.75	0.21

Table O–9. Biology

Group	Category	Spring 2021 Waves 1 & 2 (N = 100,952)	Spring 2021 Wave 1 only (N = 86,599)	Winter 2020 (N = 12,143)	Summer 2020 (N = 6,979)	Winter 2019 (N = 35,603)	Spring 2019 (N = 135,438)
Ethnicity	AIAN	0.16	0.15	0.12	0.13	0.17	0.15
	Asian	4.54	4.01	3.63	3.83	2.76	4.12
	Black	11.80	7.99	9.06	3.93	17.57	14.33
	Hispanic	10.58	8.58	8.55	6.49	13.12	11.51
	Multi-Race	3.38	3.24	3.35	2.29	3.42	3.01
	NHPI	0.08	0.08	0.08	0.04	0.09	0.09
	White	68.97	75.66	74.59	82.82	62.83	66.67
Gender	Female	48.53	48.31	48.55	50.29	47.72	49.20
	Male	50.99	51.40	50.83	49.25	52.23	50.69
Subgroup	Ec. Disadvantaged	37.67	34.13	33.76	30.51	47.11	42.77
	EL	3.08	2.05	1.68	1.98	4.18	3.74
	IEP	15.93	15.83	14.63	11.82	19.41	16.66
Mode	Online	29.52	32.70	40.35	25.89	15.23	12.70
	Paper	70.48	67.30	59.65	74.11	84.77	87.30
Locale	City	16.83	7.45	11.90	4.00	23.02	20.14
	Rural	16.42	18.39	29.42	31.34	17.16	15.93
	Suburb	55.42	62.10	49.15	49.28	48.08	52.46
	Town	10.25	11.19	8.84	15.39	9.63	10.07
Grade	7	0.00	0.00				0.00
	8	0.28	0.25	0.01	0.96		0.17
	9	43.49	43.94	12.99	32.70	6.01	38.81
	10	49.74	50.20	66.47	54.36	45.27	47.40
	11	6.09	5.31	19.38	11.55	47.62	13.30
	12	0.29	0.23	1.07	0.36	1.08	0.27

Table O–10. Literature

Group	Category	Spring 2021 Waves 1 & 2 (N = 97,909)	Spring 2021 Wave 1 only (N = 84,036)	Winter 2020 (N = 11,714)	Summer 2020 (N = 5,399)	Winter 2019 (N = 33,239)	Spring 2019 (N = 126,692)
Ethnicity	AIAN	0.17	0.17	0.14	0.09	0.19	0.15
	Asian	4.57	4.03	3.24	3.35	2.71	4.01
	Black	11.89	8.00	7.52	3.72	18.21	14.42
	Hispanic	10.53	8.42	8.96	5.54	13.49	10.79
	Multi-Race	3.32	3.19	3.17	2.35	3.15	2.84
	NHPI	0.08	0.07	0.03		0.09	0.09
	White	68.90	75.83	76.18	84.37	62.16	67.61
	Gender	Female	48.86	48.65	47.20	49.47	43.92
	Male	50.61	51.06	52.05	49.92	56.06	51.49
Subgroup	Ec. Disadvantaged	37.41	33.81	33.17	30.82	47.80	42.13
	EL	2.80	1.79	1.14	2.35	4.28	3.39
	IEP	15.82	15.61	14.35	11.19	22.01	17.08
	Mode	Online	27.80	30.45	35.15	28.89	14.59
	Paper	72.20	69.55	64.85	71.11	85.41	88.05
Locale	City	17.14	7.60	11.17	3.07	22.01	19.96
	Rural	16.91	18.97	23.30	38.25	17.12	16.13
	Suburb	54.54	61.26	57.12	44.16	50.41	52.22
	Town	10.27	11.29	7.42	14.52	8.26	10.20
Grade	8	0.02	0.02		0.02		0.03
	9	5.03	5.00	1.67	7.72	1.58	5.32
	10	88.08	89.03	70.04	72.98	36.69	80.48
	11	6.40	5.62	26.93	19.02	60.45	13.89
	12	0.32	0.24	1.31	0.26	1.28	0.25

GROUP COMPARISON ON COVARIATES BEFORE PROPENSITY SCORE MATCHING

Table O–11. Algebra Group Comparisons on Covariates Before Propensity Score Matching

Covariate	Control (N=141,934) Mean	Control (N=141,934) SD	Treatment (N=100,219) Mean	Treatment (N=100,219) SD	Z/t	Mean Diff.	Cohen's D
Asian	0.037	0.189	0.041	0.199	5.022*	0.004	0.021
Black	0.150	0.357	0.120	0.325	21.198*	-0.030	-0.088
Hispanic	0.110	0.312	0.103	0.304	5.218*	-0.007	-0.022
Multi-race	0.033	0.177	0.036	0.186	4.553*	0.003	0.019
White	0.668	0.471	0.697	0.459	15.298*	0.029	0.063
Female	0.491	0.500	0.489	0.500	1.240	-0.003	-0.005
Ec. Disadvantaged	0.438	0.496	0.388	0.487	24.584*	-0.050	-0.102
EL	0.022	0.148	0.023	0.149	0.726	0.000	0.003
IEP	0.166	0.372	0.159	0.365	4.515*	-0.007	-0.019
Accommodated Form	0.009	0.094	0.011	0.103	4.52*	0.002	0.018
Retester	0.235	0.424	0.059	0.236	115.693*	-0.176	-0.491
City	0.205	0.404	0.163	0.369	26.337*	-0.042	-0.109
Rural	0.163	0.370	0.178	0.382	9.239*	0.014	0.038
Suburb	0.522	0.500	0.550	0.498	13.587*	0.028	0.056
Town	0.099	0.299	0.102	0.302	1.853	0.002	0.008
Prior ELA Scaled Score	-0.066	0.972	0.080	0.981	-36.207*	0.146	0.150
Prior Math Scaled Score	-0.096	0.951	0.070	1.003	-41.072*	0.166	0.171
Grade	9.046	1.047	8.897	0.933	36.848*	-0.149	-0.149

* $p < .05$

Note. Highlighted values indicate an effect size larger than the absolute value of 0.10 as these represent larger than ideal values.

Table O–12. Biology Group Comparisons on Covariates Before Propensity Score Matching

Covariate	Control (N=122,940) Mean	Control (N=122,940) SD	Treatment (N=93,447) Mean	Treatment (N=93,447) SD	Z/t	Mean Diff.	Cohen's D
Asian	0.038	0.191	0.043	0.202	5.514*	0.005	0.024
Black	0.139	0.346	0.117	0.321	15.028*	-0.022	-0.065
Hispanic	0.102	0.302	0.099	0.299	1.970	-0.003	-0.009
Multi-race	0.030	0.170	0.034	0.181	5.270*	0.004	0.023
White	0.689	0.463	0.705	0.456	7.849*	0.016	0.034
Female	0.493	0.500	0.488	0.500	2.385	-0.005	-0.010
Ec. Disadvantaged	0.423	0.494	0.379	0.485	20.487*	-0.044	-0.089
EL	0.020	0.139	0.021	0.143	1.976	0.001	0.009
IEP	0.168	0.374	0.163	0.369	3.18*	-0.005	-0.014
Accommodated Form	0.011	0.102	0.012	0.110	3.797*	0.002	0.016
Retester	0.154	0.361	0.026	0.158	99.060*	-0.128	-0.440
City	0.189	0.391	0.161	0.368	16.528*	-0.027	-0.072
Rural	0.167	0.373	0.170	0.375	1.758	0.003	0.008
Suburb	0.530	0.499	0.556	0.497	11.664*	0.025	0.051
Town	0.104	0.305	0.104	0.306	0.470	0.001	0.002
Prior ELA Scaled Score	-0.047	0.971	0.085	0.975	-31.194*	0.132	0.136
Prior Math Scaled Score	-0.055	0.974	0.095	0.994	-35.174*	0.150	0.153
Grade	9.714	0.673	9.614	0.605	36.168*	-0.100	-0.155

* $p < .05$

Note. Highlighted values indicate an effect size larger than the absolute value of 0.10 as these represent larger than ideal values.

Table O–13. Literature Group Comparisons on Covariates Before Propensity Score Matching

Covariate	Control (N=114,736) Mean	Control (N=114,736) SD	Treatment (N= 91,114) Mean	Treatment (N=91,114) SD	Z/t	Mean Diff.	Cohen's D
Asian	0.037	0.188	0.043	0.204	7.878*	0.007	0.035
Black	0.140	0.347	0.118	0.322	14.633*	-0.022	-0.065
Hispanic	0.096	0.295	0.099	0.299	2.628	0.004	0.012
Multi-race	0.028	0.165	0.033	0.179	6.454*	0.005	0.029
White	0.698	0.459	0.704	0.456	3.180*	0.007	0.014
Female	0.485	0.500	0.490	0.500	2.585	0.006	0.011
Ec. Disadvantaged	0.416	0.493	0.378	0.485	17.695*	-0.039	-0.079
EL	0.019	0.137	0.020	0.140	1.505	0.001	0.007
IEP	0.172	0.377	0.161	0.368	6.234*	-0.010	-0.028
Accommodated Form	0.001	0.023	0.001	0.025	0.699	0.000	0.004
Retester	0.125	0.330	0.015	0.123	92.746*	-0.109	-0.420
City	0.189	0.391	0.164	0.371	14.267*	-0.024	-0.063
Rural	0.169	0.374	0.174	0.379	3.165*	0.005	0.014
Suburb	0.526	0.499	0.547	0.498	9.461*	0.021	0.042
Town	0.105	0.307	0.105	0.307	0.147	0.000	-0.001
Prior ELA Scaled Score	-0.042	0.986	0.091	0.982	-30.465*	0.133	0.135
Prior Math Scaled Score	-0.038	0.986	0.098	0.995	-31.051*	0.137	0.138
Grade	10.069	0.423	10.010	0.344	35.166*	-0.059	-0.152

* $p < .05$

Note. Highlighted values indicate an effect size larger than the absolute value of 0.10 as these represent larger than ideal values.

PROPENSITY SCORE MATCHING ANALYSES BY SUBGROUPS

Students in different communities have been differently impacted by the pandemic. For example, students in rural communities may have resumed in-person instruction earlier in the year than students in large urban communities. Moreover, students who are economically disadvantaged, students with IEPs, or English Learners (ELs) may have also been adversely impacted as some resources may not have been accessible during the pandemic (e.g., free or reduced-price lunches, accommodations). As such, we conducted the same analyses presented in this paper on subgroups by selecting only test-takers in each subgroup, re-estimating propensity scores, and re-matching samples. Match quality was assessed using the same methods, and differences in the population were assessed using independent t-tests for scaled scores and chi-square tests for performance level classifications. Table O–14 shows the final dataset counts for each subject and condition that were used for matching (after common support was established). Table O–15 summarizes the quality of the match by subject (Algebra I, Biology, Literature) and subgroup analysis (Economically Disadvantaged, IEP, and EL).

Table O–14. Total Counts by Subject and Condition

Subject	Group	Control (Prior)	Treatment (New)
Algebra I	Ec. Disadvantaged	62198	38870
	EL	3173	2285
	IEP	23527	15909
Biology	Ec. Disadvantaged	51987	35434
	EL	2421	1954
	IEP	20636	15208
Literature	Ec. Disadvantaged	47738	34403
	EL	2186	1820
	IEP	19681	14690

Table O–15. Group Comparisons on Covariates After Propensity Score Matching by Subject and Subgroup

Subject/ Subgroup	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen’s D
Algebra I Ec. Disadvantaged (N=38868)	Asian	0.037	0.190	0.038	0.190	0.245	0.000	0.002
	Black	0.214	0.410	0.216	0.412	0.646	0.002	0.005
	Hispanic	0.182	0.386	0.178	0.383	1.363	-0.004	-0.010
	Multi-race	0.049	0.215	0.049	0.216	0.333	0.001	0.002
	White	0.515	0.500	0.516	0.500	0.301	0.001	0.002
	Female	0.491	0.500	0.492	0.500	0.416	0.002	0.003
	EL	0.045	0.208	0.046	0.210	0.446	0.001	0.003
	IEP	0.224	0.417	0.218	0.413	2.032	-0.006	-0.015
	Accommodated Form	0.013	0.114	0.013	0.115	0.220	0.000	0.002
	Retester	0.078	0.268	0.078	0.268	0.040	0.000	0.000
	City	0.288	0.453	0.287	0.452	0.206	-0.001	-0.002
	Rural	0.170	0.375	0.166	0.372	1.296	-0.004	-0.009
	Suburb	0.422	0.494	0.427	0.495	1.509	0.005	0.011
	Town	0.107	0.309	0.107	0.309	0.105	0.000	0.001
	Prior ELA Scaled Score	-0.309	0.916	-0.298	0.913	-1.766	0.012	0.013
	Prior Math Scaled Score	-0.337	0.852	-0.333	0.855	-0.649	0.004	0.005
Grade	9.204	0.857	9.219	0.842	-2.452	0.015	0.018	
Algebra I EL (N=2126)	Asian	0.172	0.378	0.166	0.372	0.573	-0.007	-0.018
	Black	0.102	0.302	0.086	0.280	1.790	-0.016	-0.055
	Hispanic	0.624	0.484	0.648	0.478	1.657	0.025	0.051
	Multi-race	0.016	0.124	0.009	0.097	1.797	-0.006	-0.055
	White	0.084	0.277	0.089	0.285	0.655	0.006	0.020
	Female	0.471	0.499	0.466	0.499	0.338	-0.005	-0.010
	Ec. Disadvantaged	0.793	0.405	0.778	0.416	1.196	-0.015	-0.037
	IEP	0.179	0.384	0.143	0.351	3.169*	-0.036	-0.097
	Accommodated Form	0.016	0.124	0.014	0.116	0.512	-0.002	-0.016
	Retester	0.076	0.265	0.076	0.265	0.000	0.000	0.000
	City	0.616	0.486	0.569	0.495	3.121*	-0.047	-0.096
	Rural	0.027	0.162	0.041	0.198	2.543	0.014	0.078
Suburb	0.325	0.469	0.373	0.484	3.281*	0.048	0.101	

Table O–15 (continued). Group Comparisons on Covariates After Propensity Score Matching by Subject and Subgroup

Subject/ Subgroup	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen's D
	Town	0.025	0.156	0.014	0.118	2.550	-0.011	-0.078
	Prior ELA Scaled Score	-1.152	0.660	-1.118	0.674	-1.674	0.034	0.051
	Prior Math Scaled Score	-0.852	0.599	-0.883	0.551	1.749	-0.031	-0.054
	Grade	9.355	0.713	9.444	0.681	-4.136*	0.088	0.127
Algebra I IEP (N=15894)	Asian	0.015	0.121	0.015	0.123	0.321	0.000	0.003
	Black	0.157	0.364	0.154	0.361	0.728	-0.003	-0.008
	Hispanic	0.119	0.324	0.120	0.325	0.259	0.001	0.003
	Multi-race	0.040	0.197	0.043	0.203	1.263	0.003	0.014
	White	0.667	0.471	0.665	0.472	0.238	-0.001	-0.003
	Female	0.364	0.481	0.361	0.480	0.514	-0.003	-0.006
	Ec. Disadvantaged	0.541	0.498	0.532	0.499	1.507	-0.008	-0.017
	EL	0.022	0.148	0.023	0.148	0.114	0.000	0.001
	Accommodated Form	0.052	0.222	0.065	0.246	4.757*	0.013	0.053
	Retester	0.077	0.267	0.077	0.267	-0.021	0.000	0.000
	City	0.180	0.385	0.172	0.377	2.032	-0.009	-0.023
	Rural	0.176	0.381	0.180	0.384	0.880	0.004	0.010
	Suburb	0.511	0.500	0.518	0.500	1.223	0.007	0.014
	Town	0.099	0.298	0.097	0.296	0.623	-0.002	-0.007
	Prior ELA Scaled Score	-0.837	0.800	-0.825	0.805	-1.428	0.013	0.016
	Prior Math Scaled Score	-0.753	0.707	-0.746	0.712	-0.879	0.007	0.010
Grade	9.530	0.841	9.546	0.842	-1.666	0.016	0.019	
Biology Ec. Disadvantaged (N=35431)	Asian	0.038	0.191	0.038	0.191	0.079	0.000	0.001
	Black	0.213	0.409	0.212	0.409	0.202	-0.001	-0.002
	Hispanic	0.175	0.380	0.175	0.380	0.158	-0.001	-0.001
	Multi-race	0.047	0.211	0.047	0.212	0.302	0.001	0.002
	White	0.525	0.499	0.526	0.499	0.135	0.001	0.001
	Female	0.490	0.500	0.491	0.500	0.105	0.000	0.001
	EL	0.043	0.202	0.044	0.206	1.103	0.002	0.008
	IEP	0.231	0.421	0.228	0.420	0.813	-0.003	-0.006
	Accommodated Form	0.017	0.129	0.017	0.128	0.380	0.000	-0.003
	Retester	0.038	0.192	0.038	0.192	0.000	0.000	0.000
	City	0.280	0.449	0.281	0.449	0.151	0.001	0.001
	Rural	0.163	0.369	0.159	0.366	1.338	-0.004	-0.010
	Suburb	0.426	0.494	0.431	0.495	1.594	0.006	0.012
	Town	0.117	0.321	0.115	0.318	0.940	-0.002	-0.007
	Prior ELA Scaled Score	-0.311	0.905	-0.307	0.904	-0.642	0.004	0.005
	Prior Math Scaled Score	-0.325	0.853	-0.321	0.857	-0.678	0.004	0.005
Grade	9.714	0.583	9.726	0.610	-2.702	0.012	0.020	

Table O–15 (continued). Group Comparisons on Covariates After Propensity Score Matching by Subject and Subgroup

Subject/ Subgroup	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen's D
Biology EL (N=1735)	Asian	0.180	0.385	0.150	0.357	2.423	-0.031	-0.082
	Black	0.095	0.293	0.092	0.289	0.350	-0.004	-0.012
	Hispanic	0.646	0.478	0.663	0.473	1.071	0.017	0.036
	Multi-race	0.011	0.104	0.008	0.089	0.875	-0.003	-0.030
	White	0.067	0.251	0.086	0.280	2.042	0.018	0.069
	Female	0.452	0.498	0.505	0.500	3.092*	0.052	0.105
	Ec. Disadvantaged	0.813	0.390	0.802	0.398	0.819	-0.011	-0.028
	IEP	0.194	0.396	0.152	0.359	3.275*	-0.042	-0.111
	Accommodated Form	0.019	0.137	0.018	0.132	0.252	-0.001	-0.009
	Retester	0.018	0.132	0.018	0.132	0.000	0.000	0.000
	City	0.599	0.490	0.562	0.496	2.236	-0.038	-0.076
	Rural	0.022	0.148	0.027	0.161	0.769	0.004	0.026
	Suburb	0.349	0.477	0.384	0.486	2.149	0.035	0.073
	Town	0.024	0.152	0.024	0.154	0.111	0.001	0.004
	Prior ELA Scaled Score	-1.233	0.660	-1.090	0.622	-6.586*	0.144	0.224
	Prior Math Scaled Score	-0.870	0.607	-0.831	0.557	-1.997	0.040	0.068
Grade	9.796	0.527	9.889	0.550	-5.105*	0.093	0.173	
Biology IEP (N=15206)	Asian	0.013	0.115	0.015	0.121	1.021	0.001	0.012
	Black	0.155	0.362	0.149	0.356	1.342	-0.006	-0.015
	Hispanic	0.117	0.321	0.118	0.322	0.267	0.001	0.003
	Multi-race	0.037	0.188	0.043	0.202	2.762	0.006	0.032
	White	0.677	0.468	0.673	0.469	0.563	-0.003	-0.006
	Female	0.367	0.482	0.362	0.481	0.894	-0.005	-0.010
	Ec. Disadvantaged	0.545	0.498	0.532	0.499	2.163	-0.012	-0.025
	EL	0.022	0.147	0.021	0.144	0.474	-0.001	-0.006
	Accommodated Form	0.063	0.244	0.073	0.260	3.389*	0.010	0.039
	Retester	0.041	0.197	0.041	0.197	0.000	0.000	0.000
	City	0.168	0.374	0.166	0.372	0.569	-0.002	-0.006
	Rural	0.167	0.373	0.164	0.370	0.771	-0.003	-0.009
	Suburb	0.528	0.499	0.531	0.499	0.540	0.003	0.006
	Town	0.105	0.307	0.104	0.305	0.338	-0.001	-0.004
	Prior ELA Scaled Score	-0.870	0.773	-0.843	0.785	-3.051*	0.027	0.035
	Prior Math Scaled Score	-0.784	0.677	-0.764	0.689	-2.502	0.020	0.029
Grade	9.815	0.589	9.842	0.616	-3.909*	0.027	0.045	

Table O–15 (continued). Group Comparisons on Covariates After Propensity Score Matching by Subject and Subgroup

Subject/ Subgroup	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen's D
Literature Ec. Disadvantaged (N=34370)	Asian	0.039	0.193	0.039	0.195	0.532	0.001	0.004
	Black	0.213	0.409	0.213	0.410	0.102	0.000	0.001
	Hispanic	0.176	0.381	0.177	0.382	0.350	0.001	0.003
	Multi-race	0.044	0.206	0.046	0.208	0.700	0.001	0.005
	White	0.525	0.499	0.522	0.500	0.878	-0.003	-0.007
	Female	0.490	0.500	0.490	0.500	0.069	0.000	0.001
	EL	0.038	0.191	0.042	0.201	2.919*	0.004	0.023
	IEP	0.226	0.418	0.225	0.418	0.155	-0.001	-0.001
	Accommodated Form	0.001	0.023	0.001	0.024	0.164	0.000	0.000
	Retester	0.024	0.152	0.024	0.152	0.025	0.000	0.000
	City	0.283	0.450	0.286	0.452	0.879	0.003	0.007
	Rural	0.173	0.379	0.166	0.372	2.673	-0.008	-0.021
	Suburb	0.416	0.493	0.422	0.494	1.747	0.007	0.013
	Town	0.113	0.317	0.111	0.315	0.713	-0.002	-0.005
	Prior ELA Scaled Score	-0.293	0.908	-0.303	0.911	1.442	-0.010	-0.011
	Prior Math Scaled Score	-0.297	0.862	-0.304	0.868	1.101	-0.007	-0.008
Grade	10.005	0.336	10.037	0.385	-11.59*	0.032	0.089	
Literature EL (N=1579)	Asian	0.160	0.366	0.158	0.365	0.097	-0.001	-0.004
	Black	0.098	0.298	0.106	0.308	0.706	0.008	0.025
	Hispanic	0.662	0.473	0.636	0.481	1.492	-0.025	-0.053
	Multi-race	0.016	0.125	0.013	0.112	0.751	-0.003	-0.027
	White	0.063	0.244	0.083	0.276	2.119	0.020	0.075
	Female	0.445	0.497	0.495	0.500	2.852*	0.051	0.102
	Ec. Disadvantaged	0.803	0.398	0.795	0.404	0.577	-0.008	-0.021
	IEP	0.213	0.410	0.149	0.357	4.664*	-0.064	-0.167
	Accommodated Form	0.001	0.025	0.002	0.044	1.001	0.001	0.037
	Retester	0.013	0.112	0.013	0.112	0.000	0.000	0.000
	City	0.631	0.482	0.569	0.495	3.56*	-0.062	-0.127
	Rural	0.027	0.163	0.026	0.159	0.221	-0.001	-0.008
	Suburb	0.320	0.466	0.371	0.483	3.031*	0.051	0.108
	Town	0.018	0.134	0.028	0.166	1.882	0.010	0.067
	Prior ELA Scaled Score	-1.285	0.617	-1.105	0.591	-8.361*	0.180	0.298
	Prior Math Scaled Score	-0.942	0.534	-0.815	0.560	-6.564*	0.128	0.234
Grade	10.047	0.357	10.078	0.401	-2.250	0.030	0.080	

Table O–15 (continued). Group Comparisons on Covariates After Propensity Score Matching by Subject and Subgroup

Subject/ Subgroup	Covariate	Control Mean	Control SD	Treatment Mean	Treatment SD	Z/t	Mean Diff.	Cohen's D
Literature IEP (N=14687)	Asian	0.013	0.114	0.015	0.121	1.194	0.002	0.014
	Black	0.156	0.363	0.150	0.357	1.457	-0.006	-0.017
	Hispanic	0.116	0.321	0.121	0.326	1.244	0.005	0.015
	Multi-race	0.034	0.181	0.043	0.203	4.126*	0.009	0.048
	White	0.678	0.467	0.669	0.471	1.717	-0.009	-0.020
	Female	0.367	0.482	0.363	0.481	0.800	-0.005	-0.009
	Ec. Disadvantaged	0.539	0.499	0.528	0.499	1.743	-0.010	-0.020
	EL	0.021	0.144	0.021	0.143	0.325	-0.001	-0.004
	Accommodated Form	0.003	0.051	0.003	0.050	0.116	0.000	-0.002
	Retester	0.030	0.171	0.030	0.171	0.034	0.000	-0.001
	City	0.176	0.381	0.172	0.377	0.955	-0.004	-0.011
	Rural	0.171	0.377	0.181	0.385	2.282	0.010	0.027
	Suburb	0.516	0.500	0.512	0.500	0.665	-0.004	-0.008
	Town	0.103	0.305	0.098	0.297	1.551	-0.005	-0.018
	Prior ELA Scaled Score	-0.863	0.763	-0.838	0.773	-2.709	0.024	0.032
	Prior Math Scaled Score	-0.788	0.672	-0.771	0.688	-2.205	0.018	0.026
Grade	10.047	0.375	10.080	0.434	-6.993*	0.033	0.082	

* $p < .05$

Note. Highlighted values indicate an effect size larger than the absolute value of 0.10 as these represent larger than ideal values.

The match quality for each covariate was within expectation for all subjects and subgroup analyses (average effect size less than 0.05; see Table O–16). However, several covariates for each EL analysis exceeded the generally accepted evaluation criteria ($D < 0.10$).

Table O–16. Average Mean Difference and Cohen's D to Support Covariate Balance by Subject and Subgroup

Subject	Subgroup	Avg. Mean Difference	Max. Mean Difference	Avg. Cohen's D	Max. Cohen's D
Algebra I	Ec. Disadvantaged (N=38,868)	0.003	0.015	0.006	0.018
	EL (N=2,126)	0.023	0.088	0.056	0.127
	IEP (N=15,894)	0.005	0.016	0.012	0.053
Biology	Ec. Disadvantaged (N=35,431)	0.002	0.012	0.005	0.020
	EL (N=1,735)	0.031	0.144	0.066	0.224
	IEP (N=15,206)	0.008	0.027	0.017	0.045
Literature	Ec. Disadvantaged (N=34,370)	0.005	0.032	0.012	0.089
	EL (N=1,579)	0.038	0.180	0.084	0.298
	IEP (N=14,687)	0.009	0.033	0.021	0.082

Table O–17 reports McFadden’s pseudo- R^2 for the unmatched raw data, the matched raw data for each subject (McFadden, 1974). For each subject, the total pseudo- R^2 is greater than the matched sample pseudo- R^2 . Moreover, the pseudo- R^2 is very close to 0 for each subject, however, the initial values for the total unmatched population are very low, possibly indicating that the PSM did not succeed at creating two well-matched samples for comparison. On the other hand, the other match quality evidence presented in Tables O–15 and O–16 support match quality by the small mean differences and effect sizes that meet acceptable criteria.

Table O–17. McFadden’s Pseudo- R^2

Subject	Subgroup	Total (Unmatched)	Matched
Algebra I	Ec. Disadvantaged	0.0673	0.0003
	EL	0.1099	0.0164
	IEP	0.0687	0.0009
Biology	Ec. Disadvantaged	0.0541	0.0002
	EL	0.1218	0.0282
	IEP	0.0634	0.0014
Literature	Ec. Disadvantaged	0.0594	0.0018
	EL	0.1381	0.0283
	IEP	0.0729	0.0026

After conducting PSM by subgroup, the results were similar to those of the overall findings. Table O–18 displays the results from t-tests for each subject and subgroup. The findings showed that the 2021 test-takers earned significantly lower scaled scores for each subject and subgroup except Biology EL subgroup (mean difference = -1.168). In addition, larger mean differences and effect sizes were observed for Algebra and Biology than Literature.

Table O–18. T-test Results for Matched Sample by Subject and Subgroup

Subject/ Subgroup	Scaled Score	Control Mean	Control SD	Treatment Mean	Treatment SD	Mean Diff.	SE	t	Cohen's D
Algebra I Ec. Disadvantaged (N=38,868)	Module 1 SS	1471.070	60.646	1453.839	54.166	-17.231	0.412	41.777*	-0.300
	Module 2 SS	1474.053	60.589	1453.689	57.055	-20.364	0.422	48.241*	-0.346
	Total SS	1473.270	56.400	1454.759	50.868	-18.511	0.385	48.050*	-0.345
Algebra I EL (N=2,126)	Module 1 SS	1435.350	49.004	1426.466	41.369	-8.884	1.391	6.386*	-0.196
	Module 2 SS	1438.651	45.701	1420.845	42.970	-17.806	1.361	13.085*	-0.401
	Total SS	1438.540	41.781	1425.358	36.638	-13.182	1.205	10.934*	-0.335
Algebra I IEP (N=15,894)	Module 1 SS	1444.437	53.943	1432.127	47.822	-12.310	0.572	21.527*	-0.241
	Module 2 SS	1448.227	52.864	1431.673	50.956	-16.554	0.582	28.422*	-0.319
	Total SS	1447.626	48.399	1433.498	43.802	-14.128	0.518	27.286*	-0.306
Biology Ec. Disadvantaged (N=35,431)	Module 1 SS	1488.016	53.385	1473.231	49.804	-14.785	0.388	38.118*	-0.286
	Module 2 SS	1488.067	54.424	1477.603	48.564	-10.464	0.388	27.005*	-0.203
	Total SS	1488.276	49.745	1476.103	44.952	-12.173	0.356	34.176*	-0.257
Biology EL (N=1,735)	Module 1 SS	1453.571	36.583	1443.047	36.361	-10.524	1.239	8.496*	-0.289
	Module 2 SS	1450.720	37.727	1449.552	32.986	-1.168	1.203	0.971	-0.033
	Total SS	1453.320	32.591	1447.761	29.279	-5.559	1.052	5.284*	-0.179
Biology IEP (N=15,206)	Module 1 SS	1467.640	47.001	1456.588	45.562	-11.052	0.531	20.821*	-0.239
	Module 2 SS	1465.637	48.554	1459.736	44.595	-5.901	0.535	11.037*	-0.127
	Total SS	1467.515	43.389	1459.350	40.329	-8.165	0.480	16.997*	-0.195
Literature Ec. Disadvantaged (N=34,370)	Module 1 SS	1492.038	62.916	1489.132	59.723	-2.906	0.468	6.21*	-0.047
	Module 2 SS	1492.479	63.661	1481.654	63.969	-10.825	0.487	22.236*	-0.170
	Total SS	1491.967	58.521	1485.191	57.132	-6.776	0.441	15.36*	-0.117
Literature EL (N=1,579)	Module 1 SS	1440.886	48.893	1448.946	46.195	8.060	1.693	-4.76*	0.169
	Module 2 SS	1440.324	48.494	1444.938	46.548	4.614	1.692	-2.727*	0.097
	Total SS	1441.525	42.999	1447.651	41.089	6.126	1.497	-4.092*	0.146
Literature IEP (N=14,687)	Module 1 SS	1456.495	57.194	1458.100	54.683	1.605	0.653	-2.458*	0.029
	Module 2 SS	1457.319	58.089	1447.055	55.564	-10.264	0.663	15.474*	-0.181
	Total SS	1457.475	52.619	1453.174	50.107	-4.301	0.600	7.173*	-0.084

* $p < .05$

Lastly, we calculated chi-square tests for performance level classification differences. Results indicated similar findings to the t-tests (see Table O–19).

Table O–19. Chi-Square Results for Performance Level Differences

Subject/ Subgroup	Module	Performance Level	Control N	Control Proportion	Treatment N	Treatment Proportion	χ^2	p
Algebra I Ec. Disadvantaged (N=38,868)	Total	Below Basic	12101	31.1%	16073	41.4%	878.134*	0.000
	Total	Basic	15063	38.8%	15398	39.6%	6.044	0.014
	Total	Proficient	6866	17.7%	5358	13.8%	220.79*	0.000
	Total	Advanced	4838	12.4%	2039	5.2%	1249.859*	0.000
	Total	Pass	11704	30.1%	7397	19.0%	1287.676*	0.000
	Module 2	Pass	12263	31.6%	8201	21.1%	1094.519*	0.000
	Module 1	Pass	12185	31.3%	7858	20.2%	1258.809*	0.000
Algebra I EL (N=2,126)	Total	Below Basic	1224	57.6%	1482	69.7%	67.293*	0.000
	Total	Basic	730	34.3%	559	26.3%	32.666*	0.000
	Total	Proficient	118	5.6%	63	3.0%	17.483*	0.000
	Total	Advanced	54	2.5%	22	1.0%	13.734*	0.000
	Total	Pass	172	8.1%	85	4.0%	30.672*	0.000
	Module 2	Pass	185	8.7%	95	4.5%	31.013*	0.000
	Module 1	Pass	209	9.8%	112	5.3%	31.754*	0.000
Algebra I IEP (N=15,894)	Total	Below Basic	8115	51.1%	9757	61.4%	344.355*	0.000
	Total	Basic	5592	35.2%	4773	30.0%	96.099*	0.000
	Total	Proficient	1382	8.7%	1016	6.4%	60.444*	0.000
	Total	Advanced	805	5.1%	348	2.2%	187.982*	0.000
	Total	Pass	2187	13.8%	1364	8.6%	214.788*	0.000
	Module 2	Pass	2380	15.0%	1576	9.9%	186.684*	0.000
	Module 1	Pass	2368	14.9%	1472	9.3%	237.855*	0.000
Biology Ec. Disadvantaged (N=35,431)	Total	Below Basic	11984	33.8%	13982	39.5%	242.565*	0.000
	Total	Basic	10311	29.1%	12162	34.3%	223.186*	0.000
	Total	Proficient	8487	24.0%	6626	18.7%	291.351*	0.000
	Total	Advanced	4649	13.1%	2661	7.5%	602.898*	0.000
	Total	Pass	13136	37.1%	9287	26.2%	966.694*	0.000
	Module 2	Pass	14193	40.1%	10117	28.6%	1040.467*	0.000
	Module 1	Pass	12941	36.5%	9695	27.4%	684.088*	0.000
Biology EL (N=1,735)	Total	Below Basic	1160	66.9%	1207	69.5%	2.850	0.091
	Total	Basic	419	24.1%	443	25.5%	0.871	0.351
	Total	Proficient	127	7.3%	76	4.4%	13.640*	0.000
	Total	Advanced	29	1.7%	9	0.5%	10.654*	0.001
	Total	Pass	156	9.0%	85	4.9%	21.892*	0.000
	Module 2	Pass	176	10.1%	106	6.1%	18.956*	0.000
	Module 1	Pass	164	9.5%	108	6.2%	12.545*	0.000
Biology IEP (N=15,206)	Total	Below Basic	7975	52.4%	8697	57.2%	69.098*	0.000
	Total	Basic	4212	27.7%	4392	28.9%	5.235	0.022
	Total	Proficient	2124	14.0%	1563	10.3%	97.177*	0.000
	Total	Advanced	895	5.9%	554	3.6%	84.287*	0.000
	Total	Pass	3019	19.9%	2117	13.9%	190.671*	0.000
	Module 2	Pass	3358	22.1%	2378	15.6%	206.432*	0.000
	Module 1	Pass	3124	20.5%	2341	15.4%	136.822*	0.000

Table O–19 (continued). Chi-Square Results for Performance Level Differences

Subject/ Subgroup	Module	Performance Level	Control N	Control Proportion	Treatment N	Treatment Proportion	χ^2	<i>p</i>
Literature Ec. Disadvantaged (N=34,370)	Total	Below Basic	7400	21.5%	8405	24.5%	82.948*	0.000
	Total	Basic	10379	30.2%	12356	35.9%	256.788*	0.000
	Total	Proficient	14993	43.6%	12112	35.2%	505.711*	0.000
	Total	Advanced	1598	4.6%	1497	4.4%	3.454	0.063
	Total	Pass	16591	48.3%	13609	39.6%	525.326*	0.000
	Module 2	Pass	17066	49.7%	12980	37.8%	987.334*	0.000
	Module 1	Pass	17219	50.1%	15945	46.4%	94.634*	0.000
Literature EL (N=1,579)	Total	Below Basic	827	52.4%	748	47.3%	8.003*	0.005
	Total	Basic	600	38.0%	661	41.8%	4.847	0.028
	Total	Proficient	151	9.6%	168	10.6%	0.996	0.318
	Total	Advanced	1	0.1%	2	0.1%	0.333	0.564
	Total	Pass	152	9.6%	170	10.8%	0.987	0.320
	Module 2	Pass	186	11.8%	183	11.6%	0.030	0.863
	Module 1	Pass	206	13.0%	237	15.0%	2.500	0.114
Literature IEP (N=14,687)	Total	Below Basic	6063	41.3%	6712	45.7%	58.266*	0.000
	Total	Basic	5414	36.9%	5421	36.9%	0.006	0.936
	Total	Proficient	3042	20.7%	2384	16.2%	97.928*	0.000
	Total	Advanced	168	1.1%	170	1.2%	0.012	0.913
	Total	Pass	3210	21.9%	2554	17.4%	92.941*	0.000
	Module 2	Pass	3543	24.1%	2351	16.0%	301.682*	0.000
	Module 1	Pass	3579	24.4%	3488	23.7%	1.551	0.213

* $p < .05$

REFERENCES

- Allman, C. (2004). *Test access: Making tests accessible for students with visual impairments—A guide for test publishers, test developers, and state assessment personnel* (2nd ed.). Louisville, KY: American Printing House for the Blind. Retrieved from www.aph.org.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, L.W. and Krathwohl, D.R. (Eds.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives: Complete Edition*. New York: Longman.
- Bloom, B.S. (1956). *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain*. New York: Longman.
- Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, 17(1), 5–9.
- Brennan, R. L. (2004). BB-Class (Version 1.0) [Computer Software]. Retrieved from <http://www.education.uiowa.edu/casma>.
- Center for Disease Control (July 24, 2020). Health Equity Considerations and Racial and Ethnic Minority Groups. Retrieved from <https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html>.
- Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.
- Connell, B. R., Jones, M., Mace, R., Mueller, J., Mullick, A., Ostroff, E., et al. (1997). *The principles of universal design*. Raleigh: North Carolina University, College of Design.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J., & Shavelson, R. L. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391–418.
- Data Recognition Corporation. (2010). *Fairness in testing: Training manual for issues of bias, fairness, and sensitivity*. Maple Grove, MN.
- Dorans, N., Schmitt, A., & Bleistein, C. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, 29, 309–319.
- Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT verbal and mathematical sections* (Research Report 85-10). Princeton, NJ: Educational Testing Service.
- Eignor, D. R., & Stocking, M. L. (1986). *An investigation of the possible causes for the inadequacy of IRT pre-equating* (Research Report 86-14). Princeton, NJ: Educational Testing Service.
- Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, 24(3), 21–28.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley and Sons.
- Hambleton, R., & Novick, M. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10, 159–170.

- Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score theory models. *Journal of Educational Measurement*, 27(4), 345–359.
- Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices*, 10(2), 33–41.
- Hess, K. (2004). *Applying Webb's depth-of-knowledge levels in reading*. [online] available: www.nciea.org.
- Ho D.E., Imai K., King G., Stuart E.A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(8), 1–28. <https://www.jstatsoft.org/v42/i08/>.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices* (2nd ed.). New York: Springer.
- Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, 27 (1), 27–39.
- Lane, S. (1999). *Validity evidence for assessments*. Paper presented at the 1999 Edward F. Reidy, Interactive Lecture Series, Providence, RI.
- Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, 21(1), 23–30.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A bookmark approach*. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.
- Linacre, J. M. (2018a). *A user's guide to WINSTEPS MINNISTEP Rasch-model computer programs*. Chicago: Winsteps.
- Linacre, J.M. (2018b). Winsteps® (Version 4.8.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.
- Livingston, S., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, 9(3), 200–215.
- McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research*, 14, 21–38.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 104–142). New York: Academic Press.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 3–104). Washington, DC: American Council on Education.
- National Center for Education Statistics (2020a). School District Characteristic 2018–2019 [Data file]. Retrieved from <https://data-nces.opendata.arcgis.com/datasets>
- National Center for Education Statistics (2020b). The Common Education Data Standards (CEDs). Retrieved from <http://ceds.ed.gov/>

- Pennsylvania Department of Education. (2010). *Psychometric analysis report for the fall 2010 Keystone field tests*. Harrisburg, PA: PDE.
- Pennsylvania Department of Education. (2011). *Algebra I Keystone Feb 2011 item and scoring sampler*. Retrieved November 11, 2011, from www.pdesas.org/Assessment/Keystone.
- Pennsylvania Department of Education. (2011). *Biology Keystone Feb 2011 item and scoring sampler*. Retrieved November 11, 2011, from www.pdesas.org/Assessment/Keystone.
- Pennsylvania Department of Education. (2011). *Keystone Exams score report focus group findings*. Harrisburg, PA: PDE.
- Pennsylvania Department of Education. (2011). *Keystone standard setting technical report: Algebra I, Biology, and Literature*. Harrisburg, PA: PDE.
- Pennsylvania Department of Education. (2021). *19 Item and Scoring Sampler – Algebra I*. Retrieved February 9, 2021 from www.education.pa.gov.
- Pennsylvania Department of Education. (2021). *2019 Item and Scoring Sampler – Biology*. Retrieved February 9, 2021 from www.education.pa.gov.
- Pennsylvania Department of Education. (2021). *2019 Item and Scoring Sampler – Literature*. Retrieved February 9, 2021 from www.education.pa.gov.
- Pennsylvania Department of Education. (2021). *2021 Accommodations Guidelines* (PDE, revised 9/2020). Retrieved April 26, 2021 from www.education.pa.gov.
- Pennsylvania Department of Education. (2021). *2021 Accommodations Guidelines for ELs* (PDE, revised 9/2020). Retrieved April 26, 2021 from www.education.pa.gov.
- Pennsylvania Department of Education. (2021). *2021 Pennsylvania System of School Assessment: Handbook for Assessment Coordinators*. Retrieved April 26, 2021 from www.education.pa.gov.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–262). Washington, DC: American Council on Education.
- R (Version 3.6) [Computer software]. Vienna, Austria: R Core Team.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 17(1), pp. 41–55.
- SAS (Version 7) [Computer software]. Cary, NC: SAS Institute, Inc.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
- Staffa, S. & Zurakowski, D. (2018). Five steps to successfully implement and evaluate propensity score matching in clinical research studies. *Anesthesia & Analgesia*, 127(1).
- Stocking, M. L., & Eignor, D. R. (1986). The impact of different ability distributions on IRT preequating (Research Report No. 86–14). Princeton, NJ: Educational Testing Service.

- Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- U.S. Department of Education, Office of Elementary and Secondary Education. (2004, April 28). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001*. Washington, DC: Author.
- Valencia, S.W. and Wixson, K.K. (2000). *Policy-oriented research on literacy standards and assessment*. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, and R. Barr (Eds.), *Handbook of Reading Research: Vol. III*. Mahwah, NJ: Lawrence Erlbaum.
- Webb, N. L. (1997). *Criteria for alignment of expectations and tests in mathematics and science education*. Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.
- Webb, N. L. (1997; 2006). *Research monograph number 6: Criteria for alignment of expectations and assessments on mathematics and science education*. Washington, D.C.: CCSSO.
- Webb, N. L. (1999). *18: Alignment of science and mathematics standards and assessments in four states*. Research Monograph No. Madison, WI: National Institute for Science Education.
- Webb, N. L. (1999). *Research monograph No. 18: Alignment of science and mathematics standards and assessments in four states*. Washington, D.C.: CCSSO.
- Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and tests for four states: State collaborative on test and state standards (SCASS)*. Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research.
- Webb, N. L. (November, 2005). *Depth-of-Knowledge levels for four content areas*. Presentation to the Florida Education Research Association, 50th Annual Meeting, Miami, Florida.
- Webb, N. L. (2006). *Web alignment tool* [Computer Software]. Madison: Wisconsin Center of Educational Research. University of Wisconsin-Madison.
- Webb Alignment Tool (WAT) Training Manual retrieved from <http://www.wcer.wisc.edu/WAT/index.aspx>.
- Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.
- Zwick, R., & Erickson, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26, 55–66.