Technical Report for the Delaware Next Generation Science Integrative Transfer Assessments

Grades 5, 8, and Biology

Academic Year 2018–19

May 2021

Delaware
Department of Education

Prepared by Pearson

Table of Contents

Chapter 1: Introduction

Background

In 2013, Delaware adopted the Next Generation Science Standards (NGSS) as its state science standards. The goal of NGSS is to make certain all students leave Delaware schools able to apply their scientific knowledge and skills to real-world circumstances. Delaware is committed to a science assessment system that honors the principles of three-dimensional science learning while monitoring student readiness for challenging coursework in science and for college and career. The system reflects the state's mission for students to contextualize the crosscutting concepts across science core ideas and science and engineering practices.

The state's science assessment system offers three types of measures for understanding a student's progress in science.

- Embedded classroom assessments (grades 3 through 10),
- End-of-Unit Assessments (grades 3 through 10), and
- The Integrative Transfer Assessments (ITAs; grades 5, 8, and high school biology).

Embedded classroom assessments are developed by teachers to provide information on learning in real time in every grade from third grade through tenth grade. The assessments are primarily for instructional use and are therefore short and administered at the discretion of each teacher. The development of these has been supported by professional development.

End-of-Unit Assessments, aligned to instructional units in every grade from third through tenth, are administered by teachers after the completion of each instructional unit. Each End-of-Unit Assessment is meant to provide information on student learning of the NGSS content in each unit for the purposes of instruction (e.g., to determine whether additional instruction on previously instructed topics is needed, or to use as a classroom assessment for grading purposes) and evaluation (e.g., to inform curriculum adoption, adaptation, and modification) at classroom, school, and district levels. End-of-Unit Assessments are developed by vendors working with DDOE staff and informed by educator reviews for classroom administration by teachers.

The Integrative Transfer Assessment is administered to students in grade 5, grade 8, and high school biology. The Integrative Transfer Assessment is meant to capture students' learning of the content instructed during the entire year in each of the three grades in greater breadth and focus on long term transfer of science understanding. The Integrative Transfer Assessment requires students to apply their knowledge of science to grade-level-appropriate situations in order to solve unique, real-life problems centered on investigating and explaining phenomena. They are administered through an online system in a secure testing environment and student performance is used for the State high-stakes school accountability.

The ITA provides an evaluative measure in the benchmark years at the end of elementary school (grade 5), middle school (grade 8), and high school (biology). The results of the summative assessment are reported as a singular (or a total) score and the performance level. The total score is intended to make a broad statement about lasting and pervasive knowledge and skills and provide educators, parents, and the public with information on student progress towards science literacy (or the NGSS). Subscores are not provided, as the length of the assessments does not provide for reliable information that is actionable in a benchmark model of assessment. The ITAs offer a systemic measure, while End-of-Unit and Embedded assessments offer more refined measures of student achievement during instruction when adaptation and adjustment to ongoing instruction is most appropriate.

The ITA is specifically designed to evaluate the proficiency of students to meet the NGSS standards in benchmark years, including the ever-present and ever-progressing practices and crosscutting concepts in the shortest amount of time possible (to meet the needs/expectations of the field). The system of assessments allows for actionable feedback to be provided during instruction by the End-of-Unit and the Embedded assessments, from grade 3 through high school.

Administered annually in the spring, the ITAs were established to meet the requirements of the Every Students Succeeds Act (ESSA) of 2015. ESSA requires that states administer to all students annual assessments in science once in each grade span (3-5, 6-8 and HS) that are aligned to state standards.

This report provides technical details of the first ITA administration in the 2018–2019 test administration cycle, including the data collected from the 2017–2018 stand-alone field test.

Purpose and Uses

By assessing student achievement against the NGSS, the ITAs serve two important purposes. First, these summative assessments provide an accountability tool that measures overall performance as well as differing levels of defined performance across students, schools, and districts against the NGSS standards. Second, it provides stakeholders with important systemic information about what students have learned, which, if applied constructively and in combination with information for embedded classroom assessments and end of unit assessments, can foster improvement of instructional programs, classroom education, and school performance. Improved student learning is a key goal of any educational assessment program. The ITAs are intended for all students, including those in private institutions or wherever they may be taught, as set forth by the requirements of the adopted NGSS standards.

The Technical Report for the Delaware Next Generation Science ITAs provides objective information regarding technical aspects of the 2019 operational tests at grades 5, 8, and biology. It is intended to be one source of information to Delaware K-12 educational stakeholders (including testing coordinators, educators, parents, and other interested citizens) about the design, development, implementation, scoring, and technical attributes of the ITAs.

The information provided here fulfills professional and scientific guidelines for technical documentation of educational assessments. Specifically, information was selected for inclusion in this report based on ESSA requirements and the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014).

This manual provides information about the ITAs assessments regarding:

1. Validity evidence for test score interpretations and uses;
2. Test design and test construct;

3. Content of the tests;
4. Scoring and reporting the results of the tests.
5. Statistical characteristics of the test questions;
6. Identification of ineffective items;
7. Calibration of test forms;
8. Detection of potential item bias and evaluation of fairness for all test-takers;
9. Reliability of test scores;

From test development to final reporting, each of these facets of the ITAs contribute to the validity of the inferences made about the test results. This technical manual addresses these topics for the 2018-2019 testing year.

Chapter 2: Validity

Validity

The *Standards for Educational and Psychological Testing (Standards)*, issued jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) establish professional guidelines for the development and evaluation of tests and are considered the industry standard for guidance on testing. The most recent edition of the *Standards* (2014) reports:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. (p. 11)

The purpose of test validation is not to validate the test itself but to validate interpretations of the test scores for particular uses. Test validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the lifetime of an assessment. Every aspect of an assessment provides evidence in support of the validity of test score interpretations and uses (or evidence of lack of validity of those inferences and uses), including design, content specifications, item development, and psychometric characteristics. The 2019 ITA operational assessments provided an opportunity to gather evidence of validity based on both test content and on the internal structure of the tests.

This chapter explains the argument-based approach to validation around which this technical report is organized. Following is a brief outline of the components of an argument-based approach to validation. The chapter continues by identifying the claims of the ITAs and provides a framework, including studies with methodologies and results, for gathering validity evidence.

Test Score Interpretation and Use Validation

As alluded to previously, validity is a property of the proposed interpretations and uses of test score, and it refers to the degree to which evidence supports those intended interpretations and uses. The Delaware Department of Education (DDOE) must make valid and reliable decisions about student achievement in science that serves as an indicator for high-stakes school accountability. These decisions could be used to matriculate to the next grade as well as decisions by schools about remediation programs for students. In addition, student and school-level test results could be used by local educators to make decisions about curriculum and instruction that, over time, will lead to improvements in student achievement.

To support these intended uses, ITA scores must provide information that reflects what students know and can do in relation to the academic expectations defined in the academic content it measures and achievement standards applied to scores. This is the primary claim that all other claims depend upon. Through the validity evaluation process, evidence is gathered related to this claim and to the decisions that rely upon it.

In evaluating the validity of test score interpretations, test developers should also anticipate and eliminate or mitigate threats to validity. Two such threats of particular importance are construct-irrelevant variance and construct underrepresentation. The *Standards* (AERA/APA/NCME; 2014) define construct-irrelevant variance as “variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation” (p. 217). In other words, construct-irrelevant variance refers to differences in test scores due to factors other than those that the test is intended to measure. Construct underrepresentation, on the other hand, refers to “the extent to which a test fails to capture important aspects of the construct domain that the test is intended to measure, resulting in test scores that do not fully represent that construct” (AERA/APA/NCME; 2014; p. 217). Understanding and recognizing these two threats to valid interpretations of test scores will help to frame the evidence in support of the intended inferences and use of test scores.

Validity Claims and Methodology Overview

Using Kane’s (2013) framework for argument-based validation, an interpretation/use argument for the ITAs is outlined below with the claims for the validity argument following. Those claims are explicated in the chapters that follow.

As stated in Chapter 1, the results of the ITAs are used for high-stakes school accountability as the ITA measures overall performance as well as differing levels of defined performance across students, schools, and districts against the NGSS standards, and provides stakeholders with important systemic information about what students have learned.

The sections below present a summary of the validity argument evidence for the four parts of the interpretation/use argument: *scoring, generalization, extrapolation, and implication*. Much of this evidence is presented in greater detail in other chapters in this report. In fact, the majority of this report can be considered validity evidence for the ITAs (e.g., item development, performance standards, scaling, equating, reliability, item scoring, quality control). Relevant chapters are cited as part of the validity evidence given below.

Scoring Validity Evidence

Using Kane’s interpretation/use argument, the scoring inference refers to the relation of an observed performance to an assigned test score. Evidence supporting this inference should show that several assumptions are met, including “that the scoring procedures are appropriate, are applied as intended, and are free of overt bias” (Kane, 2013; p. 25). The chapters of this report provide evidence related to the development of the content being assessed, the administration procedures, and the documentation of scoring procedures.

Model Fit and Scaling

Item response theory (IRT) models provide a basis for the ITAs. IRT models are used for the selection of items to go on the test, the equating procedures, and the scaling procedures. A failure of model fit would undermine the validity of these procedures. Item fit is examined during test construction. Any item displaying misfit is carefully scrutinized before a decision is made to place the item on the test. However, the vast majority of items display good model fit.

Dimensionality

Further evidence of the fit for the IRT model comes from dimensionality analysis. Internal structure evidence shows the degree to which items and test components conform to the construct on which the proposed test score interpretations are based (AERA/APA/ NCME, 2014). ITA reports overall science scale scores for individual students at each grade level. Internal structure validity evidence identifies the degree to which the item relationships conform to the overall scores.

While the NGSS are presented as reflective of several interwoven components that address multiple dimensions, ITA test questions and sets are designed around scientific phenomena and crafted to be reflective collectively of the standards. While individual items may each measure multiple elements of the standards and dimensions, the tests are designed to be unidimensional and to measure overall NGSS primarily. Assuming this holds true, then it is appropriate to apply a unidimensional IRT model for calibrating and scaling the ITAs.

An exploratory factor analysis (EFA) was conducted on each operational ITA form to determine whether a unidimensional IRT model was appropriate. EFA is a statistical method of identifying unobservable (i.e., latent) variables that account for the covariances among items. Unlike confirmatory factor analysis, EFA does not impose a predetermined factor structure and does not require each item to load on only one factor. To check the unidimensionality of the 2019 ITAs, an EFA was conducted using all operational items on each ITA form.

A general procedure in exploratory factor analysis for evaluating dimensionality is to compare the relative size of the eigenvalues associated with each latent factor (e.g., Lord, 1980). The ratio of first to second eigenvalues is an index of the relative strength of the first latent factor. Although there is no definitive criterion for establishing unidimensionality, a rule of thumb suggests that if the first latent factor accounts for considerably more variance than the next latent factor(s), the test can be considered sufficiently unidimensional for analysis within an IRT framework (e.g., Reise & Waller, 1990; Lord, 1980).

Table 2.1 summarizes the results of the first and second factor eigenvalues of the 2019 ITA assessments. Here, the first eigenvalue is substantially larger than the second in all instances and suggests that a unidimensional IRT model is appropriate.

Figures 2.1 through 2.3 are scree plots of the eigenvalues from each ITA. The scree plots provide a visual summary of the information in Table 2.1, and show that each assessment has a step reduction in size of the eigenvalues after the first one. Again, these results support the use of a unidimensional model to fit these data.

Table 2.1 ITA EFA First and Second Eigenvalues

Grades	First Eigenvalue	Second Eigenvalue	Ratio of First to Second Eigenvalue
Grade 5	3.63	0.20	18.2
Grade 8	5.43	0.30	18.1
Biology	6.58	0.47	14.0

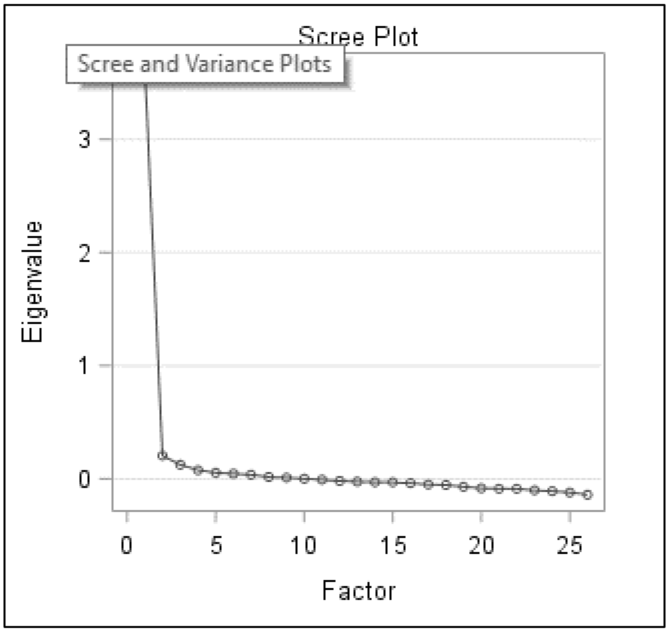


Figure 2.1. Grade 5 Scree Plot

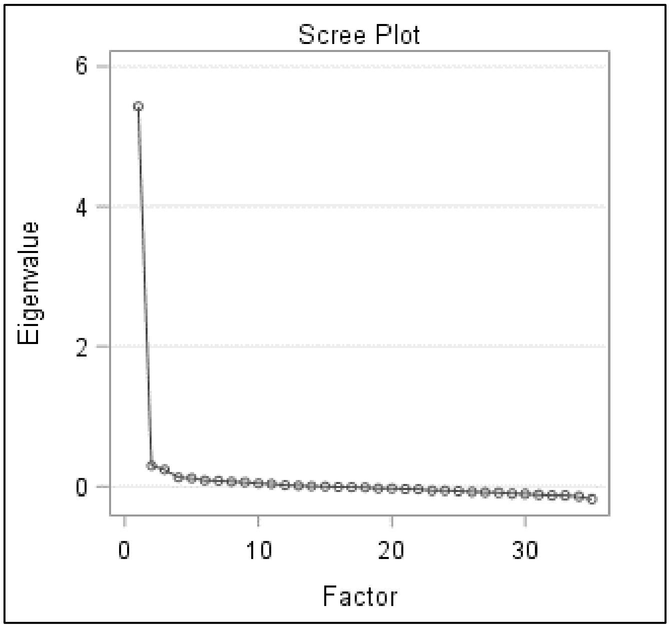


Figure 2.2. Grade 8 Scree Plot

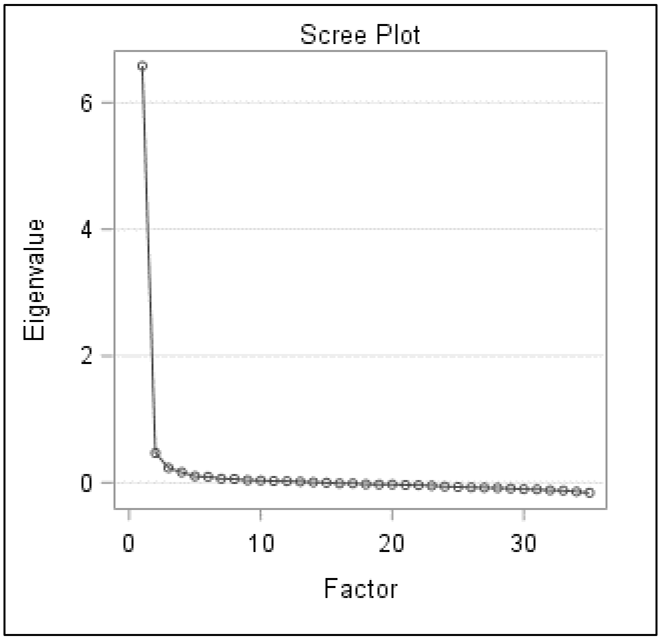


Figure 2.3. Biology Scree Plot

Model-data fit based on the Rasch model calibrations are also indicators of unidimensionality. That is, the model assumes unidimensionality as a necessary condition supporting its application. To the extent that indicators of fit suggest data do *not* appropriately fit the model as applied may be the result of multidimensionality. Discussion of model fit is presented in Chapter 8 with Rasch Infit and Outfit statistics for all ITA operational items presented in Appendix B. These statistics support the overall fit of ITA items to the Rasch model.

Generalization Validity Evidence

There are two major requirements for validity that allow generalization from observed scale scores to universe scores[1]. First, the items administered on the test must be representative of the universe of possible items. Evidence regarding this requirement includes evidence that the test adequately samples the content standards and benchmarks and that construct over-/underrepresentation and construct-irrelevant variance do not threaten the validity of inferences about test scores. These sources of evidence are reported in the sections that follow.

Validity Evidence for Content

Content validity evidence addresses whether a given assessment adequately samples from the full given domain. Where the assessment is determined to be representative in terms of the standards and in the manner intended, it is said to have strong validity evidence for content. The ITAs are designed to measure NGSS broadly and involve more complex content and synthesis of responses according to the content and three-dimensional nature of the standards. The item

[1] Universe score is defined as the expected value of a person's observed scores over all observations in the universe of generalization, which is analogous to a person's "true score" in classical test theory (Shavelson & Webb, 2006).

development, test construction, alignment study, and scoring, for example, provide content/construct-related and empirical validity evidence.

For the ITAs, test design and blueprint specifications were developed in concert between WestEd and DDOE science experts well versed in NGSS. These specifications drive item and stimulus development targets intended to effectively support the intended purposes of the ITA assessments in relation to the NGSS. As noted, the contractors and the DDOE were directly involved in item and stimulus development based on the test specifications. Items and stimuli were rigorously scrutinized during the various content reviews by Delaware educators in public schools and state colleges. These reviews examine the appropriateness of test items, difficulty, clarity, correctness of answer choices, plausibility of 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 difficult, contains only one correct (or best) answer for multiple-choice questions, and has an appropriate and complete scoring guideline for technology-enhanced items. Next, a bias and sensitivity committee must review and approve the items for language or content that may be inappropriate or offensive to students, parents, or community members, or that contain stereotypical or biased references to gender, ethnicity, or cultural background. The process of the ITA test design, development, and test construction is described in chapters 3 and 4 of this report. As documented, DDOE, WestEd, Pearson, and educator committees expend tremendous effort to ensure the ITAs are content-valid.

The DDOE and Delaware educators also developed achievement level descriptors (ALDs) for the ITAs, which provide a description of typical grade-level performance for each level of achievement in relation to the NGSS. The ALDs are descriptions of the knowledge and skills demonstrated by students in each performance category. Higher scores translate to a greater level of knowledge and skills demonstrated. There is a link between the ALDs and the knowledge and skills required to meet proficiency according to the standards. ALDs are used to relate performance on ITAs to the NGSS through the process of standard setting. Content experts and stakeholders participated in a standard setting for the ITAs in September of 2019. This committee set the cut scores that delineate the four levels of science achievement at grades 5, 8, and Biology as reported in Delaware. Evidence of these activities is presented in the context of student performance on ITAs (Chapter 9) and includes a link to the Delaware ITAs standard setting report.

Human Resources Research Organization (HumRRO) conducted an alignment study for the ITAs. The study looked at the following three research questions:

1. To what extent do the Delaware ITA operational test forms and operational test items reflect the test design and intended distributions of DCI domains, SEPs, and CCCs?
2. To what extent do Delaware ITA operational test items integrate at least two NGSS dimensions (i.e., disciplinary core idea, science and engineering practice, and/or crosscutting concept)?
3. To what extent does the set of Delaware ITA operational test items reflect the test design and intended distribution of cognitive complexity?

To answer the research questions, HumRRO developed acceptability criteria that are specific to the science summative assessments. The criteria include:

1. Items Represent Intended Content.
2. Items Represent the Multidimensional Nature of the NGSS.
3. Items Reflect Appropriate Levels of Cognitive Complexity.

The alignment results found that all the ITAs met or partially met alignment for all of criteria 2 and 3, but that criterion 1 did not meet for Biology for subcriteria 1B, which reviewed if the operational test form includes multiple SEP and CCC such that a range of student performance is addressed. The alignment study concluded that the results indicated a strong start for this new summative assessment and suggested improvements including to continue monitoring content coverage and to continue providing professional development on the NGSS and on cognitive complexity.

Content validity evidence involves explicit assumptions about the cognitive processes engaged in by the test takers. Cognitive complexity refers to the cognitive demand associated with interacting with a given item. The level of cognitive demand focuses on the type and level of thinking and reasoning required of the student. Levels of cognitive complexity for ITAs are based on “A Framework for Analyzing Cognitive Demand and Content-Practices Integration: Task Analysis Guide in Science” (Tekkumru-Kisa, Stein, and Schunn, 2015). These 3 levels of cognitive complexity, *Scripted, Guided,* and *Doing* science are described in detail in Chapter 3.

In line with the nature of NGSS, items developed and appearing on the ITAs address *Scripted* and *Guided* levels, with only limited *Doing* items.

Evidence of Controlling Random Measurement Error

Also important for content validity is the control of random measurement error. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. Reliability, the standard error of measurement (SEM), and the conditional standard error of measurement (CSEM) are discussed in chapter 10. Chapter 10 and the appendices present tables reporting the SEM, CSEM, and the coefficient alpha reliabilities by grade and broken down by demographic groups.

Evidence Based on Different Student Populations

In addition, generalization validity evidence should show that individual items are functioning similarly for different demographic subgroups within the population being measured. The ITAs are developed to assess NGSS and administered to all students irrespective of any particular demographic characteristic (as described in Chapters 4 and 6). Great care has been taken to ensure the items on ITAs are fair and 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.

This begins with item writers trained on how to avoid economic, regional, cultural, and ethnic biases 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' administration.

Differential item functioning (DIF) analyses are conducted for the purpose of identifying items that are differentially difficult for different subpopulations of individuals. Chapter 7 details the methodology used to evaluate DIF for ITA items. Though DIF analyses flag items as being differentially difficult for one group as compared to another, it does not solely provide sufficient evidence for removing the item from use. Flagged items are re-examined post administration for any potentially overlooked biases attributable to the content of those item.

Extrapolation Validity Evidence

Validity for extrapolation requires evidence that test performance supports an inference about a test-taker's actual knowledge, skills, and abilities. Although it is usually impractical or impossible to design an assessment measuring every concept or skill in the domain, it is desirable for the test to be robust enough to allow some degree of extrapolation from the measured construct. The validity argument for extrapolation can use either analytical evidence or empirical evidence. Empirical evidence of extrapolation may be provided when a suitable criterion exists. Finding an adequate criterion for a standards-based achievement test can be difficult.

Analytic Evidence

The NGSS create a common foundation to be learned by all students and define the domain of interest. As documented in chapters 3 and 4 of this report, the ITAs are designed to measure as much of the domain defined by the standards as possible.

A threat to the validity of the test can arise when the assessment requires competence in a skill unrelated to the construct being measured (i.e., construct-irrelevant variance). For example, students who are English Learners (EL) may have difficulty fully demonstrating their science knowledge if the science assessment requires fluency in English. The use of accommodation avoids this threat to validity by allowing students who are EL to demonstrate their ability on a test that limits the quantity and complexity of English language used in the items. The ITAs also allow accommodations for students with vision impairment or other special needs. The use of accommodated forms allows accurate measurement of students who would otherwise be unfairly disadvantaged by taking the standard form. Accommodations are discussed in Chapter 5 of this report. Further, the coefficient alpha reliability measures for the EL, disability and accommodation groups (see Chapter 11 and the appendices) in particular provide some evidence for the effectiveness of accommodations that would allow meaningful interpretation of results and comparisons across subgroups.

Empirical Evidence

Delaware provides extrapolation validity evidence for the ITAs by analyzing the relationship between students' performance and their performance on external similar measures. By examining this relationship, evidence can be collected to show that the relationships are consistent with those expected at the level of the construct underlying the proposed score interpretations.

To examine validity evidence based on external measures, an analysis was conducted of the relationship between 2019 ITA performance and 2019 student performance on English language arts (ELA) and mathematics assessments. A Pearson correlation was computed on the reported scale score for students who took both the ITA and the ELA or mathematics assessment. Results are reported in Tables 2.2 and 2.3. The reported correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related. All of the correlations had values between 0.65 and 0.76 indicating a moderate positive linear relationship between ITA scale scores and ELA and mathematics scale scores, respectively, which is generally consistent with what is seen of student performance for tests in different content areas. For Grade 5 and Grade 8, the relationship of ITA performance was compared to the Smarter Balanced (SBAC) ELA and Mathematics results. Biology ITA performance was compared to PSAT and SAT results.

In general, the correlation between science and mathematics is expected to be higher than that between science and ELA. Although the correlations between science and mathematics tests are moderately high, they are nearly identical to those between science and ELA. These results may be due in part to the relatively higher reading load of item clusters on the science assessments as compared with more traditional stand-alone multiple-choice items. The correlations of the science assessments with other external measures are also influenced by the reliability of the assessments themselves. Future validation work should include monitoring both the reliability of the tests over time as well as their relationship to other external measures. It is expected that as the NGSS are further incorporated into student instruction, the correlations between the ITAs and external mathematics measures will be stronger.

Table 2.2 2019 Correlations of External Measures

Test	Smarter Math		Smarter ELA	
	N	**Correlation**	**N**	**Correlation**
Grade 5 Science	10799	0.75	10769	0.76
Grade 8 Science	10147	0.75	10116	0.76

Table 2.3 2019 Correlations of External Measures

Test	PSAT Math		PSAT ELA	
	N	**Correlation**	**N**	**Correlation**
HS Biology	6600	0.65	6399	0.69
	SAT Math		**SAT ELA**	
	N	**Correlation**	**N**	**Correlation**
	399	0.69	399	0.72

Implication Validity Evidence

There are inferences made at different levels based on the ITAs. Individual student scores are reported, as well as aggregate scores for schools and districts. Inferences at some levels may be more valid than those at others.

One of the most important inferences to be made concerns the student’s achievement level, especially for accountability tests. Even if the total-correct score can be validated as an appropriate measure of the standards, it is still necessary to validate the scaling and achievement level designation procedures. Evidence of achievement level setting is presented in the context of student performance on ITAs (Chapter 9) and includes a link to the Delaware ITAs standard setting report, while Chapter 8 discusses scaling. These chapters serve as documentation of the validity argument for these processes.

At the aggregate level (i.e., school, district, or statewide), the implication validity of the accountability assessment can be judged by its impacts on the overall proficiency of students. Validity evidence for this level of inference will result from examining changes over time in the percentage of students classified as proficient. As mentioned before, there exists a potential for negative impacts on schools as well, such as increased dropout rates and narrowing of the curriculum. Future validity studies need to investigate possible unintended negative effects.

Validity evidence related to the implications of the test may also include changes in instruction due to introduction of the test. Because this was the first operational administration of the ITAs, evidence of improving student performance over time might support that the test is having a positive effect on the implementation of the NGSS. Likewise, evidence that certain parts of the curriculum are being neglected may indicate negative unintended consequences of the test.

Summary of Validity Evidence

This chapter, together with other chapters of this report, provides validity evidence supporting the appropriate inferences from Delaware’s ITA scores. In general, the validity evidence provided supports to the primary claim that the ITAs provide information that measures the overall performance at differing achievement levels of defined performance across students, schools, and districts against the NGSS standards. Validity arguments based on rationale and logic are strongly supported for the ITAs. The empirical validity evidence for the scoring and the generalization validity arguments for these assessments are also quite strong. Reliability indices, model fit, and dimensionality studies provide consistent results, indicating the ITAs are properly scored, and scores can be generalized to the universe score.

Less strong is the empirical evidence for extrapolation and implication. This is due in part to the absence of an ideal criterion for the ITAs. Empirical evidence for the extrapolation argument, such as convergent validity evidence, should be explored. Future studies are needed to verify some implication arguments. This is especially true for the inference that the state’s accountability high-stakes assessment program is making a positive impact on student learning.

Chapter 3: Test Development

Content Coverage

The ITAs are built to align with the Next Generation Science Standards (NGSS). Science instruction aligned to the NGSS requires students to engage in scientific and engineering practices in the context of disciplinary core ideas and to use crosscutting concepts to make connections across topics. The NGSS were developed and organized by grade level for kindergarten through grade 5. The middle and high school standards were organized by grade band. For the middle and high school grades, Delaware has outlined which standards should be taught each year in grades 6–10 and which standards should be assessed in each year. Some standards at the secondary level are introduced more than once, or only to a certain degree at a given grade or over the course of a grade band. Assessment boundaries for each repeated or limited performance expectation are presented in this document.

- Science and Engineering Practices (SEPs)
- Disciplinary Core Ideas (DCIs)
- Crosscutting Concepts (CCCs)

The goals for science learning are outlined in the NGSS in the form of performance expectations (PEs). PEs are statements about what students should know and be able to do at the end of instruction. Each PE combines an SEP with a DCI and a CCC.

There are eight SEPs which describe how students should engage in the practices used by scientists and engineers. Students are assessed on selected SEPs in each year (see Table 3.8). The SEPs increase in complexity and sophistication, reflecting the progression in students' capabilities to use each of the practices as they progress from the lower to the upper grades. The eight SEPs are listed here:

1. Asking Questions and Defining Problems (Q/P)
2. Developing and Using Models (MOD)
3. Planning and Carrying out Investigations (INV)
4. Analyzing and Interpreting Data (DATA)
5. Using Mathematics and Computational Thinking (MCT)
6. Constructing Explanations and Designing Solutions (E/S)
7. Engaging in Argument from Evidence (ARG)
8. Obtaining, Evaluating, and Communicating Information (INFO)

The CCCs provide a way of linking the different domains of science and have application across all domains of science. Students are assessed on selected CCCs in each year (see Table 3.9). As students' understanding of the disciplinary core ideas increases, their depth of understanding of the CCCs increases as well. Therefore, like the SEPs, the CCCs increase in complexity and sophistication across the grades. The seven CCCs are listed here:

1. Patterns (PAT)
2. Cause and Effect (C/E)
3. Scale, Proportion, and Quantity (SPQ)
4. Systems and System Models (SYS)
5. Energy and Matter (E/M)
6. Structure and Function (S/F)
7. Stability and Change (S/C)

The organization of the NGSS is structured around three domains (Physical Science; Life Science; Earth and Space Science) in the major fields of natural science, with the addition of one domain focused on Engineering, Technology, and Applications of Science (ETS). The domains are divided into core ideas, and the core ideas are further divided into supporting ideas. Each supporting idea includes a description of what students should understand at the end of instruction for the K–2, 3–5, 6–8, and 9–12 grade bands; together these make up the DCIs in the NGSS.

The DCIs are organized into the following 12 categories:

Physical Sciences
PS1: Matter and Its Interactions
PS2: Motion and Stability: Forces and Interactions
PS3: Energy
PS4: Waves and Their Applications in Technologies for Information Transfer
Life Sciences
LS1: From Molecules to Organisms: Structures and Processes
LS2: Ecosystems: Interactions, Energy, and Dynamics
LS3: Heredity: Inheritance and Variation of Traits
LS4: Biological Evolution: Unity and Diversity
Earth and Space Sciences
ESS1: Earth’s Place in the Universe
ESS2: Earth’s Systems
ESS3: Earth and Human Activity
Engineering, Technology, and Applications of Science
ETS1: Engineering Design

The performance expectations that are assessed on the ITAs at each grade level are shown in Tables 3.1 through 3.3 below.

Table 3.1. Grade 5 ITA Performance Expectations Assessed by Domain

Earth & Space Science	Life Science	Physical Science	Engineering, Technology, and Applications of Science
5-ESS1-1	5-LS1-1	5-PS1-1	3-5-ETS1-1
5-ESS1-2	5-LS2-1	5-PS1-2	3-5-ETS1-2
5-ESS2-1		5-PS1-3	3-5 ETS1-3
5-ESS2-2		5-PS1-4	
5-ESS3-1		5-PS2-1	
		5-PS3-1	

Table 3.2. Grade 8 ITA Performance Expectations Assessed by Domain

Earth & Space Science	Life Science	Physical Science	Engineering, Technology, and Applications of Science
MS ESS2-1	MS LS1-6	MS PS3-1	MS-ETS1-1
MS ESS2-4	MS LS1-7	MS PS3-2	MS-ETS1-2
MS ESS2-5	MS LS2-1	MS PS3-3	MS ETS1-3
MS ESS2-6	MS LS2-2	MS PS3-4	MS-ETS1-4
MS ESS3-2	MS LS2-3	MS PS3-5	
MS ESS3-3	MS LS2-4	MS PS4-1	
MS ESS3-4	MS LS2-5	MS PS4-2	
MS ESS3-5			

Table 3.3. High School Biology ITA Performance Expectations Assessed by Domain

Earth & Space Science	Life Science	Physical Science	Engineering, Technology, and Applications of Science
HS ESS2-6	HS LS1-1	HS PS1-4	HS-ETS1-1
HS ESS2-7	HS LS1-2	HS PS1-7	HS-ETS1-2
	HS LS1-3	HS PS3-1	HS-ETS1-3
	HS LS1-4		
	HS LS1-5		
	HS LS1-6		
	HS LS1-7		
	HS LS2-1		
	HS LS2-2		
	HS LS2-3		
	HS LS2-4		
	HS LS2-5		
	HS LS2-6		
	HS LS2-7		

Earth & Space Science	Life Science	Physical Science	Engineering, Technology, and Applications of Science
	HS LS2-8		
	HS LS3-1		
	HS LS3-2		
	HS LS3-3		
	HS LS4-1		
	HS LS4-2		
	HS LS4-3		
	HS LS4-4		
	HS LS4-5		
	HS LS4-6		

The ITAs in the Delaware Next Generation Science Assessment System are composed of item clusters and standalone (or discrete) items. These general guidelines provide descriptions of common elements of the summative assessments.

PE Bundles, Phenomena, and Stimuli

Item clusters are designed to prompt students to make sense of a phenomenon. The development of an item cluster begins with the selection of the Performance Expectation (PE) bundle and an appropriate phenomenon. PE bundles are selected from the eligible grade-level PEs based on the PEs needed to fulfill the assessment blueprint and the ability of the PE bundle to support a phenomenon that is appropriate for the assessment. A PE bundle may be within a domain or across multiple domains.

A phenomenon is defined as an observable event that occurs in the universe that can be explained through the application of the three dimensions of the NGSS. The stimulus provides the context or setting, including data sets, graphs, tables, models, and/or descriptions of investigations, in which the phenomenon is presented to students. The context should support the use of the intended specific dimensions. For example, the stimulus could describe an investigation if the intended SEP is Planning and Carrying Out Investigations. Stimuli must be scientifically correct and should be sourced from reputable sources.

A standalone item is not adequate for making sense of a phenomenon. Although each standalone item is based on a phenomenon, the item focuses on a single aspect of the phenomenon. The item requires the student to makes sense of only that aspect rather than the entire phenomenon.

Test Design

Item Types

A variety of innovative item types are available for use in the computer-based ITAs. The item type is deliberately chosen to elicit the evidence most appropriate for the PE and the dimensions being assessed by the item. The following list describes the available item types.

- Multiple choice (MC): A prompt and four answer options with one correct choice (1 point)

- Multiple select (MS): A prompt and five to seven answer options with two correct choices (1 point)
- Technology-enhanced item (TEI): Designed for online administration only; item types include graphic gap match, match table grid, hot spot, and bar graph (1 or 2 points)
- Two-part dependent (TPD): Two-part item where the answer to Part B is an explanation of, or provides evidence to support, the answer to Part A; the student must get Part A correct to get any credit for Part B; partial credit is possible only if Part A is correct; possible combinations of item types in a TPD are: MC/MC, MC/TEI, and TEI/TEI (2 points)
- Two-part independent (TPI): Two-part item where each part is scored independently, and the student may get credit for Part B even if Part A is incorrect; possible combinations of item types in a TPI are: MC/MC, MC/TEI, and TEI/TEI (2 points)
- Constructed response (CR): Open-ended item that typically requires a 1–3 sentence response, for use in item clusters only (2 points)
- Extended response (ER): Open-ended item that typically requires a 5-sentence or longer response, for use in item clusters only (4 points)

Graphics

Graphics should be purposeful and be included as needed to add clarity, present data, and/or simplify a concept through a visual representation. Graphics may also provide a backdrop for some technology-enhanced items. Graphics for the ITA are developed in color. Colors are chosen to accommodate students with the various forms of color blindness.

Item Cluster Characteristics

Each item cluster is aligned to a grade-specific performance expectation (PE) bundle and is based on a well-articulated phenomenon. Item clusters must be inclusive of all three dimensions in each of the associated PEs. It is not expected that a single item cluster will fully assess the breadth of the associated PEs. Any given PE may appear in more than one item cluster and in a variety of contexts.

The items in an item cluster share a common stimulus (or stimuli) that provides a realistic context in which to present the phenomenon. The information in the stimulus must be necessary and used along with content knowledge to answer every item. In other words, students must bring their knowledge of the standards in addition to the information in the stimulus in order to answer the items.

There are two types of item clusters in an ITA, Integrative Item Clusters (IICs) and Regular Item Clusters (RICs).

The characteristics of each type of item cluster, RIC and IIC, are described below.

- Integrative Item Clusters (IICs)
    - Multiple stimuli; each stimulus applies to a subset of the six items
    - Aligned to a PE bundle (2–3 PEs)
    - Must include one ER item and may include any of the other item types, with the exception of a CR item

- Regular Item Clusters (RICs)
    - A single stimulus with adequate data and/or information to support up to 10 items (Note: A RIC is developed and field-tested with 8 to 10 items. A RIC in the ITA will appear with 5 of those items.)
    - Aligned to a PE bundle (2–3 PEs)
    - Must include one CR item and may include any of the other item types, with the exception of an ER item

Standalone items in the ITA help to ensure that a broad representation of the PE and dimensions appear on every assessment. Each standalone item is aligned to a single PE. The item must align to at least two of the three dimensions associated with the PE. Alignment to SEPs or CCCs outside of the PE is allowed only if the item also meets the requirement to align to at least two of the dimensions of the PE. All stimulus information is contained within the item itself.

Alignment

Delaware's assessment system prioritizes the importance of students' abilities to apply scientific content and principles across contexts. The three-dimensional nature of the NGSS PEs means that judging the alignment to these PEs is more complex in nature than judging the alignment to more traditional, one-dimensional standards. An item cluster should achieve alignment to all three dimensions associated with a PE or PE bundle when all items are considered in totality. Additionally, item clusters may include one or two items that are aligned to an SEP and/or a CCC not represented in, but supportive of, the dimensions in the PE bundle. For example, in an item cluster with a PE that includes the SEP of Planning and Carrying Out Investigations, there might be an item aligned to the SEP of Asking Questions and Defining Problems.

Any single item, be it a standalone or part of an item cluster, is unlikely to thoroughly access all three dimensions of a PE. However, responding to an item should require the integration of at least two of the three dimensions specific to a PE or a PE bundle. While individual items may be two- or three-dimensional, that is, addressing SEPs, DCIs, and/or CCCs in the PE or PE bundle, it is not necessary for each item to address the full breadth of a PE or of a dimension. Instead, the items address different aspects of the PE through different combinations of the dimensions. Additionally, the degree of alignment to each dimension may vary. For example, an item should be considered as aligned to the PE if it is strongly aligned to one dimension, partially aligned to another, and not aligned to the third. The fact that an item has some degree of alignment to at least two of the three dimensions of a PE determines the overall alignment of an item, not necessarily the strength of the alignment to each dimension.

Specific details about item and item cluster alignment requirements are summarized below.

- Each individual item, whether part of an item cluster or a standalone item, must align to at least two dimensions of a PE. Extended response (ER) items are an exception to this rule. ER items must be three-dimensional. An ER should align to all three dimensions of a PE, or to three of the dimensions across a PE bundle (an SEP, a CCC, and a DCI).

- Items are limited by the assessment boundary statement associated with a PE, when present.

- Item clusters are typically developed to align to a PE bundle of 2–3 PEs. The items that compose an item cluster must, as a set of items, include alignments to all three dimensions of each PE in the PE bundle.

- Additional Science and Engineering Practices (SEPs) and Crosscutting Concepts (CCCs), beyond those specified in the PEs in the PE bundle, may be included in the alignment of items in an item cluster.

- Alignment begins at the PE level. At the dimension level, the sub-bullets of the specific dimensions represented in a PE should be the basis of alignment decisions. However, these decisions should also be informed by the progressions documents (NGSS Lead States, 2013; see, in particular, NGSS appendices E, F, and G), as well as by additional SEPs and CCCs, outside the specified PEs, that can—and, in many cases, should—be included in the alignment. Any additional alignment(s) should be captured as metadata.

Cognitive Complexity

The NGSS are structured in such a way as to expect student understanding beyond recall. The PEs have been written with high levels of cognitive complexity, incorporating knowledge with practice and explicitly identifying and utilizing unifying concepts to develop scientific explanations. A student's demonstration of basic content recognition or usage is not sufficient evidence of meaningful understanding of any PE. As a result, the traditional measures of cognitive complexity (e.g., Webb's Depth of Knowledge rating scale) are insufficient to represent the cognitive challenges embedded in the NGSS. Delaware's approach to cognitive complexity is based on "A Framework for Analyzing Cognitive Demand and Content-Practices Integration: Task Analysis Guide in Science" (Tekkumru-Kisa, Stein, and Schunn, 2015). An item's cognitive complexity is classified according to three levels, or categories, described below.

- Scripted: The item provides heavy scaffolding, or a script, that explicitly tells the student what to do. "Cookbook" lab procedures fall into the scripted category. There is a well-defined set of actions or procedures a student needs to take, usually in a given order. Students can follow those actions and reach the desired answer without knowing how or why the script leads to that answer.

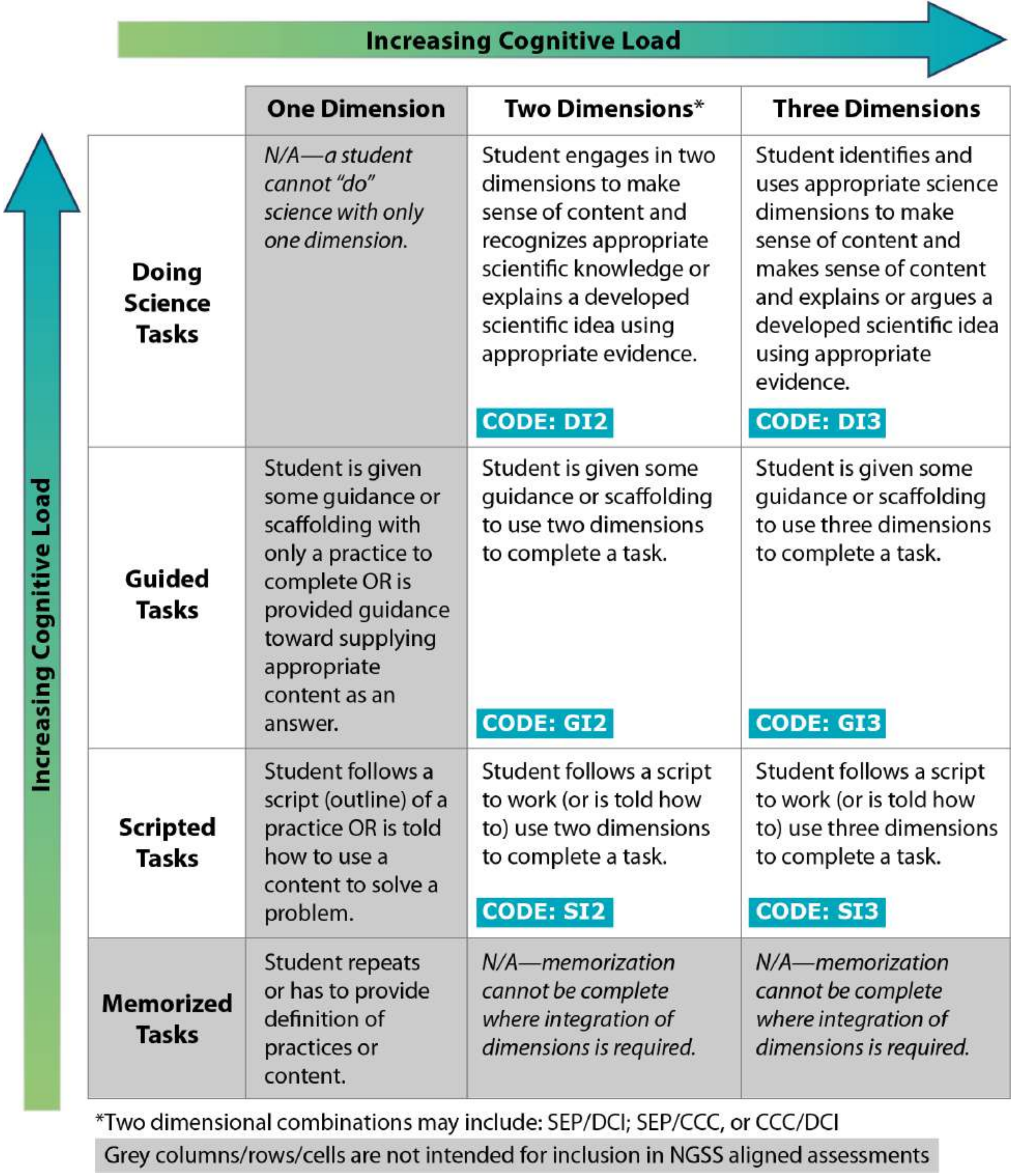
- Guided: The item provides some scaffolding, or suggested pathways, while requiring students to transfer their knowledge to a novel context. Students are expected to explain their reasoning for what they are doing. These items usually include using a model, data, and/or information to develop an explanation or argument.

- Doing Science: The item provides very little to no scaffolding and provides the opportunity for student-designed explorations. The students are required to identify which practices, or which use of practices and/or crosscutting concepts, are most appropriate to develop or deepen understanding of a scientific idea and/or explore a phenomenon. These items may engage a student in developing a model, developing an explanation, or developing an argument from their own analysis of raw data.

This approach to cognitive complexity also accounts for the number of dimensions to which an item is aligned. Every item must be an integration of at least two of the three dimensions (SEP, DCI, and CCC). Therefore, an item's cognitive complexity is a combination of the level of independence required of the student in responding to the item and the level of integration of dimensionality. For example, an item with heavy scaffolding aligned to an SEP and a DCI would have a cognitive complexity designation of Scripted Integration 2 (SI2), whereas an item with somewhat less scaffolding aligned to all three dimensions would have a cognitive complexity designation of Guided Integration 3 (GI3).

The categories of cognitive complexity are summarized in Table 3.4. The cognitive complexity of an item increases from bottom to top across Table 3.4, from Scripted to Guided to Doing Science tasks. Cognitive complexity also increases from left to right with the increased integration of the dimensions, from the integration of two dimensions (SI2, GI2, or DI2) to the integration of three dimensions (SI3, GI3, or DI3), as shown in Table 3.4. Items are developed across the range of cognitive complexity to support the goal of representing a range of cognitive complexity across each assessment.

Table 3.4. Cognitive Complexity


Increasing Cognitive Load
Increasing Cognitive Load

	One Dimension	Two Dimensions*	Three Dimensions
Doing Science Tasks	*N/A—a student cannot "do" science with only one dimension.*	Student engages in two dimensions to make sense of content and recognizes appropriate scientific knowledge or explains a developed scientific idea using appropriate evidence. CODE: DI2	Student identifies and uses appropriate science dimensions to make sense of content and makes sense of content and explains or argues a developed scientific idea using appropriate evidence. CODE: DI3
Guided Tasks	Student is given some guidance or scaffolding with only a practice to complete OR is provided guidance toward supplying appropriate content as an answer.	Student is given some guidance or scaffolding to use two dimensions to complete a task. CODE: GI2	Student is given some guidance or scaffolding to use three dimensions to complete a task. CODE: GI3
Scripted Tasks	Student follows a script (outline) of a practice OR is told how to use a content to solve a problem.	Student follows a script to work (or is told how to) use two dimensions to complete a task. CODE: SI2	Student follows a script to work (or is told how to) use three dimensions to complete a task. CODE: SI3
Memorized Tasks	Student repeats or has to provide definition of practices or content.	*N/A—memorization cannot be complete where integration of dimensions is required.*	*N/A—memorization cannot be complete where integration of dimensions is required.*

*Two dimensional combinations may include: SEP/DCI; SEP/CCC, or CCC/DCI

Grey columns/rows/cells are not intended for inclusion in NGSS aligned assessments

Design of the ITAs

To ensure that DDOE is in accordance with the federal requirements that the ITA fully align to the state's content standards, the NGSS serves as the guiding document for test development and design. Item development was a collaborative effort between DDOE, educators, WestEd, and Pearson. Classroom teachers, content specialists, and school administrators were recruited from all over Delaware for several test development committees. These committees reviewed items originally developed by contractors for ITA assessments.

The test specifications were established by the DDOE with help from WestEd and Pearson to guide the test development. The grade 5 test form consists of a core form with a total of 26 items yielding 40 raw score points, while the grade 8 and biology test forms had a core form with a total of 35 items yielding 53 raw score points.

ITA Blueprints

Depending on the grade level, an ITA is composed of one Integrative Item Cluster (IIC), either two or three Regular Item Clusters (RICs), and 10–14 Standalone Items (SAIs), as shown in the blueprints in Table 3.5 and Table 3.6.

Table 3.5. Grade 5 ITA Blueprint

RIC 01	SAI	IIC	RIC 02	SAI
5 items	5 items	6 items	5 items	5 items

Note: RIC: Regular Item Clusters; SAI: Standalone Items

Table 3.6. Grade 8 and Biology ITA Blueprint

RIC 01	SAI	IIC	RIC 02	SAI	RIC 03
5 items	7 items	6 items	5 items	7 items	5 items

Note: RIC: Regular Item Clusters; SAI: Standalone Items; IIC: Integrative Item Cluster

In addition to ITA test blueprints based on item types, ITAs also have blueprint ranges based on the overall domains, SEPs, and CCCs. Tables 3.7 to 3.9 provide these ranges that were used for building the 2019 ITAs.

Table 3.7. ITA Blueprints by Domain

Domains	Percent by Points of All Items		
	Grade 5	Grade 8	Biology
ESS	30%–46%	32%–48%	7%–15%
LS	10%–28%	17%–33%	
LS1			16%–32%
LS2			20%–36%
LS3			5%–18%
LS4			13%–29%
PS	38%–54%	27%–43%	5%–18%
ETS	0%–15%	0%–15%	5%–18%

Table 3.8. ITA Blueprints by SEP

Science and Engineering Practices	Percent by Points of All Items		
	Grade 5	Grade 8	Biology
SEP 1 - Q/P	0%	5%–15%	5%–15%
SEP 2 - MOD	24%–34%	30%–40%	14%–24%
SEP 3 - INV	16%–26%	5%–15%	6%–15%
SEP 4 - DATA	5%–15%	8%–18%	6%–15%
SEP 5 - MCT	9%–19%	4%–10%	14%–24%
SEP 6 - E/S	0%	8%–18%	21%–31%
SEP 7 - ARG	16%–26%	12%–22%	11%–21%
SEP 8 - INFO	5%–15%	0%	3%–15%

Table 3.9. ITA Blueprints by CCC

Crosscutting Concepts	Percent by Points of All Items		
	Grade 5	Grade 8	Biology
CCC 1 - PAT	5%–15%	5%–15%	5%–15%
CCC 2 - C/E	10%–20%	15%–25%	19%–29%
CCC 3 - SPQ	23%–43%	5%–15%	5%–15%
CCC 4 - SYS	18%–28%	5%–15%	9%–19%
CCC 5 - E/M	10%–20%	25%–35%	19%–29%
CCC 6 - S/F	0%	5%–10%	3%–15%
CCC 7 - S/C	0%	10%–20%	9%–19%

Field Test Design

The ITAs follow an embedded field-test design. For all grade levels, the 2019 ITAs had 8 different field-test forms. Each of the 8 forms had the same core form of operational items but had a unique set of matrixed field test items. The 8 forms are spiraled across the state population of students taking the ITA. With approximately 10,000 students per grade, having 8 forms allows for approximately 1,250 students taking each field test item.

For the grade 5 ITA, this set of field test items consisted of 6 or 8 items, depending on whether the form was field testing a 6-item IIC or a 5-item RIC plus 3 standalone items. For the grade 8 and biology ITAs, the matrix set of field test items consisted of 6 or 10 items, depending on whether the form was field testing a 6-item IIC or a 5-item RIC plus 5 standalone items. When possible, each RIC was tested across 2 forms so that in case items were lost due to poor performance the RIC would still survive. The total number of 2019 embedded field test items by item type is presented in table 3.10.

Table 3.10. 2019 Embedded Field Test Items

Grade	Total Number of Field Test Items	IICs		RICs		Standalone Items
		Clusters	Items	Clusters	Items	
Grade 5	53	2	12	3	24	17
Grade 8	67	2	12	3	26	29
Biology	65	3	18	3	22	25

Note: IIC: Integrative Item Cluster; RIC: Regular Item Clusters

Chapter 4: Test Construction

The 2019 operational ITAs were created in line with the test design and blueprint presented earlier. The process of selecting items for the core forms was an iterative process primarily involving WestEd content experts, DDOE, and Pearson psychometricians.

Initial Build

WestEd content specialists, Pearson psychometricians, and DDOE content staff worked jointly on the preliminary test build. The test development team selected the “best” items from the available item bank from a content perspective, to meet the ITA test construction guidelines.

Statistical Guidelines for Selection of Items

Several classical item statistics were used to evaluate the quality of individual items within the item bank during the test construction process. These statistics include:

For dichotomously scored items

- p-value for item difficulty
- point-biserial correlation for item discrimination
- percent choosing each item option (i.e., distractor) for multiple choice items
- item option point-biserial for multiple choice items
- Mantel-Haenszel differential item functioning flags and levels

For polytomously scored items

- mean score for item difficulty
- item-total correlation for item discrimination
- item score distribution
- standardized mean difference (SMD) DIF statistics flags and levels

Items were flagged for further review when their stand-alone field test statistics failed to meet certain statistical criteria, with flagged items avoided for selection for the operational test forms. The statistical criteria included:

- Extremely high or low p-value, or item mean with respect to range:

  If greater than 0.90 or less than 0.20

- Extremely low point-biserial or item-total correlation:

  If less than 0.20 (Note that items with point-biserial item-total correlations less than 0.10 are extremely flawed and not acceptable for operational forms)

- Highly attractive multiple-choice item option (distractor):

  If an item option percentage greater than 40%, or an item option point-biserial is greater than the point-biserial.

- Item shows differential item functioning

  If the DIF index is significant (i.e., Category C).

Differential Item Functioning

Differential item functioning (DIF), is a statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups. In examining DIF, the student group of interest is the focal group and the group to which performance on the item is being compared is the reference group (a detailed description of DIF is presented in Chapter 7).

For the ITA DIF analyses, the reference groups were White for ethnicity, and male for gender. The focal groups were females, and African-American and Hispanic ethnic groups. DIF analyses were also conducted to compare performance of English learners with non-English learners and students with disabilities with students without disabilities.

Items were flagged into one of three categories based on the magnitude of their DIF statistics:

- Category A: no or negligible DIF
- Category B: slight or moderate DIF, and
- Category C: moderate to large values of DIF. These items which exhibit significant DIF, are of primary concern.

Category C items were avoided for selection for the operational test forms. All items exhibiting DIF underwent additional content review in order to determine if content judgments support the DIF flag. If content reviewers determined that the statistical DIF flag was the results of bias against one or more subgroups, the item was marked for revision or removal from the item pool.

Content Review

To better identify potentially “dependent” items (i.e., items in which the answer to one question may influence how students perform on another) during content review, the entire draft test form was reviewed to identify and address potential issues of cluing. It was important that all items selected for use within and across item sets do not clue one another. Note that although the items within regular item clusters (RICs) and integrative item clusters (IICs) share one or more common stimuli, each individual item within the cluster is intended to be independent of the others.

Each item was reviewed for accuracy, clarity, and appropriateness of content. The test was also reviewed for coherency, diversity of content and flow. Additionally, the test development team verified the following:

- the accuracy of item-level content classifications

- the accuracy of scoring keys
- equal representation of scoring keys
- the appropriateness of the proposed item sequence (e.g., no more than 3 items with same key in row)
- diversity of subject matter within stimulus

When determining the order in which items should be presented several factors were considered:

- Item keys – Several selected response items having the same key should not be presented adjacent to each other on a form.
- Similarity of passages or stimulis – To the extent possible the subject matter, length and reading difficulty was varied across the test.

After content review was completed, the content team determined whether the initial build needed to be revised. If not, the form was sent to psychometrics for review and then to DDOE for their review. This iterative process continued until content experts, DDOE, and psychometrics finalized and approved each respective core form.

Field Test Form Assembly

After operational forms were approved, the test development team assigned newly developed items to field test forms for field-testing. Factors that were considered in determining how to assign items to forms are outlined below. The number of items associated with a given item set varied slightly from one form to another in some instances.

Several factors were considered when assigning items/passages to forms:

- *Cueing/Clueing.* Field-test items were evaluated against the given core form to ensure they did not clue the answer to other field test items on the form OR any of the operational items.
- *The type of items represented on each form.* Ideally a mix of item types appeared on each form. Similarly, multiple standards and objectives were represented.
- *The number of items associated with a given item set.* Item sets were field-tested with enough items to allow for attrition. Each field-tested item set was placed on two different field test forms with its own set of items. (For example, the same item set stimulus appeared on one form with six items and on a second form with another six items).
- *The distribution of keys and the number of items having the same key placed adjacent to one another.* Similar to operational forms – the key distribution and placement was considered when selecting/sequencing items.
- *Stimulus passage difficulty/reading load.* The mix of stimulus passage difficulties and lengths on a given form was considered.

Chapter 5: Test Administration

The overall test window for the Delaware DeSSA Next Generation Science (NGSS) assessment is established by the Delaware Department of Education (DDOE). For each given grade level (grade 5, 8, and Biology), all testing takes place according to the approved schedule established by the state.

Testing should be scheduled by districts and schools to allow for the completion of each grade-specific test in one day. Although there is no mandated testing time, more time can be provided to students who need certain time-based accommodations.

For Delaware DeSSA NGSS, the testing schedule for 2019 was as follows:

- Test Materials arrive in Schools February 25–March 1, 2019
- Paper Test Window April 1–30, 2019
- Online Test Window March 5–May 30, 2019

The 2019 *Online Test Administration Manual for DeSSA Social Studies and Next Generation Science* (TAM) provides needed information regarding policies and procedures for the DeSSA NGSS assessments.

Test Format

Each test consists of Selected Response (SR), Constructed Response (CR) items, and Technology Enhanced (TE) items (online only), based on shared stimuli, as well as stand-alone items. The online version of the test also includes interactive stimuli and may also contain videos.

The paper version of Delaware DeSSA NGSS is used for students with certain accommodations. Each student uses a test book containing all test items and response areas. The responses are then transcribed into an online version of the paper test by the Test Administrator (TA).

Practice Tests

Online practice tests are available to allow students to become familiar with the online testing environment. Students or guest users are allowed to take practice tests using either the secure DeSSA secure browser or another supported internet browser. Not all content standards are represented on the practice tests, and they are not intended to be used to predict performance on the ITAs. Online test accommodations or supports, including large print, alternate background color, and text-to-speech, are also available on the practice tests.

Testing Accommodations

Testing accommodations for students with disabilities (i.e., students having an Individualized Education Program [IEP] or a 504 Plan) or students who are English Learners (EL) (i.e., students who have an EL Plan) have to be approved and documented according to local policies or state policies set by the DDOE Office of Assessment.

Note that students with significant cognitive disabilities are assessed with an alternate assessment. Eligibility for the alternate assessment is determined by the student's IEP.

Braille Test Books and Transcription

The Delaware DeSSA NGSS assessment is administered to students requiring Braille Test Books. For Braille Test Books, student responses have to be transcribed into TestNav after testing. An eligible TA transcribes the student's responses into TestNav exactly as given by the student.

Text-to-Speech Text and Graphics

For Delaware DeSSA NGSS, all students are provided the option of using text-to-speech capability for test items. For students who additionally need descriptions of graphics for items, this accommodation is provided in TestNav pending approval by DDOE. Text-to-Speech and Graphics also has to be selected for the student in PearsonAccess^{next}.

Language-Based Accessibility Features

For those students requiring language-based accomodations for the Delaware DeSSA NGSS based on a documented need in an IEP, 504, or EL plan, this is provided through separate test form assignments for American Sign Language (ASL), Spanish, or the translation of key terms in 4 different languages (Arabic, Haitian Creole, Korean, and Mandarin Chinese) in TestNav. The form assignments also need to be selected for the student in PearsonAccess^{next}. The ASL and Spanish test forms allow for the entire content of the tests to be delivered in American Sign Language or Spanish, while the Translation of Key Terms allows for designated words in test items to be displayed in a student's native language.

Administrative Procedures for Students with IEP, 504 Plan, or EL Plan Permitting use of a Human Scribe

For individual students who receive the use of a human scribe to transcribe their test answers, the individual providing the scribing marks the responses indicated by the student in TestNav, using the student's login credentials on their testing ticket. This accommodation can also be used in combination with any of the test forms (ASL, Braille, Spanish, Text -to-Speech and Graphics, and Translation of key terms).

Administrative Procedures for Students with IEP, 504 Plan, or EL Plan Permitting Paper-and-Pencil Transcription

A student whose IEP, 504 Plan, or EL Plan permits a paper-and-pencil assessment has his/her responses transcribed at the school level by an eligible TA TestNav using the student's login credentials. This accommodation also involves assigning a different test form for the student in PearsonAccess^{next}.

Table 5.1 presents the number of students using embedded accommodations in 2019.

Table 5.1. 2019 Students Using Embedded Accommodations by Grade

Grades	Total Students	Braille	Text to Speech Graphics	ASL	Spanish	Key Term Translation	Human Scribe	Paper Pencil
Grade 5	10840	0	9	3	31	10	100	12
Grade 8	10245	0	26	5	29	1	8	0
Biology	10072	0	8	1	22	3	1	5

Non-Embedded Accommodations

If indicated in a student's IEP or 504 Plan, other accommodations can also be provided for them on the Delaware DeSSA NGSS assessments: frequent breaks, small group testing, specified area or seating, or use of scribe (for temporary injuries). None of these accommodations require different test form assignments in PearsonAccess^{next}, although the "temporary" scribe accommodation requires DDOE approval.

Test Security

The security of Delaware DeSSA NGSS assessment instruments and the confidentiality of student information are vital to maintaining the validity, reliability, and fairness of the results.

All DeSSA Next Generation Science test items and test materials are secure and must be appropriately handled. Secure handling protects the integrity, validity, and confidentiality of assessment items, prompts, and student information. Any deviation in test administration must be reported as a test security incident to ensure the validity of the assessment results. Please refer to the *DeSSA Test Security Manual* for additional information.

Test Security Plans

Each district and charter school shall adopt and enforce a plan setting forth procedures to ensure the security of all state assessments as outlined below. This plan must encompass all public schools in the district, including district-sponsored charter schools. The plans must be submitted to the DDOE test security coordinator.

To protect the security of the state assessments, each district and charter school must establish the plan to be consistent with the procedures outlined in the *DeSSA Test Security Manual*, and it must address the following criteria:

- Identification and training of personnel authorized to have access to the tests or the testing system;
- Identification and training of personnel authorized to proctor assessments;
- Procedures for test administrators to follow when monitoring students during test sessions;
- Procedures for monitoring test materials before, during, and after testing;
- Procedures to verify the identity and eligibility of students taking an assessment;
- Procedures to report any alleged violation in test administration or test security;

- Procedures that set forth actions taken in response to a reported violation;
- Procedures for communication of test security procedures

Appropriate Assessment Practices

The *DeSSA Test Security Manual* provides a compilation of appropriate assessment practices. These practices are used to determine whether or not a specific practice related to the assessment is consistent with the principles of performing one's duties with integrity, honesty, and fairness to all. Each district and charter school must ensure all staff members have training and knowledge of these appropriate assessment practices and must monitor the practices of all staff to ensure compliance. The list of appropriate assessment practices includes details on:

- Communication;
- Training;
- Assessment preparation;
- Assessment administration;
- Overall assessment security;
- Physical security; and
- Electronic security;

Security of the Testing Environment and Secure Handling of Printed Materials

The *DeSSA Test Security Manual* describes security requirements for the test environment during various stages of testing. The test environment refers to all aspects of the testing situation while students are testing and includes what a student can see, hear, or access (including access via technology).

The manual also describes the secure handling of printed materials, including procedures for the destruction of printed materials and scratch paper, and the use of scratch paper on performance tasks.

Administration Monitoring

The *DeSSA Test Security Manual* provides guidelines for the monitoring of NGSS assessments by school district personnel, and provides detailed procedures for districts to respond to testing improprieties, irregularities, and breaches.

Chapter 6: Scoring Procedures

This chapter provides details about human scoring of constructed response (CR) and extended response (ER) items. The processes described included benchmarking activities to define the range of performance within each score point, training procedures for scorers and their supervisors, and the scoring and monitoring procedures employed to ensure continual quality of scores.

Benchmarking

Benchmarking is the activity of identifying student responses to define the range of performance levels within each score point on a given scoring rubric. Ultimately, the purpose is to arrive at consensus scores according to the standards established by the rubric so that training sets can be built that accurately reflect those standards.

Pearson’s scoring staff conduct Benchmarking in Dover, Delaware. To help ensure that decisions remain consistent, there are three benchmarking committees, one for each grade. Each grade-level committee is comprised of one scoring director/content facilitator and two to six Delaware educators. A Pearson Content Specialist and the DDOE Science Assessment Associate are also in attendance. Pearson begins the process with a brief review of the purpose of benchmarking and the rubric, as well as other documentation of standard evaluation criteria that facilitate a common understanding of the standards and intentions of the DDOE.

Each benchmarking committee systematically reviews a sample of student responses for each item, determining and recording consensus scores. Reaching consensus scores on a sufficient number of student responses allows for construction of effective training materials for each item. These responses accurately represent the range of student performance levels described in the rubrics, as interpreted by the committee members and the DDOE.

The general process for the review of benchmarking material is:

1. A brief discussion is held for each item to gain further insight into the prompt, rubric wording, and potential student approaches.

2. The committee then reviews the Set A responses selected by Pearson as “potential anchor papers.” These responses reflect the entire range of scores and help the committee define the major lines between score points. Then the committee review Set B which help define the minor lines between score points. Papers in Set A and B are in random score point order.

3. Five to ten responses at a time are then assigned to all the attendees to read individually. Each committee member reads each response and assigns their individual scores on their copy of the matrix. The scoring directors record all committee members’ scores on the consensus sheet/matrix before any discussion begins.

4. The committee discusses each response so that scoring directors can take adequate notes for training purposes, but discussion is more extensive on responses that do not have immediate consensus. The discussion always refers to the rubric and all scores are justified with the rubric in mind. A consensus score is reached by the teacher committee

members. The scoring directors note any discussion points during the review of each response.

5. Upon the completion of the first item, the process is repeated for subsequent items.

DDOE and Scoring Services staff meet at the end of each day to:

1. Review and compare the scoring of items to confirm the consistency of scoring.
2. Finalize consensus scores.
3. Discuss the committee work and any scoring issues from the day.
4. Sign and date the matrix (consensus sheet) to certify the scores are recorded accurately.

Scorer Training

Students' responses to CR and ER items are individually read and scored by Pearson. Using DDOE-approved training materials, Pearson scoring directors and supervisors train readers to score the science open-ended portion of the test. Scorers attend all training and prove they have internalized the project standards by qualifying on item-specific content. Only qualified scorers are allowed to score the DE NGSS.

All scorers complete training and qualifying in order to score the test. To maintain security of test items, student responses, data, and employees, the following safeguards are employed:

Pearson allows only controlled access to the facility.

- Scoring personnel sign a Confidentiality and Acknowledgement agreement when hired in which they agree not to use or divulge any information concerning test items, scoring guides, or individual student work.
- All staff display Pearson identification badges at all times while in the scoring facility.
- Pearson allows no recording or photographic equipment in the scoring area without the consent of Pearson or the DDDE.

Supervisors and scorers for the science test are selected based on their ability to commit to the duration of the project and to the professional standards of scoring, including their willingness to complete the entire training program. Pearson strives to hire only scorers that have science experience. Scorers are required to demonstrate an understanding of the scoring criteria and to meet the project's qualification standards (acceptable scores on qualifying sets).

The training includes the following information:

1. Overview of Pearson
2. Overview of DE NCSS
3. Reader Bias Training
4. Training goals and objectives

5. Item Training
6. Overview of how to use the ePEN2 scoring system

Supervisor Training

Prior to scorer training, scoring directors train supervisors on the items their team scores. Content training for supervisors follows the same steps as scorer training. Scoring supervisors complete training for all items in the upfront supervisor training window. Supervisors receive training on backreading, providing feedback to scorers, scoring issue documentation, condition codes, resolution scoring, and scorer documentation. Supervisors also receive training on the supervisor tools in the image-based scoring system.

Scorer Training

Scoring directors train one item per scoring group for operational scoring. When scoring on an item is complete, scoring directors train scorers on a new item. Scorers are required to qualify on each new item. Each scoring group scores 2-3 items.

The training process for each item consists of the following materials:

1. Scoring Guide (which includes the rubric, the item, and item stimulus for the constructed response and extended response items), the anchor set, and anchor annotations.
2. 2 practice sets
3. 2 qualifying sets

Led by their scoring directors, scorers review their first item, and discuss the anchors. Scorers then take the first practice set in the image-based scoring system, and assign scores to these sample responses. Scorer performance on practice set 1 is recorded in reports in the image-based scoring system. Once a scorer completes the set, he/she then reviews the true scores and annotations for the given practice set; if they have any questions about the scores or annotations in the practice set, the scoring director is available to answer those questions. The same process occurs for the second practice set. If scorer performance or discussion of practice sets indicates any need for review or retraining with the Scoring Director, it occurs at that time. When all scorers complete those practice sets, everyone moves on to qualification sets.

Finally, scorers complete the two qualification sets, each consisting of 10 sample student responses. Scoring directors and scoring supervisors monitor scorers' progress on each qualification set through online reports. If scorer performance on qualification set 1 indicates any need for review and discussion with the Scoring Director, it occurs at that time. The scores achieved on these qualification sets determine if a trainee understands and can apply the scoring criteria. The table below shows the Qualification, inter-rater reliability (IRR), and Validity thresholds. (These statistics are described in subsequent sections of this chapter.)

Table 6.1 Qualification, Inter-rater Reliability (IRR), and Validity Standards

Item Type by Score Points	Qualification (take 2 sets, must pass 1)	IRR	Validity
0–4	70% perfect and 90% perfect plus adjacent agreement	65%	65%
0–3	70% perfect and 90% perfect plus adjacent agreement	70%	70%
0–2	80% perfect and 90% perfect plus adjacent agreement	80%	80%

Qualified scorers receive training on how to identify responses (alerts and condition codes) that need to be sent to scoring directors or scoring supervisors, as well as how to navigate and use the image-based scoring system. Training on the types of responses that may receive condition codes occurs after scorer qualification. Scorers are trained to recognize these types of responses and to forward them to scoring directors, but scorers do not assign condition codes themselves. Blanks are auto-scored.

Scoring directors and supervisors are responsible for assigning condition codes.

Scoring and Monitoring

All scoring is computer-based, with a 10 percent second scoring for operational items. Scorers begin scoring each item immediately after qualification. Scorers do not know whether a response has received a previous score.

There are three generic rubrics used to score CR items based on the maximum points earnable; 0 to 2, 0 to 3, and 0 to 4. For responses scored by two scorers, the first score is the score of record where scores are adjacent (one-point difference). Backread scores override first scorers. Resolution reads are required where there is a two-point or greater difference between two readers. In such cases, the "expert" third score applied by scoring directors or supervisors override the scores of the previous two readers.

Field test scoring consists of approximately 1200-1500 responses per item and is 10 percent second scored.

The following highlights the quality measures that scoring services staff takes to ensure accurate scoring of DE NGSS.

Backreading

Backreading is one of the primary responsibilities of scoring directors and scoring supervisors and starts at the beginning of scoring. It is an immediate source of information on scoring accuracy. It alerts scoring directors and scoring supervisors to misconceptions at the team level, allowing them to quickly calibrate or retrain scorers. Backreading continues throughout the scoring of the project. Approximately five percent of the scored responses will be reviewed through backreading. To help ensure that students receive accurate scores, scores assigned in the backreading queue will override scores assigned in the first or second scoring queue.

Findings from backreading result in any or all of the following:

- The supervisor clarifies the issue(s).
- Scorers review training materials.
- Supervisor backreads the scorers' work more extensively.
- Supervisory staff gives scorers further training.
- Supervisor monitors reports for improvement.

If a scorer's inter-rater reliability and/or validity statistics fall below the expected rate (see Table 6.1), scoring supervisors increase backreading on the scorer. If a scorer has low backreading agreement, an intervention log is opened for that scorer. This log provides documentation of the steps taken to retrain the scorer and is signed by the scorer. The scoring director determines whether the same issue or trend is being experienced by several scorers and determines the need for a calibration set.

General Calibration

Calibration sets are administered as project leadership deems necessary. Calibration provides a way to proactively promote accuracy by exploring project- or item-specific issues, score boundaries, or types of responses particularly challenging to score consistently. General calibration sets consist of 2-3 papers, address a single issue, and are administered online.

Scorers who fall below acceptable standards may be required to complete a targeted calibration before being dismissed from the project.

Validity

Pre-scored validation responses are used to verify that scorers are applying the same standards throughout the project, and early indications of reader drift from the standards are watched for closely. Validity papers are presented blind; scorers cannot distinguish them from live responses. Validity papers are prepared by item and administered on a regular schedule (at least 3 percent of responses). Validity papers are interspersed with and indistinguishable from unscored student responses. All validity papers are approved by the DDOE. Exact agreement rates and exact plus adjacent agreement rates for the validity responses for the general online forms are presented in Table 6.2 at the end of this chapter. All items met the validity standards presented in table 6.1.

Inter-Rater Reliability

This reliability statistic allows scoring leadership to monitor individual and group scoring agreement. The statistic reflects a level of agreement between two scorers' scores to the same student response. Monitoring allows scoring supervisory staff to target individuals for increased backreading, feedback, and—if necessary—retraining. Readers with less than expected IRR (see Table 6.1) are monitored closely and their work is backread at a higher rate. Exact agreement rates and exact plus adjacent agreement rates for the inter-rater reliability (IRR) for the general online forms are presented in Table 6.2 at the end of this chapter. All items met the IRR standards presented in table 6.1.

Frequency Distribution

The number or percentage of scores assigned at each score point of a given rubric. This is calculated at the scorer and item levels. Anomalous scoring trends are evaluated in conjunction with validity and other statistics which allow for intervention as needed with the individuals involved to ensure that individual drift has not occurred. Frequency distribution reports are available to the DDOE.

Validity Reports

Validity reports are used to identify struggling scorers (scorer below the validity requirement and/or significantly below the room average) or room drift (as a group, the scorers are scoring an item incorrectly or inconsistently). These reports are also used to determine whether a scorer is misunderstanding a particular issue. Validity as Review is an extension of the validity process whereby select validity responses are annotated and used to provide feedback to scorers. If a validity response is scored incorrectly, it subsequently appears on the scorer's screen with the true score, the score they assigned, and an annotation explaining the true score. In this way, this quality monitoring tool serves an immediate, valuable secondary function: that of automated real-time feedback.

If struggling scorers or room drift is identified, scoring directors and scoring supervisors will follow the same procedure described in backreading. All reports are monitored daily by the scoring director(s), the content specialist, and the project manager.

Hand-Scoring Results

Table 6.2 presents IRR and Validity statistics for each operational hand-scored item. Exact IRR percentages ranged from 79.7% to 89.3%, and exact Validity percentages ranged from 89.0% to 95.6%. These percentages are well above the acceptable minimums for IRR and Validity specified in Table 6.1. As denoted in the table, statistics were calculated separately for each trait for multi-trait items.

Table 6.2 2019 ITA Hand Scoring Statistics

Single / Multi Trait	Item Name	n	Trait	Score Points	Human 1st Score Count	Human 2nd Score Count	SYS (Auto-Scores)	Validity Read Count	Exact Validity %	Exact + Adj. Validity %	Exact IRR %	Exact + Adj IRR %	SP 0 %	SP 1 %	SP 2 %	SP 3 %
Multi-Trait	Grade 5 Item 1	11,112	Part A	0,1,2	11,097	1,109	18	386	94.3%	100.0%	86.9%	99.3%	80.0%	16.7%	2.8%	
			Part B	0,1,2	11,097	1,109	18	386	93.3%	100.0%	86.1%	99.4%	80.3%	16.6%	2.4%	
	Grade 8 Item 1	10,526	Part A	0,1,2	10,472	1,046	61	364	95.6%	100.0%	85.5%	98.8%	48.3%	30.7%	19.4%	
			Part B	0,1,2	10,472	1,046	61	364	92.0%	100.0%	86.1%	98.5%	80.9%	13.4%	3.5%	
Single-Trait	Grade 5 Item 2	11,112	Overall	0,1,2	11,085	1,107	32	391	98.2%	100.0%	89.3%	99.6%	47.7%	41.3%	10.3%	
	Grade 8 Item 2	10,528	Overall	0,1,2	10,502	1,049	30	364	89.0%	98.9%	84.4%	97.9%	76.9%	14.6%	7.7%	
	Biology Item 1	10,363	Overall	0,1,2	10,312	1,031	57	344	92.7%	100.0%	85.5%	99.3%	36.1%	48.2%	10.9%	
	Biology Item 2	10,362	Overall	0,1,2,3	10,322	1,032	45	341	92.4%	99.1%	79.7%	98.0%	34.5%	16.9%	28.2%	16.8%
								Totals	**93.5%**	**99.8%**	**85.5%**	**98.9%**				

Note: “Item 1” and “Item 2” refer to the first and second hand-scored item on each assessment, not their sequence on the operational test.

Chapter 7: Classical Item Analysis

This section describes the results of the classical item analysis conducted for data obtained from the 2019 ITA administration. A set of classical item statistics were computed for each operational and field test item. The following statistics and associated flagging rules were used to identify items that were not performing as expected. Appendix A presents classical item analysis summaries for the 2019 ITA operational tests.

Classical Item Difficulty Indices (P-Value and Average Item Score)

Item difficulty offers an index of how easy or hard a given test question is to answer correctly or to earn a given score point for items scored according to a rubric. Item difficulty statistics are used by test developers to help construct test forms that contain a range of items from easy to hard. For items that appear to be unexpectedly difficult, this may indicate students' lack of familiarity with the item type or students' limited opportunity to learn the content represented in the item and are worth further review.

For dichotomously scored items (items scored correct or incorrect), item difficulty is indicated by its p-value, which is the proportion of test takers who answered that item correctly. The possible range for p-values is from 0.00 to 1.00. Items with high p-values are easy items and those with low p-values are difficult items. Dichotomously scored items were flagged for further review if the p-value was above 0.90 (i.e., too easy) or below 0.20 (i.e., too difficult).

For polytomously scored items (items scored according to a rubric with multiple points awarded), difficulty is indicated by the item mean score (IMS). The IMS can range from 0.00 to the maximum total possible points for an item. To facilitate interpretation, the IMS values for polytomously scored items are often expressed as percentages of the maximum possible score, which are equivalent to the p-values of dichotomously scored items. The desired p-value range for polytomously scored items is also 0.20 to 0.90; items with values outside this range were flagged for review.

P-values and item means for the operational items are reported in Appendix A. For grade 5, p-values ranged from 0.104 to 0.873, with a mean of 0.410. For grade 8, p-values ranged from 0.142 to 0.696, with a mean of 0.399. For Biology, p-values ranged from 0.139 to 0.840, with a mean of 0.478. These statistics suggest that the ITAs overall were relatively difficult.

Table 7.1 presents the total number of items flagged for p-value in 2019.

Table 7.1. 2019 Flagged Items for P-value

Grades	Flagged Operational items	Total Operational Items	Flagged Field Test Items	Total Field Test Items
Grade 5	2	26	9	53
Grade 8	2	35	14	67
Biology	1	35	6	65

Flagged items are subjected to additional content review to ensure that the item does not contain flaws and is appropriate for the test. Every item flagged for p-value in 2019 was due to the item being very difficult (i.e. p-value below 0.20).

Item-Total Score Correlation

This statistic describes the relationship between test takers' performance on a specific item and their performance on the total test. The item-total correlation is usually referred to as the item discrimination index. For ITA item analysis, the total score on the assessment was used as the total test score for both operational and field test items. The point-biserial correlation was calculated for both selected response items and constructed response items as an estimate of the correlation between an observed continuous variable and an unobserved continuous variable hypothesized to underlie the variable with ordered categories (Olsson, Drasgow, and Dorans, 1982). Item-total correlations can range from -1.00 to 1.00. Desired values are positive and larger than 0.20. Negative item-total correlations indicate that low ability test takers perform better on an item than high ability test takers, an indication that the item may be potentially flawed.

Item-total correlations for the operational items are reported in Appendix A. For grade 5, item-total correlations ranged from 0.081 to 0.593, with a mean of 0.393. For grade 8, correlations ranged from 0.173 to 0.668, with a mean of 0.400. For Biology, correlations ranged from 0.111 to 0.710, with a mean of 0.429.

Table 7.2 presents the total number of items flagged for low item-total correlations in 2019. Flagged items are subjected to additional content review to ensure that the item does not contain flaws and is appropriate for the test.

Table 7.2. 2019 Flagged Items for Low Item-Total Correlations

Grades	Flagged Operational items	Total Operational Items	Flagged Field Test Items	Total Field Test Items
Grade 5	4	26	5	53
Grade 8	1	35	16	67
Biology	3	35	11	65

Percentage of Students Choosing each Response Option or Earning each Score Point

Selected response items refer primarily to single-select multiple-choice items. These items require that the test taker select a single response from a number of answer options. These statistics for single-select multiple-choice items indicate the percentage of students who select each of the answer options. Also included are the percentage of students that omit the item. These statistics give an indication of whether the items are functioning well as a whole. Anomalies can indicate problems with item functioning, such as multiple correct answers or non-functioning distractors.

Constructed response items are scored according to rubrics in determining the number of points to award a given response. For these items and other non-MC items, the statistics indicate the percentage of students who earn each possible score point. The percentage of students omitting the items are also indicated.

Differential Item Functioning

Differential item functioning (DIF) analyses were conducted using the data obtained from the operational ITAs. If an item performs differentially across identifiable subgroups (e.g., gender, ethnicity, English learners (ELs), or students with disabilities (SWD)) when students are matched on ability, this may indicate an issue with fairness or that the item may be measuring something other than the intended construct (i.e., possible evidence of DIF). It is important, however, to recognize that item performance differences flagged for DIF might be related to actual differences in relevant knowledge or skills (item impact) or statistical Type I error. As a result, DIF statistics are used to identify potential biases. Subsequent reviews by content experts and bias/sensitivity committees are required to determine the source and meaning of performance differences.

This section provides information about differential item functioning (DIF) analyses used for the 2019 ITAs for both operational and field test items. The reference group was either male, Caucasian students, non ELs, or non SWDs, and the focal group was either female, African-American students, Hispanic students, ELs, or SWDs. It is the business rule of Pearson to conduct DIF analyses when the following sample size requirements are met:

- at least 100 students in the reference group
- at least 100 students in the focal group
- at least 400 students in the combined group (reference plus focal)

For example, if there are 150 students each in the reference and focal groups, DIF analyses will not be conducted, because there are only 300 students in the combined group.

The Mantel-Haenszel (MH) DIF statistic was calculated for selected-response items and for dichotomously-scored constructed-response items. The Mantel-Haenszel chi-square statistic is computed as

$$MH - \chi^2 = \frac{\left(\sum_k F_k - \sum_k E\left(F_k\right)\right)^2}{\sum_k Var(F_k)},$$

where F_k is the sum of scores for the focal group at the *k*th level of the matching variable (Zwick, Donoghue, & Grima 1993). Note that the MH statistic is sensitive to *N* such that larger sample sizes increase the value of chi-square.

In addition to the MH chi-square statistic, the MH delta statistic (ΔMH) was computed. Educational Testing Service (ETS) first developed the ΔMH DIF statistic. To compute the ΔMH DIF, the MH alpha (the odds ratio) is first computed

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^{K} N_{f1k} N_{r0k} / N_k}$$

Where N_{r1k} is the number of correct responses in the reference group at ability level *k*, N_{f0k} is the number of incorrect responses in the focal group at ability level *k*, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level *k*, and N_{r0k} is the number of incorrect responses in the reference group at ability level *k*. The ΔMH DIF is computed as

$$\Delta \text{MH DIF} = -2.35\ln(\alpha_{MH}).$$

Positive values of ΔMH DIF indicate items that favor the focal group whereas negative values of ΔMH DIF indicate items that favor the reference group.

For polytomously scored constructed-response items, the standardized mean difference (SMD) (Dorans & Schmitt, 1991; Zwick, Thayer & Mazzeo, 1997; Dorans, 2013), in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959), is used to identify items with DIF. This statistic compares the means of the reference and focal groups, adjusting for differences in the distribution of the reference and focal group members across the values of the matching variable.

$$SMD = \sum_k P_{Fk} m_{Fk} - \sum\nolimits_k P_{Fk} m_{Rk}$$

where

$P_{Fk} = \frac{n_{F+k}}{n_{F++}}$, the proportion of the focal group members who are at the *k*th level of the matching variable,

$m_{Fk} = \frac{1}{n_{F+k}} x(\sum_t y_t n_{Ftk}))$, the mean item score of the focal group members at the *k*th level, and

m_{Rk} = the analogous value for the reference group.

The SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights for the reference group are applied to make the weighted number of the reference group students the same as in the focal group within the same ability.

Classification of DIF statistics

Based on the DIF statistics and significance tests, items are classified into three categories and assigned values of A, B, or C (Zieky, 1993). Category A items contain negligible DIF, Category B items exhibit slight to moderate DIF, and Category C items possess moderate to large DIF values. Positive values indicate that, conditional on the total score, the focal group has a higher mean item score than the reference group. In contrast, negative DIF values indicate that, conditional on the total test score, the focal group has a lower mean item score than the reference group. The flagging criteria for dichotomously scored items are presented in Table 7.3; the flagging criteria for polytomously scored constructed response items are provided in Table 7.4.

Table 7.3. DIF Categories for Dichotomous Selected Response and Constructed Response Items

DIF Category	Criteria
A (negligible)	The Mantel Chi-square is not significantly different from zero, or the absolute value of ΔMH DIF is less than one.
B (slight to moderate)	1. The Mantel Chi-square is significantly different from zero but not from one, and the absolute value of ΔMH DIF is at least one; OR 2. The Mantel Chi-square is significantly different from one, but the absolute value of ΔMH DIF is less than 1.5. Positive values are classified as "B+" and negative values as "B-".
C (moderate to large)	The Mantel Chi-square is significantly different from one, and the absolute value of ΔMH DIF is at least 1.5. Positive values are classified as "C+" and negative values as "C-".

Table 7.4. DIF Categories for Polytomous Constructed Response Item

DIF Category	Criteria
A (negligible)	Mantel Chi-square *p value* > 0.05 or |*SMD/SD*| ≤ 0.17
B (slight to moderate)	Mantel Chi-square *p value* < 0.05 and | *SMD/SD* | > 0.17
C (moderate to large)	Mantel Chi-square *p value* < 0.05 and | *SMD/SD* | > 0.25

Note: SMD: Standardized Mean Difference; SD: total group standard deviation of item score.

Flagging Items for DIF

Items are flagged into one of three categories based on the magnitude of their DIF statistics:

- Category A: no or negligible DIF
- Category B: slight or moderate DIF, and
- Category C: moderate to large values of DIF. These items which exhibit significant DIF, are of primary concern.

2019 DIF results

Appendix B presents DIF results for operational items appearing on the 2019 ITAs. No 2019 operational items were flagged for category B or C DIF for any of the ITAs for Female vs. Male, Black vs. White, or Hispanic vs. White. There was one B-, one B+, and 1 C- item for non-E vs. EL, and there was one B- item for Non-SWD vs. SWD. There were few items with any DIF because any items that showed DIF when field tested during the 2018 standalone field test were avoided during 2019 test construction. Any items that have DIF on the operational tests are reviewed for potential bias before being included on any future tests. If content review determines that flagged items are performing differently due to bias against one or more subgroups, those items are avoided for future operational test forms and marked for revision before additional field-testing.

A summary of the 2019 DIF results for the 2019 embedded field test items is presented in table 7.5. DIF results by item are not reported for the field test items.

Table 7.5. DIF Results for 2019 Field Test Items

Grades	Total Number of Field Test Items	Female vs. Male			Black vs. White			Hispanic vs. White			Non-EL vs. EL			Non-SWD vs. SWD		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Grade 5	53	49	4	0	46	6	1	48	4	1	46	7	0	47	4	2
Grade 8	67	63	3	1	63	3	1	59	7	1	67	0	0	61	6	0
Biology	65	61	3	1	62	3	0	63	2	0	65	0	0	59	6	0

Chapter 8: Calibration, Scaling, and Equating

This section describes calibration, scaling, and equating procedures that took place for the Spring 2019 ITA operational administration. As this administration marks the first operational administration of the ITAs, these procedures include calibration of operational forms to set a base scale and equating of field-test items.

Measurement Models

The Rasch model (Rasch, 1980) and its polytomous extension, the Partial Credit model (PCM) (Masters, 1982) are the item response theory models used to develop and calibrate the 2019 operational ITAs. These measurement models are regularly used to construct test forms, for scaling and equating, and to develop and maintain large item banks in large scale K-12 testing programs. The PCM reduces to the Rasch model for items with only two response categories, such as multiple-choice items. For an item involving *m_i* score categories, the general expression for the probability of scoring *x* on item *i* is given by:

$$P_{xi} = exp\sum_{j=0}^{x}\left(\theta - b_{ij}\right) / \sum_{k=0}^{m_i}\left[exp\sum_{j=0}^{k}\left(\theta - b_{ij}\right)\right]$$

where $x = 0, 1, ..., m_i$, and by definition,

$$\sum_{j=0}^{m}(\theta - b_{ij}) = 0$$

The above equation gives the probability of scoring *x* on the *i*-th test item as a function of ability (θ) and the difficulty (b_{ij}) of the m_i steps of the task. According to this model, the probability of an examinee scoring in a particular category (step) is the sum of the logit (log-odds) differences between θ and b_{ij} of all the completed steps, divided by the sum of the differences of all the steps of a task.

ITA operational items for each respective grade were calibrated according to the Rasch and PCM concurrently and can be found in Appendix C. The following information is provided:

- Item type
- Rasch item difficulty estimate (b_i); for polytomous items, this value is the average of the item step difficulty estimates (b_{ij})
- Standard error (SE) of Rasch item difficulty estimate (b_i)
- Mean-square infit
- Rasch step difficulty estimate (or structure measure estimate, b_{ij}) for polytomous items

The following formula shows how structure measure estimate (b_{ij}) is calculated from both b_i and F_{ij} directly obtained from a run of Winsteps:

$$b_{ij} = b_i + F_{ij},$$

Where b_{ij} = structure measure estimate, b_i = item difficulty estimate, and F_{ij} = structure calibration estimate (i.e., step difficulty estimate).

Finally, the following formulas show how the standard error (SE) of item difficulty estimate (b_i) and structure measure estimate (F_{ij}) were derived (Wright & Masters, 1982):

$$SE(b_i) = 1 \Bigg/ \sqrt{\sum_{n=1}^{N} \left[\sum_{k}^{m_i} k^2 p_{nik} - \left(\sum_{k}^{m_i} k p_{nik} \right)^2 \right]}$$

$$SE(F_{ij}) = 1 \Bigg/ \sqrt{\sum_{n=1}^{N} \left[\sum_{k=0}^{j} p_{nik} - \left(\sum_{k=j+1}^{m_i} p_{nik} \right)^2 \right]}$$

where

$$P_{nix} = exp \sum_{j=0}^{x} (\theta_n - b_{ij}) \Bigg/ \sum_{k=0}^{m_i} \left[exp \sum_{j=0}^{k} (\theta_n - b_{ij}) \right]$$

$x = 0, 1, \dots, m_i$, and

$k = 1, 2, \dots, m_i$.

Fit Statistics for the Rasch Model

Fit statistics are used for evaluating the goodness-of-fit of a model to the data. Fit statistics are calculated by comparing the observed and expected trace lines obtained for an item after parameter estimates are obtained using a particular model. *WINSTEPS* provides fit statistics called mean-squares that show the size of the randomness or amount of distortion of the measurement system. Infit mean-squares are not influenced by outliers.

In general, mean-squares near 1.0 indicate little distortion of the measurement system, while values less than 1.0 indicate observations are too predictable (redundancy, model overfit). Values greater than 1.0 indicate unpredictability (unmodeled noise, model underfit). For the ITAs, items with infit mean-square values above 1.3 are flagged as having poor model fit and will be avoided on future assessments.

Calibration

The Rasch family of item response theory models were used to establish the operational base scales for the ITAs. Since all forms at each grade level share a single a set of common operational items, a single calibration was conducted of all ITA operational items by grade using a single *WINSTEPS* run (WINSTEPS version 3.73; Linacre 2000). *WINSTEPS* uses joint maximum likelihood estimation (JMLE) as described by Wright and Masters (1982) for determining item parameter estimates.

Equating

The 2019 ITAs set the base scales using the operational items. Starting in 2020 through a non-equivalent groups anchor test design (NEAT), anchor items from the 2019 ITAs will serve as linkage to the 2019 base scale. In 2019, the only equating that was conducted was to place the field-test items on the 2019 base scale.

The field-test items were placed on the 2019 scale by using the mean/sigma method (Marco, 1977; Kolen & Brennan, 1995). The mean/sigma equating for field-test items used the set of operational items as anchor items. In addition to base-scale calibration (BS) of operational items, a separate calibration (FT) was conducted of all operational and field-test items. The linear transformation constants needed to place the field-test items on the base scale was determined using the item difficulty estimates of the operational anchor items across the two calibrations.

Using the mean/sigma method, the scaling coefficients slope (X) and intercept (Y) are determined by matching the mean and standard deviations of the anchor items across the forms:

$$X = \frac{\sigma_{b_{BS}}}{\sigma_{b_{FT}}}; \text{ and}$$

$$Y = \mu_{b_{BS}} - X * \mu_{b_{FT}},$$

Where $\sigma_{b_{BS}}$ and $\mu_{b_{BS}}$ are, respectively, the standard deviation and mean of the item difficult parameters of the anchor items for the base-scale calibration, and $\sigma_{b_{FT}}$ and $\mu_{b_{FT}}$ are the standard deviation and mean of the item difficulty parameter of the anchor items for the field-test calibration. After the slope and intercept parameters are determined the rescaled item difficulty parameter b_j^* on the base scale is calculated for the field-test items by linear transformation $b_j^* = X * b_j + Y$, where b_j is the initial item difficulty parameter from the FT calibration.

Scaling

For each ITA, scale scores are linear transformations of the underlying IRT-based (theta) metric where level 2 and level 3 scale score cuts for each ITA are set to 575 and 600, respectively, with the level 4 scale score cut and the lowest and highest obtainable scale scores (LOSS and HOSS) free to vary and determined by the linear scaling transformation that places the level 2 and 3 theta cut scores to 575 and 600. The following linear transformation was used for transforming the underlying Rasch theta scales to the final operational ITA reporting scales:

$$SS = a * \theta + b$$

where the slope (a) and intercept (b) are listed in table 8.1, and θ is the person ability estimate on the base theta scale. Table 8.1 summarizes the scaling constants used for ITA scale score reporting. Table 8.1 also presents all of the ITA cut scores on the IRT (θ) metric as a result of the standard setting held in summer of 2019 (see Chapter 9), as well as on the reporting scale score metric.

Table 8.1. ITA Scaling Values

Grades	Slope (a)	Intercept (*b*)	2019 LOSS	2019 HOSS	Theta Cut Scores			Scale Cut Scores		
					Level 2	Level 3	Level 4	Level 2	Level 3	Level 4
Grade 5	60.386	598.007	321	894	-0.381	0.033	1.274	575	600	676
Grade 8	41.391	599.296	388	822	-0.587	0.017	1.133	575	600	646
Biology	34.388	598.556	405	784	-0.685	0.042	1.129	575	600	637

Chapter 9: Student Scores, Achievement Standards, and Student Performance

Score Interpretation

To help provide appropriate interpretation of the 2019 ITA test scores, two types of scores were created: scale scores and achievement levels with descriptions. As presented in the previous chapter, it was decided that scale scores are reported on a scale ranging from 300 to 900, with 575 designating Level 2 and 600 designating level 3.

In addition to the use of scale scores for reporting results, the ITAs also report on achievement levels. The ITA policy-level achievement level descriptors are presented in table 9.1.

Table 9.1. Policy-level ALDs for ITA Assessments

Achievement Level	Policy-level Achievement Level Descriptors
Level 4	Students in Achievement Level 4 show mastery and thorough understanding of the Delaware Content Standards beyond what is expected at the grade level.
Level 3	Students in Achievement Level 3 show mastery and adequate understanding of the Delaware Content Standards at grade level.
Level 2	Students in Achievement Level 2 show a partial or incomplete understanding of the fundamental skills and knowledge articulated in the Delaware Content Standards.
Level 1	Students in Achievement Level 1 show minimal understanding and evidence of an inability to apply the fundamental skills and knowledge articulated in the Delaware Content Standards

Based on the Delaware regulations (Title 14 Education – Delaware Administrative Code, p. 4)

Scale Scores

As explained in the proceeding section, the 2019 ITAs yield scale scores that range between 300 and 900. As a result of calibration, scaling, and future equating, the scale scores from operational base forms are comparable over time within the same grade, but not across grade levels. Generally, the only inferences that can be appropriately drawn from scale scores are that higher scale scores represent higher performance on the ITAs.

Achievement Levels and Descriptions

The ITAs tests were designed as criterion referenced tests in that they offer indicators of student performance in relation to a set of achievement descriptions premised on the Next Generation Science Standards. Achievement level descriptions (ALDs) describe what students at each of the four levels generally know and can do. These ALDs provide more specific information about students' knowledge and abilities than the policy-level ALDs listed in Table 9.1. In addition, range ALDs were developed for each standard and performance level on the assessments. The range ALDs describe evidence of achievement and how skill changes may be demonstrated across achievement levels. The determination of what ITA scale score values reflect each of the thresholds between achievement levels was determined in the summer of 2019 as a result of standard setting. A description of this process and the more detailed ALDs and range ALDs themselves can be found in the *Delaware System of Student Assessment Science and Social Studies Achievement Level Setting Technical Report*.

Table 9.2 provides scale score ranges for each of the ITA achievement levels by grade.

Table 9.2. ITA Scale Score Ranges by Achievement Level and Grade

Achievement Level	Scale Score Range		
	Grade 5	Grade 8	Biology
Level 1	300-574	300-574	300-574
Level 2	575-599	575-599	575-599
Level 3	600-674	600-645	600-636
Level 4	675-900	646-900	637-900

Student Performance

Tables 9.3, 9.4, and 9.5 present performance information for grades 5, 8, and biology of the 2019 operational ITA administration. Results are presented overall for mean scale score and percentage of students being classified into each of the achievement levels. Additionally, results are also broken out by subgroup.

Table 9.3. 2019 Grade 5 ITA Scale Score and Achievement Level Summary Results

	Scale Scores			% Within Each Achievement Level			
Group	N	Mean	SD	Level 1	Level 2	Level 3	Level 4
Overall	10840	572.9	46.6	51	23	25	2
Female	5300	573.7	45.4	50	24	25	2
Male	5540	572.2	47.8	51	22	25	2
Hispanic\Latino	2047	562.0	40.5	60	24	15	1
Not-Hispanic\Latino	8793	575.5	47.6	48	23	27	2
American Indian or Alaskan Native	29*						
Asian	396	608.0	48.3	22	19	52	7
Black or African American	3297	552.5	41.5	70	19	11	0
Native Hawaiian or Other Pacific Islander	12*						
White	4541	589.4	44.7	35	26	37	3
Multiple Indication	518	574.8	46.1	52	22	25	2
Economic Disadvantage	3901	554.8	40.6	69	18	12	0
English Learner	1306	547.1	35.4	77	18	6	0
Students with Disability	1914	539.4	40.1	81	13	6	0

*Results not reported for subgroups with less than 30 students.

Table 9.4. 2019 Grade 8 ITA Scale Score and Achievement Level Summary Results

	Scale Scores			% Within Each Achievement Level			
Group	N	Mean	SD	Level 1	Level 2	Level 3	Level 4
Overall	10245	579.4	32.2	44	29	25	2
Female	5105	580.4	30.7	42	31	25	2
Male	5140	578.3	33.6	45	27	26	2
Hispanic\Latino	1812	570.5	29.3	56	28	16	1
Not-Hispanic\Latino	8433	581.3	32.5	41	29	27	2
American Indian or Alaskan Native	51	577.9	34.2	25	13	11	2
Asian	388	602.9	31.4	16	28	48	8
Black or African American	3188	567.5	29.6	60	26	14	0
Native Hawaiian or Other Pacific Islander	11*						
White	4408	589.3	30.9	30	31	36	3
Multiple Indication	387	582.7	31.3	39	32	28	1
Economic Disadvantage	3246	567.4	30.0	60	26	14	0
English Learner	487	553.3	25.5	82	16	2	0
Students with Disability	1540	554.0	25.6	80	15	4	0

*Results not reported for subgroups with less than 30 students.

Table 9.5. 2019 Biology ITA Scale Score and Achievement Level Summary Results

	Scale Scores			% Within Each Achievement Level			
Group	N	Mean	SD	Level 1	Level 2	Level 3	Level 4
Overall	10072	594.7	31.3	27	28	37	7
Female	4938	591.4	29.1	22	29	41	7
Male	5134	592.2	33.1	32	27	34	8
Hispanic\Latino	1646	588.5	28.3	33	32	32	4
Not-Hispanic\Latino	8426	596.0	31.7	26	27	38	8
American Indian or Alaskan Native	60	601.6	26.2	18	32	40	10
Asian	415	612.9	30.3	12	16	53	19
Black or African American	3041	583.1	29.6	40	31	27	2
Native Hawaiian or Other Pacific Islander	12*						
White	4609	602.6	30.4	18	26	44	11
Multiple Indication	289	600.3	31.1	19	27	44	10
Economic Disadvantage	2617	582.8	29.3	40	32	25	3
English Learner	436	566.7	22.4	66	27	7	0
Students with Disability	1346	570.5	27.3	61	27	11	1

*Results not reported for subgroups with less than 30 students.

Student performance is also presented in Appendix E with scale score frequency distributions. For all grades, the range of student performance was captured by the 2019 tests, as there were no floor or ceiling affects. The distributions were all positively skewed, with very few students who achieved the highest possible scale scores. This indicates that the 2019 tests were difficult in general for the students. Although the Biology distribution also shows a slight positive skew, it also appears to be bimodal, with one peak near the level 2 cutscore (575) and another near the level 3 cutscore (600).The plots of conditional standard error of measurement (CSEM) provided in Appendix D show that while there is relatively high measurement errors at the extremes of each score scale, it is minimized near the achievement level cut scores (particularly at the Level

3 cut score) and in the region where most students tended to score (i.e., in the 500 to 700 scale score range) on each assessment.

Student performance across demographic subgroups was similar across grades. Several groups showed lower performance as compared with the overall student population. In particular, economically disadvantaged students, English learners, and students with disabilities had lower mean scale scores and higher percentages of students in the lower achievement levels than the overall student population for every grade level, as did students who were Black or African-American. The purpose of the DIF analyses conducted at the item level (see Chapter 7) is to identify items on which subgroup performance differences may be due to factors other than the construct being measured by the test. However, these trends at the overall test level should continue to be monitored to ensure that the ITAs provide an equal opportunity for all students to demonstrate their knowledge and abilities.

Chapter 10: Reliability

Reliability

Reliability coefficients are usually forms of correlation coefficients and must be interpreted within the context and design of the assessment and of the reliability study. The estimates of reliability reported here are internal consistency measures, which are derived from analysis of the consistency of the performance of individuals on items within a test (internal consistency reliability). Therefore, they apply only to the test form being analyzed.

Internal Consistency

The equation displayed below is the formula for the most common index of reliability, namely, Cronbach's coefficient alpha (α; Cronbach, 1951). In this formula, the s_i^2 denotes the variance for the *k* individual items; s_{sum}^2 denotes the variance for the sum of all items.

$$\alpha = \frac{k}{k-1} * \left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_{sum}^2}\right)$$

Standard Error of Measurement

The standard error of measurement (SEM) is commonly used in interpreting and reporting individual test scores and score differences on tests (Harvill, 1991). The SEM is calculated using both the standard deviation and the reliability of test scores, as follows:

$$SEM = \sigma_X \sqrt{1 - P'_{XX}}$$

Where P'_{XX} is the reliability estimate (for example, coefficient alpha) and σ_x is the standard deviation of raw scores on test *X*. A standard error provides some sense of the uncertainty or error in the estimate of the true score using the observed score.

Coefficient alpha and SEM by raw score were calculated by core operations form for grade 5, 8, and biology and are shown in Table 10.1.

Table 10.1 2019 Coefficient Alpha and SEM by Raw Score

Grade	N	Coefficient Alpha	SEM
Grade 5	10788	0.79	3.06
Grade 8	10184	0.84	3.48
Biology	10028	0.88	3.44

The grade 5 value of 0.79, while acceptable, is not as high as the upper grades. This may be due in part to the relatively shorter test length required for the grade 5 assessment, which contained 26 operational items, as compared with grade 8 and Biology assessments, which each contained 35 items. In general, adding items to a test will increase reliability, so it is unsurprising that the longer tests had higher reliability values. It is also true that the NGSS are more complex as compared with previous science standards, and these standards were still in the process of being implemented in Delaware. To that end, the items on the ITAs are quite difficult, which may have reduced total score variability and, therefore reliability. The score frequency distributions in

Appendix E show positively skewed distributions for grades 5 and 8, with the majority of students scoring at the lower end of the score scale. It is expected that as the NGSS continue to be adopted, reliability will increase as students become more familiar with the content, shifting those distributions upward.

Tables 10.2 through 10.4 provides coefficient alpha and SEM by raw score breakdowns by subgroup for total test.

Table 10.2. 2019 Grade 5 ITA Coefficient Alpha and SEM by Subgroup

Group	N	Coefficient Alpha	SEM
Female	5282	0.78	3.08
Male	5506	0.80	3.03
Hispanic\Latino	2012	0.72	2.98
Not-Hispanic\Latino	8776	0.80	3.07
American Indian or Alaskan Native	28*		
Asian	395	0.79	3.18
Black or African American	3291	0.73	2.92
Native Hawaiian or Other Pacific Islander	12*		
White	4532	0.78	3.13
Multiple Indication	518	0.79	3.05
Economic Disadvantage	3874	0.72	2.94
English Learner	1271	0.60	2.85
Students with Disability	1892	0.70	2.78

*Results not reported for subgroups with less than 30 students.

Table 10.3. 2019 Grade 8 ITA Coefficient Alpha and SEM by Subgroup

Group	N	Coefficient Alpha	SEM
Female	5074	0.82	3.50
Male	5110	0.85	3.44
Hispanic\Latino	1779	0.80	3.36
Not-Hispanic\Latino	8405	0.84	3.50
American Indian or Alaskan Native	51	0.86	3.43
Asian	388	0.84	3.65
Black or African American	3172	0.78	3.35
Native Hawaiian or Other Pacific Islander	11*		
White	4398	0.83	3.57
Multiple Indication	385	0.81	3.52
Economic Disadvantage	3223	0.79	3.33
English Learner	455	0.61	3.07
Students with Disability	1526	0.67	3.11

*Results not reported for subgroups with less than 30 students.

Table 10.4. 2019 Biology ITA Coefficient Alpha and SEM by Subgroup

Group	N	Coefficient Alpha	SEM
Female	4920	0.86	3.50
Male	5180	0.89	3.40
Hispanic\Latino	1624	0.87	3.34
Not-Hispanic\Latino	8404	0.88	3.44
American Indian or Alaskan Native	60	0.81	3.88
Asian	415	0.85	3.73
Black or African American	3029	0.86	3.32
Native Hawaiian or Other Pacific Islander	12*		
White	4600	0.87	3.50
Multiple Indication	288	0.89	3.07
Economic Disadvantage	2607	0.86	3.29
English Learner	414	0.78	2.99
Students with Disability	1334	0.84	3.06

*Results not reported for subgroups with less than 30 students.

Tables 10.2 through 10.4 present fairly consistent coefficient alphas across gender and ethnicity, but also show lower coefficient alphas for some groups, particularly English learners (ELs). While this highlights the need for further examination into ELs and the accessibility of the test, these results can be explained at least in part by homogeneity of the EL group as compared to other groups. As illustrated by the standard deviations in tables 9.3 through 9.5, the variability of the test scores of the EL group is considerably less than other groups. When the variability in total test score decreases, then the alpha coefficient is smaller. Because the SEM includes both reliability and variability in its calculation, some of the low reliability observed within subgroups appears to be due in part to the low variability, as the SEMs for these same subgroups are as low as—or lower—than subgroups with higher coefficient alpha values.

While the SEM provides an estimate of the average test score error for all students regardless of their individual proficiency levels, the SEM can vary across the range of student proficiencies. For this reason, it is useful to report not only a test-level SEM estimate but also individual score-level estimates. Individual score-level SEMs are commonly referred to as conditional standard errors of measurement.

Conditional Standard Error of Measurement

Like the SEM, the conditional standard error of measurement (CSEM) reflects the amount of variance in a score resulting from random factors other than what the assessment is designed to measure, but it provides an estimate conditional on proficiency. In other words, the CSEM provides a measurement error estimate at each score point on an assessment.

IRT methods for estimating score-level CSEM are used because test- and item-level difficulties for ITAs are calibrated using the Rasch and PCM measurement models, as described above. By using CSEMs that are specific to each scale score, a more precise error band can be placed around each student's observed score. The CSEM is calculated by first calculating the information function for each item, and then summing them across all items on the test, which results in the test information function. The CSEM at each ability level is the inverse of the square root of the information function at that ability level. CSEMs are converted to the scale score metric by multiplying them by the scaling transformation slope.

Appendix D provides CSEM values for all ITA scale scores and plots of CSEM values by scale score. The dashed lines on the plots indicate the Level 2, Level 3, and Level 4 cut scores. For the 2018–2019 school year, CSEM values for the ITAs in the middle of the scale score ranges where most students performed were approximately 19–20 scale score points for the grade 5 test and 10–12 scale score points for the grade 8 and Biology tests. These scale scores translate to approximately 3 raw score points in the middle range. At the extreme low and high scale score the CSEM can be up to 3 times the value of the middle of the scale score range. While these values can seem quite large, the scale scores at these extremes are a large distance from any of the cut scores and very few students perform at these score points, so the effect of these large values is limited.

Reliability of Classifications Consistency and Accuracy

Reliability of classification estimates the proportion of students who are accurately classified into proficiency levels. There are two kinds of classification reliability statistics provided here: *decision accuracy* and *decision consistency*. Consistency refers to the percentage of students who are classified into the same performance levels if they took two parallel forms of a test, while accuracy refers to the percentage of students who are correctly classified into their true performance levels based on their observed scores on a test. Classification consistency and accuracy are two related but different concepts; high consistency does not necessarily lead to high accuracy, and vice versa. To better understand the classification quality, an analysis of the consistency and accuracy of student classifications into performance levels is conducted based on results of tests for which performance standards have been previously established.

The classification consistency index developed for IRT models (Lee, 2010) is used here. The basic idea is to estimate the probability (P1) of classifying into each achievement level conditional on each test raw score based on an IRT model. For an achievement level and a raw score, the probability (P2) that the raw score is classified into the same achievement level on two parallel forms is just the square of the above probability for one test (P1). Across all performance levels, the probability (P3) that a raw score is consistently classified on two parallel forms is the sum of the above probabilities for two tests and one achievement level (P2). The consistency index for a test is then the sum of the above probabilities (P3) over all raw scores weighted by the observed percentages of students on each raw score.

The method recommended by Rudner (2001, 2005) is adapted here for computing classification accuracy. Under an IRT model, for an estimated theta proficiency score associated with a raw test score, the true proficiency score is expected to be normally distributed with a mean of the estimated theta and an estimated standard deviation of the CSEM. The estimated proficiency score cut for each achievement level is also available. Then, for each raw score point in an achievement level, the probability of correctly classifying into this level can be estimated. The accuracy index is just the sum of these probabilities across all test raw scores weighted by the observed percentages of students on each raw score point.

Table 10.5 provides information about the accuracy and the consistency of the classifications made on the basis of the scores on the 2019 grades 5, 8 and biology ITAs. Information is provided for the classification based on the exact level, and for the classification of students above and below the Level 3 cut score. The data in Table 10.5 suggest the classification of students as Level 3 or Higher (i.e., “At Standard” or above) or Level 2 or Lower (i.e., “Below Standard” or lower) was more than 90% accurate for each assessment, and more than 80% of those classifications would be the same if students were to take a parallel version of these assessments.

Table 10.5. 2019 ITA Classification Accuracy and Consistency Results by Grade

	Decision Accuracy		**Decision Consistency**	
Grade	**Exact Level**	**Level 3 or Higher vs. Level 2 or Lower**	**Exact Level**	**Level 3 or Higher vs. Level 2 or Lower**
Grade 5	76.6	90.3	64.8	83.0
Grade 8	77.9	91.0	65.3	84.0
Biology	78.0	90.8	64.9	83.4

Chapter 11: Quality-Control Procedures

Quality control is a critically important element of every phase of ITA development, administration, and score reporting in ensuring the accuracy of student-, school- and district-level data. Pearson has developed and refined a set of quality procedures to help ensure that all DDOE's testing requirements are met or exceeded. These quality-control procedures are detailed in the paragraphs that follow. In general, Pearson's commitment to quality is incorporated in both task-specific quality standards applied to processing functions and services as well as a network of systems and procedures that coordinate quality steps *across* functions and services.

Quality Control for Test Construction

Following a legally sanctioned test development process (Smisko, Twing, & Denny, 2000), items are selected and placed on particular test forms that are as comparable as possible with respect to content and statistical characteristics. The process is an iterative process involving content experts, psychometricians, and DDOE. The goals are to create forms that meet blueprint and statistical targets using the highest quality items (in terms of content and statistical characteristics) that result in the most comparable test forms. Once an initial core is selected, all responsible parties evaluate and recommend improvements until final *best* core forms have been affirmed and moved to production. Throughout the process, standard checklists are used to ensure all steps are followed.

Quality Control for Printed Documents

Pearson follows a meticulous set of internal quality standards to ensure high-quality printed products. Specific areas of responsibility for staff involved in materials production include monitoring all materials-production schedules to meet contract commitments, overseeing the production of test materials, coordinating detailed printing and post-printing specifications, outlining specific quality control requirements for all materials, and conducting print reviews and quality checks. The quality production and printing processes follow printers' reviews and quality checks. Project management and print procurement staff work closely with the printers duringthe production phase. Press proofs are checked to ensure high-quality printing and to verify adherence to specifications. The printing staff randomly pull documents throughout the print run for additional 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, operational ITAs that will be administered are used. In addition to the UAT, Production Validation (PV) testing occurs. Pearson publishes the ITAs in a production environment and recommends test scenarios. The tests are completed and scoring deliverables are generated during the PV period. The validation process includes confirmation of the tests published and the scoring deliverables. Approvals 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 that include fixes made and/or system upgrades. Any patches are tested and approved before being deployed to the field. Deployments are scheduled outside of the regular testing window timeframes.

Quality Control for Test-Form Calibration and Equating

Test-form calibration and equating are the processes that enables fair and equitable comparisons between test administrations across years. Pearson uses several quality-control procedures to ensure this equating is performed accurately.

1. Pearson performs a statistical "key check" analysis for the multiple-choice (MC) item type to ensure the appropriate scoring key is being used.
2. Pearson performs an "adjudication" analysis for the 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 as intended. Along with the key check analysis, the adjudication process ensures that item calibration is based on accurate item scores.
3. Starting in 2020, for the first year of operational equating, an anchor item stability analysis will be conducted in order to determine whether the Rasch item parameters have shifted over time. Items which have shifted will be investigated and a resolution whether to keep or remove an item within an equating protocol will be made.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, D.C.: Author.

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

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach. (ETS Research Report No. 91-47).* Princeton, NJ: Educational Testing Service.

Dorans, N. J. (2013). *ETS Contributions to the Quantitative Assessment of Item, Test and Score Fairness (ETS R&D Science and Policy Contributions Series, ETS SPC-13-04).* Princeton, NJ: Educational Testing Service.

ESSA (2015). *Every Student Succeeds Act of 2015*, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*, 181-189.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer-Verlag.

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1–17.

Linacre, J. M., & Wright, B. D. (2000). *A user's guide to WINSTEPS: Rasch-model computer program.* Chicago, IL: MESA Press.

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

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22,* 719-748.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160. doi: 10.1111/j.1745-3984.1977.tb00033.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

National Research Council. (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* Board on Science Education, Division of Behavioral and Social Sciences and Education, Committee on a Conceptual

Framework for New K-12 Science Education Standards. Washington, DC: The National Academies Press.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982), The polyserial correlation coefficient, *Biometrika, 47*, 337–347.

Orlando, M. (2004, June). *Critical issues to address when applying item response theory (IRT) models.* Paper presented at the The Drug Information Association, Bethesda, MD.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*(1), 45–58.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14). Available online: http://pareonline.net/getvn.asp?v=7&n=14.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13.

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.

South Carolina Department of Education. (2001). *Technical documentation for the 2000 Palmetto achievement challenge tests of English language arts and mathematics (Technical Report)*. Columbia: South Carolina Department of Education.

Tekkumru-Kisa, Stein, and Schunn. (2015). A framework for analyzing cognitive demand and content-practices integration: Task analysis guide in science. *Journal of Research in Science Teaching, 52*(5), 659-685

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices, 31*, 2–13.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA PRESS.

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development*. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337–348). Hillsdale, NJ: Erlbaum

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items (ETS Research Report RR-97-05).* Princeton, NJ: Educational Testing Service.