ISASP IOWA STATEWIDE ASSESSMENT of STUDENT PROGRESS

Technical Manual

Version 2.0

Prepared at The University of Iowa by

Catherine Welch and Stephen Dunbar

Iowa Testing Programs

Table of Contents

LIST OF TABLES ..... **IV**

LIST OF FIGURES ..... **V**

PURPOSE ..... **VI**

CHAPTER 1 : OVERVIEW **1-1**

IOWA EDUCATORS 1-2
IOWA DEPARTMENT OF EDUCATION 1-3
PEARSON 1-3
HUMAN RESOURCES RESEARCH ORGANIZATION 1-3
TECHNICAL ADVISORY COMMITTEE 1-3

CHAPTER 2 : TEST DEVELOPMENT **2-1**

TEST DEVELOPMENT PROCEDURES 2-1
TEST SPECIFICATIONS 2-1
Mathematics Test Specifications *2-4*
Science Test Specifications *2-6*
Item Types *2-7*
Cognitive Complexity *2-8*
Statistical Specifications *2-9*
DEVELOPMENT OF ITEMS AND TESTING MATERIALS 2-10
REVIEW PROCESSES 2-10
Content Review *2-11*
Fairness Review *2-11*
FIELD TESTING 2-15
DATA REVIEW 2-15
Item-Level Statistics Reviewed *2-15*
Fairness Review Summaries *2-15*
Analysis of Differential Item Functioning *2-17*
FORMS ASSEMBLY 2-20

CHAPTER 3 : TEST ADMINISTRATION AND ACCOMMODATIONS **3-1**

ELIGIBILITY FOR ASSESSMENTS 3-1
ADMINISTRATION TO STUDENTS 3-2
SECURE TESTING MATERIALS 3-2
Iowa Statewide Assessment of Student Progress Security *3-3*
FEATURES AND ACCOMMODATIONS 3-4
Research Base for Features and Accommodations *3-4*
Accommodations Use Monitoring *3-9*

CHAPTER 4 : REPORTS **4-1**

DESCRIPTION OF SCORES 4-1
Scale Scores *4-1*
Achievement Levels *4-1*
Domain Scores *4-2*
Iowa Percentile Ranks *4-2*
DESCRIPTION OF REPORTS 4-3
Individual Student Reports *4-3*
APPROPRIATE SCORE USES 4-4
Individual Students *4-5*
Groups of Students *4-5*
Cautions for Score Use *4-6*

CHAPTER 5 : PERFORMANCE STANDARDS 5-1

PROCESS FOR DEVELOPING PERFORMANCE LEVEL DESCRIPTIONS 5-1
Overview 5-1
General Performance Level Definitions 5-1
Implementation 5-2
Finalized Performance Level Descriptors 5-3
STANDARD SETTING 5-3
Process 5-4
Facilitator Training 5-4
Committee Panelist Composition 5-5
General Method 5-5
Vertical Articulation 5-6
Cut Scores 5-6
Results for ISASP Assessments 5-8
State Approval 5-8

CHAPTER 6 : SCALING AND EQUATING 6-1

RATIONALE 6-1
MEASUREMENT MODELS 6-2
Two-Parameter Logistic Model 6-2
DEVELOPMENT OF THE ISASP SCALE SCORE REPORTING METRIC 6-6
Scale Attributes and Interpretive Guidance for ISASP Scale Scores 6-9
DOMAIN SCORES FOR THE ISASP ASSESSMENTS 6-10
EQUATING AND LINKING THE ISASP ASSESSMENTS 6-11
Rationale 6-11
Pre-Equating 6-12

CHAPTER 7 : VALIDITY 7-1

TEST VALIDITY EVIDENCE 7-1
Evidence Based on Test Content 7-2
Evidence Based on Response Processes 7-2
Evidence Based on Relations to Other Variables 7-5
Evidence of Comparability across Modes of Administration 7-7
Construct Comparability 7-8

CHAPTER 8 : RELIABILITY 8-1

ESTIMATING RELIABILITY 8-1
SCORING SYSTEM 8-7
EVALUATION OF HUMAN SCORERS AND INTELLIGENT ESSAY ASSESSOR (IEA) 8-10
CLASSIFICATION CONSISTENCY AND ACCURACY 8-13

CHAPTER 9 : QUALITY-CONTROL PROCEDURES 9-1

QUALITY CONTROL FOR TEST CONSTRUCTION 9-1
QUALITY CONTROL FOR NON-SCANNABLE DOCUMENTS 9-1
QUALITY CONTROL FOR ONLINE TEST DELIVERY COMPONENTS 9-2
QUALITY CONTROL IN SCALING, EQUATING, AND LINKING IN THE ISASP PROGRAM 9-3

REFERENCES R-1

List of Tables

Table 2.1. Iowa Core Domain Coverage in ELA by Grade 2-3
Table 2.2. Complexity Ranges for ISASP Reading Texts 2-3
Table 2.3. Complexity Ranges for ISASP Writing Texts 2-4
Table 2.4. Iowa Core Domain Coverage in Mathematics – Grades 3–8 2-5
Table 2.5. Iowa Core Domain Coverage in Mathematics – Grades 9–11 2-6
Table 2.6. Iowa Core Domain Coverage in Science by Grade 2-7
Table 2.7. Examples of Technology-Enhanced Item Types 2-8
Table 2.8. ISASP Cognitive Level Descriptions 2-9
Table 2.9. Percentage of ELA Items by DOK Level 2-9
Table 2.10. Percentage of Mathematics Items by DOK Level 2-9
Table 2.11. Percentage of Science Items by DOK Level 2-9
Table 2.12. ISASP Fairness Procedures in Test Development 2-11
Table 2.13. Fairness Ratings for Reading 2-16
Table 2.14. Fairness Ratings for Language/Writing 2-16
Table 2.15. Fairness Ratings for Mathematics 2-17
Table 2.16. Fairness Ratings for Science 2-17
Table 2.17. Comparison Groups for Differential Item Functioning Analysis 2-18
Table 2.18. DIF Classification Categories for Dichotomous Items 2-19
Table 2.19. DIF Classification Categories for Polytomous Items 2-19
Table 3.1. Support and Accommodations for ISASP 3-5
Table 4.1. ELA Cut Score Ranges for ISASP Performance Levels 4-1
Table 4.2. Mathematics Cut Score Ranges for ISASP Performance Levels 4-2
Table 4.3. Science Cut Score Ranges for ISASP Performance Levels 4-2
Table 4.4. ISASP Reports 4-3
Table 4.5. Appropriate Uses of ISASP Results 4-5
Table 5.1. General Performance Level Descriptors for ISASP 5-2
Table 5.2. Cut Score Ranges for ISASP Performance Levels 5-7
Table 6.1. Vertical Scale Parameters for the ISASP Scale Score Distributions 6-7
Table 7.1 Goodness-of-Fit of Bifactor Internal Structure Models for ISASP Domain Scores Based on the Iowa Core Standards 7-5
Table 7.2. Correlations Between Student Standard Scores on the 2019 ISASP and 2018 *Iowa Assessments* 7-6
Table 7.3. Correlations Between Student Standard Scores on the 2019 ISASP and FAST CBMreading 7-7
Table 7.4. Correlations Between Student Standard Scores on the 2019 ISASP and FAST aReading 7-7
Table 7.5. Goodness-of-Fit Statistics for Structure Models of Measurement Metric Invariance for ISASP Domain Scores for Computer-Based and Paper-Based Test Administrations 7-8
Table 7.6. Number of C-DIF Flagged Items for Mode 7-9
Table 8.1. Estimates of Reliability and Standard Errors of Measurement for 2019 ISASP 8-3
Table 8.2. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Reading Assessment 8-5
Table 8.3. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Language/Writing Assessment 8-5
Table 8.4. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Mathematics Assessment 8-5
Table 8.5. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Science Assessment 8-6
Table 8.6. CSEM of Theta by Achievement Level 8-6
Table 8.7. Analysis of scorer recruitment, training, retention/ dismissal 8-9
Table 8.8. Interrater Reliability – Human-Human Scoring 8-11
Table 8.9. Interrater Reliability – IEA-Human Scoring 8-12
Table 8.10. Science 8-13
Table 8.11. Reading 8-13
Table 8.12. Classification Consistency and Accuracy for Not-Yet-Proficient and Proficient Designations 8-14
Table 9.1. Summary of Technical Analysis and Ongoing Maintenance 9-4

List of Figures

Figure 6.1. Two-Parameter Item Response Functions for a Mathematics Item from ISASP Grade 5.....6-3
Figure 6.2. Two-Parameter Graded Response Model Category Response Functions for a Constructed-Response Mathematics Item from ISASP Grade 8 6-5
Figure 6.3. Two-Parameter Graded Response Model Expected Item Score Function for a Constructed-Response Mathematics Item from ISASP Grade 8 6-6
Figure 6.4. Relative Frequency Distributions of ISASP Scale Scores in Reading, Language/Writing, Mathematics, and Science – Grades 3–11.....6-8
Figure 6.5. Cumulative Frequency Distributions of ISASP Scale Scores in Reading, Language/Writing, Mathematics, and Science – Grades 3–11 6-9
Figure 6.6. Raw Score to Iowa Scale Score Transformation – ISASP Grade 8 Mathematics.....6-10
Figure 6.7. Test Characteristic Curves for the 2019 and 2021 ISASP Reading Assessments in Grade 6 6-14
Figure 6.8. Test Characteristic Curves for the 2019 and 2021 ISASP Mathematics Assessments in Grade 8 6-15

Purpose

Technical documentation for the Iowa Statewide Assessment of Student Progress (ISASP) is organized into two documents. The *ISASP Technical Manual* addresses the development and measurement characteristics in chapters that outline the construction of the assessment, statistical analysis of the results, and meaning of scores on these tests. The *ISASP Annual Statistical Reports (ISASP ASR-2019, ISASP ASR-2021*, and continuing in subsequent years*)* provide data summaries of various aspects of technical quality that pertain to the reliability, validity, and technical adequacy of the ISASP program and its assessments.

This manual does not include all the information available regarding the assessment program in Iowa. Additional information can be found at the Iowa Department of Education (IDOE).

Iowa Testing Programs (ITP) is committed to following generally accepted professional standards when creating, administering, scoring, and reporting test scores. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014) is one source of professional standards. This document will be referenced throughout this manual as the *Standards.*

Chapter 1: Overview

Iowa Code Chapter 256 "Department of Education," subsection 7 was amended effective July 1, 2018, to specify that the Iowa Statewide Summative Assessment of Student Progress (ISASP) administered by school districts for purposes of the core academic indicators shall be the summative assessment developed by Iowa Testing Programs (ITP) within the University of Iowa's College of Education.

Iowa Code Chapter 256.7(21)(2b) identifies the purpose of the assessment as "accurately describe student achievement and growth for purposes of the school, the school district, and state accountability systems; provide valid, reliable and fair measures of student progress toward college or career readiness."

Iowa Code requires that the ISASP be available for administration in both paper-and-pencil and computer-based formats as well as in Spanish for assessments in Mathematics and Science.

The ISASP assesses students in English Language Arts (ELA) and Mathematics in grades 3–11. The Science assessment is administered to students in grades 5, 8, and 10. The assessments in paper-and-pencil and computer-based formats include multiple-choice and technology-enhanced items, constructed-response items, and open-ended essay questions.

The major claims of the ISASP include statements regarding measures of student achievement on the Iowa Core Standards with respect to readiness and growth. Major claims include:

Student Achievement on the Iowa Core

- Students demonstrate their understanding of the Iowa Core Standards in ELA
- Students demonstrate their understanding of the Iowa Core Standards in Mathematics
- Students demonstrate their understanding of the Iowa Core Standards in Science

Readiness

- Students demonstrate progress toward college and career readiness in ELA in the areas of Reading, Language, and Writing.
- Students demonstrate progress toward college and career readiness in Mathematics.
- Students demonstrate progress toward college and career readiness in Science.

Growth

- Students demonstrate growth from grade to grade in ELA in the areas of Reading, Language, and Writing.
- Students demonstrate growth from grade to grade in Mathematics.
- Students demonstrate growth across grade bands (3–5, 6–8, and high school) in Science.

A major requirement for the ISASP to achieve its claims involves implementing a test design that supports the measurement of the content knowledge and skills students need to ultimately graduate from high school prepared for their postsecondary pursuits. To that end, the ISASP is designed to measure students'

understanding of the Iowa Core Standards. The Iowa Core Standards are derived from the Common Core State Standards, which are widely used across the nation to guide states in their preparation of students for college and careers.

The test design specifications outline the content domains assessed by the ISASP assessments. These domains reflect the breadth of content outlined in the standards documents. The test blueprints further detail the number of operational items that test forms will include in each content domain, as well as the levels of cognitive complexity, or depth of knowledge, at which the content will be measured.

The assessments are aligned with the Iowa Core Standards and provide a clear and accurate assessment of student learning outcomes. The ISASP is administered between March and May annually.

Iowa Testing Programs

ITP is designated by Iowa Code as the designer and developer of the ISASP. ITP has been part of the University of Iowa's College of Education for over 90 years. From its inception, ITP's mission has been to serve Iowa schools' assessment needs by developing high-quality instruments and working with schools to promote valid use of results. In fulfilling this mission, ITP pursues:

- research that improves the practice of educational measurement,
- design and development of assessments that provide information for a variety of purposes and audiences; and
- outreach that delivers assessment results and supports use of assessment information by local, state, national, and international audiences.

ITP faculty teach undergraduate and graduate courses in the University of Iowa College of Education's Educational Measurement and Statistics program and train hundreds of professionals across the country and around the world as leaders in the field of large-scale assessment. ITP faculty and staff conduct research on new testing initiatives and develop new approaches to assessment and reporting on student achievement.

Organizations and Groups Involved

The ISASP is developed in close collaboration with Iowa educators and students; additional groups and organizations are also involved with the ISASP assessment program. Each of the major contributors listed below serves a specific role, and their collaborative efforts contribute significantly to the program's success.

Iowa Educators

Iowa educators—including classroom teachers and instructional coaches from K–12 and higher education, curriculum specialists, administrators, and members of the best practice networks (working groups of expert teachers in specific content areas)—play a vital role in all phases of the test development process. Committees of educators provide expert feedback and verification in critical test development areas. Specific examples of their involvement include reviewing test specifications, providing and verifying alignment, drafting and/or reviewing items, participating in scoring activities for open-ended items, drafting and reviewing performance level descriptors, and participating in the standard setting study.

Iowa Department of Education

The Iowa Department of Education (IDOE) works with ITP to coordinate communication about the ISASP to Iowa schools, Area Education Agencies, advisory boards, and membership organizations. The IDOE works closely with ITP to define the training modules to be delivered to the state of Iowa and to set policy for critical aspects of the ISASP program related to compliance with federal law. In addition, ITP and the IDOE worked collaboratively with Pearson, ITP's test delivery partner, to design and implement a standard-setting process in accordance with Peer Review Guidance.

Pearson

Iowa Code Chapter 256.7 specifies that the ISASP program be administered by ITP's designee. Pearson was awarded a five-year contract for the administration, scoring, and reporting for the ISASP in July 2018.

Human Resources Research Organization

Human Resources Research Organization (HumRRO) is a separate contractor working with ITP to complete quality assurance checks associated with elements of the ISASP. HumRRO conducted quality alignment checks during test development and assembly of the ISASP. HumRRO has also conducted alignment studies for ITP to evaluate the congruence between the items on the ISASP assessments and the skills specified in the Iowa Core.

Technical Advisory Committee

To support the ISASP, ITP conducts two Technical Advisory Committee meetings annually. The Technical Advisory Committee includes five nationally recognized specialists in critical aspects of large-scale assessment and accountability:

- John Poggio, University of Kansas, online testing, accessibility, and accommodations
- Joan Herman, National Center for Research on Evaluation, Standards, and Student Testing, UCLA, test interpretations and use
- Tim Davey, Educational Testing Service, adaptive testing, item pools/management, scaling, linking
- Walter D. Way, College Board, adaptive testing, IRT scaling, scoring algorithms
- Erica L. Landl, Center for Assessment, alignment, accessibility, validation, and use

Chapter 2: Test Development

The assessments that make up the Iowa Statewide Assessment of Student Progress (ISASP) are the result of an extended, iterative process during which test materials are developed and administered to statewide samples to evaluate the materials' measurement quality and appropriateness.

Test Development Procedures

The following steps provide an overview of the process that was followed to develop the ISASP. More detailed descriptions of these processes are provided in the sections that follow.

1. **Test Specifications.** Test specifications outline the distribution of content, skills, and cognitive levels across a test form, the statistical specifications, and the item types and formats.
2. **Development of Items and Testing Materials.** Using the *Standards*, the Iowa Core Standards, and the test specifications, item developers develop items, stimuli, tasks, and scoring rubrics.
3. **Review Processes.** All testing materials are reviewed for a variety of purposes and with respect to a variety of audiences. The review processes include content, sensitivity and fairness, language accessibility, and universal design considerations.
4. **Field Testing.** Items that successfully pass the review process are administered as part of the assessment process. Data describing student performance on the items are used to estimate item difficulty, item discrimination, test reliability, and differential item functioning.
5. **Data Review.** Committees of Iowa Testing Programs (ITP) staff and external consultants, including the Technical Advisory Committee, evaluate the item-level statistics and calibration results to identify those items eligible for inclusion in a final ISASP form.
6. **Forms Assembly.** Items that are eligible after the review processes are complete make up the pool to be used in the final forms assembly process. Forms are built to match the test specifications in terms of content representativeness and statistical specifications.

Test Specifications

Criterion-referenced assessments like the ISASP are based on Iowa's academic content standards in English Language Arts (ELA), Mathematics, and Science. The Iowa Core standards specify what students are expected to know and be able to do by the end of each grade. The standards for ELA and Mathematics were adopted by the Iowa State Board of Education on July 29, 2010. On August 6, 2015, the Iowa State Board adopted the Iowa Core Science Standards. These standards collectively provide the content foundation for the ISASP test specifications.

The content and cognitive emphasis of the Iowa Core Standards should be reflected in the overall test specifications. To that end, the ISASP assessments have been designed to mirror the rigor of the Iowa Core Standards while providing scores that are accurate and informative. Two different approaches have been used to help achieve this balance. First, a mix of various item formats was selected to best measure the types of knowledge, skills, and abilities reflected in the Iowa Core. Second, test specifications were designed to reflect the appropriate content and skills coverage as defined by the Iowa Core while at the same time providing a technically sound, efficient, and informative assessment design.

English Language Arts Specifications

The ELA assessments of ISASP have been designed and developed to support the following claims with respect to student performance.

- Students demonstrate their understanding of the Iowa Core Standards in ELA.
- Students demonstrate progress toward college and career readiness in ELA in the areas of Reading, Language, and Writing.
- Students demonstrate growth from grade to grade in ELA in the areas of Reading, Language, and Writing.

The ELA section of the ISASP includes separately administered, untimed assessments in Reading and Language/Writing for use during the last 12 weeks of the academic year, as specified in Iowa Code. These summative assessments measure student achievement, college and career readiness, and growth as reflected in the grade-level scope and sequence defined by the Iowa Core Standards.

In the Reading assessments, students experience a dynamic, multi-layered interaction with text and questions that measure their understanding of text features and meaning. As the assessments progress from grade 3 to grade 11, the text complexity of reading passages increases (materials increase both in length and structural complexity/difficulty), and the number of higher-order thinking and interpretive items increases. Students are asked both to identify key ideas and details and to interpret, evaluate, and integrate them with ideas expressed in other print material presented in context with the main passage. Approximately two-thirds of the questions at each test level require the reader to construct either inferential or evaluative meaning. In addition to integrating knowledge and key details, students must make judgments about the author's craft and technique, or about large ideas and applications based on the text. These comprehension tasks authentically reflect what educators know about the complex nature of the reading process.

The Language/Writing assessment is composed of separately administered parts that focus on both the editing and composing aspects of the writing process as defined by the Iowa Core. In the first section, students read drafts of texts composed for a particular purpose and audience (narrative, expository, persuasive) and answer questions about language structure and writing technique in areas such as style, word choice, linguistic conventions, and related aspects of the use of language to express thoughts and ideas. These questions assess components of the writing process students are taught in the classroom. This section presents both elementary and high school students with an "edit and revise" format, where they consider possible revisions to consider and make choices about what needs to be revised in the written drafts. This section of the Language/Writing assessment is composed of multiple-choice and technology-enhanced items.

The second section of this assessment is devoted to evidence-based writing that requires students to integrate ideas from source materials provided to the student (e.g., visual material and/or written texts across a broad range of subject matter) into their writing as support or illustration. Each writing task presents students with a real-world writing situation and designates a specific intended audience and purpose. A variety of task types (modes of discourse such as opinion/argument, explanatory, or narrative) are used over the grade 3 to grade 11 range, and task types are drawn from those specified in the Iowa Core Standards at each grade level.

Domain Coverage of the Iowa Core in ELA

Table 2.1 provides the coverage targets in the Iowa Core of the domains that are assessed and reported for the ELA assessments of the ISASP in grades 3 through 11.

Table 2.1. Iowa Core Domain Coverage in ELA by Grade

Iowa Core	Grade						
ELA Domain	3	4	5	6	7	8	9–11
Key Ideas and Details	45–55%	45–55%	45–55%	45–55%	45–55%	45–55%	40–55%
Craft and Structure	30–38%	30–38%	32–44%	34–44%	34–44%	34–44%	35–43%
Integration of Knowledge and Ideas	10–20%	10–20%	10–20%	10–20%	9–18%	9–18%	10–20%
Conventions of Standard English / Knowledge of Language	35–45%	35–45%	35–45%	38–48%	38–48%	38–48%	36–46%
Vocabulary Acquisition and Use	5–15%	5–15%	5–15%	5–15%	5–15%	5–15%	4–12%
Text Types and Purposes	22–32%	22–32%	22–32%	25–35%	25–35%	25–35%	25–35%
Production and Distribution of Writing	7–15%	7–15%	7–15%	5–15%	5–15%	5–15%	6–14%
Research to Build and Present Knowledge	7–15%	7–15%	7–15%	5–15%	5–15%	5–15%	6–14%

Note: Texts include Literature and Informational (including History/Social Studies and Science/Technical)

Text Complexity

Text complexity through quantitative and qualitative measures is addressed for all texts in the Reading and Writing sections of the ELA assessments. Quantitative measures are aspects of text complexity that are unlikely to be evaluated by a subject matter expert reliably, and therefore calculations made with computer software are used. The quantitative measures relevant to passage development for the ISASP are the Flesch-Kincaid index of readability and Lexile® scores. ITP follows the Iowa Core Standards guidelines for Flesch-Kincaid and Lexile grade band ranges for all assessments. Tables 2.2 and 2.3 present the quantitative text complexity ranges in each grade for ISASP Reading and ISASP Writing tests, respectively.

Table 2.2. Complexity Ranges for ISASP Reading Texts

	Grade				
Reading Text Complexity	**3**	**4–5**	**6–8**	**9–10**	**11**
Flesch-Kincaid	1.98–5.34	4.51–7.73	6.51–10.34	8.32–12.12	10.34–14.2
Lexile®	420–820	740–1,010	925–1,185	1,050–1,335	1,185–1,385

Table 2.3. Complexity Ranges for ISASP Writing Texts

	Grade			
Writing Text Complexity	**3**	**4–5**	**6–8**	**9–11**
Flesch-Kincaid	1.98–5.34	4.51–7.73	6.51–10.34	6.51–10.34
Lexile®	420–820	740–1,010	925–1,185	925–1,185

Qualitative measures of text complexity are best determined by subject matter experts who can evaluate the use, organization, language appropriateness, and the likely level of understanding of the target reader. The qualitative measures used as part of passage development in the ISASP program are documented in a passage review checklist and evaluated by a minimum of two independent subject matter experts during the development process. Finally, an overall determination of specific grade appropriateness is conducted using both quantitative and qualitative scores.

This process is also followed for evaluating the complexity of texts that appear in the ISASP Mathematics and Science assessments to ensure reading load is appropriate for the grade and does not introduce construct-irrelevant variance in students' scores on those assessments.

In addition to the information included in this manual, detailed tables of test specifications are posted on the ISASP portal to allow educators to obtain a thorough understanding of the domain coverage of each grade-level assessment in ELA with respect to the Iowa Core Standards. Detailed test specifications for the ISASP ELA assessments for grades 3 to 11 can be found at http://iowa.pearsonaccessnext.com/test-prep/.

Mathematics Test Specifications

The ISASP Mathematics assessments have been designed and developed to support the following claims with respect to student performance:

- Students demonstrate their understanding of the Iowa Core Standards in Mathematics.
- Students demonstrate progress toward college and career readiness in Mathematics.
- Students demonstrate growth from grade to grade in Mathematics.

The ISASP Mathematics tests place an important emphasis on understanding, discovery, and quantitative thinking in Mathematics. As a result, these summative assessments provide educators, parents, and students with meaningful information that reflects each student's ability to meet challenging Mathematics content standards. They measure student achievement, college and career readiness, and growth as reflected in the grade-level scope and sequence defined by the Iowa Core Standards.

Domain Coverage of the Iowa Core in Mathematics

Tables 2.4 and 2.5 provide the coverage targets in grades 3–8 and 9–11, respectively, in the Iowa Core of the domains that are assessed and reported for the Mathematics assessments of the ISASP.

Table 2.4. Iowa Core Domain Coverage in Mathematics – Grades 3–8

Iowa Core	Grade					
Math Domain	**3**	**4**	**5**	**6**	**7**	**8**
Operations and Algebraic Thinking	31–37%	16–22%	13–18%			
Number and Operations in Base Ten	14–20%	19–24%	23–28%			
Number and Operations – Fractions	11–14%	22–27%	23–28%			
Measurement and Data	23–29%	19–24%	18–23%			
Geometry	11–14%	11–16%	13–18%	12–17%	18–22%	19–23%
Ratios and Proportional Relationships				14–19%	18–22%	
The Number System				21–26%	20–24%	9–13%
Expressions and Equations				29–33%	22–27%	28–32%
Statistics and Probability				12–17%	11–16%	15–19%
Functions						19–23%

Table 2.5. Iowa Core Domain Coverage in Mathematics – Grades 9–11

Iowa Core	Grade		
Math Domain	**9**	**10**	**11**
Geometry	11–17%	26–31%	17–23%
Statistics and Probability	11–17%	11–17%	11–17%
Functions	17–23%	14–20%	20–26%
Algebra	29–34%	17–23%	20–26%
Numbers and Quantity	17–23%	17–23%	17–23%

In addition to the information included in this manual, detailed tables of test specifications are posted on the ISASP portal to allow educators to obtain a thorough understanding of the domain coverage of each grade-level assessment in Mathematics with respect to the Iowa Core Standards. Detailed test specifications for the ISASP Mathematics assessments for grades 3–11 can be found at http://iowa.pearsonaccessnext.com/test-prep/.

Science Test Specifications

The ISASP Science assessments have been designed and developed to support the following claims with respect to student performance:

- Students demonstrate their understanding of the Iowa Core Standards in Science.
- Students demonstrate progress toward college and career readiness in Science.
- Students demonstrate growth across grade bands (3–5, 6–8, and high school) in Science.

Science standards cut across three dimensions as outlined by the Iowa Core Standards in Science–Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts. ISASP items are developed to align to at least two and up to three of these grade-level dimensions. Taken as a whole, the items comprising the Science test forms provide coverage of all three dimensions. They measure student achievement, college and career readiness, and growth in Science as reflected in the grade-level scope and sequence defined by the Iowa Core Standards.

The ISASP Science tests are administered in grades 5, 8, and 10. The following guidelines are used when preparing the science test specifications each year:

- For grade 5, a sample of standards are identified from the 3rd, 4th, and 5th grade Iowa Core science standards.
- For grade 8, a sample of standards are identified from the 6th, 7th, and 8th grade Iowa Core science standards.
- For grade 10, a sample of standards are identified from the high school Iowa Core science standards based on their appropriateness for 9th or 10th grade students.

From these identified standards, Iowa educators develop phenomenon-based science experiments, scenarios, and/or descriptions, along with the associated test questions. Each test item is aligned to the three-dimensional standards corresponding to the performance expectation within the identified standard. Items

selected for each operational test fulfill specific measurement criteria, represent grade-level appropriate standards, and comprise the science domains outlined in the Iowa Core science standards. A balance of item types, standards, and domains appear on the grades 5, 8, and 10 science tests each year.

Domain Coverage of the Iowa Core in Science

Table 2.6 provides the coverage targets in the Iowa Core of the domains that are assessed and reported for the Science assessments of the ISASP in grades 5, 8, and 10. For each Iowa Core domain, the content-related claim referenced in the previous section is made based on student performance.

Table 2.6. Iowa Core Domain Coverage in Science by Grade

Iowa Core	Grade		
Science Domain	**5**	**8**	**10**
Life Science	30–38%	31–39%	31–39%
Physical Science	36–44%	34–42%	31–39%
Earth and Space Sciences	25–33%	22–30%	25–33%

As in other content areas, detailed tables of test specifications are posted on the ISASP portal to allow educators to obtain a thorough understanding of the domain coverage of each grade-level assessment in Science. This information can be found at http://iowa.pearsonaccessnext.com/test-prep/

Item Types

Measuring the depth and breadth of the current Iowa Core Standards requires a balanced and layered approach that incorporates a range of tasks and stimulus materials. Included in this approach is careful consideration of the item type being utilized for the skill or ability being assessed. For example, selected-response items are excellent for efficiently evaluating student knowledge and understanding of a variety of concepts and content included within the Iowa Core. However, additional assessment formats are needed to measure those skills that are not easily assessed by this traditional format. The use of multiple item types expand and improve the measurement of student understanding and proficiency overall. The rigor of the current Iowa Core is mirrored in the ISASP assessments by employing a robust suite of traditional and nontraditional item types, including the following:

Short constructed- and extended constructed-response items: These items challenge students to draw upon higher-order thinking and cognitive processes to generate their own responses. Short constructed-response items may require the student to solve a multistep problem in the Mathematics assessment, write an objective summary of a reading passage, or make and support with evidence a claim based on the results of a scientific experiment. Extended constructed-response items require the student to draft a well-developed essay in response to a writing prompt. Short constructed-response items are designed to be answered in 5–7 minutes of testing time; extended constructed-response items are designed to be answered in about 30 minutes of testing time.

Technology-enhanced items (TEIs): This online item type requires students to engage in tasks designed to use complex thought processes. These items take advantage of the latest computer-based technologies. They

may include response interface features such as hot spots, drag-and-drop, point-and-click, cloze, and graphing. TEIs in the Mathematics assessments can use additional online tools such as equation editors. All TEIs are machine scored. Some specific examples of technology-enhanced items types are included in Table 2.7.

Table 2.7. Examples of Technology-Enhanced Item Types

Item Type	Description
Drop-down Item	This item type allows students to make a selection from a drop-down menu.
Fill-in Item	This item type allows students to type in a text-based response using a keyboard (virtual or physical).
Open-ended Item	This item type allows students to type in an extended text-based response using a keyboard (virtual or physical).
Order Item	This item type allows students to order options into a sequence.
Hot Spot Item	This item type allows students to select one or more regions on a graphic or image to identify their choice.
Graphing Item	This item type allows students to manipulate, create, or edit line graphs, scatterplots, function graphs, pie charts, and bar graphs.
Equation Editor	This item type allows students to create equations (including those with symbols, fractions, etc.) that can be machine scored.

Selected-response items: These items are efficient to administer and offer strong technical properties. These items can be written to address varying levels of cognitive complexity to measure students' skills and knowledge. All selected-response items in grades 3–11 have four options, except for items in Mathematics in grades 9–11, which have five options.

Cognitive Complexity

The depth of knowledge (DOK) required to answer items on the ISASP assessments should be consistent with what is required by the Iowa Core Standards for the standard being assessed. To ensure this consistency, all items are reviewed for cognitive demand to ensure that what students are expected to know and do is consistent between the ISASP and the Iowa Core. Additionally, during the development process itself, the item-level DOKs are written to meet or exceed the DOK levels specified for each standard in the Iowa Core. Table 2.8 describes these levels.

The result is an assessment with a full range of item complexity. Tables 2.9 to 2.11 provide the percentage of items at each of the three DOK levels by grade by test area.

Table 2.8. ISASP Cognitive Level Descriptions

Cognitive Level	Description
Essential Competencies (DOK 1)	This level of complexity involves recalling information such as facts, definitions, terms, or simple one-step procedures.
Conceptual Understanding (DOK 2)	This level of complexity requires engaging in some cognitive processing beyond recalling or reproducing a response. A conceptual understanding item requires students to make some decisions as to how to approach the problem or activity and may require them to employ more than a single step.
Extended Reasoning (DOK 3)	This level of complexity requires problem-solving, planning, and/or using evidence. These items require students to develop a strategy to connect and relate ideas to solve the problem, and the problem may require that the student use multiple steps and draw upon a variety of skills.

Table 2.9. Percentage of ELA Items by DOK Level

	Grade								
Reading	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
DOK 1	15–30%	15–30%	15–25%	15–25%	15–25%	15–25%	10–25%	10–25%	10–25%
DOK 2	40–55%	40–55%	40–55%	40–55%	40–55%	40–55%	40–55%	40–55%	40–55%
DOK 3	30–40%	30–45%	30–45%	30–45%	30–45%	30–45%	33–50%	33–50%	33–50%
Language/ Writing									
DOK 1	5–15%	5–15%	5–15%	5–15%	5–15%	5–15%	5–15%	5–15%	5–15%
DOK 2	20–35%	20–35%	20–35%	20–35%	20–35%	20–35%	20–35%	20–35%	20–35%
DOK 3	60–70%	60–70%	60–70%	60–70%	60–70%	60–70%	60–70%	60–70%	60–70%

Table 2.10. Percentage of Mathematics Items by DOK Level

	Grade				
Mathematics	**3**	**4**	**5**	**6–8**	**9–11**
DOK 1	20–35%	20–35%	20–35%	20–35%	20–35%
DOK 2	45–65%	45–65%	45–65%	45–65%	45–65%
DOK 3	10–25%	10–25%	10–25%	10–25%	10–25%

Table 2.11. Percentage of Science Items by DOK Level

	Grade		
Science	**5**	**8**	**10**
DOK 1	0–10%	0–10%	0–10%
DOK 2	60–75%	50–65%	50–65%
DOK 3	25–40%	30–45%	40–55%

Statistical Specifications

To ensure support for claims that make inferences about student achievement, readiness, and growth, item-level statistics based on both classical and item-response theory (IRT) obtained through field testing are used to assemble test forms. For classical statistics the selection of items is limited to those that have p-values within an acceptable range (.20 to .90) and discrimination indices (point-biserial correlations) greater than

.20. For IRT statistics of item characteristics, the selection of items is based on parameter estimates from the 2-parameter logistic (2PL) model (see Chapter 6 of this manual) where the discrimination estimate exceeds .4 and the difficulty estimate is between -3.0 and 3.0. Items that fail to meet these specifications after field testing may be revised and field tested again or eliminated from the item pool.

Development of Items and Testing Materials

Sound item development is critical for providing quality and consistency across forms of the ISASP assessments. Items and stimulus/item sets (reading passages, graphs, maps, tables, etc. that support a group of items) are created according to the test specifications. The content domains, number of items per domain, cognitive levels, and item types are defined in the test specifications and serve as a basis for item writing. The initial development of items and related testing materials is the first critical step in an extensive, iterative process of drafting, rewriting, editing, aligning, and reviewing items. Only at the end of this extensive process are items considered eligible for inclusion on an ISASP form.

Item writers for the ISASP program are educators who are knowledgeable about the Iowa Core Standards and about Iowa students. ITP works with Iowa educators to identify, select, and contract with individuals for item writing assignments. Hundreds of Iowa educators have contributed to the item writing process for ISASP. These individuals are representative of the state teacher population and have extensive experience with students who are representative of Iowa's student test-taking population in terms of geographic region, demographics, and district size.

ITP content specialists periodically convene item writing workshops and train educators on sound item writing practices. Specific guides for writing test materials for each ISASP content area summarize general item writing principles and provide support resources for item writers. Educators are assigned to write items in the content areas and grade levels that best align with the Iowa Core Standards consistent with their expertise and experiences. ITP employs procedures for item writers that protect the security of the assessment materials as well as the confidentiality of the item writing assignments using secure file transfer protocols and nondisclosure agreements.

To assist in the evaluation of open-ended items, writers who are developing such items also draft scoring criteria simultaneously. The scoring criteria are used to evaluate each item's alignment to the Iowa Core Standards as well as understand the cognitive demands required by the item given the rubric by which it will be scored. The complementary process for writers of selected-response items is that the item writer is expected to provide substantive rationales for the keyed response as well as distractors. Reviewers use these rationales in alignment and other validation activities during the item development process.

Item production goals ensure an "overage" of items across assessment areas so that the pool of available items for each ISASP assessment contains far more material than is needed to build each form. This overage allows content experts to discard those items that do not survive internal and external item review or post field test data review.

Review Processes

After items are written, content specialists review them individually and collectively for issues related to content fidelity, accuracy, fairness, universal design, and alignment to the Iowa Core Standards. The goal of these reviews is to ensure items are accurate, fair, and accessible to all student groups in the diverse population of test takers.

Content Review

Once the items have been reviewed internally, ITP convenes panels of Iowa educators to review the items and associated stimuli (reading passages, tables, graphs, maps, etc.). After a formal training session in the review process, educators evaluate the items for grade level alignment, content relevance, and accuracy. Since these external reviewers have not been involved in the development process up to this point, they provide an objective "cold read" of test materials for potential concerns and unintended interpretations. A main goal of the content review is to confirm that the items are aligned at the appropriate grade level, content standard, and cognitive level. ITP development staff processes the information obtained for each item and determines whether further editorial work is needed. This review focuses on any edits made to the items throughout the process and again checks for content accuracy, fairness, and universal design.

Fairness Review

For review purposes, the term "fairness" can be defined as the extent to which test scores are valid for different groups of test takers. Fairness does not require that all groups have the same average scores. Fairness requires any existing differences in scores to be construct-relevant and therefore valid. An item would be unfair if the source of the difficulty were not a valid aspect of the item. For example, an item would be unfair if a test item contains unnecessarily complex language that acts as a more significant barrier for students who are not native speakers of English than for students who are native English speakers. However, an item could be considered fair if the group difference in difficulty reflected real and relevant differences in the groups' levels of mastery of the Iowa Core. The fairness review process, as well as differential item functioning information, is intended to identify aspects of items that, based on a reviewer's judgment, might produce differences in performance.

Careful consideration of the issues related to fairness is required at each step of the test development process for the ISASP. Table 2.12 (Welch & Dunbar, 2022) summarizes the various steps at which fairness issues are addressed in the development of ISASP assessments.

Table 2.12. ISASP Fairness Procedures in Test Development

Test Development Stage	Considerations for Valid and Fair Interpretations
Articulation of test purpose and constructs to be measured	Delineation of the construct to be measured Review of the curricular standards for fairness or accessibility issues
Defining intended test taker population	Delineation of the test taker population with respect to characteristics such as age, geographic location, grade level, ethnicity, and socioeconomic status
Test specifications	Inclusion of educators who are representative of the test taker population for which the achievement test is designed
Item development	Inclusion of item writers who represent the groups of test takers for which inferences will be made Item writers should reflect widely diverse focuses. Deliberate inclusion of item writers who contribute diversity to the item-writer pool will help ensure the ability to reflect diversity in developed test materials. Inclusion of educators who have experience teaching the core content that the achievement test is designed to measure
Alignment	Inclusion of aligners who represent the groups of test takers for which inferences will be made

	Deliberate inclusion of alignment experts who contribute diversity to the overall pool Inclusion of educators who have experience teaching the core content that the achievement test is designed to measure to a diverse and representative sample of test takers
Item review	Inclusion of reviewers who represent the groups of test takers for which inferences will be made Judgmental review to identify potential fairness problems Reviewers recruited from a variety of sources Reviewers provided guidance and training about what to look for when reviewing items Reviewers made aware of potential fairness issues with respect to cultural stereotyping, irrelevant characteristics of an item, sensitive topics, and offensive language Reviewers made aware of the principles of universal design
Pilot testing	Test takers who are representative of the total test taking population All delivery modes pilot tested on a representative sample All item formats pilot tested on a representative sample
Field testing	Proportional representativeness of the test taker population for which the achievement test is designed
Generation of item-level and test-level statistics	Disaggregated item-level statistics to allow for comparison of performance across groups of interest (differential item functioning (DIF), reliability estimates, precision estimates, relationships between domains)
Assembly of forms/pools	Balance of forms with respect to content using items that have successfully cleared previous steps in the test development process
Review of forms/pools	Inclusion of reviewers who represent the groups of test takers for which inferences will be made Judgmental review to identify potential fairness problems Reviewers recruited from a variety of sources Reviewers provided guidance and training about what to look for when reviewing items Reviewers made aware of potential fairness issues with respect to cultural stereotyping, irrelevant characteristics of an item, sensitive topics, and offensive language Reviewers made aware of the principles of universal design
Linking, equating, and scaling	Special studies designed to collect evidence for any post-administration adjustments or links should be designed to select samples that are representative of the total test taker population. Samples should be large enough to represent the diverse characteristics of the test taking population.

Reviewers follow guiding principles of fairness as they consider each item, including suggested revisions to avoid construct-irrelevant variance and to allow all students the same opportunity to show what they know. Specifically, to make items accessible to all groups of students, reviewers are asked to consider whether items contain the following:

Unnecessarily difficult language. It is best practice to keep testing language simple and direct. The test should use accessible language. While the use of accessible language is particularly important for test takers who have limited English skills, it is beneficial for all test takers when linguistic competence is not relevant to the construct the test intends to measure.

Unfamiliar language/vocabulary. The test should use language that is common. Items should avoid words or phrases that are associated with a particular social class.

Regionalisms. Test language should not require knowledge of words, phrases, or concepts more likely to be known in some regions of the United States than in others, unless the words, phrases, or concepts are important for valid measurement. It is best practice to use words and phrases that are understood across regions.

Jargon. Items should not contain specialized language used by particular groups that is difficult for others to understand. Test language should avoid technical terms relating to finance, politics, specific professions, cultures, or regions.

Emotional topics. Test content that is unnecessarily controversial, offensive, or upsetting should be avoided when possible. It is best practice to avoid topics that may evoke feelings of discomfort, fear, sadness, or anxiety in test takers.

Stereotypes. Test content should be respectful of all people in all groups of the population. Stereotypes attempt to classify or group people based on a single aspect, such as age, race, ethnicity, religion, income level, geographic region, or gender. Some stereotypes are blatant and easy to eliminate, while others are less obvious and require careful reading of the material and attentiveness to cultural sensitivity. Fairness and sensitivity are not limited to specific groups commonly mentioned in fairness discussions. It is important to avoid biased language and stereotypes for any group.

After receiving training in the principles outlined above, a fairness committee evaluates each item and stimulus through a formal review before the items are field tested (an additional fairness review also occurs after forms are constructed). Committee members are educators who represent potentially affected groups relevant to test score interpretation and use, including those based on race, ethnicity, gender, socioeconomic status, and English language learners. After the review is complete, items may be revised based on the feedback received, or they may be removed from the potential item pool. After items are field tested, fairness is examined further through DIF analyses. Additional detail concerning the specifics of review committees and DIF results are presented later in this chapter.

Universal Design Review

The principles of universal design for the ISASP assessments provide guidelines for the test development and review processes to help ensure that no test taker is unduly disadvantaged owing to a special need, incomplete language mastery, or membership in any demographic or educational group. Universal design

was a guiding principle in the creation of the publishing specifications that determine the appearance of the materials as they are experienced by students in both paper-and-pencil and online formats. Aspects of universal design including ease of navigation of test materials; clarity of typeface, graphics, and page layout; and respect for the diversity of the test-taking population in the content of the materials are evaluated during this review process.

External Alignment Review

ITP contracted with the Human Resources Research Organization (HumRRO) to conduct an external alignment study for the ISASP to establish and document evidence of consistency among the test blueprints, items, and the Iowa Core Standards. The alignment study included evaluations of assessments in ELA and Mathematics in grades 3–11, and of assessments in Science in grades 5, 8, and 10. The assessments were evaluated using an approach derived from the methodology established by the Council of Chief State School Officers (CCSSO, 2013). The evaluation of the Science assessments was further informed by criteria outlined by Achieve. The approach convened teachers and content experts to confirm the standards alignment and cognitive complexity levels of items the item writers identified (captured in item metadata in the content management system used by ITP), and to rate all items on several other indicators of item quality.

The alignment study was conducted in two phases. During Phase 1, 34 Iowa educators representing 23 districts and eight regions of the state were convened as a panel for a two-day workshop during which they reviewed test items. Panelists were experienced Iowa educators with expertise in the content area and grade span for which they reviewed items. Panelists were organized into three groups each for ELA and Mathematics (3–5, 6–8, and 9–11; six groups total), and two groups for Science (5/8 and 10).

During this phase of the alignment process, panelists provided independent ratings for items but ultimately reached a consensus rating for each item based on group discussion. Data from Phase 1 were used to edit or replace items prior to the finalization of the 2019 operational test forms for the ISASP. Phase 2 of the study convened a subsample of the Iowa educators who participated in Phase 1 along with one nationally recognized subject-matter expert for each content area. During Phase 2, revised and replacement items were rated using the same process implemented in Phase 1.

This study provided substantial evidence to support the content validity of the 2019 and 2021 ISASP assessments in ELA, Mathematics, and Science. Across the grade/subject tests, a large majority of items were rated as measuring content outlined in the Iowa Core Standards. With a small number of exceptions, the number of aligned items fell within the ranges of items specified in the test blueprints. Finally, the majority of items were determined to have been written at a level of cognitive complexity that is at or above the range specified for the aligned content standard (ELA and Mathematics) in the Iowa Core, or that test forms contain an appropriate distribution of cognitive complexity at the item level (Science).

Field Testing

Once items have passed through the review processes described above, ITP collects data on the performance of the items by conducting a field test to determine how well the items perform in an actual testing situation. In the ISASP program, field-test items are embedded in multiple forms of operational tests through a random assignment of items to students across the state during a live ISASP administration. This approach leverages a large representative sample of students throughout the state to constitute equivalent groups of test takers who respond to the field-test items. Adequate numbers of field-test items are administered annually to replenish the item pool. Responses from the field-test items do not contribute to a student's scores on the operational test. The specific locations of the embedded items within the assessment are not identified for test takers.

Prior to the first operational administration of the ISASP, field-test items were administered to Iowa test takers as part of the previous statewide assessment program (the *Iowa Assessments*) that was used by the IDOE for No Child Left Behind (NCLB) and Every Student Succeeds Act (ESSA)accountability through the spring of 2017. Field test materials were administered as a standalone set of items during the same testing window as the administration of the *Iowa Assessments*. Extensive studies conducted by ITP compared the item-level statistics, both classical and IRT-based, from those items field tested prior to ISASP to those embedded within the ISASP assessments. Item-level statistics were stable across administration conditions.

Data Review

The item data collected during the field test are analyzed. This analysis determines not only whether the items meet statistical specifications but also whether they are appropriate measures of students' knowledge and the extent to which the items will contribute to the test's overall reliability.

Item-Level Statistics Reviewed

Several statistical analyses, based on classical test theory and item response theory, are completed and documented to assist in the review of the field test data in preparation for the assembly of final forms. The classical test theory statistics include the difficulty value for each item, the item-total correlations, item fairness reviews (see next section), differential item functioning indices (see final section in this chapter), distractor analysis, and latency data. The item response theory statistics include the estimation of parameters for each dichotomous item as well as category boundary parameters for polytomous items.

Fairness Review Summaries

Although the items are reviewed extensively for fairness throughout the test development process (as referenced in Table 2.12), a specific review addressing only issues related to fairness occurs in preparation for the assembly of final forms. A committee of reviewers was recruited based on ethnic, racial, and gender diversity, as well as diversity of the student population with which members have experience teaching. Reviewers received training on fairness guidelines described previously, including handouts and a PowerPoint presentation that they could access as they were conducting their reviews.

The post-field-test reviews took place within Pearson's ABBI item banking system. This allowed for

maximum test and item security, as well as allowing reviewers to experience all item types in the testing environment as experienced by test takers. Comments and ratings of the items were securely recorded within the ABBI platform. The reviewers were instructed to use the following categories for ratings:

- Approved: The item is approved as is, with no changes.
- Approved with edits: The item has a small issue but can be approved following edits to fix the issue.
- Rejected: The item has inherent flaws that cannot be fixed. The item should be removed from the item pool.

Tables 2.13 to 2.16 show the item ratings obtained for the 2019 ISASP item pool. The reviewers' ratings for each item were used to determine the final review category for the item. Items with positive ratings were considered eligible for inclusion in the acceptable item pool used during forms assembly.

Table 2.13. Fairness Ratings for Reading

Grade	Accepted	Accepted with Edits	Rejected	Total Reviewed	Percent Accepted
3	28	0	0	28	100%
4	29	0	0	29	100%
5	30	0	0	30	100%
6	31	0	0	31	100%
7	32	0	0	32	100%
8	32	0	0	32	100%
9	28	0	0	28	100%
10	28	0	0	28	100%
11	28	0	0	28	100%
Total	**266**	**0**	**0**	**266**	**100%**

Table 2.14. Fairness Ratings for Language/Writing

Grade	Accepted	Accepted with Edits	Rejected	Total Reviewed	Percent Accepted
3	22	3	0	25	88.0%
4	26	0	0	26	100.0%
5	27	0	0	27	100.0%
6	25	3	0	28	89.3%
7	28	1	0	29	96.6%
8	29	0	0	29	96.6%
9	29	1	0	30	96.7%
10	30	0	0	30	100.0%
11	30	0	0	30	100.0%
Total	**246**	**8**	**0**	**254**	**96.9%**

Table 2.15. Fairness Ratings for Mathematics

Grade	Accepted	Accepted with Edits	Rejected	Total Reviewed	Percent Accepted
3	32	3	0	35	91.4%
4	35	2	0	37	94.6%
5	39	1	0	40	97.5%
6	41	1	0	42	97.6%
7	45	0	0	45	100.0%
8	44	3	0	47	93.6%
9	35	0	0	35	100.0%
10	34	1	0	35	97.1%
11	35	0	0	35	100.0%
Total	**340**	**11**	**0**	**351**	**96.9%**

Table 2.16. Fairness Ratings for Science

Grade	Accepted	Accepted with Edits	Rejected	Total Reviewed	Percent Accepted
5	32	0	0	32	100%
8	32	0	0	32	100%
10	40	0	0	40	100%
Total	**104**	**0**	**0**	**104**	**100%**

Analysis of Differential Item Functioning

DIF analyses were conducted on items as an additional fairness check. DIF analyses identify items that function differently for two groups of examinees with the same total test score. In many cases, one group will be more likely to answer an item correctly on average than another group when the groups are not matched on a covariate related to the construct measured by the test. These differences might be due to construct-relevant contrasts in the levels of knowledge and or skills between the groups. For example, if members of one group tend to take more advanced classes or attend higher-performing schools than members of another group, then the performance of the two groups might differ on some items. DIF analyses attempt to control for these group differences and help identify items that might unfairly favor one group over another when the groups are matched on a relevant covariate. It is important to note that items may show DIF because they are measuring an aspect other than the intended construct, but they may also show DIF because of differences in knowledge or because of false positives in the DIF hypothesis testing procedure. Items that were identified as potentially problematic by DIF methods were then presented for additional review by experts with respect to the relevant focal group.

Specific item-level comparisons of performance were made for gender, race, ethnicity, free and reduced

lunch (FRL) status, Individual Education Program (IEP) status, and English Language Learner (ELL) status. The reference group and focal group comparisons can be found in Table 2.17.

Table 2.17. Comparison Groups for Differential Item Functioning Analysis

Group Type	Reference Group	Focal Group
Gender	Male	Female
Race	White	African American
Ethnicity	Non-Hispanic White	Hispanic
Free Reduced Lunch (FRL)	Not FRL-eligible	FRL-eligible
Individual Education Plans (IEP)	No IEP	IEP
English Language Learners (ELL)	Non-ELL	ELL

Because ISASP includes both dichotomous and polytomous items, two different DIF analysis procedures were used. For dichotomous items, the DIF statistic *MH D-DIF* was calculated (Holland & Thayer, 1988). The *MH D-DIF* statistic expresses the difference between the focal and reference groups after conditioning on the total test score; this difference is reported on the delta scale. To obtain the *MH D-DIF* test statistic, the Mantel-Haenszel estimate of the conditional odds ratio was first calculated. This conditional odds ratio is defined as:

$$\hat{\alpha}_{MH} = \frac{\sum_k N_{R1k} N_{F0k} / N_k}{\sum_k N_{R0k} N_{F1k} / N_k} \qquad (2\text{-}1)$$

where

N_{R1k} = number in reference group at score level *k* who answered the item correctly,
N_{F0k} = number in focal group at score level *k* who answered the item incorrectly,
N_{R0k} = number in reference group at score level *k* who answered the item incorrectly,
N_{F1k} = number in focal group at score level *k* who answered the item correctly, and
N_k = total number in both comparison groups at score level *k*.

This value was then transformed into the appropriate DIF test statistic using the transformation *MH D-DIF* = –2.35 ln ($\hat{\alpha}_{\text{MH}}$).

A positive value of *MH D-DIF* indicates that, conditional on same total score, the item is more difficult for the reference group, whereas a negative value indicates that it is more difficult for the focal group. Based on the magnitude of *MH D-DIF*, items were classified into one of three categories: A, B, or C, based on the criteria in Table 2.18 from Dorans and Holland (1993).

Table 2.18. DIF Classification Categories for Dichotomous Items

DIF Category	Description
A ***(negligible)***	The absolute value of the *MH D-DIF* is not significantly different from zero or is less than 1.0.
B ***(slight to moderate)***	The absolute value of the *MH D-DIF* is significantly different from zero but not from 1.0 and is at least 1.0; OR the absolute value of the *MH D-DIF* is significantly different from 1.0 but is less than 1.5.
C ***(moderate to large)***	The absolute value of the *MH D-DIF* is significantly different from 1.0 and is at least 1.5.

For polytomous items, the standardized mean difference (SMD) procedure (Dorans & Kulick, 1983; Dorans & Kulik, 1986; Zwick & Thayer, 1996) was utilized. Like the *MH D-DIF* approach, the SMD also expresses the difference between the focal and reference groups' performances on an item while conditioning on the total test score. The SMD is defined as:

$$\text{SMD} = \sum_k p_{FK} m_{FK} - \sum_k p_{FK} m_{RK} \qquad (2\text{-}2)$$

where k = score level on the raw score scale of the subtest,
p_{FK} = proportion of the focal group at level k,
m_{FK} = mean item score for the focal group at level k, and
m_{RK} = mean item score for the reference group at level k.

Conditional on total score, a positive value of SMD indicates that the item is more difficult for the focal group and a negative value indicates that it is more difficult for the reference group. The SMD was then divided by the standard deviation to obtain an effect size. Based on the magnitude of the effect size, items were classified into one of three categories: A, B, or C, as described in Table 2.19.

Table 2.19. DIF Classification Categories for Polytomous Items

DIF Category	Description
A ***(negligible)***	The chi-square is not significant ($p \geq .05$), or the absolute value of the effect size is less than or equal to .17.
B ***(slight to moderate)***	The chi-square is significant, and the absolute value of the effect size is over .17 and less than or equal to 0.25.
C ***(moderate to large)***	The chi-square is significant, and the absolute value of the effect size is over .25.

Summaries of the DIF results for the ISASP Reading, Language/Writing, Mathematics, and Science tests can be found in the *ISASP ASR-2019* and *ISASP ASR 2021*. The summary contains the total number of items on the test, the total number of items with C-DIF, and the number of items with C-DIF for each group comparison. The overall percentages of items flagged were small. Content development staff reviewed each flagged item for evidence of potential bias that might have produced the DIF results. Flagged items were also reviewed by experts for each of the comparisons, such as experts on ELL. All flagged items were cleared subsequently through the review process. If an item had been identified as measuring something other than the intended construct, then the item would have been removed from the operational form. This process served as an additional statistical check against the results of the fairness and sensitivity reviews.

Forms Assembly

Items that ITP has determined are available to appear on operational test forms become part of a pool of items that are eligible for selection during forms construction. To ensure the final subject area test has adequate content coverage while at the same time being meaningful to students of varying achievement levels, the items within a typical subject area's item pool are chosen to be diverse regarding skill alignment, cognitive level alignment, and difficulty. Items are then pulled from the item pool into test forms. During this process, careful attention is paid to item selection so that the final tests follow the predetermined test specifications and meet psychometric targets for difficulty and discrimination. Additional information on forms construction is provided in Chapter 6 with respect to the concept of pre-equating.

Chapter 3: Test Administration and Accommodations

Students from all Iowa public and state-accredited schools in the specified grade levels must participate in the Iowa Statewide Assessment of Student Progress (ISASP). Students from all Iowa independently accredited and non-accredited nonpublic schools in the specified grade levels may opt to participate in the ISASP.

Statewide assessments, such as the ISASP, are annual, summative measures of student achievement that are used to evaluate student learning and skills. The ISASP is one approach for measuring how Iowa students are performing on the Iowa Core. Although the ISASP is just one measure of a student's achievement, participation from all students is important to understand and interpret the results.

Information from the ISASP is used in several ways. The State of Iowa uses the aggregated test scores to report to the public and to the US Department of Education how Iowa students are performing on the Iowa Core. Educators and policymakers use information from the ISASP to make decisions about resources and support to be provided. Parents use this information to make decisions about how best to prepare their students. School performance results from the ISASP are less interpretable if students do not participate in the assessment.

Eligibility for Assessments

Mathematics: Grades 3–11

General education students and students in special populations—(e.g., English Language Learners (ELLs) and students with disabilities (SWDs able to do so)—take the Mathematics ISASP to fulfill their Mathematics requirement.

English Language Arts (ELA): Grades 3–11

General education students—and SWDs able to do so—take the ISASP Reading and the ISASP Language and Writing tests to fulfill their ELA requirement.

Science: Grades 5, 8, and 10

General education students—and SWDs able to do so—in grades 5, 8, and 10 take the Science ISASP to fulfill their Science requirement.

Instructions to educators for ordering tests, assigning students to online test sessions, and navigating the online testing platform (TestNav) are presented in the *Test Administrator Manual*, updated periodically and accessed at http://iowa.pearsonaccessnext.com/manuals/.

Administration to Students

Mathematics

The grades 3–11 Mathematics ISASP are available for either online or paper administration, with Braille and large-print forms available for students requiring an accommodated form. For grades 3–11, the test is intended to be administered in one session (60 minutes recommended) on one day. The test may be administered on the same day as other subject areas of the ISASP, or on its own day, at district (Local Education Agency, or LEA) discretion.

The grades 3–11 2019 Mathematics ISASP online and paper accommodated versions were administered any time within the 12-week testing window (March 4–May 31). The testing window in 2021 was March 15 through May 28.

English Language Arts (ELA) -- Reading and Language/Writing

The grades 3–11 Reading ISASP are available for either online or paper administration, with Braille and large-print forms available for students requiring an accommodated form. For grades 3–11, the test is intended to be administered in one session (60 minutes recommended) on one day. It was also recommended that the Language/Writing test be given on the same day. The test may be administered on the same day as other subject areas of the ISASP, or on its own day, at LEA discretion. The grades 3–11 Reading ISASP online and paper accommodated versions were administered during the same testing windows as the 2019 and 2021 Mathematics assessments.

The grades 3–11 Language/Writing ISASP are available for either online or paper administration, with Braille and large-print forms available for students requiring an accommodated form. For grades 3–11, the test is intended to be administered in one session (120 minutes recommended) on one day. It was also recommended that the Reading test be given on the same day. The test may be administered on the same day as other subject areas of the ISASP, or on its own day, at LEA discretion. The grades 3–11 Language/Writing ISASP online and paper accommodated versions were administered during the same testing windows as the 2019 and 2021 Mathematics assessments.

Science

The grades 5, 8, and 10 Science ISASP are available for either online or paper administration, with Braille and large-print forms available for students requiring an accommodated form. For grades 5, 8, and 10, the test is intended to be administered in one session (60 minutes recommended) on one day. The test may be administered on the same day as other subject areas of the ISASP, or on its own day, at LEA discretion.

The grades 5, 8, and 10 Science ISASP online and paper accommodated versions were administered during the same testing windows as the 2019 and 2021 Mathematics assessments.

Secure Testing Materials

The recovery of testing materials after each administration is critical. All secure materials, including test booklets, must be returned to preserve the security and confidential integrity of items that will be used on future tests.

Iowa Testing Programs (ITP) directed its testing contractor to assign secure test booklets to school districts by unique barcoded security numbers. School districts completed packing lists to assist the testing contractor in determining whether secure materials are missing. The testing contractor scanned incoming barcodes to determine whether all secure materials have been returned from each school and district. School districts are responsible for ensuring the confidentiality of all testing materials and their secure return. ITP and the testing contractor contacted any district with unreturned secure materials.

The Iowa Department of Education (IDOE) has established a comprehensive policy and practice regarding test security for all statewide assessments (including ISASP) which govern the conduct of persons involved with test administration. These security procedures are documented in the *State of Iowa Test Security Manual*, updated periodically and accessed at: https://educateiowa.gov/pk-12/student-assessment-pk-12/

Iowa Statewide Assessment of Student Progress Security

Mathematics

The grades 3–11 Mathematics ISASP are delivered online, with paper, Spanish, Braille, or large-print forms available for students requiring one of those forms. For the computer-delivered assessments, there are no secure materials to return, but districts are asked to collect and securely dispose of student testing tickets and any scratch paper used. For students taking paper-based accommodated forms, secure materials include large-print test books and Braille test books. All used and unused test books, answer folders, and accommodated materials must be returned to ITP's testing contractor.

Reading

The grades 3–11 Reading ISASP are delivered online, with paper, Braille, or large-print forms available for students requiring one of those forms. For the computer-delivered assessments, there are no secure materials to return, but districts are asked to collect and securely dispose of student testing tickets and any scratch paper used. For students taking paper-based accommodated forms, secure materials include large-print test books and answer books and Braille test books. All used and unused test books, answer folders, and accommodated materials must be returned to ITP's testing contractor.

Language/Writing

The grades 3–11 Language/Writing ISASP are delivered online, with paper, Braille, or large-print forms available for students requiring one of those forms. For the computer-delivered assessments, there are no secure materials to return, but districts are asked to collect and securely dispose of student testing tickets and any scratch paper used. For students taking paper-based accommodated forms, secure materials include large-print test books and answer books and Braille test books. All used and unused test books, answer folders, and accommodated materials must be returned to ITP's testing contractor.

Science

The grades 5, 8, and 10 Science ISASP are delivered online, with paper, Spanish, Braille, or large-print forms available for students requiring one of those forms. For the computer-delivered assessments, there are no secure materials to return, but districts are asked to collect and securely dispose of student testing tickets and any scratch paper used. For students taking paper-based accommodated forms, secure materials include large-print test books and answer books and Braille test books. All used and unused test books, answer folders, and accommodated materials must be returned to ITP's testing contractor.

Districts were instructed to return **all** materials—used and unused—to ITP's testing contractor.

Features and Accommodations

Some students use features or accommodations in order to fully demonstrate their knowledge and skills on statewide tests. Such features and accommodations allow students to participate in the testing program without being disadvantaged by a disability or lack of English language proficiency. The available features and accommodations are documented in Sections 1 and 2 of the *Iowa Statewide Assessment of Student Progress (ISASP) Accessibility and Accommodations Manual,* which is updated annually and available on the ISASP portal.

Universal features allow all students to tailor aspects of the testing experience to their needs or preferences. Features include accessibility tools available in online assessments and general test-taking practices. The use of a universal feature may remove the need for an accommodation, depending on the student's disability. The choice to use a universal feature is made at the student level.

Designated features are available to any student when indicated in advance and assigned by a teacher but do not require the student to have an Individualized Education Program (IEP) or 504 plan. They allow students to further tailor aspects of the testing experience to their needs or preferences. The use of a designated feature may remove the need for an accommodation, depending on the student's disability. The decision to make designated features available to students is made at the school and district level.

Accommodations are changes in the way that a test is administered that reduce or eliminate the effects of a disability. Accommodations are only available to students with an IEP or 504 plan. Districts are responsible for ensuring that accommodations do not compromise test security, difficulty, reliability, or validity and are consistent with a student's IEP or 504 plan. All needed accommodations must be documented annually in the IEP or 504 plan prior to testing.

Students who are identified as ELLs may use linguistic features, such as Spanish-language versions of the Mathematics and Science tests. A limited number of accommodations may also be considered linguistic features for students who are ELLs.

The decision to use a particular support or accommodation with a student should be made on an individual basis. This decision should take into consideration the needs of the student as well as whether the student routinely receives the accommodation during classroom instruction. Not every support or accommodation is appropriate or permitted for every subject area.

IDOE has established a comprehensive policy and practice regarding test accessibility and accommodations for all statewide assessments (including ISASP) which govern the assignment and usage of supports and accommodations. These procedures are documented in the *Iowa Statewide Assessment System Accessibility Manual,* updated periodically and accessed at: https://educateiowa.gov/pk-12/student-assessment-pk-12.

Research Base for Features and Accommodations

Abedi and Ewers (2013) provide a compilation of expert judgments on key issues related to the use of accommodations for students with disabilities and/or English language learners. This research was used to organize and define the features and accommodations available for the ISASP tests. A summary of features and accommodations is provided in Table 3.1.

Table 3.1. Support and Accommodations for ISASP

Support/Accommodation	Research and Recommendations
Assistive technology **The category of assistive technology includes devices that range from very commonplace supports to sophisticated technologies.** **Supports available to all students include materials commonly used during instruction such as pencil grips, place markers, line guides, color and masking overlays, highlighters, low-vision aids (e.g., magnifiers, large monitor screen sizes), whisper phones, and audio amplification devices. Many of these supports are provided as tools in the online testing interface.** **Assistive technologies identified as accommodations for SWDs include talking calculators and devices such as computer tablets that serve as calculators or for note-taking. Generally, internet access must be disabled and students' computer use must be monitored. This accommodation generally requires an individual or small group test administration.**	**According to Blaskey, Scheiman, Parisi, Ciner, Gallaway, and Selznick (1990); Cormier, Altman, Shyyan, and Thurlow (2010); Iovino, Fletcher, Breitmeyer, and Foorman (1996); Johnson, Kimball, Brown, and Anderson (2001b); Robinson and Conway (1990), Salend (2009); and Scarpati, Wells, Lewis, and Jirka (2011):** **Although most assistive technologies have not undergone experimental research, there is no evidence these accommodations unfairly advantage students. In addition, official studies confirm that the use of assistive technologies either greatly benefits or has little to no negative impact on students. Therefore, their use is supported.** **In the case of audio amplification and magnifying equipment, all students benefit.** **Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.**

Support/Accommodation	Research and Recommendations
American Sign Language and signed English interpretation *Test content* **IEP teams may indicate sign language interpretation of the Mathematics and Science scripts (see human read-aloud) for students who are deaf or hard-of-hearing. Interpreters may access the script up to 48 hours prior to test administration and are required to review it in order to prevent cueing test answers.** *Test directions* **Sign language interpretation of the scripted test monitor and student directions may be provided to students who are deaf or hard-of- hearing.**	**According to Johnson, Kimball, and Brown (2001a) and Russell, Kavanaugh, Masters, Higgins, and Hoffmann (2009):** **Research calls into question the capabilities and qualifications of on-site sign language interpreters, especially when interpreters are unfamiliar with the tested subject and its technical terms; the inability for interpreters to gain access to and prepare for the assessment prior to testing further complicates the issue.** **According to Russell et al. (2009):** **The obstacles and limitations presented by televised recordings of asigned test may be overcome by computer programs.** **According to Johnson et al. (2001a):** **It is difficult to assess [whether] students gain an unfair advantage, as the signing of a test is "an accommodation of an accommodation."** **According to Ray (1982), Sullivan (1982), and Thurlow and Bolt (2001):** **Experts agree that sign language interpretation of test directions, which is used in most states, levels the playing field for deaf and hearing-impaired students. Signed test directions give these students the same opportunity to participate in and score as well on the assessments as general education students.** **Research supports the effectiveness of the accommodation and recommends its use, although there are concerns about its implementation. The risk of the accommodation giving students an unfair advantage is low.**
Audio presentation of Mathematics and Science assessments *Text-to-speech* **Iowa provides two types of text-to-speech support for online Math and Science assessments. Text-to-speech and other read-aloud methods are allowed for grades 6–11 Reading assessments, but not grades 3–5.** **General text-to-speech is available to all students who choose to use it. Only text in the stem and answer options is read aloud. Tables, graphs, labels, etc. generally are not read, but exceptions are made if they contain a relatively large amount of text.** **Accommodated text to speech is available as an accommodation for SWDs and as a linguistic support for ELLs. All text in stems, answer options, tables, charts, graphs, labels, etc. are read aloud and positional descriptions are provided, if appropriate.**	**According to: Acosta, Rivera, and Shafer-Willner (2008); Barton (2002); Bolt and Thurlow (2004); Brown (2007); Burch (2002); Castellon-Wellington (2000); Calhoon, Fuchs, and Hamlett (2000); Christensen, Braam, Scullin, and Thurlow (2011); Cormier et al. (2010); Dolan, Hall, Banerjee, Chun, and Strangman (2005); Elbaum (2007); Fuchs, Fuchs, Eaton, Hamlett, and Karns (2000); Helwig, Rozek-Tedesco, and Tindal (2002); Johnson et al. (2001b); Kopriva, Emick, Hipolito-Delgato, and Cameron (2007); Pennock-Roman and Rivera (2011); Pennock-Roman and Rivera (2012); Sato, Rabinowitz, Worth, Gallagher, Lagunoff, and McKeag (2007); Tindal, Heath, Hollenbeck, Almond, and Harniss (1998); and Wolf, Kim, Kao, and Rivera (2009):** **Collective research provides varied conclusions as to the effectiveness of this accommodation. Although results vary across grades, subjects, disability type, and level of proficiency in a subject or skill, the overall consensus confirms SWDs benefit from this accommodation.**

Support/Accommodation	Research and Recommendations
Human read-aloud **Mathematics and science texts are available as a read-aloud accommodation for SWDs and as a linguistic support for ELLs. All text in stems, answer options, tables, charts, graphs, labels, etc. are read aloud and positional descriptions are provided, if appropriate.**	**According to Wolf et al. (2009):** **On a math test, [ELLs] who are unfamiliar with read-aloud on assessments do not benefit, but [ELLs] familiar with read-aloud support on assessments greatly benefit.** **Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.**
Braille **IEP teams may select Braille test booklets for students who are blind or partially sighted and are competent Braille readers. As of 2016–17, Braille materials are provided in Unified English Braille format.**	**According to Bennett, Rock, and Kaplan (1987); Bennett, Rock, and Novatkoski (1989); Bolt and Thurlow (2004); Coleman (1990); Thurlow and Bolt (2001); and Thurlow, House, Boys, Scott, and Ysseldyke (2000):** **Although Braille tests require more time to complete and may make certain types of test questions more difficult, research recommends the use of the accommodation. Most, but not all, states use Braille tests.** **Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.**
Extended testing time **Iowa's accountability tests are sectioned and untimed. Testing may be split over multiple days with one or more sections completed on a given day. Taking a single test section over multiple days or sessions is allowable as an accommodation for SWDs and as an indirect linguistic support for ELLs.**	**According to Crawford and Tindal (2004); DiCerbo, Stanley, Roberts, and Blanchard (2001); Fletcher et al. (2009); Thurlow and Bolt (2001); and Walz, Albus, Thompson, and Thurlow (2000):** **Research is divided on whether extending testing time over multiple days is effective. Some studies revealed that SWDs in lower grades and students with low level reading abilities benefited. In other studies, SWDs benefited little or did not benefit at all and general education students benefited. Experts recommend the accommodation, which is used in most states, be used thoughtfully and carefully and only when absolutely needed.** **Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.**
Handheld calculator for Mathematics **Iowa's online math tests have built-in calculators. SWDs who need to use a handheld calculator test using paper materials.**	**According to Bouck and Bouck (2008); Fuchs et al. (2000); Russell (2006); and Shaftel, Belton-Kocher, Glasnapp, and Poggio (2006):** **Calculators, which often are automatically included for math tests, are widely used by all students. Although research is divided on whether the accommodation provides a significant benefit to students, the use of the accommodation is strongly supported.** **Research supports the effectiveness of the accommodation and recommends its use. There is no risk the accommodation gives students an unfair advantage.**
Large-print test book **IEP or 504 plan teams may select large-print test booklets for students with low vision or for SWDs who need to take a paper test and a standard font test booklet is not available.**	**According to Beattie, Grise, and Algozzine (1983); Bennett, Rock, and Jirele (1987); Brown (2007); Burk (1998); Grise, Beattie, and Algozzine (1982); Perez (1980); Thurlow and Bolt (2001); and Wright and Wendler (1994):** **Much of the research concludes that large-print tests, which are used in most states, offer little benefit. However, select studies strongly**

Support/Accommodation	Research and Recommendations
	indicate that students with visual impairments and specific learning disabilities significantly benefit from this accommodation. **Research supports the effectiveness of the accommodation and recommends its use. There is no risk the accommodation gives students an unfair advantage.**
Mathematics manipulatives; abacus **SWDs who use manipulatives or an abacus for Mathematics take the test using paper materials.**	**According to Elliot, Kratochwill, McKevitt, and Malecki (2009):** **Experts are uncertain of the effectiveness and fairness of Mathematics manipulatives but support the accommodation's use.** **Despite uncertainties, research supports the use of the accommodation. The risk of the accommodation giving students an unfair advantage is moderate.**
Scribe **SWDs may dictate to a scribe who enters student responses into an online or paper test form. It is also possible for students to record their responses for later transcription by a scribe.**	**According to Thurlow and Bolt (2001):** **Experts recommend that SWDs, including students who use American Sign Language, submit answers via computer, whenever possible, rather than relay answers to a scribe.** **According to Fuchs et al. (2000), Koretz and Barton (2004), Koretz and Hamilton (2000), MacArthur and Graham (1987), and Tippets and Michaels (1997):** **SWD research is limited, especially regarding the impact a disability has on test taking. A body of research suggests, however, that SWDs benefit from the use of scribes. Certain factors, such as type and difficulty of test and whether other accommodations are in place, should also be considered.** **Research supports the effectiveness of the accommodation and recommends its use. The risk of the accommodation giving students an unfair advantage is low.**

Accommodations Use Monitoring

Iowa uses a data audit system—as well as selected field audits—to monitor the use of accommodations on its assessments. At a state level, data are reviewed for all accommodations for students who are (1) receiving special education or identified as disabled under Section 504 of the *Rehabilitation Act of 1973* and (2) ELLs.

Data Audit

The data collection is intended to provide IDOE with the information about districts' use of accommodations on state assessments. This information allows IDOE to analyze the accommodation data to draw conclusions about the use and overuse of accommodations and will inform future policy decisions and training needs regarding the use of accommodations.

Chapter 4: Reports

After each test administration of the Iowa Statewide Assessment of Student Progress (ISASP), a number of reports are provided. These reports and files contain individual student assessment scores with demographics. Summary reports are also created that provide test results aggregated at school, district, and state levels. The reports focus on two types of scores: scale scores and achievement levels. This chapter provides an overview of the types of scores reported and a brief description of each type of report. Also provided in this chapter are guidelines for proper use of scores and cautions about misuse.

Information about student performance is provided on ISASP Individual Student Reports and summary reports for schools, districts, and the state. This information may be used in a variety of ways. Interpretation guidelines were developed and published as a component of the release of public data and also contained in the Reports Overview Training available for Iowa educators.

Description of Scores

Scale Scores

Scale scores are statistical conversions of raw scores that maintain a consistent metric across test forms and permit comparison of scores across all test administrations within a particular grade and subject. Because scale scores adjust for different form difficulties, they can be used to determine whether a student met the standard or achievement level in a manner that is fair across forms and administrations. Schools can also use scale scores to compare the knowledge and skills of groups of students within a grade and subject across years. These comparisons can be used in assessing the impact of changes or differences in instruction or curriculum. Characteristics of the ISASP Scale Scores are described in Chapter 6.

Achievement Levels

To help parents and schools interpret scale scores, achievement levels are reported. The scale score determines each achievement level, also referred to as performance levels. The range for an achievement level is set during the standard setting process. Cut score ranges for each of the ISASP performance levels by grade in ELA, Mathematics, and Science are provided in Tables 4.1 to 4.3.

Table 4.1. ELA Cut Score Ranges for ISASP Performance Levels

	Performance Levels		
Grade	Not-Yet-Proficient	Proficient	Advanced
3	345 to 397	398 to 446	447 to 510
4	350 to 413	414 to 477	478 to 540
5	355 to 436	437 to 512	513 to 590
6	360 to 455	456 to 540	541 to 640
7	370 to 474	475 to 568	569 to 680
8	385 to 493	494 to 593	594 to 720
9	410 to 504	505 to 617	618 to 750
10	435 to 529	530 to 641	642 to 780
11	460 to 560	561 to 659	660 to 800

Table 4.2. Mathematics Cut Score Ranges for ISASP Performance Levels

	Performance Levels		
Grade	Not-Yet-Proficient	Proficient	Advanced
3	345 to 389	390 to 442	443 to 510
4	350 to 408	409 to 475	476 to 540
5	355 to 428	429 to 502	503 to 590
6	360 to 449	450 to 531	532 to 640
7	370 to 468	469 to 574	575 to 680
8	385 to 489	490 to 605	606 to 720
9	410 to 512	513 to 625	626 to 750
10	435 to 536	537 to 653	654 to 780
11	460 to 558	559 to 674	675 to 800

Table 4.3. Science Cut Score Ranges for ISASP Performance Levels

	Performance Levels		
Grade	Not-Yet-Proficient	Proficient	Advanced
5	355 to 451	452 to 541	542 to 590
8	385 to 507	508 to 608	609 to 720
10	435 to 544	545 to 655	656 to 780

Domain Scores

The domain score is the percent of points earned for items that constitute a specific domain of the Iowa Core. These scores can be interpreted only in reference to the total number of points possible on a subject-area test or within a domain. They cannot be compared across tests or administrations. State of Iowa mean performance by domain is provided as a point of reference for schools to interpret.

Iowa Percentile Ranks

Conversion tables for educators to find the Iowa percentile rank associated with ISASP Scale Scores (ISS) assessments can be found at:
https://iowa.pearsonaccessnext.com/resources/reports/IowaPercentileRanksfor2019.pdf and
https://iowa.pearsonaccess.com/resources/reports/IowaPercentileRanksfor2021.pdf

The conversion tables are organized by grade, and each grade-level table includes columns containing the Iowa Percentile Rank (IPR) of each ISS.

The ISS to IPR conversions were obtained directly from the statewide distributions of ISASP scale scores in the Spring 2019 and 2021 administrations of the ISASP. In each test area and grade, these distributions were obtained from the final scoring file created for the purpose of reporting scores to Iowa schools.

Although the ISASP was developed to align with the Iowa Core Standards and provide standards-based information for students and their parents, the IPRs support other administrative uses of ISASP results.

Description of Reports

PAN is Pearson's secure website used for all test administration preparation, setup, and reporting tasks for ISASP and the location for all operational test results.

Authorized district and school users in the Coordinator user role can sign in to PAN to access and download published PDF score reports, request and order printed score reports from Pearson, and download student data files to upload into the district and school reporting systems. All reports and student data files are in the Published Reports tab in PAN. Table 4.4 provides a summary of the types of reports available for ISASP. Each of these three reports is discussed in greater detail in the text that follows.

Authorized users in PAN have different views and access based on the user role. Not all users in PAN have access to all reports. The District Coordinator (District Assessment Coordinator) has access to all reports at both the district and school levels. In comparison, the School Coordinator (School Assessment Coordinator) will only have access to the reports at their given school.

Table 4.4. ISASP Reports

Report	Description
Individual Student Report	Student scores and achievement levels in subjects taken
Student Roster	All student scores at a school grouped by grade, subject, and performance level
Achievement Level Summary	Chart comparing the percentage of students at each achievement level at a school compared to both the district and the state averages.

Individual Student Reports

ISRs provide information on a student's overall performance in each subject measured. This report provides scale scores as well as achievement-level designations associated with the student's performance level. Performance within a domain is also reported for each student as is overall performance over time. Parents can use the information presented in these reports to help them understand their child's achievement.

The ISR is the official and final record of individual student results provided for student, parent, and teacher use. For each area tested, the ISR contains the following information:

1. **Performance Meter** – The graphic under each subject test heading displays a visualization of the student's achievement level for that test. The cut score ranges, unique for each grade level and subject, are displayed below the graphic.
2. **Scale Score** – The scale score is a score converted from the student's raw score that allows for comparisons across grades and years. The ELA scale score is a total derived from the combination of the Reading and Language/Writing scale scores.
3. **Achievement Level** – The achievement level reports the student's performance on the test. There are three levels: Advanced, Proficient, and Not Yet Proficient. It provides a general explanation of what the student knows and is able to do.
4. **Description of Performance** – The description under each subject test heading is an explanation of the student's understanding of the content specific to grade level.

5. **Iowa Core Domains** – The Iowa Core Domains are grade level and content specific areas of focus that are tested for the subject.
6. **Percent Correct** – The bar graph next to the Iowa Core Domains provides the percentage of points the student earned by domain.

Student Roster

Student Rosters are provided to schools by grade and subject area, organized by level of achievement. Rosters contain the following information:

1. **School and District Information** – The top of the page includes the class name, school name, district name, and grade level. The class name is the test session name students were assigned to in PAN. If there was no specific test session name set up, class name will be listed as the grade level.
2. **Achievement Levels** – The students will be listed alphabetically by last name under each achievement level heading. The cut score ranges, unique for each grade level and subject, are also provided in the heading.
3. **Scale Scores** – The scale score for each student is provided. For ELA, there are ELA Total, Reading, and Language/Writing scale scores.
4. **No Score Available** – If a registered student was marked complete but did not respond to any items, this student will be listed under the No Score Available heading. Students who only took one of the ELA tests would also appear here.

Achievement Level Summary

The Achievement Level Summary report provides the following information:

1. **School and District Information** – The top of the page will include the school name, district name, and grade level.
2. **Bar Graphs** – The bar graph shows the percentages by achievement levels at the school, district, and state level.
3. **Percentages by Achievement Levels** – Each bar shows the achievement level distribution for the school, district, and state. The percentages for each bar may not add to 100 percent due to rounding of each achievement level.

Appropriate Score Uses

ISASP has been designed, developed, and researched to support a variety of important educational purposes. These purposes involve the collection and use of information that describes either the individual student or groups of students. The underlying validity and reliability of the ISASP was thoroughly documented throughout the entire development and implementation process and after each year of operational administration.

Table 4.5 identifies the three primary purposes that are supported by ISASP. Appropriate interpretations and uses of the results from the proposed assessment support a broad range of educational discussions and decisions including accountability, instructional improvement, and school-based performance.

Table 4.5. Appropriate Uses of ISASP Results

Purpose	Intended Interpretation	Example Uses of Results
Measure achievement	Measure student achievement on the Iowa Core Standards in ELA, Mathematics, and Science Determine the degree to which students have acquired the essential skills and concepts of the Iowa Core Standards	Identify students at risk for poor learning outcomes Place students into appropriate programs Inform students and parents Inform and improve instruction Support accountability Evaluate programs Support professional development opportunities
Monitor growth	Describe change in student performance over time	Identify students not showing growth over time Support accountability
Indicate readiness	Monitor student achievement toward college and career readiness standards	Compare student achievement with established benchmarks for readiness Inform course planning Inform students and parents

Individual Students

Scale scores determine whether a student's performance has met or fallen short of the proficiency criterion level. Test results can also be used to compare the performance of an individual student with the performance of a similar demographic group or to an entire school, district, or state group.

Domain scores provide information about student performance in more narrowly defined Core content areas. For example, domain scores can provide information to help identify areas in which a student may be having difficulty. When an area of possible weakness has been identified, supplementary data should be collected to further define the student's instructional needs.

Finally, individual student test scores must be used in conjunction with other performance indicators to assist in making placement decisions. All decisions regarding placement and educational planning for a student should incorporate as much student data as possible.

Groups of Students

Test results can be used to evaluate the performance of student groups. The data should be viewed from different perspectives and compared to district and state data to gain a more comprehensive understanding of group performance. For example, the average scale score of a group of students may show they are above the district and/or state average, yet the percentage of students who are proficient in the same group of students may be less than the district or state percentages. One perspective is never sufficient.

ISSs can also be used to compare the performance of different demographic or program groups (within the same subject and grade). Average performance on a domain can help identify areas where further diagnosis may be warranted for a group of students.

Test results can also be used to evaluate the performance of student groups over time. Average ISSs can be compared across test administrations within the same grade and subject area to provide insight into whether student performance is improving across years. In making longitudinal comparisons, it is important to

recognize that new testing programs cannot be directly compared to previous testing programs. For example, results from the 2019 ISASP cannot be directly compared to previous administrations of the *Iowa Assessments*, because the ISASP was developed to different test specifications.

Test results for groups of students may also be used when evaluating instructional programs; year-to-year comparisons of average scores or the percentage of students considered proficient in the program will provide useful information. Considering test results by subject area and by domain may be helpful when evaluating curricula, instruction, and their alignment to standards because ISASP was designed to measure content areas within the Iowa Core. Generalizations from test results may be made to the specific content domain represented by the domains being measured on the test.

Cautions for Score Use

Test results can be interpreted in many ways and used to answer many different questions about a student, educational program, school, or district. As these interpretations are made, there are always cautions to consider.

When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error. Some score fluctuations would be expected if the same student tested across occasions using equivalent forms of the test. This effect is partly due to day-to-day fluctuations in a person that can affect performance and partly a consequence of the specific items contained on a particular test form the student takes. Chapter 8 describes measures that provide evidence indicating that measurement precision on ISASP is within the expected range. Nevertheless, measurement error must always be considered when making score interpretations.

Chapter 5: Performance Standards

This chapter summarizes the process for developing Performance Level Descriptions (PLDs) and results of setting performance levels for the Iowa Statewide Assessment of Student Progress (ISASP) for grades 3–11 English Language Arts (ELA), grades 3–11 Mathematics, and grades 5, 8, and 10 Science.

Process for Developing Performance Level Descriptions

PLDs are statements of the standards: the knowledge, skills, and abilities required for each performance level. PLDs are used both in the standard setting process and in reporting assessment results. PLDs are the key factor of importance for setting performance level cut scores; they are the statements of the standards used in standard setting. The PLDs give meaning to the scale score values in the form of cut scores used for reporting student achievement at each level. Typically, the results are reported as the percentage of students performing at or above each cut score (Proficient and Advanced) and the percentage of students performing within each level. It is imperative that the PLDs articulate clearly and concisely the essential knowledge and skills that students must demonstrate in their performance on the assessment. They must provide logical consistency in their calibration across levels within each grade and within each level across grades.

Overview

The goal of this project was to develop descriptions for grades 3–8, 9, 10, and 11 in ELA (including both Reading and Language/Writing), grades 3–8, 9, 10, and 11 in Mathematics, and grades 5, 8, and 10 in science. Two cut scores will be set, one for Proficient and one for Advanced, to demarcate three performance levels: Not Yet Proficient, Proficient, and Advanced, for each subject and grade. The process implemented for developing PLDs for ISASP was modeled on the procedures used for the National Assessment of Educational Progress (NAEP). It is a thorough methodology that focuses on content expertise of professional assessment development staff, as well as content expertise of outstanding teacher-educators coupled with an extensive solicitation of reviews and recommendations from education stakeholders throughout the state. A former leader of NAEP standard setting worked with Iowa Testing Programs (ITP) staff to design and implement this procedure. The iterative process provided several points of review and evaluation to strengthen the quality of the PLDs and to further add to the evidence in support of a valid process of developing the PLDs to be used in the standard setting process and for reporting ISASP assessment results relative to the performance standards.

General Performance Level Definitions

Having a general definition of performance is helpful in developing PLDs to provide a criterion for calibrating the PLDs. The general definitions state the meaning of Proficient and Advanced performance in terms that apply to any subject and grade in the ISASP program for which standards (cut scores) are to be set. ITP leadership staff discussed the need for general performance level definitions with state Department of Education leadership to reach agreement of the definitions in Table 5.1. The titles and descriptions of the performance levels were defined to be part of a cohesive assessment system.

Table 5.1. General Performance Level Descriptors for ISASP

Level	General Performance Level Descriptors
Advanced	Students performing at the Advanced level demonstrate thorough competency over the knowledge, skills, and abilities that meet the requirements for their grade level associated with academic readiness for college and careers in the subject.
Proficient	Students performing at the Proficient level demonstrate adequate competency over the knowledge, skills, and abilities that meet the requirements for their grade level associated with academic readiness for college and careers in the subject area.
Not Yet Proficient	Students performing at the Not Yet Proficient level have not yet demonstrated the knowledge and skills to be classified as Proficient.

Implementation

The process of developing the PLDs is intense and requires time. The persons charged with drafting the descriptions must have expertise in the content area and a clear understanding of the academic abilities of students in the grade levels. They must be intimately familiar with and well-versed in the curriculum standards guiding the assessment framework and in the test specifications. Assessment Development staff of ITP who worked to develop the ISASP worked together to draft the PLDs for each level (Proficient and Advanced) for each grade and content area. These staff members have the qualifications needed for the task, and their experience in developing the ISASP increased the efficiency of the process.

To further increase the efficiency of the process, a review and evaluation of the PLDs used in other states known to have good assessment programs based on the Common Core curriculum were undertaken to identify PLDs that could serve as a model for Iowa. These PLDs from other state assessment programs were used as a starting point for the development of PLDs for the ISASP.

PLDs are based on the curriculum standards (the Iowa Core), and not on the actual items developed for the assessment. This is an important point because different test forms will be developed in future years to meet the test specifications for ISASP, and different items will be developed to measure the same knowledge, skills, and abilities on each assessment that are specified in the Iowa Core for each grade and subject. So, both the assessment and the PLDs are based on the Iowa Core.

The development team reviewed the initial drafts of PLDs for appropriate calibration between the two performance levels within each grade and subject, and across the grades within each of the performance levels. The expert consultant in NAEP standard setting also reviewed them to check for any concerns regarding the word choices in the statements of what students should know and be able to do within and across grades. Following an initial check of the PLDs, they were sent out for a broad-based vetting process. ITP worked with the Iowa Department of Education (IDOE) to identify teachers, curriculum directors, leaders of professional content organizations (e.g. the Iowa chapter of NCTM, IRA, etc.), and leaders of state teachers' unions. Including a wide variety of key stakeholders in the vetting process is important to the overall process because this helps to increase the sense of "buy-in" among the stakeholders, thereby

increasing the probability of acceptance and approval of the assessment results based on these statements of the standards.

Information regarding the purpose of the review and vetting was provided, along with links to website information and instructions for the review process, key questions for response, and the PLDs to be evaluated. Information regarding the educational background and experience of respondents was collected. Reviewers were asked to submit their responses by April 15, 2019. Responses were collected and reviewed in order to determine whether there were major concerns or systematic concerns so that the responses could be categorized to highlight those clearly. A total of 74 responses were received, but only 48 respondents submitted comments and recommendations.

Teacher panels were recruited from throughout the state to participate with the ITP development team in the final review and process of modifications based on the comments collected through the vetting process and based on the evaluations of the members of the expert panels. The goal was to recruit three teachers for each subject and grade span to provide their expertise at this stage of the process. The teachers recruited for this work were known to have a high level of expertise regarding the curriculum standards and familiarity with Iowa state assessments. A total of 28 panelists were recruited to meet with ITP staff and the consultant on April 27, 2019.

A scribe was assigned to each panel group to take notes, and the content facilitators each recorded recommended edits and modifications recommended by panel members. The ELA groups completed edits of the PLDs during the allocated time, but additional work was needed for the Science PLDs and those for Mathematics in grades 9 and 10. The ITP staff worked with the notes from the panel meeting to make recommended edits and modifications, which were then shared with members of the expert panels for their review and evaluation. Comments and further edits recommended were exchanged and the finalized PLDs were shared with all members of the content review panels by May 15, 2019.

Finalized Performance Level Descriptors

The finalized version of PLDs was presented on May 23, 2019, to IDOE for review and input prior to use in the standard setting process scheduled for July 22–26, 2019. The finalized PLDs for all subjects and grades can be found in the *ISASP Math, English, and Science Tests Standard Setting Technical Report*.

Standard Setting

Once an assessment is administered, various groups—including students, parents, educators, administrators, and policymakers—want to know how students performed on the assessment and how to interpret that performance. By establishing levels associated with different student performance on the assessment, a frame of reference is developed for interpreting student scores. For a criterion, standards-based assessment, such as the ISASP program, performance on the assessment is compared to a set of predefined content standards. The standards communicated within the Iowa Core Standards define a set of knowledge, skills, and abilities the students taking the assessment are expected to demonstrate upon completion of each course or grade. The cut scores established through the standard setting represent the level of competence students are expected to demonstrate on the assessment to be classified into each performance level.

In order to classify student performance into the different performance levels, the following components are generally required: 1) General Performance Level Descriptors, 2) PLDs, and 3) cut scores. General Performance Level Descriptors and PLDs, described in the previous section, were established prior to the standard setting meeting and approved by ITP for use during the standard setting meeting. Cut scores,

which represent the lowest boundary of each performance level on the scale, were established during the standard setting. The process of establishing cut scores and recommending performance standards for the ISASP assessments was in line with national best practice for standard setting. A summary of the process is presented below; additional details and results can be found in the *ISASP Math, English, and Science Tests Standard Setting Technical Report*.

Process

The ISASP standard setting process consisted of three steps: pre-meeting development, standard setting meetings, and vertical articulation. Prior to the meeting, various tasks were completed, including the development of materials for the panelists, preparation of the Pearson standard setting website for panelists and facilitators, presentation materials for the facilitators, and development of data analysis sources and procedures. For the standard setting meeting, committees of panelists worked with grade- and subject-specific content and referenced borderline descriptions to make recommendations for cut scores that define the different performance levels for each assessment. In vertical articulation the recommended cut scores for each assessment were reviewed for reasonableness and alignment of performance level expectations across grades by select members of the committee.

Facilitator Training

The sessions were facilitated by a psychometrician from Pearson with knowledge and experience leading standard setting meetings. The facilitator was responsible for ensuring appropriate processes were followed throughout all sections of the meeting and that panelists had a solid understanding of the tasks they were asked to complete.
All facilitators underwent an extensive program of training to prepare them for leading the set of standard setting meetings. The facilitator training included:

- Use of the Pearson standard setting website—Because the standard setting website was used as a facilitation tool during the meeting, facilitators needed to become familiar with the use of the platform. Specific guidelines for modeling the website and providing access to the panelists were discussed.
- ISASP Assessments—The facilitators were provided an overview of the ISASP assessment program, including the content areas assessed, different item types, scoring rules, performance levels, and scaling design.
- Standard setting process—The facilitators participated in a walkthrough of the standard setting meeting agenda with a focus on specific issues for these meetings, such as time management, the use of the online platform, and communicating feedback information.
- Training slides and presentation notes—The facilitators were introduced to the standard setting training slides before the meetings. Notes in the standard setting training slides provided the facilitators with specific guidance throughout the presentation, including when specific language was to be used during the panelist training.

A general facilitator training was conducted on June 28, 2019, for all facilitators. Subject-specific facilitator training meetings were held for 60 minutes each on July 15, 16, and 18, 2019, to prepare the facilitators to address distinctive aspects of the subject specific meeting. A final training and discussion were held on-site July 21, 2019, the day before the standard-setting meetings commenced, to address any final topics. There was also an additional discussion on July 23, 2019, for the facilitators of the grades 9-11 math and ELA

committees, since they started mid-week. At the end of each day during the standard setting meetings, a debriefing was held to discuss concerns, positives, and the materials and procedures for the next day.

Content experts from ITP were available, as observers, to assist panelists with content and policy questions during the standard setting meetings. A staffing plan was provided to ITP prior to the standard setting meetings to communicate the psychometric and support staff scheduled to attend.

Committee Panelist Composition

All panelists for the standard setting committees were selected by ITP to represent educators and key stakeholders from across Iowa who had knowledge of and experience working with student groups within the populations administered the ISASP assessments. The selection process of committee panelists involved considerations intended to create a sample as representative of the state as possible, including demographic variables (gender, race, etc.), geographic representation, and background (educational experience, education, etc.). ITP placed an emphasis on educators who had relevant content knowledge as well as experience with a variety of student groups.

There was a total of 182 participants at the standard setting meetings. The panelists were divided into 15 breakout committees. Each committee focused on establishing cut score recommendations for one grade (e.g., grade 9 ELA) or grade-band (e.g., grades 3 and 4 ELA). The tables in Appendix C summarize the characteristics and experience of the panelists in each committee, including demographic information, current positions in education, experience working with various types of student populations, and the types of districts they represent.

The panelists in each committee were assigned to table groups. The table groups were selected prior to the meeting to ensure that, to the greatest extent possible, the panelists at each table were representative of the committee. The panelists were placed into table groups to facilitate discussions during the standard setting meeting and ensure each panelist had the opportunity to fully engage in the process.

Prior to the standard setting meeting, an individual was selected from each table group to serve as a table leader. The table leaders assisted the process facilitator during the meeting by facilitating the table discussions, encouraging all panelists to participate, and ensuring the discussion remained relevant to the meeting. To assist the table leaders in understanding and fulfilling their role during the meeting, a table-leader training was held during the first day of the standard setting, so table leaders were informed of the expectations for facilitating group discussions and participating in the articulation meeting.

General Method

From July 22 to July 26, 2019, after the first year of operational administration, a standard setting committee meeting was conducted to provide cut score recommendations for the ISASP assessments for ELA, Mathematics, and Science. The committees comprised individuals including teachers and non-teacher educators. The participants were selected for the standard setting committee to provide content and grade-level expertise during the committee meeting and be representative of the state teaching population, including geographic region, gender, ethnicity, educational experience, community size, and community socioeconomic status.

The Extended Modified (Yes/No) Angoff standard setting method was used at the standard setting meeting (Davis & Moyer, 2015; Plake, Ferdous, Impara & Buckendahl, 2005). This is a content- and item-based method that leads participants through a standardized process in which they consider expectations of student

performance, as defined by the borderline descriptions, and the individual items administered to students to recommend cut scores for each performance level. The committees used the standardized process for each grade and subject, which resulted in cut score recommendations.

The process started with participants experiencing the assessment for the respective grade of their review committee from the Spring 2019 administration through an online testing environment similar to the one used to administer items. Based on their experience with the test items and a review of the borderline descriptions, participants reviewed each item on the test and answered the following question for each performance level:

"How many points would a student performing at the borderline of the [specific] performance level likely earn if they answered the question?"

The cut score recommendation for each individual participant was the expected raw score a student performing at the borderline of the respective performance level would likely earn, calculated as the sum of the individual item judgments. For the purposes of the standard setting, "likely" was defined as two out of three students at the borderline of the performance level. Each recommended cut score from the standard setting committee is the median of the recommendations from the individual participants in the committee.

Vertical Articulation

The table leaders from each standard setting committee convened in an articulation panel for each subject area. The purpose of the articulation meeting was to review and evaluate the reasonableness of the cut score recommendations from the standard setting committees within each subject.

After an introduction to the purpose of articulation, the panelists were guided through a process where they considered the cut score recommendations from the standard setting committees of their subject area and, if necessary, made changes to the recommendations. Panelists reviewed the PLDs and recommended cut scores for the ISASP assessments within their content area. Panelist then compared the student impact for the different performance levels, based on the committees' Round 3 recommendations. The final result of each articulation committee was a set of recommended cut scores.

Panelists from the science breakout sessions came together on the morning of Wednesday, July 24, 2019, to participate in their articulation meeting. The facilitator for the science articulation was Eric Moyer, Ph.D. Panelists from the ELA and mathematics breakout sessions participated in separate articulation meetings on the morning of Friday, July 26, 2019. The facilitator for the ELA articulation was Jennifer Galindo, Ph.D. and the facilitator for the mathematics articulation was Eric Moyer, Ph.D.

Cut Scores

The cut scores recommended for adoption for the ISASP assessments for ELA, Mathematics, and Science are shown in Table 5.2. This table shows the scale score ranges corresponding to each performance level. The cut scores for the performance levels are the lowest cut score within each range.

Table 5.2. Cut Score Ranges for ISASP Performance Levels

Subject	Grade	Performance Levels		
		Not-Yet-Proficient	Proficient	Advanced
English Language Arts	3	345 to 397	398 to 446	447 to 510
	4	350 to 413	414 to 477	478 to 540
	5	355 to 436	437 to 512	513 to 590
	6	360 to 455	456 to 540	541 to 640
	7	370 to 474	475 to 568	569 to 680
	8	385 to 493	494 to 593	594 to 720
	9	410 to 504	505 to 617	618 to 750
	10	435 to 529	530 to 641	642 to 780
	11	460 to 560	561 to 659	660 to 800
Mathematics	3	345 to 389	390 to 442	443 to 510
	4	350 to 408	409 to 475	476 to 540
	5	355 to 428	429 to 502	503 to 590
	6	360 to 449	450 to 531	532 to 640
	7	370 to 468	469 to 574	575 to 680
	8	385 to 489	490 to 605	606 to 720
	9	410 to 512	513 to 625	626 to 750
	10	435 to 536	537 to 653	654 to 780
	11	460 to 558	559 to 674	675 to 800
Science	5	355 to 451	452 to 541	542 to 590
	8	385 to 507	508 to 608	609 to 720
	10	435 to 544	545 to 655	656 to 780

Results for ISASP Assessments

Results for the 2019 ISASP administration can be found in the *ISASP Math, English, and Science Tests Standard Setting Technical Report*. Related to performance standards, the results provide the percentage of students who would be classified into each performance level based on the recommended cut scores. The percentage of students in a performance level is not directly comparable across grades and subjects. The population of students tested is different for each assessment. Performance levels from different tests are not comparable because the cut scores for these tests are criterion-referenced—they are based on content-specific expectations of what students should know and be able to do. Results for future administrations of the ISASP are available in the *ISASP Annual Statistics Reports*.

State Approval

The cut score recommendations from the standard setting process were presented to IDOE and the Iowa State Board of Education (ISBE) for consideration and approval. ITP worked with IDOE to provide the ISBE with additional supporting information about the assessment and impact of the cut score recommendations. The ISBE adopted the cut score proficiency recommendations for the ISASP assessments on September 12, 2019.

Chapter 6: Scaling and Equating

The Iowa Statewide Assessment of Student Progress (ISASP) is a system of standards-based, summative accountability assessments. The tests are designed and developed through an evidence-centered approach to support interpretation and use in terms of the Iowa Core Standards (Iowa Core) adopted by the Iowa State Board of Education and used by Iowa educators in the design of curricula and instruction. For each subject and grade level, the content standards specify the subject matter students should know and the skills they should be able to perform. In addition, as described in Chapter 5, performance standards are defined to specify how much of the content standards students must demonstrate mastery of to be designated Proficient or Advanced with respect to performance.

Building tests to content standards ensures the tests assess the same constructs from one year to the next. Small differences from one year to the next in overall test difficulty or in other test characteristics are the reason it is desirable to report test scores on a score scale whose properties remain constant over time. The scale scores for the ISASP tests serve that primary purpose. The ISASP scale score (ISS) metric also accommodates the psychometric requirement of equating scores on different test forms or in transitioning the testing program from fixed test forms to adaptive testing, in which there are no test forms per se but rather items selected to increase the precision of measurement for each individual. Procedures used to develop the ISS metric guarantee the equity of performance standards from one year to the next. These procedures create derived scores through the process of scaling (which is addressed in the first part of this chapter) and the equating of test forms (which is described later in the chapter).

Rationale

Scaling of tests is the process in which student performance is associated with a number. The simplest way to score a single test is to calculate the total number of points earned, the raw score, which can also be reported as a percent of total points. The raw score, however, can be interpreted only in terms of a particular set of items. When new test forms are used in subsequent administrations of a testing program, raw scores have little value in and of themselves. To achieve comparability in reported scores, some type of metric derived from raw scores must be used. In the ISASP program, the ISS is that metric; it allows for direct comparisons of student performance between administrations.

The ISS metric for ISASP is a vertical scale that spans the full performance continuum on each subject-area assessment from grades 3–11. The scale means between grades are spaced in intervals based on the relationships of between- and within-grade standard deviations such that the overlapping grade-level distributions yielded a pattern of growth across grades that was consistent with the research literature on vertical scales (Kolen, 2006; Petersen, Kolen & Hoover, 1989; Tong & Kolen, 2007). The scale parameters determined in this way were used to define the relationship between raw scores on the 2019 ISASP assessments and the subject-area ISS. Because the ISASP assessments are standards-based, summative accountability assessments, the result of the scaling process should be a metric that readily associates each ISS with an achievement level that represents the degree to which students meet the standards of a given grade and achievement level. For the ISASP assessments in English Language Arts (ELA), Mathematics, and Science, the final scaling results lead to a designation of Not Yet Proficient, Proficient, and Advanced as approved by the Iowa State Board of Education.

Measurement Models

The ISASP program uses both classical test theory (CTT) and item-response theory (IRT) as modeling frameworks for the psychometric foundations of the assessments. Procedures based on CTT are used primarily to establish the psychometric integrity and scaling of test scores, whereas procedures based on IRT are used in scaling items and establishing comparability of test forms assembled to test specifications and aligned to the Iowa Core.

CTT derives its psychometric strength from its simple model for observed scores, namely

$$X = T + E, \tag{6-1}$$

where X is an observed test score, T is the corresponding true score, and E is a random error of measurement. From this simple model of test scores useful results for characterizing both reliability (as the ratio of true score variance to error variance) and the standard error of measurement can be obtained. CTT also provides a basis for developing distributions of scale scores that remove the influence of measurement error (Petersen et al., 1989) without relying on item-level assumptions of unidimensionality and local independence of item scores. Because of the simplicity of its model for test scores, CTT is often described as a robust psychometric modeling framework. In the ISASP program, CTT results are used both to document the technical quality of the assessments and to validate model-based results that utilize IRT, a psychometric model with strong assumptions about individual test items.

IRT is used in the ISASP program for a variety of purposes. It is a general theoretical framework that models data resulting from an interaction between students and test items. The advantage of using IRT models in the scaling of items is that all the items measuring performance in a particular content area can be placed on the same scale of difficulty. Placing items on the same scale across years facilitates the creation of equivalent forms each year as well as transitions to tailored or adaptive assessments.

IRT encompasses a number of related measurement models. The models used in the ISASP program share in common the item attributes of difficulty and discrimination. These are the 2-parameter logistic model (2PL; Lord & Novick, 1968; Lord, 1980) for use with items that are scored right or wrong (i.e., dichotomous items) and the graded-response model (GRM; Samejima, 1969, 1972; Thissen & Steinberg, 1986) for use with items scored with ordered categories (i.e., polytomous items scored with a rubric). The 2PL model for dichotomous items is used to scale multiple-choice, gridded-response, and technology-enhanced items. The GRM is designed for scaling items associated with multiple scores or points, for example, constructed response (CR) items and essays.

Two-Parameter Logistic Model

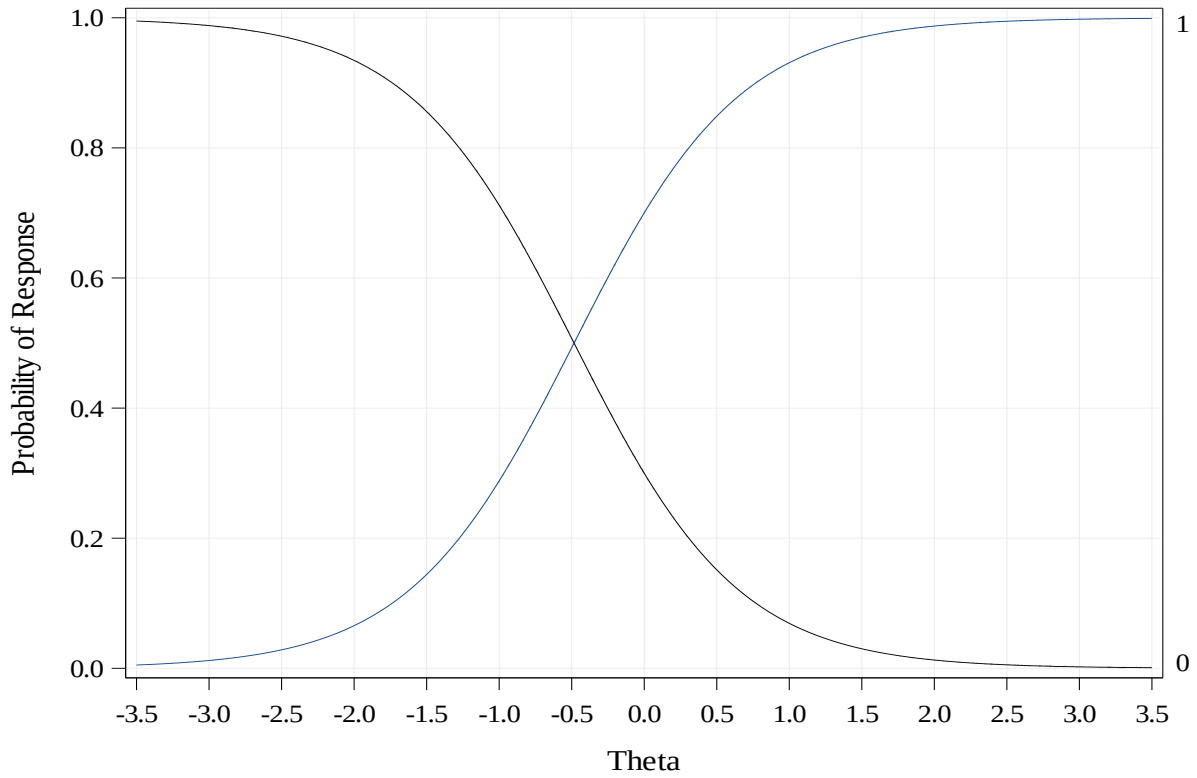
The 2PL model can be written as the following mathematical equation, where the probability of a correct response for person i taking item j, $P_{ij}(\theta_i)$, is given by:

$$P_{ij}(\theta_i) = \frac{\exp[1.7\alpha_j(\theta_i - \beta_j)]}{1 + \exp[1.7\alpha_j(\theta_i - \beta_j)]} \tag{6-2}$$

Equation 6-2 is the model that describes probabilistically what happens when students answer questions on a test. Terms in the equation represent attributes of students and items. The student attribute of ability or achievement (sometimes called the IRT scale value) is represented by the symbol θ_i for each individual i, and the item attributes of discrimination and difficulty are represented by the symbols α_j and β_j, respectively, for each item j. By IRT convention, θ and β are expressed in a common, standardized IRT metric with a mean of zero and standard deviation of one for each grade level. In this way, the model places student ability and item difficulty on the same scale. Values of α measure the strength of the relationship between an item and the ability or achievement construct measured by the assessment, with higher values indicating a stronger relationship. Figure 6.1 provides a graphical display of equation 6-2, the item characteristic curve of the 2PL model for an item from the ISASP grade 5 Mathematics assessment.

In Figure 6.1, the blue curve displays the probability of a correct response to this item, P_{ij}, as a function of θ_i, whereas the black curve displays the complementary probability of an incorrect response, $1 - P_{ij}$. For the item in Figure 6.1, α is 1.8 and β is -0.5. The value of β (-0.5) signifies that a student at that level on the achievement construct has a 50 percent chance of answering this math item correctly. The value of α (1.8) is proportional to the slope of the blue curve in Figure 6.1 at the value of β. For this item, the relatively high value of α means that the chance of a correct answer goes up rapidly with only small increases in a student's IRT scale value, θ. Students with an IRT scale value 2 or more units below the midpoint have a near 0 percent chance of answering this item correctly, whereas students 2 or more units above have a chance approaching 100 percent.

Figure 6.1. Two-Parameter Item Response Functions for a Mathematics Item from ISASP Grade 5



The IRT model for polytomous items in ISASP, items worth more than one point such as CR items and student essays, is the GRM. The GRM shares the same IRT scale metric of the 2PL model, θ, as well as the same item discrimination value, α. It differs from the 2PL model in that the single difficulty value of the 2PL is replaced by multiple values, β_{jk}, that represent the locations of the boundaries between scores of 0 versus 1, 1 versus 2, 2 versus 3, and so on for as many scores, K, as are possible on the CR item. For a CR item worth K = 2 points and scored 0-1-2, there are two score or category boundaries and therefore two values of β, β_{j0} for the boundary between 0 and 1 point and β_{j1} for the boundary between 1 point and 2 points.

The GRM used in the ISASP program can be written as the expression in equation 6-3. In equation 6-3, $P_{ij,k}(\theta)$ represents the probability of a response for person i on item j worth k or more points, which is given by:

$$P_{ij,k}(\theta) = \frac{\exp[1.7\alpha_j(\theta_i - \beta_{jk})]}{1 + \exp[1.7\alpha_j(\theta_i - \beta_{jk})].} \quad (6\text{-}3)$$

The fact that the GRM in equation 6-3 employs a dichotomous 2PL model for each pair of adjacent scores accounts for the multiple β values instead of the single β value in the dichotomous case. Instead of capturing overall item difficulty as in the dichotomous 2PL model, the polytomous GRM model uses the category boundary parameters to provide a measure of the relationship between the response functions of adjacent score categories. The GRM of equation 6-3 gives rise to differences of the form

$$P_{ij,k}(\theta) - P_{ij,k+1}(\theta), \quad (6\text{-}4)$$

which can be used to obtain the chance that a student with a given IRT scale value obtains k points on a CR item. Equation 6-4 illustrates why in their taxonomy of IRT models, Thissen and Steinberg (1986) refer to the GRM as a difference model. Figure 6.2 provides a graphical display of the category characteristic curves of the GRM model for a 2-point item from the ISASP grade 8 Mathematics assessment.

In Figure 6.2, each curve represents the conditional probability of obtaining a score of zero (black), one (blue), or two (green) on this item. For the math item in Figure 6.2, the value of α is 1.5 and the values of β_{j0} and β_{j1} are –0.1 and 1.6, respectively. The score boundary parameter β_{j0} of –0.1 is the IRT scale value at the crossing point of the "zero" (black) response function and the "one" (blue) response function. Similarly, β_{j1} equal to 1.6 is the scale value at the crossing point of the response functions for score points one (blue) and two (green). This math item has a reasonable spread of score category boundary parameters, 1.6 minus –0.1, or 1.7 IRT scale units, which is an indication of a well-constructed item. Boundaries that are too close together may indicate the score categories are not distinguishing students in an effective manner. IRT scaling of CR items with the GRM provides this convenient check to support item development, test assembly, and scoring of polytomous items.

Figure 6.2. Two-Parameter Graded Response Model Category Response Functions for a Constructed-Response Mathematics Item from ISASP Grade 8

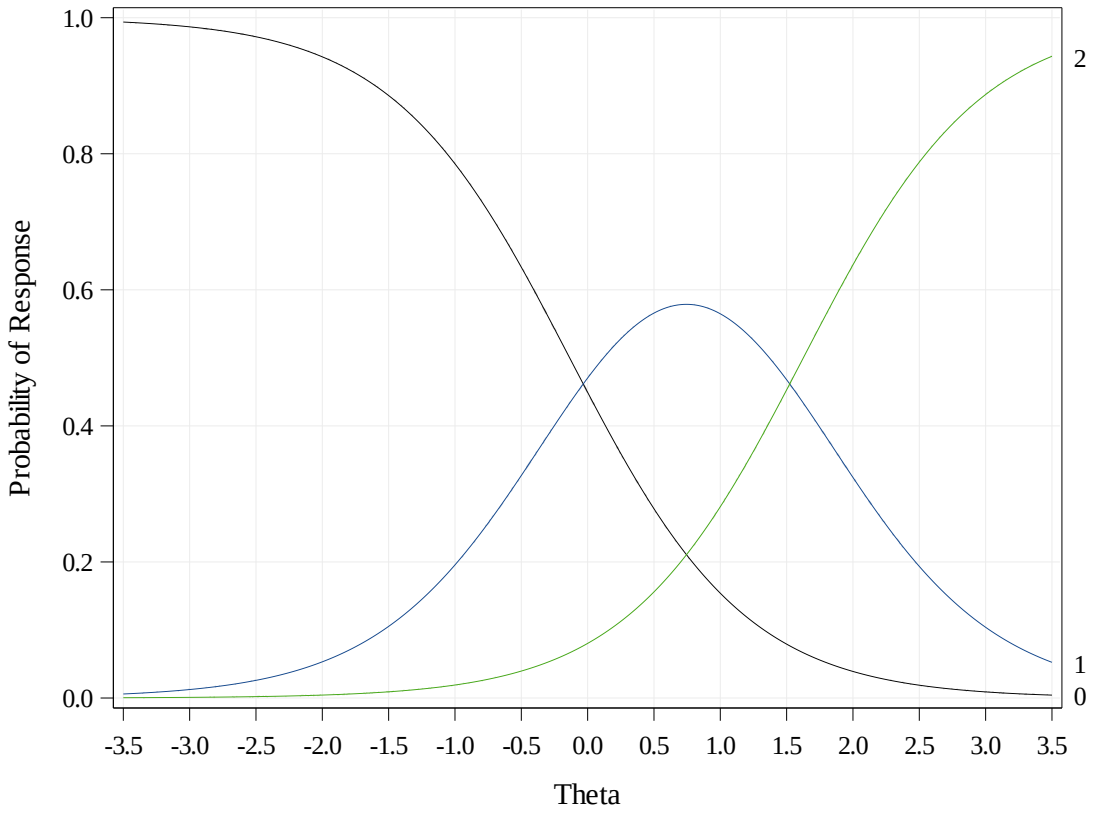
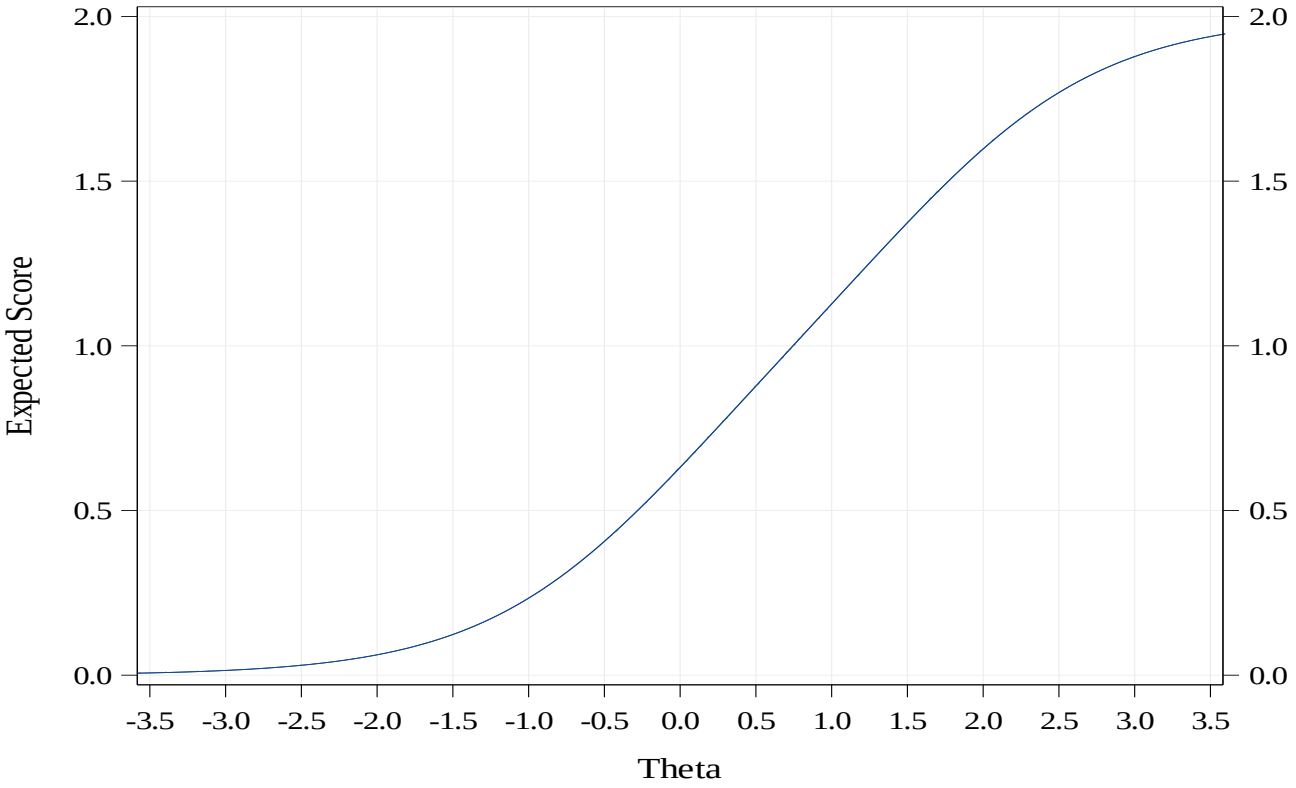


Figure 6.3 displays the expected score for the math item in Figure 6.2 as a function of an individual's ability or achievement level on the construct measured by the assessment. The figure shows that a group of students near the middle scale value of 0 would be expected to earn less than 1 point on average for this math item, whereas a group of students who are 2 standard deviations above the midpoint ($\theta = +2$) would be expected to average closer to 1.5 points on this item.

Figure 6.3. Two-Parameter Graded Response Model Expected Item Score Function for a Constructed-Response Mathematics Item from ISASP Grade 8



Development of the ISASP Scale Score Reporting Metric

Determination of a reporting metric for standards-based assessments that (1) spans grades 3–11 and (2) is suitable for interpretations of student growth requires information about grade-to-grade relationships with respect to both content standards and student performance (Kolen, 2006). The item pool used to develop the ISASP assessments was developed to be aligned to grade-level content of the Iowa Core. Over a period of five years leading up to the first ISASP administration, field-test items were administered to representative samples of students throughout the state as part of their participation in regular testing. During this period, the Iowa Core in grades 3–11 had been fully implemented in Iowa schools. Each field-test item was administered to students in multiple grades, and data from the field test samples were examined by content area (Reading, Language/Writing, Mathematics, and Science) for consistency with the IRT requirements of unidimensionality and local independence. Dimensionality analyses are reported elsewhere in this manual and in the *ISASP ASR-2019* and *ISASP ASR-2021)*.

Calibrations of items for the IRT models used in the ISASP program were performed using the computer program IRTPRO 4.20 (Cai, Thissen & du Toit, 2018). The program estimates item discrimination and difficulty for multiple-choice items with the 2PL model and discrimination and category boundary parameters for polytomously scored CR items with the GRM. Within-grade item calibrations in each content area served as the basis for a series of transformations aimed at determining grade-to-grade relationships in item performance for the purpose of informing the vertical scaling process used to develop the ISASP scale

score (ISS) in each content area (Stocking & Lord, 1983). More specifically, the constants developed to transform IRT item parameters to a common scale across grades indicated that a vertical scale metric suitable for capturing moderately increasing variability across grades 3–11 would be needed. Consistent with the IRT scale transformations, within-grade standard deviations of the ISASP vertical scale were set to increase by a factor of about 10 percent for each successive grade over the grade span of 3–11. The vertical scale metric itself was centered at grade 7, midway between grade 3 and grade 11, with a mean of 500 and a standard deviation of 50. Within-grade means were spaced on average 22 scale score points apart; within-grade standard deviations decreased by 10 percent below grade 7 and increased by the same amount from grades 7 to 9. After grade 9, the within-grade standard deviations remained the same because there was no consistent evidence from calibrations of increasing variability in the remaining high school grades. Table 6.1 provides the parameters of the ISASP scale score distributions used to develop the raw score to scale score conversion tables for the ISASP program.

Table 6.1. Vertical Scale Parameters for the ISASP Scale Score Distributions

Grade	Mean	Standard Deviation
3	409	28.7
4	432	34.4
5	454	41.3
6	476	45.5
7	500	50.0
8	521	55.0
9	544	60.5
10	568	60.5
11	593	60.5

Figures 6.4 and 6.5 display the within-grade frequency and cumulative frequency distributions of the ISSs for the assessments in Reading, Language/Writing, Mathematics, and Science. These distributions are the result of the application of the scale parameters in Table 6.1 to empirical ISASP distributions from the 2019 administrations. They follow a pattern across grades that is similar to the patterns observed in the vertical scales for other federally approved statewide and consortium-based assessment programs (cf. Smarter Balanced Assessment Consortium, 2016). The steadily increasing means from grade to grade establish the expected change for grade cohorts in each test area. The scale scores thus lend themselves to use in several potential ways to quantify student growth. For example, they can be used as a direct measure of year-to-year change on a common metric in a change-score growth model (Castellano & Ho, 2013). They can also serve as the foundation of growth measurement via student growth percentiles (Betebenner, 2009) or via empirically derived distributions of change between adjacent ISASP administrations. The distributions in these figures also exhibit the property of steadily increasing variability across grades found in other vertical scaling studies in K-12 assessment (Petersen et al., 1989; SBAC, 2016; Snow & Lohman, 1989).

Figure 6.4. Relative Frequency Distributions of ISASP Scale Scores in Reading, Language/Writing, Mathematics, and Science – Grades 3–11

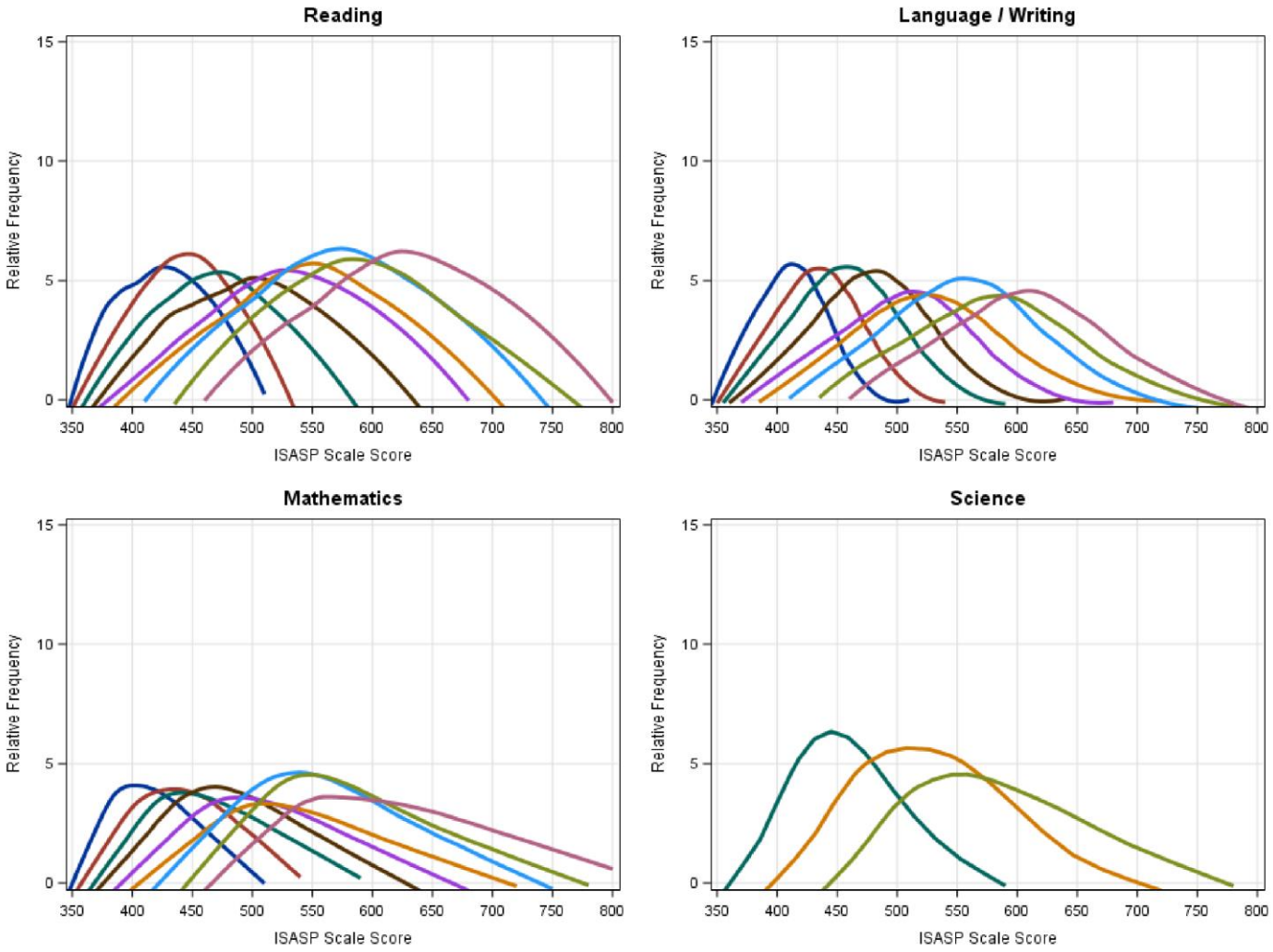
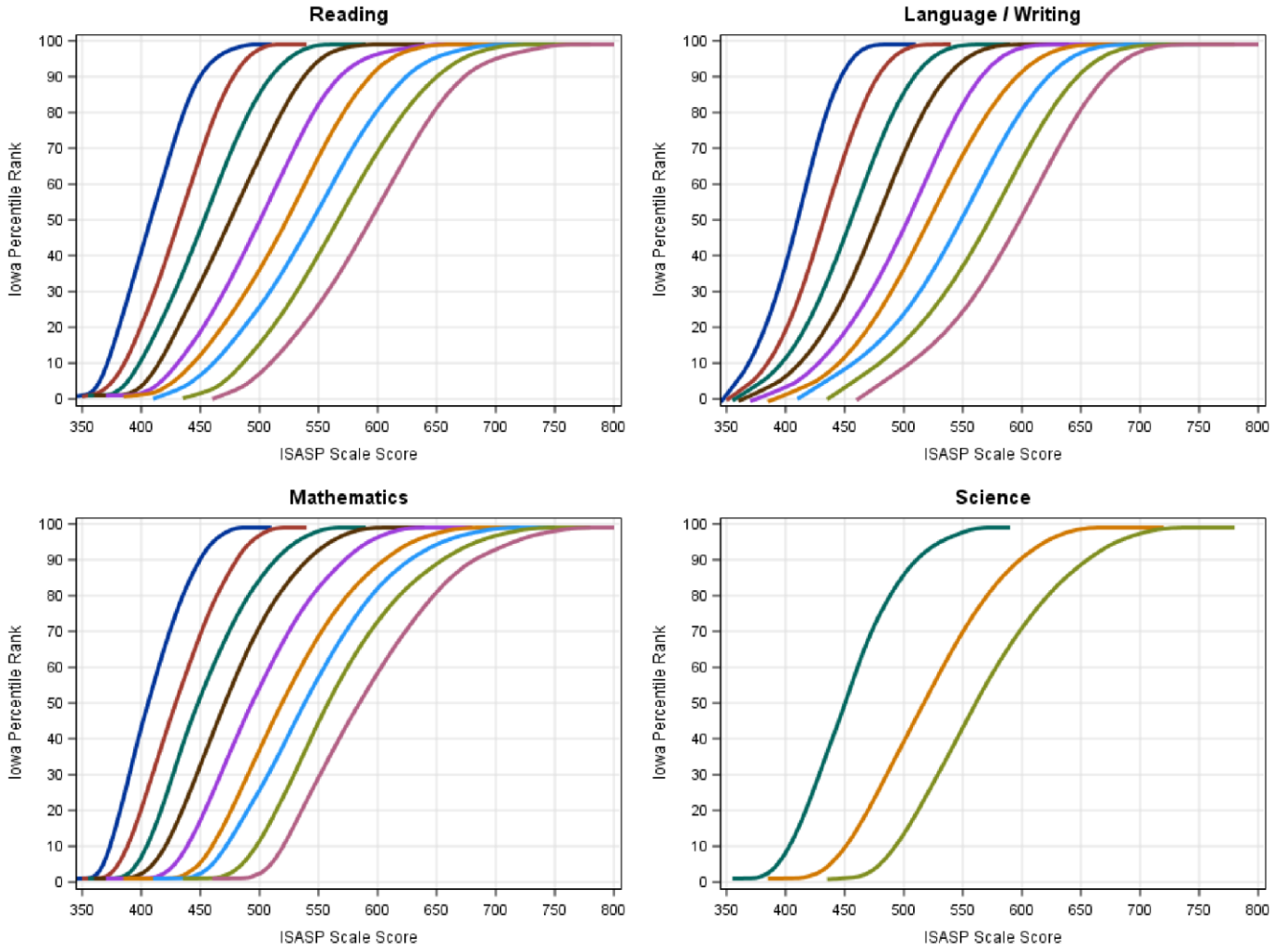


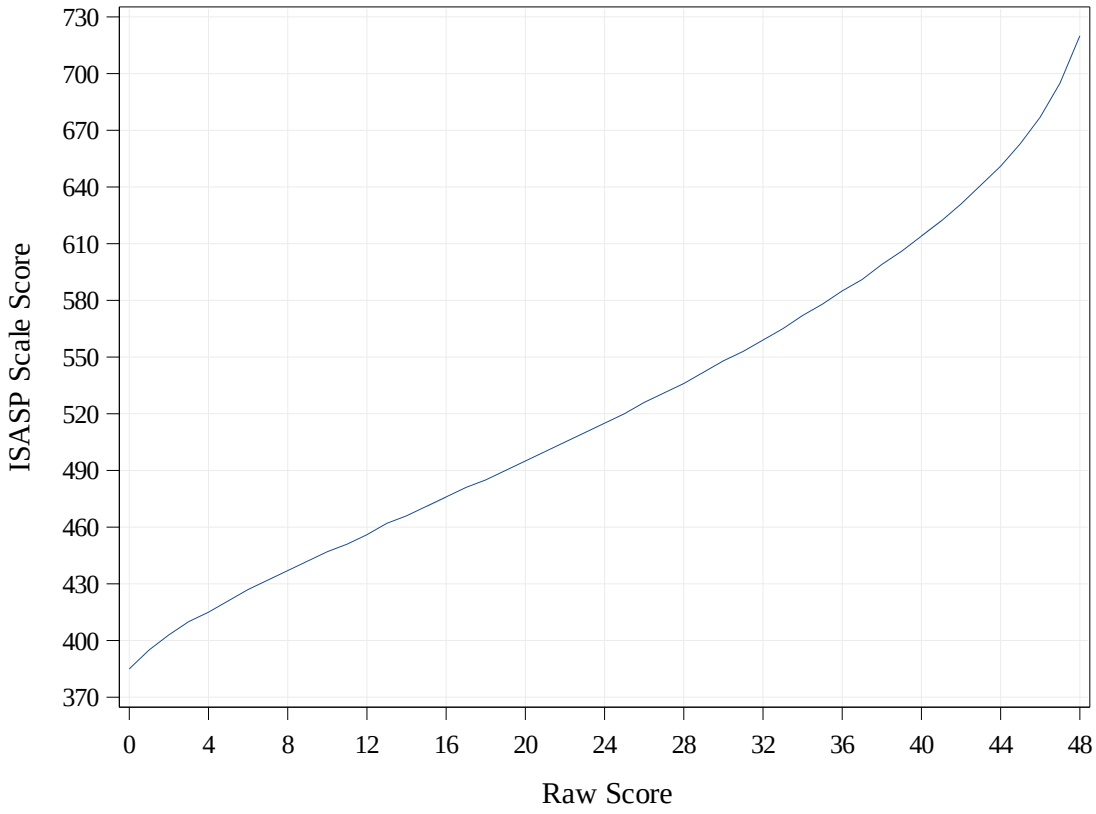
Figure 6.5. Cumulative Frequency Distributions of ISASP Scale Scores in Reading, Language/Writing, Mathematics, and Science – Grades 3–11



Scale Attributes and Interpretive Guidance for ISASP Scale Scores

The ISASP raw-to-scale score conversions are nonlinear transformations from one metric to the other. Figure 6.6 shows this transformation for the 2019 ISASP Mathematics assessment in grade 8. In this example, in the middle of either scale the transformation is generally linear such that a single point increase or decrease in the raw-score metric represents about a 5-point change in the scale score metric. This relationship is maintained over the range where the graph in Figure 6.6 follows a straight line.

Figure 6.6. Raw Score to Iowa Scale Score Transformation – ISASP Grade 8 Mathematics



The primary functions of the scale score are (1) to provide a consistent metric for translating ISASP scores into the standards-based achievement levels adopted by the Iowa State Board of Education of Not Yet Proficient, Proficient, and Advanced; (2) to establish a common metric for scores on ISASP forms across years that can be responsive to changes in assessment design such as adaptive testing; (3) to support future efforts related to metrics for student growth; and (4) to determine how far students are from the various proficiency levels without depending upon the changing raw scores across forms. Additionally, the Iowa Department of Education (IDOE) and Iowa schools may use the ISS in summary fashion for purposes of the Iowa School Performance Profile and for Local Education Agency program evaluation across the years. For example, it will be appropriate to compare the average grade 5 scale score in Reading for next year to the grade 5 average for this year. Explanations for why the differences exist, of course, will depend on factors specific to individual schools.

Domain Scores for the ISASP Assessments

In addition to the ISSs for tests in Reading, Language/Writing, Mathematics, and Science, percent correct (or percent of total points) scores are reported for the content domains of the Iowa Core Standards. These scores are discussed in greater detail in Chapter 4 of this manual. For example, in Reading, individual student reports include percent correct scores in Key Ideas and Details, Craft and Structure, and Integration of Knowledge and Ideas. The purpose of these domain scores is purely *descriptive*, that is, to provide individual students with a somewhat more detailed report of the points they earned on the assessment. Teachers and school administrators will be provided with state-level domain scores in the percent-correct

metric so that differences between observed domain scores and state averages can be used as an aid to interpretation. Teachers and administrators are cautioned, however, to not use domain scores for making judgments about student performance. Likewise, these scores play no role in state or federal accountability interpretations relative to the Iowa School Performance Profile system of the IDOE. They are intended only to describe how each student did on the test with respect to the reported domains of the Iowa Core.

Conversion Tables, Frequency Distributions, and Descriptive Statistics

ISASP ASR-2019 and *ISASP ASR-2021* provide tables for converting raw scores to derived scale scores for the fixed form assessments that constitute the ISASP program in 2019 and 2021. The *ISASP ASR-2019* and *ISASP ASR-2021* also provide tables of frequency distributions and summary statistics for scale scores by grade and subject under the section Frequency Distribution Reports.

Equating and Linking the ISASP Assessments

Equating and linking are procedures that allow test scores to be compared across years. The procedures are generally thought of as statistical processes applied to the results of a test. However, successful equating and linking requires attention to comparability throughout the test development and assembly processes. This section provides some insight into these procedures as they are applied to the ISASP.

Rationale

To maintain the same performance standards across different administrations of a particular test for linear, fixed-form tests, it is necessary for every test to be of comparable difficulty to the previous version. In a summative, standards-based assessment program for accountability, comparable difficulty should be maintained from administration to administration at the total score level, specifically the scores on which accountability decisions rest. Maintaining test form difficulty across administrations is achieved through careful test assembly followed by a statistical procedure called equating. Equating is used to transform the scores of a new test form to the scale of a previously administered test form. Although equating is often thought of as a purely statistical process, a prerequisite for successful equating of test forms is that the forms are built to the same content and psychometric specifications. Without strict adherence to test specifications, the constructs measured by different forms of a test may not be equivalent, thus compromising comparisons of scores across test administrations.

A combination of pre-equating and post-equating quality control checks are used in the ISASP program to assure comparable test scores across administrations. In the pre-equating stage, item-parameter estimates from a prior administration (either field-test or operational) are used to construct new forms of subject-area tests with difficulty levels like those of previous administrations. This approach is possible because of the embedded field-test design that allows for linking field-test items to the operational form. In the post-equating quality control check, item statistics used for test assembly are compared to the same quantities estimated with operational assessment data and used to monitor the comparability of scale scores on the pre-equated forms.

ISASP uses a pre-equating design for all subject-area tests. One of the benefits of online testing is on-demand reporting to support local districts in utilization of results. In a pre-equating design, all items are placed on the base scale prior to an operational administration and the banked item parameters can be used for scoring. The pre-equating design is fully described in the sections that follow.

Pre-Equating

The intent of pre-equating is to produce a test that is psychometrically equivalent to those used in prior years. The pre-equating process calibrates all new field-test items to the base scale, which results in a bank of items used for scoring student responses on the same base scale. In this way, each item is placed on the same metric as the metric of prior years, so the metric is maintained across years. For the ISASP, each new assessment is constructed from a pool of items for which parameters have been equated to the base scale. New items are equated to the base scale during field-test analyses described below.

Test Development, Assembly, and Review for Fixed-Form Assessments

Test construction for ISASP fixed-form assessments in all subject areas begins by selecting the operational items for an administration. Using the items available in the item pool, content specialists and psychometricians from Iowa Testing Programs (ITP) construct new forms by selecting items that meet the content specifications of the Iowa Core in the subject tested and targeted psychometric properties. Psychometric properties targeted include test difficulty, precision, and reliability captured through IRT. The test assembly process is an iterative one, involving ITP faculty and staff, ITP's test delivery contractor, and teams of Iowa teachers with classroom experience who participate on item and test review panels. Because the IRT item parameters for each item in the item bank are maintained on scale, direct comparisons of test characteristic curves and test information functions can be made to ascertain whether a newly assembled test has similar psychometric properties to those of other years. Having all items on the same scale allows the psychometricians to create raw score-to-scale score lookup tables to be used for scoring purposes.

Psychometricians and content staff review the newly constructed test to ensure specifications and difficulty levels have been maintained. Although every item on the test has been previously scrutinized by Iowa educators and curriculum experts for alignment to Iowa Core—a match to test specifications' content limits, grade-level appropriateness, developmental appropriateness, and bias—ITP reexamines these factors for each item on the new test. The difficulty level of the new test form is also evaluated, and items are further examined for their statistical quality, range of difficulties, and spread of information. Test development staff members also review forms to ensure a wide variety of content and situations are represented in the test items, to verify that the test measures a broad sampling of student skills within the content standards, and to minimize "cueing" of an answer based on the content of another item appearing in the test. Additional reviews are designed to verify that keyed answer choices are the only correct answer to an item and that the order of answer choices on the test form varies appropriately. Such quality control steps are also described in Chapter 9 of this manual.

If any of these procedures uncovers an unsatisfactory item, the item is replaced with an item in the item bank and the review process begins again. This process for reviewing each newly constructed test form helps ensure each test will be of the highest possible quality. See Chapter 2 for additional information about test development in the ISASP program.

Anchor Items

To enhance the integrity of the pre-equating design used for the ISASP program, each newly assembled subject-area test includes a set of anchor items, that is, a set of items that were administered in the previous year at the same grade level. Anchor items have item parameters from the operational administration that were used in the scoring of that year's test. The inclusion of item parameters for the common anchor items in the pre-equating process helps to ensure comparability of scores across forms. It should be noted that anchor items for a given year's test forms appear in the same locations as they did in their operational form.

Field-Test Items

When a newly constructed item has survived committee reviews (passage review for Reading, scenario review for Science, and new item review and bias and sensitivity review for Mathematics, Reading, and Science tests), the item is ready for field-testing. For each subject-area test, field-test items are embedded in operational forms, and field-test sets are randomized is such a way that a representative sample of Iowa students in each grade responds to each field-test item. The field-test items are arranged in blocks and, each student is administered only one set of items. For the stimulus-based tests in Reading, Language/Writing, and Science, items appear in testlets along with their respective stimuli, which are placed at pre-selected positions within the test. For example, in a particular grade's ISASP Mathematics administration, there may be anywhere from 20 to 48 different forms (depending on subject area and stimulus-based versus discrete item format) containing the same operational test items in addition to a randomly assigned set of field-test items. The field-test items do not count toward an individual student's score.

In online administrations of fixed forms, forms (e.g., operational plus field-test items) are assigned randomly to students. For example, for grade 5 Science, with a statewide enrollment of approximately 36,000 students, approximately 1,800 students would respond to each of 20 field-test forms. This design provides a diverse and representative sample of student performance on each field-test item. In addition, because students do not know which items are field-test items, no differential motivation effects are expected. To control for fatigue and start-up effects, all field-test items are placed in similar positions on each test form. For the fixed-form paper-based administration of ISASP subject-area tests, there is one operational form that contains a single set of new field-test items.

Critical to the success of a pre-equating design is the degree to which all aspects of the test assembly process, from content and alignment reviews to data collection to calibration and psychometric evaluation, produce test characteristic curves (TCCs) of alternate test forms that are as similar as possible. Figures 6.7 and 6.8 provide examples of the TCCs examined for psychometric comparability during the assembly of the 2021 ISASP forms for the Reading and Mathematics assessments. These curves illustrate alternate test forms virtually identical in overall difficulty. The parts of the IRT scale where they differ slightly are about one-half of a standard deviation away from the cut-scores for Proficient and Advanced performance. These curves are examined as a routine part of the test assembly process as discussed in Chapter 2 of this manual. All TCCs relevant to the assembly of the ISASP 2021 operational tests in ELA, Mathematics, and Science and to the development of the ISASP item pools are provided in *ISASP ASR-2019* and *ISASP ASR-2021*.

Figure 6.7. Test Characteristic Curves for the 2019 and 2021 ISASP Reading Assessments in Grade 6

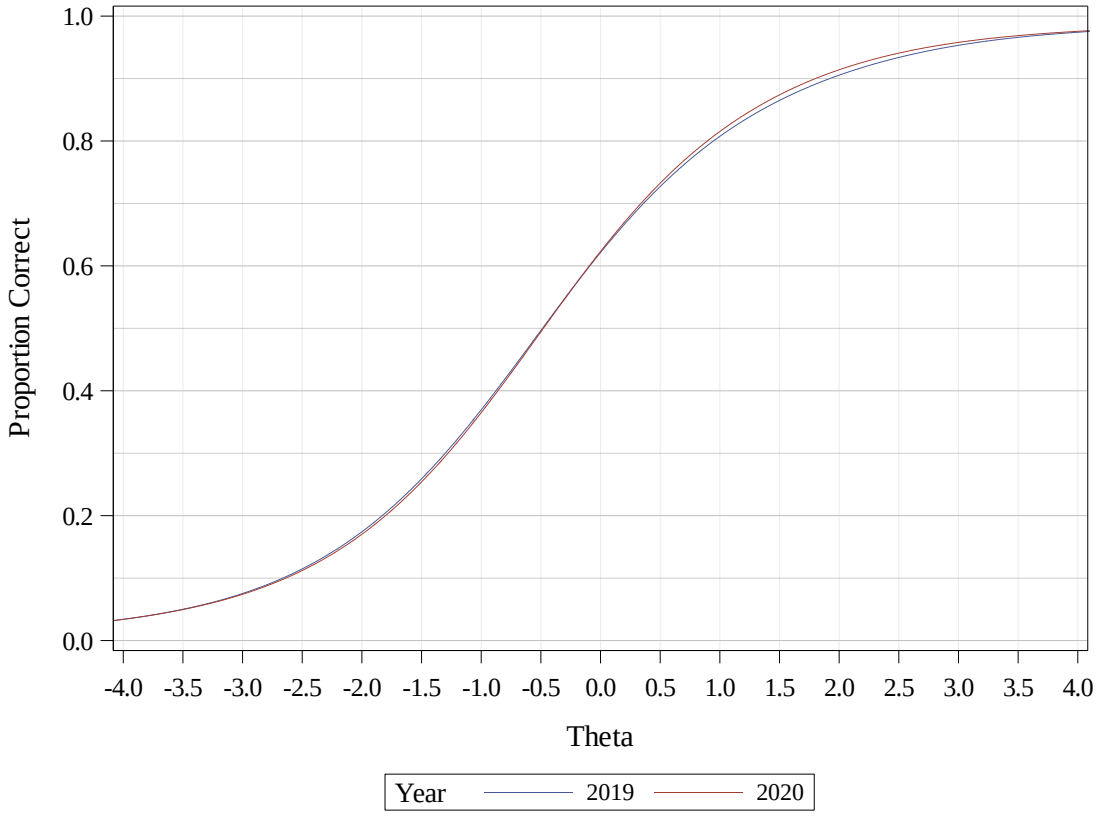
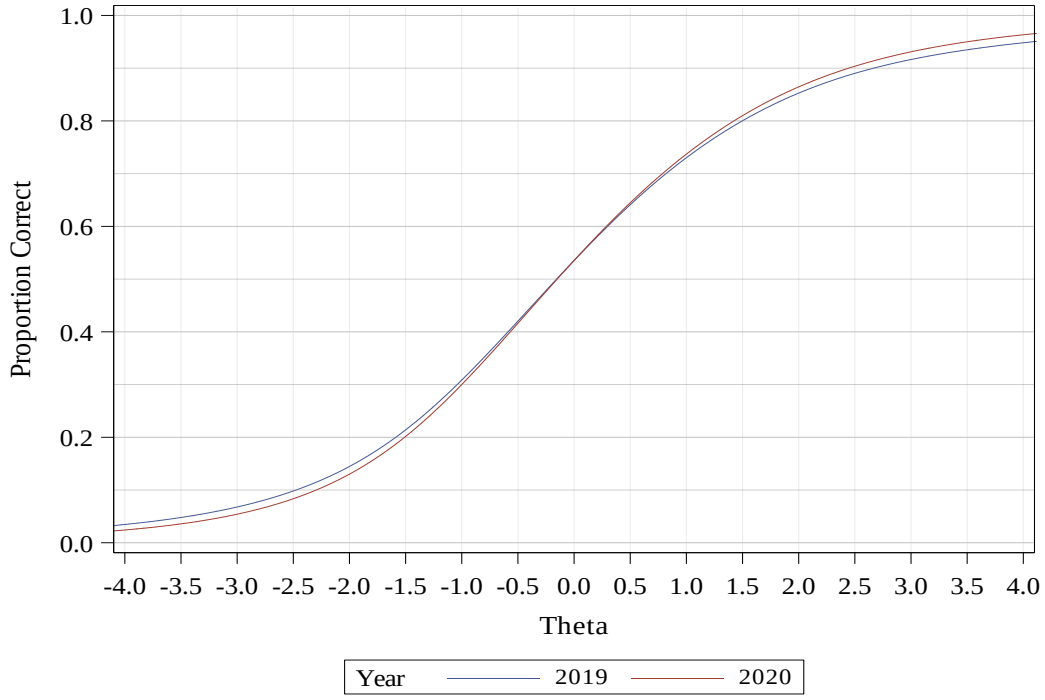


Figure 6.8. Test Characteristic Curves for the 2019 and 2021 ISASP Mathematics Assessments in Grade 8



Chapter 7: Validity

The term *validity* refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (*Standards*, 2014). Validation can be described as the process of collecting evidence to support inferences from assessment results. A primary consideration in validating test scores is determining whether the test measures what it purports to measure: the construct. When a particular individual characteristic is inferred from an assessment result, a generalization, or interpretation in terms of a construct, is being made. For example, problem-solving can be an example of a construct. An inference that students who master the mathematical reasoning portion of an assessment are "good problem-solvers" implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate this is a reasonable and valid use of the scores. During the process of evaluating whether the test measures the construct of interest, several threats to validity must be considered. For example, the test may be differentially more or less difficult for a particular demographic group relative to another group, test scores may have lower than desirable levels of reliability, students may not be properly motivated to perform on the test, or the test content may not span the entire range of the construct to be measured. Any of these threats to validity could compromise the interpretation of test scores.

Beyond ensuring the test is measuring what it is supposed to measure, it is also important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in Chapter 4, "Reports," and Chapter 6, "Scaling."

Demonstrating that a test measures what it is intended to measure and that interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning, and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity has been the subject of considerable research, thought, and debate in the educational measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and as a result, the ways in which test validity and validation are defined have evolved. This chapter summarizes validity evidence for the Iowa Statewide Assessment of Student Progress (ISASP) assessments and is based on the *Standards* (2014).

Test Validity Evidence

The *Standards* (2014) refer to "types of validity evidence, rather than distinct types of validity." The four broad categories of validity evidence mentioned in the *Standards* that are relevant to the ISASP assessments are: evidence based on test content, evidence based on response processes, evidence based on internal structure, and evidence based on relationships with other variables. Taken together, a combination of these types of validity evidence can be used to create a validity argument (Cronbach, 1988; Kane, 2006). It is important to note that the types of validity evidence selected for a given assessment must be relevant to the selected measure, so not every form of validity evidence applies to every assessment.

Evidence Based on Test Content

Content validity evidence addresses whether the test adequately samples the relevant domain of material it purports to measure. If a test is made up of a series of tasks that form a representative sample of a particular domain of tasks, then the test is said to have evidence of content validity. For example, a content-valid test of mathematical ability should be composed of items that allow students to demonstrate their mathematical ability. One way to evaluate the content validity of an assessment such as the ISASP is to evaluate the alignment of the standards with test content.

Generally, achievement tests such as the ISASP assessments are constructed in a way to ensure they have strong evidence of content validity. As documented in Chapter 2, educator committees expend tremendous effort to ensure ISASP assessments are content-valid. Although content validity evidence has limitations and cannot serve as the only evidence for validation, it is an important piece of evidence for the validation of ISASP assessments.

To ensure the content is aligned with the construct, the development of the items is based on *test specifications* for each subject and grade that is being assessed. Rigorous processes have been put in place to align items and test forms with the standards while developing items for ISASP assessments. As a result, each ISASP assessment is developed with content-related validity evidence in mind.

The test specifications as described in Chapter 2 identify eligible test content and provide item count targets for various item properties such as content domains, standards, domains, item types, and depth of knowledge levels. These targets are codified into *test specifications*, which provide direction to item writers, psychometricians, and content specialists from Iowa Testing Programs (ITP) to ensure that all relevant content is sufficiently covered by the assessment. This coverage is one piece of evidence for the content validity of the test.

The items are developed based on the test specifications. The items are rigorously scrutinized during the content review, including extensive reviews both internally and externally. This review checks for the appropriateness of test items, difficulty, clarity, correctness of answer choices, plausibility of the distractors, and fairness of the items and tasks. Then the items must be reviewed and approved by the content review committees, which assure that each item appropriately measures the intended content, is appropriate in difficulty, contains only one correct (or best) answer for multiple-choice questions, and, if an open-ended item, has an appropriate and complete scoring guideline. Next, the items are approved by a bias and sensitivity committee, which reviews the item for language, or content, that may be inappropriate or offensive to students, parents, or community members, or that contains stereotypical or biased references to gender, ethnicity, or culture.

The Human Resources Research Organization (HumRRO), an external independent contractor, conducted an alignment study to provide evidence that the ISASP tests were aligned with its respective set of test specifications. Specific areas of interest included both how much and what type of content is covered by the assessment, as well as whether students are asked to demonstrate knowledge at the same level of rigor as expected in the Iowa Core.

Evidence Based on Response Processes

Validity evidence based on response processes involves explicit assumptions about the cognitive processes in which the test takers engage. Analyses of the response processes of test takers provide evidence concerning the fit between the construct and the nature of the performance or response required of the test takers (*Standards*, 2014). The test specifications discussed previously include the item targets for each of the three depth of knowledge (DOK) levels for all ISASP tests. DOK, or cognitive complexity, refers to the cognitive demand associated with an item. The level of cognitive demand focuses on the type and level of thinking and reasoning

required of the student when interacting with a particular item. Levels of cognitive complexity for ISASP are based on Norman L. Webb's (1997) DOK levels. Three levels of cognitive complexity are represented on the ISASP. Additional detail concerning the balance of cognitive levels can be found in Chapter 2.

For ISASP, each item is developed to strictly adhere to one of the first three DOK levels and is reviewed internally by content teams. Qualified teachers and community members, who interact with students in the classroom, review and verify the DOK levels of each field-test item. Of particular concern is the development of items that contain no irrelevant information that may interfere with how the item is interpreted or scored. The test specification review committees, who have experience working with students and their cognitive processes daily, determine what proportions of the test that should be devoted to items at each of the first three levels of DOK.

Evidence Based on Internal Structure

Internal structure validity evidence shows the degree to which items and test components conform to the construct on which the proposed test score interpretations are based. For instance, an ISASP Mathematics test may be broken into several domains such as data analysis, algebra, geometry and measurement, and numbers and operations. Internal structure validity evidence identifies the degree to which the item relationships conform to the individual subscales and overall mathematics scale.

One type of evidence for internal structure provided for all ISASP assessments is a dimensionality analysis using the method of principal components. This approach to dimensionality identifies a number of components that best explain the relationships among the items as they are used to define domain scores (see Chapters 2 and 6). It is common for educational assessments to measure more than one dimension, but generally these tests at the same time organize content into clusters based on the Iowa Core. Each of the ISASP assessments is designed to measure a multifaceted composite of knowledge and skills appropriate for the subject and grade. This composite of knowledge and skills is expected to be composed of components that are separate and identifiable in terms of content but highly correlated, such that the measured composite can be considered as a unidimensional construct, thus permitting the use of unidimensional item-response theory (IRT) models.

A Principal Component Analysis of domain scores is annually conducted on the ISASP assessments, and results for all grades and subjects of the ISASP can be found in the *ISASP Annual Statistical Report (ISASP ASR-2019* and *ISASP ASR-2021)* under the section heading "Dimensionality Reports" located on the ISASP portal. Dimensionality results reported in the *ISASP ASR-2019* and *ISASP ASR-2021* include the ratio of the first to the second eigenvalue, the ratio of the second to the third eigenvalue, and the proportion of variance accounted for by the first eigenvalue. Various rules of thumb have been proposed in the research literature to help interpret these measures. Various authors (e.g., Kaiser, 1960; Morizot, Ainswrorth & Reise, 2007) give the rule of thumb that if the ratio of the first to second eigenvalue exceeds a value of three, unidimensionality is indicated. Other authors argue that when the ratio of the second to third eigenvalue is markedly smaller than the ratio of the first to second, evidence of unidimensionality for internal structure is strengthened (Cattell, 1966). As shown in analyses reported in the *ISASP ASR-2019* and *ISASP ASR-2021,* ISASP eigenvalue ratios generally always satisfy these criteria, implying the tests are unidimensional.

Regarding the percent of variance accounted for by the first dimension, since the first principal component explains the maximum variance, then the percentage of total variance explained by the first principal component is often regarded as an index of essential unidimensionality. The higher percentage of total variance the first principal component accounts for, the closer the test is to essential unidimensionality. ISASP tests generally show the first eigenvalue accounting for 55 to 75 percent of the total variance. Both the eigenvalue ratios and the proportion of variance accounted for by the first dimension provide support for reporting a single scale score of each of the ISASP's core assessments (i.e., Reading, Language/Writing, Mathematics, and Science). In

addition, analyses reported in the *ISASP ASR-2019* and *ISASP ASR-2021* provide support that the ISASP scale score in English Language Arts (ELA) represents an essentially unidimensional composite.

In addition to the Principal Component Analysis, the unidimensional composite for the ISASP fixed form assessments can be investigated at the item level through the item-total correlation. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation is the correlation between an item and the total test score excluding that item. Conceptually, if an item has a high item-total correlation (i.e., .30 or above), then students who performed well on the test tended to answer the item correctly and students who performed poorly on the test tended to answer the item incorrectly, that is, the item discriminated well between high-scoring and low-scoring students. Assuming the total test score represents the extent to which a student possesses the skills or knowledge being measured by the test, high item-total correlations indicate the items on the test require proficiency in these skills or knowledge to be answered correctly. The *ISASP ASR-2019* and *ISASP ASR-2021* present summary data on item-total correlations in tables under the section heading "Item Statistics Reports." For ISASP assessments, mean item-total correlations are generally high and the ranges of these statistics typically show minimum values at or above .30.

Additionally, to provide further evidence of the internal structure of the test, correlations among the total test score and domain scores are provided. These correlations quantify the relationships among content strands of the Iowa Core and the overall test score. The overall test score is represented by the ISASP scale score. These correlations demonstrate that the content domains comprising the overall test are highly related (as demonstrated through high correlations) to the overall test while also distinct in the factors they are measuring. Put another way, high correlations are indicative that the assessment is measuring one underlying construct. As can be referenced in the correlation tables in the *ISASP ASR-2019* and *ISASP ASR-2021*, there are high correlations between the scale score and the domain scores for each of the grades, while there are moderate-to-high correlations among the domain scores themselves. The correlation tables are provided for ISASP Reading, Language/Writing, Mathematics, and Science assessments under the section heading "Internal Consistency Reports."

The dimensionality analysis examines the number of factors measured by the items, the item-total correlations investigate the consistency of students' performance on an item to their overall test scores, and the correlations among the total scale score (or raw score for fixed-form tests) and the domain provide evidence that the domain scores are highly related to the total test score, but less related to each other. Together, these three pieces of evidence collectively demonstrate the structure of the test can be measured using a unidimensional composite.

To further characterize the internal structure of the ISASP assessments, procedures for covariance structure modeling of item parcels defined by content domains of the Iowa Core were implemented for each grade and content area. These confirmatory factor models were specified to be consistent with the domain structure of the Iowa Core and the reported domain scores. For example, in the grade 8 assessment in Mathematics, domain scores are reported in categories labeled The Number System, Expressions and Equations, Functions, Geometry, and Statistics and Probability. Items comprising each of these domains were assembled into item parcels, and the covariances among the item parcels were fit to a CFA model with the Bifactor structure (Reise, 2012). In the Bifactor model, each variable loads on a general dimension that reflects the construct measured by the assessment and on one additional dimension that represents the domain of the Iowa Core to which the items align. For the grade 8 Mathematics assessment, for example, the Bifactor model has 6 dimensions, one general mathematics factor shared by all the items and five domain factors shared by items within each domain. The goodness-of-fit results of these analyses of internal domain structure of the ISASP assessments are summarized in Table 7.1.

The goodness-of-fit statistics of the Bifactor models presented in Table 7.1 provide strong evidence of validity with respect to internal structure. The combination of domain factors corresponding to the multiple dimensions of the Iowa Core and the general dimension reflecting the common construct defined by the content domains supports the breakout reports feedback to teachers and students as well as the standards-based achievement level reports for purposes of accountability.

Table 7.1 Goodness-of-Fit of Bifactor Internal Structure Models for ISASP Domain Scores Based on the Iowa Core Standards

ISASP Test	Model Fit	3	4	5	6	7	8	9	10	11
Reading	CFI	.997	.987	.998	.999	.998	.998	.998	.995	.997
	SRMS	.009	.016	.006	.005	.006	.006	.007	.010	.008
	RMSEA									
Language/ Writing	CFI	.888	.931	.924	.904	.873	.915	.922	.927	.907
	SRMS	.078	.061	.069	.065	.085	.080	.063	.062	.079
	RMSEA									
Mathematics	CFI	.992	.932	.990	.990	.992	.983	.987	.992	.988
	SRMS	.012	.035	.013	.014	.012	.016	.015	.012	.016
	RMSEA									
Science	CFI	-	-	.980	-	-	.987	-	.988	-
	SRMS	-	-	.016	-	-	.014	-	.014	-
	RMSEA	-	-	-	-	-	-	-	-	-

Evidence Based on Relations to Other Variables

The *Standards* (2014) highlight that often, the interpretation or use of a particular measure can be validated by comparison to other measures of the same or a related construct. Criterion validity relies upon the demonstration of a relationship between the test and an external criterion measure. If the test is intended to measure mathematic performance, for example, then scores from the test should correlate substantially with measures that require mathematical performance to achieve a high score. Given that the ISASP has just completed one administrative year, opportunities to examine the relationship with other assessments are limited. However, as a more complete research agenda is being planned to expand evidence based on relations to other variables, two assessments are presented below as preliminary evidence.

Correlations Between the ISASP and the Iowa Assessments

Validity evidence supports the interpretation and use of test scores for a particular purpose. Assessment information is not considered valid or invalid in any absolute sense. Rather, the information is considered valid for a particular use or interpretation and invalid for another. A comprehensive approach to the collection of validity is an integral part of any assessment. Concurrent validity evidence is one critical piece of validity evidence; it summarizes the degree of similarity between two assessments taken at approximately the same time during the school year.

The evidence is presented in the form of correlations between scores on the ISASP and the *Iowa Assessments*. The *Iowa Assessments* were the previous state achievement tests used in Iowa for federal accountability. The

Iowa Assessments are aligned to college and career readiness standards and designed to identify students' strengths and weaknesses, monitor growth, and predict future performance. They have a longstanding history, including strong validity evidence related to predicting college readiness and correlations with ACT scores.

For these comparisons, students' scores from the 2019 ISASP were matched to their scores on the 2018 *Iowa Assessments* and correlations were calculated. Specifically, the correlations compare the following tests: Reading to Reading, English Language Arts (ELA) to Reading, Mathematics to Mathematics, and Science to Science, on the 2019 ISASP and the 2018 *Iowa Assessments*, respectively. The student match rate was above 95 percent per grade. ELA on the 2019 ISASP was compared to Reading on the 2018 Iowa Assessments because the *Iowa Assessments* do not provide a test comparable to the Language/Writing 2019 ISASP. The correlations among tests are .75 and above, except for Science. Because Grade 2 was not required on the *Iowa Assessments*, there was not a suitable random sample to provide the correlations for Grade 3. The strong correlations seen in Table 7.2 confirm the expected relationships between similar tests on the ISASP and the *Iowa Assessments.* This supports evidence of convergent validity, as the general constructs of Reading, Mathematics, and Science are defined similarly on both tests.

Table 7.2. Correlations Between Student Standard Scores on the 2019 ISASP and 2018 *Iowa Assessments*

Grade	Reading	ELA	Mathematics	Science
4	0.76	0.75	0.76	.
5	0.77	0.76	0.78	0.70
6	0.78	0.79	0.79	.
7	0.77	0.80	0.83	.
8	0.78	0.80	0.82	0.72
9	0.76	0.79	0.81	.
10	0.75	0.79	0.77	0.75
11	0.75	0.78	0.78	.

Correlations Between the ISASP and FAST Assessments

FAST™ Curriculum-Based Measurement for Reading (FAST™ CBMreading) is a universal screening test in reading given to all students in the state of Iowa. Students read aloud for one minute from grade- or instructional-level passages. The words read correctly per minute is generated from this assessment. Validity coefficients that describe the relationship between ISASP tests and FAST CBMreading for students in grade 3 are provided in Table 7.3 for the 2018–2019 and 2017–2018 years.

A sample of Iowa educators choose to administer a computer adaptive test called FAST aReading. This assessment measures multiple reading skills over a 30-minute testing session. Table 7.4 provides validity coefficients that describe the relationship between ISASP tests and the FAST aReading.

Table 7.3. Correlations Between Student Standard Scores on the 2019 ISASP and FAST CBMreading

	Sample Size	Reading	Language/Writing	ELA	Mathematics
2017–2018	26,922	0.65	0.63	0.69	0.53
2018–2019	26,279	0.64	0.65	0.70	0.54

Table 7.4. Correlations Between Student Standard Scores on the 2019 ISASP and FAST aReading

	Sample Size	Reading	Language/Writing	ELA	Mathematics
2017–2018	8,148	0.75	0.66	0.76	0.64
2018–2019	9,717	0.74	0.70	0.78	0.65

Validity Evidence for Different Student Populations

In addition, internal structure evidence should show that individual items are functioning similarly for different demographic subgroups within the population being measured. ISASP measures the Iowa Core Standards that are taught to all students. In other words, the tests have the same content validity for all students because what is taught to all students is measured for all students. Great care has been taken to ensure the ISASP items are representative of the content domain expressed in the content standards. Special attention is given to find evidence that construct-irrelevant content has not been inadvertently included in the test, as such content could result in an unfair advantage for one group versus another. Both judgmental and statistical methods are used to identify and remove such items from use, to mitigate their impact on any of the demographic subgroups that make up the population of the state of Iowa.

As described in Chapter 2, this begins with item writers trained on how to avoid economic, regional, cultural, and ethnic bias when writing items. After items have been written, they are reviewed by a bias and sensitivity committee, which evaluates each item to identify language or content that might be inappropriate or offensive to students, parents, or other community members or that contain stereotypical or biased references to gender, ethnic, or cultural groups. The bias and sensitivity committee accepts, edits, or rejects each item for use prior to the items' initial (field-test) administration.

Differential item functioning (DIF) analyses are conducted for the purpose of identifying items that are differentially difficult for different subpopulations of individuals. Refer to Chapter 2 for more details about DIF and the method used to flag items that function differently. Though DIF analyses flag items as being differentially difficult for one group as compared to another, they do not solely provide sufficient evidence for removing the item from use. Flagged items are examined during "Data Review" meetings that take place after the initial (field-test) administration of each item. Items are removed from use only when the data review committee identifies a concrete reason for the DIF, such as bias or sensitive content.

These multiple reviews are a critical component of the item and test development process. They support the validity of the test for the diverse populations that make up the state of Iowa.

Evidence of Comparability across Modes of Administration

When different administration modes (paper vs online) are permitted within an assessment program like ISASP, there is a desire to maintain comparability or score equivalence (which ensures scores across modes can be

treated similarly). Multiple criteria for evaluating the comparability of psychometric properties between online and paper/pencil assessments can be used. Provided below is evidence that evaluates comparability from the construct perspective, test specifications perspective, DIF perspective, and technical characteristics.

Construct Comparability

To address the question of construct comparability with respect to paper-based and online administrations of the ISASP assessments, methods for the evaluation of measurement invariance were used (Meredith, 1993; Stark, Chernyshenko & Drasgow, 2006). For each ISASP assessment, the models for internal domain structure described previously and summarized in Table 7.1 were applied separately to data from paper-based and computer-based administrations to examine measurement invariance by mode of administration. Models of measurement invariance vary with respect to the degree of invariance they specify. A model of weak, or configural, invariance specifies that the underlying constructs measured by the domains of the Iowa Core are defined in the same way (i.e., have the same internal structure) for paper-based and computer-based administrations, but the relative contributions of the domains (e.g., Key Ideas and Details, Craft and Structure, and Integration of Knowledge and Ideas) to the overall construct (e.g., Reading) may be different. A stronger type of measurement invariance, referred to as metric invariance, adds to the configural property equal contributions of the domains by mode of administration, which has implications for comparability in scoring in addition to comparability in terms of construct representation.

Table 7.5 summarizes the results of the goodness-of-fit statistics for models of metric invariance. These results were used to evaluate the degree to which the paper-based and computer-based administrations of the ISASP assessments show evidence of construct comparability (i.e., that the internal structure is not influenced by construct-irrelevant variance due to mode of administration). Although there is slight variation in the values of these statistics, they show an extremely high degree of consistency, leading to the conclusion that metric invariance with respect to administration mode is a property shared by all the ISASP assessments in the 2019 administration.

Table 7.5. Goodness-of-Fit Statistics for Structure Models of Measurement Metric Invariance for ISASP Domain Scores for Computer-Based and Paper-Based Test Administrations

ISASP Test	Model Fit	3	4	5	6	7	8	9	10	11
Reading	CFI	.999	.998	.999	.999	.999	.997	.999	.997	.999
	SRMS	.005	.025	.008	.007	.002	.032	.007	.028	.007
	RMSEA	.010	.050	.020	.013	.000	.062	.018	.065	.017
Language/ Writing	CFI	.989	.989	.991	.988	.991	.994	.994	.991	.991
	SRMS	.020	.010	.014	.019	.020	.018	.018	.014	.021
	RMSEA	.063	.065	.059	.069	.060	.058	.050	.063	.068
Mathematics	CFI	.998	.995	.998	.997	.999	.994	.999	.999	.999
	SRMS	.008	.015	.015	.018	.011	.020	.011	.010	.010
	RMSEA	.025	.040	.026	.028	.017	.045	.014	.018	.023
Science	CFI	-	-	.999	-	-	.999	-	.999	-
	SRMS	-	-	.006	-	-	.004	-	.013	-
	RMSEA	-	-	.012	-	-	.008	-	.032	-

*Note: For the Language/Writing assessments, measurement invariance models used separate scores from the Language and Writing sections to define the latent variables representing the underlying construct for the assessment.

The online and paper/pencil version of the ISASP tests were identical in terms of content and cognitive test specifications. That is, the distribution of items by DOK level did not vary across modes, nor did the distribution of items by content domain vary across modes. In the online version, when technology-enhanced (TE) items were used, items were selected that minimized the difference in item-level statistics (a-, b-, and c-parameters as well as classical difficulty and discrimination) between the TE and the paper/pencil version of the same item.

Differential Item Functioning

All items on the 2019 and 2021 ISASP forms were analyzed for DIF with online and paper/pencil test takers representing the reference and focal groups. The results suggest that a minimum number of items were identified as functioning differently between the two groups. These results are summarized in Table 7.5. The complete results of this DIF analysis are provided in the *ISASP ASR-2019* and *ISASP ASR-2021*.

Table 7.6. Number of C-DIF Flagged Items for Mode

Grade	Reading	Language/ Writing	Mathematics	Science
3	1	3	1	
4	2	1	0	
5	2	0	2	1
6	1	0	1	
7	0	2	1	
8	2	0	3	2
9	2	1	1	
10	1	0	0	3
11	1	0	2	

Technical Characteristics

Reliability estimates were generated for both the online and paper/pencil versions of the ISASP for the total samples as well as by groups of test takers. The *ISASP ASR-2019* and *ISASP ASR-2021* Summary Statistics for each test provide this information. Coefficient alphas were consistent across mode by groups of test takers.

Chapter 8: Reliability

Chapter 8 reviews several different estimates of reliability. These estimates can help users make informed judgments about the consistency of the Iowa Statewide Assessment of Student Progress (ISASP) scores. Specifically, this section addresses reliability coefficients and standard errors of measurement (SEM). Several approaches to the assessment of reliability and sources of variance in observed scores are presented, as well as standard errors of measurement for select score levels, also known as conditional SEMs. In addition, interrater reliability, classification consistency, and classification accuracy estimates are provided. Overall, this chapter presents evidence that demonstrates the high reliability of ISASP scores. Reliability is essential for assessing student learning outcomes. A soundly planned, carefully constructed and comprehensive large-scale assessment represents an accurate and dependable measure of student achievement available to parents, teachers, and school officials.

Definition of Reliability

Reliability is the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations in performance due to chance. Thus, reliability is the consistency of the scores across conditions that can be assumed to differ at random. In statistical terms, the variance in the distributions of test scores, a measure of the differences among individuals, is partly due to real differences in the knowledge, skill, or ability being tested ("true variance") and partly due to random differences in the measurement process ("error variance"). Reliability is an estimate of the proportion of the total variance that is true variance, or

$$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_E^2} \tag{8-1}$$

When defining reliability, it is helpful to first examine classical test theory. Classical test theory states that one's observed score is a combination of true score and error (see equation 6-1). That is, individuals are assumed to have a true score on a test and the true score is the actual amount of knowledge of the content being measured by the test. It is also assumed that observed scores contain a certain amount of measurement error. Good test practices demand that care is taken to ensure consistency in administration, scoring, and testing conditions to help reduce the measurement error.

Estimating Reliability

There are several ways of estimating reliability, including test-retest, alternate forms, and internal consistency methods. The primary type of reliability reported in this technical manual is an internal consistency measure, coefficient alpha, which is derived from analysis of individuals' consistency of performance across items within a test. Coefficient alpha was chosen as it is the most common measure of internal consistency and requires only one administration of the test.

Coefficient alpha, α, was developed by Lee Cronbach in 1951 (Cronbach, 1951). This statistic is appropriate when the test is relatively homogenous. See the Dimensionality Reports in the *ISASP ASR-2019* and *ISASP ASR-2021* for documentation on the homogeneity of the ISASP.

The formula for coefficient alpha is:

$$\alpha = (\frac{N}{N-1})(1 - \frac{\Sigma_{i=1}^{N} S_{Y_i}^2}{S_X^2}) \quad (8\text{-}2)$$

where

N= number of items on the test,

i references the specific item,

$S_{Y_i}^2$ = the sample variance of the i^{th} item, and

S_X^2 = the observed score sample variance for the test.

In numerical value, the reliability coefficient is between 0.00 and 1.00; for standardized assessments it is generally between .60 and .95. The closer the coefficient approaches the upper limit, the greater the freedom of the scores from the influence of factors that affect student performance and obscure real differences in achievement. That is, the higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain very similar scores upon repeated testing occasions (if the test takers do not change in their level of the knowledge or skills measured by the test). Other things being equal, the more items a test includes, the higher the internal consistency. This ready frame of reference for reliability coefficients is deceptive in its simplicity, however.

The *ISASP ASR-2019* and *ISASP ASR-2021* provide means, standard deviations, and reliability for Reading, Language, ELA, Mathematics, and Science under the section "Summary Statistics Reports" for each grade. The reliability estimates are all in the ranges expected for standardized tests. Note that the ELA reliability estimates reflect a composite reliability calculated using the Reading and Language alpha estimates and assume the variance of the tests were the same. Because reliability can vary within subgroups, the *Standards* (2014) call for estimates of reliability for each subgroup, as feasible. Therefore, separate estimates of reliability are also provided for female, male, African American, Hispanic, American Indian/Alaskan Native, Native Hawaiian/Pacific Islander, White, and multiracial groups, as well as by status of individual education plans, English Language Learner, and free and reduced lunch.

Standard Error of Measurement

Test reliability is also commonly described by the Standard Error of Measurement (SEM). The SEM is defined as the standard deviation of measurement errors associated with observed test scores for a specified group of test takers. In classical test theory, the SEM is a function of the observed score variance, σ_X^2, and the test reliability, $\rho_{xx'}$. Standard error of measurement is calculated using the following formula:

$$SEM = \sqrt{\sigma_X^2(1 - \rho_{xx'})}. \quad (8\text{-}3)$$

The formula for computing the SEM demonstrates how the estimate of reliability and the SEM are related. A SEM band can be placed around a student's scale score on the ISASP and would result in a range of values most likely to contain the student's observed scale score upon replication.

Table 8.1 provides reliability estimates and SEMs for the ISASP. These estimates were generated from the total testing population in the spring of 2019.

Table 8.1. Estimates of Reliability and Standard Errors of Measurement for 2019 ISASP

Grade	Reliability Index	Reading	Language/ Writing	English Language Arts	Mathematics	Science
3	Reliability	0.88	0.84	0.92	0.88	.
	SEM	10.4	11.3	7.6	9.9	.
4	Reliability	0.86	0.83	0.91	0.88	.
	SEM	12.9	14.1	9.6	11.9	.
5	Reliability	0.87	0.83	0.91	0.89	0.80
	SEM	14.9	16.9	11.4	13.7	18.3
6	Reliability	0.88	0.84	0.92	0.87	.
	SEM	15.9	18.1	12	16.4	.
7	Reliability	0.88	0.87	0.93	0.88	.
	SEM	18.2	18	12.7	17.3	.
8	Reliability	0.87	0.87	0.93	0.90	0.80
	SEM	20	19.8	13.7	17.4	24.5
9	Reliability	0.85	0.86	0.92	0.86	.

	SEM	23.7	22.2	15.9	22.1	.
10	Reliability	0.85	0.87	0.92	0.87	0.87
	SEM	23.5	21.8	16	21.7	21.9
11	Reliability	0.87	0.88	0.93	0.91	.
	SEM	22.5	20.7	15.2	18.9	.

Conditional Standard Error of Measurement

The SEM measures the net effect of all factors leading to inconsistency in student test scores and to inconsistency in score interpretation. It is reported as the typical amount by which a student's observed score may range from one testing occasion to another. The conditional SEM (CSEM) gives similar information, but rather than gauging the typical range, it provides a range that is tailored to a specific level of achievement (Feldt & Brennan, 1989; Haertel, 2006). The CSEM is interpreted similarly to the SEM, as described above, but for a specific score or score range.

CSEMs based on a single test administration were estimated using several procedures identified by previous studies to yield similar results (e.g., Brennan & Lee, 1997; Feldt & Qualls, 1998). Because the methods agreed closely, only the results of the Feldt & Qualls (1998) procedure are reported for the ISASP. See the Frequency Distribution Reports in the *ISASP ASR-2019* for the CSEM of scale scores for each grade and subject.

The CSEMs provide direct evidence of the precision of ISASP scores across the full performance continuum. Tables 8.2 to 8.5 contain CSEMs at five score points on each ISASP assessment over a range that includes the middle 80 percent of the statewide score distribution. The tables also include the overall SEM described previously and the standard deviation of the ISASP scale score (ISS) for each grade. As can be seen from the table, the CSEMs describe a range of measurement precision above and below the overall SEM. Relative to the standard deviation of the ISS at each grade, the CSEMs demonstrate that the expected variability around a given observed score (a measure of measurement precision) is markedly less than the variability of the ISSs scores overall. The CSEMs in Tables 8.2 to 8.5 are roughly one-fifth to two-fifths of a within-grade standard deviation. CSEMs of this magnitude compare favorably to those observed in other fixed-form assessments of general student achievement (Dunbar & Welch, 2014; Minnesota Department of Education, 2018). Their magnitudes also exhibit the same pattern over the score scale as that observed in some standards-based, adaptive assessments used for state accountability. The CSEMs reported by the Smarter-Balanced Assessment Consortium (SBAC), for example, for adaptive assessments in ELA and Mathematics average about three-tenths of a within-grade standard deviation (see Smarter Balanced Summative Assessments, 2014, 2017).

Table 8.2. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Reading Assessment

Reading		CSEM					
Grade	**SEM**	**P_{10}**	**P_{25}**	**P_{50}**	**P_{75}**	**P_{90}**	**SD**
3	10.4	9.3	10.0	12.0	13.0	17.5	28.7
4	12.9	12.5	14.0	15.6	16.6	17.9	34.4
5	14.9	12.5	14.5	16.4	18.1	19.2	41.3
6	15.9	13.9	15.8	17.5	19.2	23.0	45.5
7	18.2	14.9	17.0	19.8	21.8	32.1	50.0
8	20.0	17.8	20.8	23.1	25.0	28.7	55.0
9	23.7	20.8	23.8	26.3	28.5	35.8	60.5
10	23.5	19.5	22.9	26.0	28.2	30.9	60.5
11	22.5	18.8	22.5	25.0	27.1	38.2	60.5

Table 8.3. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Language/Writing Assessment

Language/Writing		CSEM					
Grade	**SEM**	**P_{10}**	**P_{25}**	**P_{50}**	**P_{75}**	**P_{90}**	**SD**
3	11.3	11.6	14.1	14.7	14.5	13.9	28.7
4	14.1	15.8	16.4	17.4	17.0	15.7	34.4
5	16.9	19.7	21.1	21.0	20.5	19.3	41.3
6	18.1	20.6	22.1	22.5	21.9	20.6	45.5
7	18.0	19.1	20.5	21.4	20.1	19.1	50.0
8	19.8	20.7	22.4	22.7	21.5	19.1	55.0
9	22.2	25.3	28.0	28.1	26.9	25.0	60.5
10	21.8	22.8	24.5	24.5	23.2	21.0	60.5
11	20.7	22.0	25.5	24.8	23.2	21.4	60.5

Table 8.4. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Mathematics Assessment

Mathematics		CSEM					
Grade	**SEM**	**P_{10}**	**P_{25}**	**P_{50}**	**P_{75}**	**P_{90}**	**SD**
3	9.9	8.0	9.3	10.6	11.3	12.1	28.7
4	11.9	9.4	11.2	12.3	13.3	14.4	34.4
5	13.7	11.7	12.6	14.5	15.9	17.4	41.3
6	16.4	14.3	15.9	17.6	19.0	20.3	45.5
7	17.3	14.5	16.3	17.9	19.7	21.1	50.0
8	17.4	15.0	15.9	17.9	20.2	21.0	55.0
9	22.1	14.2	20.7	24.0	26.3	27.7	60.5
10	21.7	16.9	19.5	22.4	24.9	26.9	60.5
11	18.9	13.2	15.9	19.3	21.5	27.2	60.5

Table 8.5. Conditional Standard Errors of Measurement at Selected Percentiles of the ISASP Science Assessment

Science	CSEM						
Grade	**SEM**	**P_{10}**	**P_{25}**	**P_{50}**	**P_{75}**	**P_{90}**	**SD**
5	18.3	16.3	18.5	20.6	22.0	23.5	41.3
8	24.5	20.9	23.5	26.4	28.6	31.0	55.0
10	21.9	17.3	19.6	23.3	25.3	26.6	60.5

As described in Chapter 6, Scaling and Equating, the item-response theory (IRT) calibrations of items also provide evidence of the precision of measurement at a particular achievement level. For the 2019 and 2021 ISASP assessments, IRT test information was examined for each form to ensure that each test could adequately capture student achievement across the full performance continuum. The inverse of the IRT information function for each ISASP assessment gives another type of CSEM, sometimes called an examinee-level CSEM, based on IRT formula scoring or the maximum likelihood estimate of the underlying achievement construct. Graphical displays of the ISASP test information functions and examinee-level CSEMs are provided in the *ISASP ASR-2019* and *ISASP ASR-2021* under the sections labeled Form Reports.

The table below summarizes the IRT CSEM values for the Not-Yet-Proficient, Proficient and Advanced achievement levels used in the ISASP program for the Iowa School Performance Profile and federal accountability. These CSEMs are based on the results of the 2021 ISASP in grades 3 through 11 and indicate substantial measurement precision at each achievement level for all grades and tests. Because the IRT CSEMs are based on fixed forms of the ISASP assessments, the values in the table represent lower bounds for the CSEMs that will be obtained when the ISASP moves to adaptive testing in the 2022 program year.

Table 8.6. CSEM of Theta by Achievement Level

	Mathematics			ELA			Science		
Grade	**NP**	**P**	**A**	**NP**	**P**	**A**	**NP**	**P**	**A**
3	0.08	0.08	0.14	0.06	0.07	0.10			
4	0.08	0.08	0.25	0.06	0.08	0.15			
5	0.08	0.08	0.17	0.06	0.08	0.21	0.11	0.12	0.35
6	0.06	0.08	0.20	0.06	0.07	0.21			
7	0.08	0.09	0.25	0.07	0.08	0.18			
8	0.07	0.07	0.18	0.05	0.07	0.21	0.14	0.15	0.24
9	0.10	0.09	0.22	0.07	0.08	0.27			
10	0.11	0.10	0.18	0.07	0.08	0.17	0.11	0.13	0.33
11	0.07	0.07	0.14	0.05	0.07	0.27			

Interrater Reliability

The Iowa Statewide Assessment of Student Progress (ISASP) assesses student's understanding of core content domains in English Language Arts (ELA), Mathematics, and Science. ISASP assessments include multiple-choice and technology-enhanced items, constructed-response items, and open-ended essay response questions. Constructed-response items and open-ended essay response questions utilize human and machine scoring.

For the ISASP program, constructed-response items occur in Science and the Reading portion of ELA, and open-ended essay questions occur in the Writing portion of ELA. Scores for these types of items on ISASP can be assigned by humans, machine scoring engines, or both. Interrater reliability is the reliability of the scoring process for all types of constructed response items. This document will outline the rigorous process followed to ensure accurate and reliable scores on these items.

Interrater reliability is estimated from the agreement between individual raters (scorers). The interrater reliability coefficient answers the question, "How consistent would the scores of these test takers be over replication of scoring of the same responses by different raters?" Rater agreement or consistency is critical for valid test score interpretation of assessments requiring human raters to rate constructed responses. When two trained raters independently assign the same score (or rating) to a test taker's item response, there is evidence that the scoring rubric is being applied consistently. Double scoring substantially increases the reliability of the scoring process. Double scoring is used to monitor and evaluate the accuracy of rating.

Interrater reliability is evaluated empirically by three different statistics: a) percentage perfect agreement between two raters, b) percentage adjacent agreement, and c) correlation between two raters.

Scoring System

ISASP utilizes a rigorous scoring system for the constructed-response items with an extensive schedule of quality checks throughout the process. Scorer training for the constructed-response items and writing prompts in each grade are completed using Pearson's training system, which follows the same general process used in scoring the ISASP selected-response items.

Pearson uses industry-standard scanning technology to capture, parse, and send data to the scoring processes (machine, human, or automated). Pearson's quality management systems are ISO certified; processes are replicable across facilities with accurate and predictable results. During each step, Pearson monitors quality, using accurate imaging and scanning systems it has used on behalf of national testing programs, like Iowa Testing Programs (ITP), for decades.

Selected-Response

Pearson's machine scoring systems score selected-response (SR) items. The quality assurance group verifies that student item answers correspond to the response recorded in the database in pre-score and scored student data files. Pearson verifies SR scoring against ISASP requirements, the test map, and item keys. Pearson then validates individual student's derived scores per level of the test. This process includes reviewing all score-value-related fields—raw scores, object scores, strand scores, performance levels, pass/fail indicators, attempt rules, and scale scores—against the tables from the psychometric team.

Constructed Response

Pearson uses innovative human-automated scoring process to score online constructed response items through their engine, Intelligent Essay Assessor (IEA). In real time, responses are scored by humans and the automated scoring engine, and results are evaluated for reliability. Before the test administration, Pearson first trains the engine using 1,000 field test responses scored by humans; these responses are also all second scored to ensure accuracy in training the engine. Next, Pearson trains the engine using human-scored operational responses.

Pearson's content staff conducts in-person rangefinding in Iowa and uses approved materials reflecting committee decisions to build training sets. Using Pearson's proven process for hiring scorers, all scorers will meet ITP requirements. Trainees who fail qualification will be dismissed; trainees might be dismissed during scoring if their performance falls below ITP requirements. Pearson conducts field-test training and scoring in Iowa City, close to ITP. Operational training is online with human-distributed scoring for all paper responses and human-automated scoring for online responses. All Writing prompts will receive 10 percent human second scoring.

ISASP Writing Online Training

Scoring for the Writing prompts utilizes Pearson's ISASP Writing Online Training tool. The training consists of 10 modules that all readers complete prior to qualifying to score. The first nine modules are the same for all ISASP readers irrespective of the grade or content to be scored. These modules introduce readers to Pearson, general scoring concepts and principles, Pearson's electronic scoring system, and specifics related to the ISASP scoring project. The tenth module is item specific and focuses on a specific item assigned to the reader. After completing these modules, readers complete practice and qualification scoring in ePEN. Following this, there are additional modules that cover how to handle unusual responses and other general topics.

Pearson ensures accurate scoring through training readers on the ISASP standards, anchor responses, practice scoring writing samples, qualification scoring in ePEN, and calibration scoring. Scores are monitored through validity measures, interrater reliability, backreading, and scoring rate and pacing. Readers are provided feedback throughout the process.

Following are brief descriptions of the content covered in each of the modules. In addition, Table 8.7 provides a summary of the number of scorers hired and qualified by grade and subject.

Modules 1 – 4 : Introduction and Policy Overview
An introduction to the training modules, Pearson's history, human resource policies, and quality policy.

Module 5: What is Scoring
An introduction to assessments as the measure of educational constructs, purposes of assessments, standards utilized in testing, scoring versus grading, types of scoring, and the importance of accuracy in scoring.

Module 6: How to Ensure Accurate Scoring
An overview of Pearson's processes to train readers and monitor their responses. The processes include use of anchor responses, many opportunities to practice scoring, qualification scoring in ePEN. In addition, readers scores are monitored through validity measures, interrater reliability, backreading, and scoring rate and pace.

- Readers are provided access to the prompts which include instructions and may also include quotes, pictures, reading selections, and charts.
- Readers are instructed to refer frequently to the rubric (scoring guide) and anchors to ensure they are

scoring accurately. Anchors are carefully selected samples that clarify the rubric and acceptable range of responses within each score point.

- Practice responses allow readers to practice assigning accurate responses without affecting test-takers and to receive feedback. Once readers complete practice scoring, they move to qualification responses. Successfully meeting quality requirements is necessary to begin scoring.

Module 7: What Can Impact Accurate Scoring

Readers are trained to avoid being influenced by assumptions about the test-taker, by the appearance of the response, or by the scoring conditions. Maintaining consistency and accuracy in scoring related to these factors helps to ensure fairness in scoring.

Module 8: Using ePEN2 to Score Responses

Readers are introduced to ePEN2, which is Pearson's electronic Performance Evaluation Network designed with functions and features to assist in marking and scoring responses. It allows for real time monitoring to aid in achieving accurate scoring.

Module 9: Scoring ISASP

An introduction to the ISASP scoring project, covering topics such as the importance of security and confidentiality, contact information for Pearson scoring support team, and an overview of ISASP assessments, scoring standards, and testing conditions.

Module 10: Training Specific to a Particular Writing Item

This training provides the rubric, anchors, and practice student responses to the particular Writing item in order to train readers on scoring. Multiple examples at each score point across the rubric are provided. Once readers complete practice scoring, they move to qualification responses. Successfully meeting quality requirements is necessary to begin scoring.

Table 8.7. Analysis of scorer recruitment, training, retention/ dismissal

Grade	Subject	Total Hired	Qualified & Scored	Quality Warnings	Released for Quality
03-05	Reading	11	8	3	1
06-08	Reading	15	13	11	1
09-11	Reading	10	7	1	0
05, 08	Science	7	5	0	0
10	Science	12	10	0	0
03-05	Writing	21	15	5	1
06-08	Writing	39	21	5	0
09-11	Writing	60	26	5	0

Note: The table represents numbers of scorers only. Scoring directors and supervisors for each group also performed some scoring.

Table 8.7 displays the number of scorers hired and qualified by grade and subject. It also provides information on how many scorers were flagged through an automated quality management process as presenting validity agreement below quality standards. Interventions occurred for each of these readers, where Supervisors and Directors reviewed exemplar papers and the scoring training. The number of readers ultimately locked out of an

item through this process is shown in the "Released for Quality" column. In Reading and Science, readers locked out of one item for quality were not necessarily disqualified from scoring other items. In Writing, readers locked out of an item did not continue scoring writing.

Evaluation of Human Scorers and Intelligent Essay Assessor (IEA)

In order to evaluate the interrater reliability for both human-human scorers and IEA-human, the human-human agreement on a set of responses were generated as well as the IEA-human agreement for constructed response items in Science, Reading, and Writing. For Science and Reading the percentage of exact score agreement was high at upwards of 92% exact agreement for both human-human and IEA-human on Science items across all grades and 83% and above exact agreement on Reading items across all grades. This verifies the reliability of human ratings and that IEA was scoring similarly to humans.

Percentage of exact score agreement becomes a more stringent criterion as the number of item score points in the rating scale increases. For the essay component of the Writing test, the rating scale ranges from 0 to 5 on each of four analytic rubrics. The percentage of exact agreement and the percentage of disagreement by 1 scale score point (adjacent agreement) were considered when evaluating the differences between ratings on each essay prompt. The results of the interrater reliability analyses for humans are presented in Table 8.8. Table 8.9 gives interrater results where one of the ratings comes from machine scoring.

In the first year of administration of the ISASP, the percentage of adjacent agreement for human-human and IEA-human were all very high. The percentage of exact agreement for human-human scoring ranged from 55-63 across all grades. To increase these values on future administrations, additional measures were included to improve the accuracy and reliability. The results from the second year of administration are provided below. The percentage of exact agreement human-human scoring increased greatly in the second year of administration, particularly for Grades 9-11 (Table 8.8). This process will continue to be adjusted and monitored throughout scoring to ensure additional gains in exact agreement between human-human and IEA-human ratings.

Table 8.8. Interrater Reliability – Human-Human Scoring

	Research to Build and Present Knowledge			Production and Distribution of Writing			Text Types and Purposes			Conventions of Standard English / Knowledge of Language		
Grade	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**
3	64.3	97.8	.76	64.7	97.2	.76	67.2	97.5	.75	67.6	98.1	.75
4	61.6	96.3	.75	59.0	96.1	.73	59.5	96.7	.72	65.2	97.7	.74
5	70.9	98.7	.88	65.8	98.4	.85	68.1	98.7	.85	73.5	99.0	.87
6	64.3	98.4	.81	69.0	98.1	.84	64.9	98.1	.81	70.8	97.8	.84
7	65.6	96.7	.86	61.8	96.7	.86	67.6	96.7	.88	65.6	97.5	.87
8	62.3	97.8	.84	64.5	98.7	.85	65.5	98.7	.86	62.0	99.0	.84
9	85.9	99.6	.93	90.0	100.0	.96	89.2	99.6	.95	86.7	99.2	.92
10	96.2	100.0	.98	85.6	99.4	.97	96.9	99.4	.97	95.6	100.0	.97
11	96.3	100.0	.98	96.9	100.0	.99	97.5	100.0	.99	96.3	100.0	.98

Table 8.9. Interrater Reliability – IEA-Human Scoring

	Research to Build and Present Knowledge			Production and Distribution of Writing			Text Types and Purposes			Conventions of Standard English / Knowledge of Language		
Grade	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**	**Percent Perfect Agree**	**Percent Adjacent Agree**	**Correlation**
3	65.2	99.7	.62	71.6	99.5	.64	67.3	99.6	.64	72.1	99.6	.68
4	66.7	98.2	.67	61.9	97.1	.67	68.3	98.9	.69	65.0	97.8	.69
5	68.2	99.1	.76	67.5	99.0	.76	71.1	99.5	.80	71.5	99.7	.79
6	66.7	99.2	.75	67.9	99.3	.76	69.1	99.5	.78	68.1	99.7	.77
7	67.4	99.4	.79	67.4	98.8	.76	69.6	99.4	.81	69.8	99.5	.80
8	64.0	99.9	.79	71.3	99.4	.82	70.1	99.7	.84	71.9	99.5	.83
9	73.7	99.8	.82	75.0	99.8	.83	71.5	99.7	.81	72.2	99.7	.78
10	72.5	99.5	.83	73.9	99.6	.87	70.4	99.7	.85	75.2	99.6	.83
11	73.5	99.5	.85	74.3	99.8	.87	74.0	99.8	.87	72.1	99.7	.84

When evaluating the performance of IEA for the purpose of scoring reading and science short constructed-response items, the human-human agreement on a set of responses was compared it to the IEA-human agreement on that set of responses at each score point. Tables 8.10 and 8.11 below provide examples of those comparisons different grade levels.

Table 8.10. Science

Grade	N	Human-Human Agreement				IEA-Human Agreement			
		Exact	0	1	2	Exact	0	1	2
5	742	94%	98%	88%	94%	92%	97%	75%	98%
8	701	97%	99%	86%	97%	93%	94%	70%	96%

Table 8.11. Reading

Grade	N	Human-Human Agreement				IEA-Human Agreement			
		Exact	0	1	2	Exact	0	1	2
5	588	83%	76%	92%	73%	84%	80%	90%	63%
8	543	87%	90%	78%	94%	83%	83%	82%	83%

Classification Consistency and Accuracy

When scores are used to classify students into different achievement levels the *Standards* (2014) call for classification consistency and accuracy to be reported. Classification consistency refers to the extent to which observed classifications of examinees would be the same across replications of the testing procedure. Classification accuracy refers to the extent to which the observed classifications of examines would agree with their true classification. For the ISASP program, the levels of achievement levels are Not Yet Proficient, Proficient, and Advanced.

Classification consistency and accuracy were estimated using statistical methods developed by Livingston & Lewis (1995). These methods use information from the administration of one test form (i.e., distribution of scores, the minimum and maximum possible scores, the cut points used for classification, and the reliability coefficient) to estimate both classification consistency and accuracy. Kim & Lee (2019) and Wan, Brennan & Lee (2007) have found the Livingston-Lewis procedure to perform well compared to several other methods for estimating consistency and accuracy from a single test administration, including the compound multinomial model. Note that accuracy indices are always larger than the consistency indices because classification consistency is affected by random variation in each of the two classifications. For classification accuracy, only the observed-score classification is affected by random variation; the true-score classification is not by definition.

The Livingston-Lewis procedure was developed to handle mixed-format tests by estimating an effective test length. Like earlier methods (Hanson & Brennan, 1990), true scores are assumed to take the form of a four-

parameter beta distribution. Based on the first four moments of the observed score distribution, the exact form of the true score distribution may be estimated by a method proposed by Lord (1965). This true score distribution is used to estimate classification accuracy by comparing it to the observed score distribution. The defined true score distribution is also used to estimate a score distribution for an alternate form to estimate classification consistency. Table 8.12 provides consistency statistics by grade.

Table 8.12. Classification Consistency and Accuracy for Not-Yet-Proficient and Proficient Designations

	ELA		Mathematics		Science	
Grade	Consistency	Accuracy	Consistency	Accuracy	Consistency	Accuracy
3	90.1	93.0	86.5	90.4		
4	89.2	92.4	87.2	90.9		
5	88.7	92.0	86.6	90.4	80.4	85.9
6	89.4	92.5	85.3	89.6		
7	90.6	93.4	86.0	90.0		
8	90.1	93.0	87.2	90.9	81.7	87.0
9	91.2	93.8	84.8	89.2		
10	90.6	93.4	84.5	89.0	85.1	89.4
11	91.0	93.6	88.3	91.7		

The method implemented in the program BB-CLASS (Brennan, 2004) was used to estimate classification consistency and accuracy for the 2019 and 2021 ISASP programs using two (Not Yet Proficient vs. Proficient/Advanced) and three (Not Yet Proficient, Proficient, Advanced) achievement levels. See the *ISASP ASR-2019 AND ISASP ASR 2021*for classification consistency and accuracy tables. For each subject, the overall classification consistency and accuracy results are reported first. Classification accuracy tables for three groups are also reported in the *ISASP ASR-2019*, where the column marginals give the observed proportions in each level and the rows give the true proportions in each level. With more than two categories, the false positive rate is defined as the sum of the upper off-diagonal; similarly, the false negative rate is defined as the sum of the lower off-diagonal elements.

Chapter 9: Quality-Control Procedures

The Iowa Statewide Assessment of Student Progress (ISASP) and its associated data play an important role in the state accountability system as well as in many local evaluation plans. Therefore, it is vital that quality-control procedures are implemented to ensure the accuracy of student-, school-, and district-level data and reports. Iowa Testing Programs (ITP) and its delivery partner have developed and refined a set of quality-control procedures to ensure that all testing requirements of ITP and the Iowa Department of Education are met or exceeded. These procedures are detailed in the paragraphs that follow. In general, the commitment of ITP and its test delivery partner to quality is evidenced by initiatives in three major areas:

1. Task-specific quality standards integrated into individual processing functions and services
2. A network of systems and procedures that coordinates quality across processing functions and services
3. Technical analysis and ongoing maintenance of psychometric procedures and characteristics of assessment materials and scoring functions

Quality Control for Test Construction

Test construction for the ISASP follows the legally sanctioned, industry-standard best practice test-development process used by ITP and its delivery partner as described in Chapter 2, "Test Development," of this document (Schmeiser & Welch, 2006; Smisko, Twing & Denny, 2000). Following these processes, items were selected for the 2019, 2021 and 2022 test forms to maximize content alignment to the Iowa Core and ensure quality in psychometric specifications. Following each test administration, items are selected and placed on a particular pre-equated test form to provide the capability of assembling a strictly parallel form in the ensuing year both in terms of content and psychometric characteristics. This work also supports the development and maintenance of item pools in each content area for the transition to adaptive testing. Item development target are established for each test to ensure pool sizes are sufficient to support both fixed form and adaptive assessments. Current development targets specify 150 to 180 items per test undergo field testing in each grade annually. Item and test form statistical characteristics from the base forms administered in 2019 are used as targets when constructing the test forms for a subsequent year. Once a set of items has been selected for a pre-equated form, ITP test development staff reviews content and suggests replacement items as needed for a variety of reasons (e.g., alignment fidelity, fairness and sensitivity, cuing, etc.). Successive changes are made, and the process iterates until a final pre-equated form is assembled and ready for external fairness reviews. Similarly, the baseline raw score-to-scale score tables are used as the target tables to ensure that the pre-equated test form (i.e., the form under construction) matches all psychometric specifications. This form is provided to ITP's delivery partner for form construction and digital publishing, as outlined in a subsequent section of this chapter. Examination of post-administration item and test characteristics, including results from IRT calibrations, indicated that the procedures followed in assembling the 2021 ISASP forms achieved the goals of the pre-equating design and supported all scoring and reporting functions.

Quality Control for Non-Scannable Documents

The ISASP program follows a meticulous set of internal quality standards to ensure high-quality printed and digitally presented products. Specific areas of responsibility for staff involved in materials production include monitoring all materials-production schedules to meet test administration commitments and

schedules, overseeing the production of test materials, coordinating detailed printing and post-printing specifications both digital and paper-based, outlining specific quality control requirements for all materials, and conducting digital/print reviews and quality checks. The quality production and printing processes follow printers' reviews and quality checks in both digital and print formats. Project Management and Print Procurement staff work closely with the compositors and printers during the production phase. ITP and its delivery partner check digitized proofs and press proofs to ensure high-quality publishing and to verify adherence to printing specifications. For printed test materials, the printing staff randomly pull documents throughout the print run for quality control inspections.

Quality Control for Online Test Delivery Components

Each release of every online test delivery goes through a complete testing cycle, including regression and performance testing. The system goes through User Acceptance Testing (UAT). During UAT, ISASP tests that are administered in that program year are used.

In addition to the UAT, the ISASP program also conducts Production Validation (PV) testing. The delivery partner publishes the tests in a production environment and runs recommended test scenarios. The tests are completed and scoring deliverables are generated during the PV period. These include preliminary student detail reports and the student data files. The validation process includes confirmation of the tests published and the scoring deliverables. Approvals by both ITP and its delivery partner are required at the close of the PV period prior to the opening of the testing window.

For changes required during the testing window, a patch build is implemented. The release notes are provided, which include the fixes made and/or system upgrades. The patch is tested and approved before it is scheduled to be deployed to the field. Only patch builds that are relevant to the ISASP program are applied to its pipeline. All deployments are scheduled outside of the regular testing window timeframes.

ITP and its delivery partner continually seek to improve quality control processes for online test delivery. One example is the was the introduction of enhancements to data collection and storage systems in 2021 to provide additional prevention and/or detection of potential anomalies. Such measures helped address a situation that occurred during test administrations in other states using the same systems used in the ISASP program. Additional enhancements were implemented in the 2021 administration that leveraged risks to online administrations that have occurred in other assessment programs.

Test administration enhancements to the online systems have consolidated the scoring and saving of data into a single task. This has prevented scoring systems from getting out of sync with the scoring database.

Quality Control in Scaling, Equating, and Linking in the ISASP Program

Quality control steps in the scaling, equating, and linking processes are designed to ensure the integrity of reported scoring within and across administration years. Multiple quality control processes were implemented in the development of the ISASP vertical scales, and the achievement levels associated with ISASP scale scores in ELA, Mathematics, and Science. Raw-score frequency distributions developed from the student data file (SDF) were replicated independently by three data analysts before they were used in the scale development process. An iterative process was used in scale development in which provisional raw to scale score transformations were used to score the SDF grade by grade, evaluate resulting between- and within-grade variability, and examine expected growth tables implied by the provisional transformation until a suitable transformation was obtained for each grade.

For each ISASP administration, additional quality control steps validate the implementation of the conversion tables used to transform raw scores to Iowa scale scores and the links between ISASP scale scores and the achievement levels of Not Yet Proficient, Proficient, and Advanced. These steps involve the scoring of the Student Data File (SDF) independently by ITP and its delivery partner at the item level, domain level, and test level. These processes are design to provide validation check for both observed scores and scores that are derived from IRT-based methods.

For each ISASP administration, the validated SDF is then used to confirm the accuracy of the reports (Individual Student Reports, Achievement Summary Reports, Historical and Longitudinal Reports, and District Data Files). Samples of students are drawn for each grade and subject from multiple districts and private schools across Iowa. The public districts sampled in the most recent administration included large (multiple high schools), mid-sized, and small rural districts. Comparing the data elements in the SDF to the printed reports and the Data Files, ITP staff confirms that the calculation of the ELA score is the expected combination of the Reading score and the Language/Writing score, for example, and that the mapping of ISASP Scale Scores to the corresponding proficiency levels in all subjects is accurate and that all other report elements are displayed accurately.

Test-form equating is the process that enables fair and equitable comparisons both across test forms within a single year and between test administrations across years, whether the equating occurs during test assembly in a pre-equating design or after test administration in a post-equating design. ITP and its delivery partner use several quality-control procedures to ensure this equating is accurate.

1. ITP and its delivery partner perform independent “key check” analyses for the multiple-choice item type to ensure the appropriate scoring key is being used.
2. ITP, together with its delivery partner, performs an “adjudication” analysis for all technology-enhanced (TE) item types. The adjudication process includes a check of all responses given by students in the current administration to ensure all possible responses are scored appropriately. This analysis includes possible adjustments to the scoring algorithm for TE items for which novel-yet-correct responses were identified.
3. ITP’s delivery partner employs industry-standard procedures that are used in the National Assessment of Educational Progress for monitoring hand-scoring of constructed-response and extended constructed-response item types. In addition, procedures used to train the artificial intelligence (AI) scoring engine for these item types include the evaluation of AI-scored constructed-response items with respect to the correlations with hand-scored results. These methods are implemented during range-finding activities prior to the test administration as well as post administration for monitoring operational scoring, performing hand-scoring checks on AI scored materials, and general quality control of the AI scoring process.

4. For all assessments, a drift analysis is conducted to determine whether the item-response theory (IRT) item parameters have shifted over time. The drift analysis involves post-administration calibration of items to provide updated parameter estimates that can be compared to those used for operational scoring. Items that have shifted are investigated, and a resolution to determine appropriate scoring is made. The criterion for identifying non-ignorable item drift is overlap in the 50 percent confidence intervals for two calibrations of the same item.
5. Drift analyses provide updated item parameters for purposes on item pool maintenance. In general, ISASP item pools include results from the most recent post-administration calibrations available.
6. For pre-equated forms administered in the 2021 ISASP program, data from districts testing early in the assessment window were sampled to monitor pre-equated scores.

Table 9.1 provides a summary of procedures related to technical analysis and on-going maintenance for the ISASP Program.

Table 9.1. Summary of Technical Analysis and Ongoing Maintenance

Maintenance Category	Processes and Procedures
Test Construction	• Content Review • Fairness and Sensitivity Review • Alignment Review • Universal Design Check • Target Assembly to Base Form Test Characteristic Curves and Test Information Functions
Non-Scannable Documents	• Production Schedule Guidelines • Print and Digital Publishing Specifications • Digital and Print Reviews • Print-run Inspection Checks
Online Delivery Components	• Regression and Performance Testing • User Acceptance Testing • Production Validation • Scoring Validation • Patch-Build Deployment
Item Pool Maintenance, Scaling and Equating	• Replicated Independent Score Validation • Regression Testing of All Reported Scores • Independent Key Checks • Adjudication Checks on Constructed-Response and TEI Scoring • Post-Administration Calibration Checks • AI Score Validation and Hand-Scoring Flagged Responses • IRT Drift Analysis • Updated Item Parameter Estimates

References

Abedi, J., & Ewers, N. (2013). *Smarter Balanced Assessment Consortium: Accommodations for English language learners and students with disabilities: A research-based decision algorithm.* Prepared for SBAC by the University of California, Davis.

Acosta, B. D., Rivera, C., & Shafer-Willner, L. (2008). *Best practices in state assessment policies for accommodating English language learners: A Delphi Study.* Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

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

Barton, K. (2002). Stability of constructs across groups of students with different disabilities on a reading assessment under standard and accommodated administrations (Doctoral dissertation, University of South Carolina, 2001). *Dissertation Abstracts International*, 62/12, 4136.

Beattie, S., Grise, P., & Algozzine, B. (1983). Test modifications and minimum competency test performance of learning-disabled students. *Learning Disability Quarterly, 6*, 75–77.

Bennett, R., Rock, D., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and non-handicapped groups. *The Journal of Special Education, 21*(3), 9–21.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24*(1), 41–55.

Bennett, R. E., Rock, D. A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille Edition. *Journal of Educational Measurement, 26*(1), 67–79.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. doi:10.1111/j.1745-3992.2009.00161.x

Blaskey, P., Scheiman, M., Parisi, M., Ciner, E., Gallaway, M., & Selznick R., (1990). The effectiveness of Irlen filters for improving reading performance: A pilot study. *Journal of Learning Disability, 23*(10), 604–612.

Bolt, S. K., & Thurlow, M. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and Special Education, 25*(3), 141–154.

Bouck, E., & Bouck, M. (2008). Does it add up? Calculators as accommodations for sixth grade students with disabilities. *Journal of Special Education Technology, 23*(2), 17–32.

Brennan, R.L. (2004). Manual for BB-CLASS: A Computer Program that uses the Beta-Binomial Model for Classification Consistency and Accuracy (Version 1.1), (CASMA Research Report No. 9). Iowa City: University of Iowa.

Brennan, R.L., & Lee, W. C. (1997). *Conditional standard errors of measurement for scale scores using binomial and compound binomial assumptions* (Iowa Testing Programs Occasional Paper No. 41). Iowa City, IA: University of Iowa.

Brown, D. W. (2007). *The role of reading in science: Validating graphics in large-scale science assessment.* Unpublished dissertation.

Burch, M. (2002). Effects of computer-based test accommodations on the math problem-solving performance of students with and without disabilities (Doctoral dissertation, Vanderbilt University, 2002). *Dissertation Abstracts International, 63/03*, 902.

Burk, M. (1998). *Computerized test accommodations: A new approach for inclusion and success for students with disabilities.* Paper presented at Office of Special Education Program Cross Project Meeting "Technology and the Education of Children with Disabilities: Steppingstones to the 21st Century."

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Calhoon, M., Fuchs, L., & Hamlett, C. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23*, 271–282.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Castellon-Wellington, M. (2000). *The impact of preference for accommodations: The performance of English language learners on large-scale academic achievement tests.* (CSE Technical Report No. 524). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Christensen, L.L., Braam, M., Scullin, S., & Thurlow, M. L. (2011). *2009 state policies on assessment participation and accommodations for students with disabilities (Synthesis Report 83).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Coleman, P. J. (1990). Exploring visually handicapped children's understanding of length (math concepts*).* (Doctoral dissertation, The Florida State University, 1990). *Dissertation Abstracts International,* 51, 0071.

Cormier, D. C., Altman, J. R., Shyyan, V., & Thurlow, M. L. (2010). *A summary of the research on the effects of test accommodations: 2007-2008* (Technical Report 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Council of Chief State School Officers (2013). *Criteria for procuring and evaluating high-quality assessments*. Washington, DC: Author. Retrieved from:

http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

Crawford, L., & Tindal, G. (2004). Effects of a student read-aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality, 12*(2), 71–88.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1988). Five perspectives on validity argument. In ,(Eds.), *Test validity*, (3–17). Hillsdale, NJ: Lawrence Erlbaum.

Davis, L. L., & Moyer, E. L. (2015). *PARCC performance level setting technical report.* Available from Partnership for Assessment of Readiness for College and Careers (PARCC), Washington, D.C.

DiCerbo, K., Stanley, E., Roberts, M., & Blanchard, J. (2001). *Attention and standardized reading test performance: Implications for accommodation.* Paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC.

Dolan, R., Hall, T., Banerjee, M., Chun, E., & Strangman, N. (2005). Universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *The Journal of Technology, Learning, and Assessment, 3*(7). Retrieved from http://napoleon.bc.edu/ojs/index.php/jtla/article/view/1660.

Dorans, N., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach 1* (ETS Research Report Series, 1) (i-14). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355–368.

Dunbar, S. B., & Welch, C. J. (2014). *Measuring Growth with the Iowa Assessments* (ITP Research Series 2014.1). Iowa City, IA: University of Iowa. Retrieved from https://itp.education.uiowa.edu/ia/documents/Measuring-Growth-with-the-Iowa-Assessments.pdf

Elbaum, B. (2007). Effects of an oral testing accommodation on the mathematics performance of secondary students with and without learning disabilities. *The Journal of Special Education, 40*, 218–229.

Elliot, S., Kratochwill, T., McKevitt, B., & Malecki, C. (2009). The effects and perceived consequences of testing accommodations on math and science performance assessments. *School Psychology Quarterly, 24*(4), 224–239.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: American Council on Education and Macmillan.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education, 11*(2), 159–177.

Fletcher, J., Francis, D., O'Malley, K., Copeland, K., Mehta, P., Caldwell, C., … Vaughn, S. (2009). Effects of a Bundled Accommodations package on high-stakes testing for middle school students with reading disabilities. *Exceptional Children, 75*(4), 447–463.

Fuchs, L., Fuchs, D., Eaton, S., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review, 29*(1), 65–85.

Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76*(1), 35–40.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345–359. doi:10.1111/jedm.1990.27.issue-4.

Helwig, R., Rozek-Tedesco, M. A., & Tindal, G. (2002). An oral versus a standard administration of a large-scale mathematics test. *Journal of Special Education, 36(1),* 39–47.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In *Test validity,* 129–145. Hillsdale, NJ: Lawrence Erlbaum.

Iovino, I., Fletcher, J., Breitmeyer, B., & Foorman, B. (1996). Colored overlays for visual perceptual deficits in children with reading disability and attention deficit/hyperactivity disorder: Are they differentially effective? *Journal of Clinical and Experimental Neuropsychology, 20*(6), 791–806.

Johnson, E. S., Kimball, K., & Brown, S. (2001a). American Sign Language as an accommodation during standards-based assessments. *Assessment for Effective Intervention, 26*(2), 39–47.

Johnson, E., Kimball, K., Brown, S., & Anderson, D. (2001b). A statewide review of the use of accommodations in large-scale, high-stakes assessments. *Exceptional Children, 67*(2), 251–264.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and*

Psychological Measurement, 20, 141–151.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 17–64). New York, NY: American Council on Education/Praeger.

Kim, S. Y., & Lee, W. (2019). Classification consistency and accuracy for mixed-format tests. *Applied Measurement in Education, 32*(2), 97–115. doi.org/10.1080/08957347.2019.1577246.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 156–186). New York: American Council on Education/Praeger.

Kopriva, R., Emick, J., Hipolito-Delgado, C., & Cameron, C. (2007). Do proper accommodation assignments make a difference? Examining the impact of improvised decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11–20.

Koretz, D., & Barton, K. (2004). Assessing students with disabilities: Issues and evidence. *Educational Assessment, 9*(1-2), 29–60.

Koretz, D., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis, 22*(3), 255–272.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197. doi:10.1111/jedm.1995.32.issue-2.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239–270.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

MacArthur, C. A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of Special Education, 21(3),* 22–42.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Minnesota Department of Education (2018). *Technical manual for Minnesota's MCA and MTAS assessments* (Technical Report). Retrieved from https://education.mn.gov/MDE/dse/test/mn/Tech/

Morizot, J., Ainsworth, A. T., & Reise, S. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F.

Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). New York, NY: Guilford.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10–28.

Pennock-Roman, M., & Rivera, C. (2012). Smarter Balance Assessment Consortium: Summary of literature on empirical studies of the validity and effectiveness of test accommodations for ELLs: 2005-2012. Prepared for Measured Progress by The George Washington University Center for Equity and Excellence in Education.

Perez, J. V. (1980). Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation (Doctoral dissertation, University of South Florida, 1980). *Dissertation Abstracts International, 41*, 0206.

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). New York: American Council on Education/Macmillan.

Plake, B. S., Ferdous, A. A., Impara, J. C., & Buckendahl, C. W. (2005). *Setting multiple performance standards using the yes/no method: An alternative item mapping method*. Presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Ray, S. R. (1982). Adapting the WISC-R for deaf children. *Diagnostique, 7*, 147–157.

Reise, S. P. (2012) The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696.

Robinson, G., & Conway, R. (1990). The effects of Irlen colored lenses on students' specific reading skills and their perception of ability: A 12-month validity study. *Journal of Learning Disabilities, 23*, 621–626.

Russell, M. (2006). *Technology and assessment: The tale of two interpretations.* Greenwich, CT: Information Age Publishing.

Russell, M., Kavanaugh, M., Masters, J., Higgins, J., & Hoffmann, T. (2009). Computer based signing accommodations: Comparing a recorded human with an avatar. *Journal of Applied Testing Technology, 10*(3). Retrieved from http://www.testpublishers.org/Documents/090727Russelletal.pdf

Salend, S. (2009). Using technology to create and administer tests. *Teaching Exceptional Children, 41*(3), 40–51.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph*, No. 18.

Sato, E., Rabinowitz, S., Worth, P., Gallagher, C., Lagunoff, R., & McKeag, H. (2007). Guidelines for ensuring the technical quality of assessments affecting English language learners and students with disabilities: Development and implementation of regulations. (Assessment and Accountability Comprehensive Center Report). San Francisco, CA: WestEd.

Scarpati, S., Wells, C., Lewis, C., Jirka, S. (2011). Accommodations and item-level analyses using mixture differential item functioning models. *Journal of Special Education, 45*(1), 54–62.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 307–353). New York: American Council on Education/Praeger.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.

Smarter Balanced Assessment Consortium (2016, January). *2013-2014 Technical Report*. Los Angeles, CA: Author. Retrieved from https://portal.smarterbalanced.org/library/2013-14-technical-report.pdf/

Smarter Balanced Summative Assessments (2014). *Testing Procedures for Adaptive Item-Selection Algorithm*. Retrieved from https://portal.smarterbalanced.org/library/en/testing-procedures-for-adaptive-item-selection-algorithm.pdf

Smarter Balanced Summative Assessments (2017). *Smarter Balanced Cut Score Validation*. Retrieved from https://portal.smarterbalanced.org/library/en/smarter-balanced-cut-score-validation-final-report.pdf

Smisko, A., Twing, J. S., & Denny, P. L. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education, 13*(4), 333–342.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., pp. 263–331). New York, NY: American Council on Education/Macmillan.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Sullivan, P. M. (1982). Administration modifications on the WISC-R Performance Scale with different categories of deaf children. *American Annals of the Deaf, 127*(6), 780–788.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577. https://doi.org/10.1007/BF02295596

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy (Synthesis Report 41).* Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 Update (Synthesis Report 33).* Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children, 64*(4), 439–450.

Tippets, E., & Michaels, H. (1997, April). *Factor structure invariance of accommodated and non-accommodated performance assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227–253.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Wan, L., Brennan, R. L., & Lee, W. (2007*). Estimating classification consistency for complex assessments* (CASMA Research Report No. 22). Iowa City, IA: University of Iowa.

Webb, N. L. (1997). *Criteria of alignment of frameworks, standards and student assessments for mathematics and science education* (Research Monograph No. 6). Madison, WI: Council of Chief State School Officers and National Institute for Science Education, University of Wisconsin.

Welch, C. J., & Dunbar, S. B. (2022). Key concerns about fairness in testing: Practical applications in achievement testing. In *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of 2014 Standards.* Washington, DC: AERA.

Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report 766). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wright, N., & Wendler, C. (1994, April). *Establishing timing limits for the new SAT for students with disabilities.* Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21*(3), 187–201.