New Meridian

Technical Report

New Jersey Graduation Proficiency Assessment

Spring 2024

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The purpose of this technical report is to describe the operational administration of the New Jersey Graduate Proficiency Assessment (NJGPA) for the 2023–2024 academic year, including test form construction, test administration, item scoring, student characteristics, classical item analysis results, reliability results, evidence of validity, item response theory (IRT) calibrations and scaling, performance level-setting procedure, growth measures, and quality control procedures. Throughout this technical report, only New Jersey student data is included in all analyses, descriptions, and data summaries.

1.1. Background

In May 2021, the New Jersey State Board of Education approved for publication in the New Jersey Register a notice of substantial changes to proposed amendments regarding N.J.A.C. 6A:8, Standards and Assessment. The notice included five new amendments, including the provision that students must take a state graduation proficiency assessment in grade 11. The New Jersey Graduation Proficiency Assessment was administered for the first time in the spring of 2022 to students of the class of 2023.

On Tuesday, July 5, 2022, Governor Murphy signed P.L.2022, c.60 (ACS for A-3196/S-2349), which required the State Board of Education to administer the NJGPA as a field test for the class of 2023. Beginning with the class of 2024, a passing score on the NJGPA is the First Pathway for a student to satisfy the New Jersey Graduation Testing requirements.

1.2. Purpose of the NJGPA

Referring to the NJGPA, New Jersey State statute § 18A:7C-6.1 indicates that “The test shall measure those basic skills all students must possess to function politically, economically, and socially in a democratic society.” There are two test components of the NJGPA: the English Language Arts Graduation Proficiency (ELAGP) component and the Mathematics Graduation Proficiency (MATGP) component. Two core operational forms are administered each year for each component: a computer-based test (CBT) administered online and a paper-based test (PBT) form to support students needing paper-based accommodations. Each core operational test form is constructed with reference to assessment blueprints. Score comparability across the core operational forms is assured by including mostly previously administered items, which have been calibrated to item bank measurement scales.

NJGPA ELA forms are based on the grade 10 ELA standards and evidence statements. Operational ELA forms consist of two units, each designed to be completed in 90 minutes. The ELA forms each contain 20 items worth a total of 74 points. Additionally, the ELA forms each contain a Literary Analysis Task (LAT) and a short passage set in unit 1, and a Research Simulation Task (RST) in unit 2.

NJGPA mathematics operational forms consist of items from Algebra I and Geometry. Operational math forms contain 30 items, worth a total of 55 points, administered in two 90-minute units. The MATGP forms utilize type I, II, and III items.

1.3. Overview of the Technical Report

This technical report presents the results of analyses for the New Jersey Graduation Proficiency Assessment (NJGPA) operational administration of the spring 2023 forms. In this document, the term "operational items" refers to the scorable items that contribute to students' raw scores and scale scores. The term "field-test items" refers to the nonscorable items that were newly developed to collect statistics for future test administrations, calibrated by using the field-test spring 2023 administration data. The report begins by providing explanations of the test form construction process, test administration, and scoring of the test items. Subsequent chapters of the report present descriptions of student characteristics and results of statistical analyses, including classical test theory statistics, item response theory (IRT) scaling and parameters and differential item functioning (DIF).

The technical report contains the following chapters:

Chapter 2 – Test Development

This chapter describes the test design and procedures followed during the development of operational test forms and characteristics for the 2023 forms.

Chapter 3 – Test Administration

This chapter presents the operational administration schedule, information regarding test security and confidentiality, accessibility features and accommodations, testing irregularities and security breaches, and administration quality control procedures.

Chapter 4 – Item Scoring

This chapter explains the key-based and rule-based processes for machine-scored items, as well as the training and monitoring processes for human-scored items.

Chapter 5 – Standard Setting

This chapter describes the NJGPA performance levels and the processes followed to establish the performance thresholds for each subject area.

Chapter 6 – Item Analysis

This chapter describes the classical item-level statistics calculated for the operational test data, the flagging criteria used to identify items that performed differently than expected, and the results of these analyses are presented in this section.

Chapter 7 – IRT Analysis, Calibration and Scoring

This chapter presents the information related to the item response theory (IRT) models and the descriptive statistics of the item parameters. The development of the reporting scales and conversion tables, and scale score distributions, are also presented. Note that all tests delivered in 2023 employed a pre-equated model, in which previously estimated item parameters are used to generate scoring tables.

Chapter 8 – Student Demographics and Differential Item Functioning

This chapter describes the rules for inclusion of students in analyses, distributions of students by subject, mode and gender, and distributions of demographic variables of interest. Also, methods for conducting differential item functioning analyses as well as corresponding flagging criteria are described. This is followed by definitions of the comparison groups and subsequent results for the comparison groups.

Chapter 9 – Reliability

This chapter presents the results of scale score reliability and internal consistency reliability analyses and corresponding standard errors of measurement for content area, mode (CBT or PBT) for all students, and subgroups of interest, is provided in this section. This is followed by reliability of classification (i.e., decision accuracy and decision consistency).

Chapter 10 – Validity

This chapter explains the validity evidence based on analyses of the content of the tests and the concurrence of data measuring related properties of the same variables. Correlations between subscores are reported by content area and mode (CBT or PBT) for all students.

Chapter 2. Test Design and Development

2.1. Overview of the Test

The ELAGP blueprint was derived from the Grade 10 NMC summative assessment blueprints with input from New Jersey ELA educators. The MATGP blueprint is similar to the NJSLA Algebra I and Geometry blueprints in item type and quantity and draws exclusively from those two item banks. Core forms are constructed to meet the blueprint and psychometric properties outlined in the test construction specifications.

The NJGPA was administered in either a computer-based test (CBT) or a paper-based test (PBT) format. ELAGP focused on reading comprehension skills and writing effectively when using and analyzing sources. MATGP focused on applying skills and concepts and included multi-step problems that require abstract reasoning and modeling of real-world problems. Each assessment was comprised of multiple units, and one of the mathematics units was split into calculator and non-calculator sections.

2.2. ELA Claims and Subclaims

The ELA GP has one master claim: New Jersey High School Graduation Readiness in ELA. Under the master claim, the ELAGP assessment is comprised of two major claims (Reading Complex Text and Writing) as well as five subclaims: Vocabulary Interpretation and Use, Reading Literature, Reading Informational Text, Written Expression, and Knowledge of Language and Conventions. The form composition of the ELAGP assessment is detailed in Table 2.2.1. As described in the ELA Content Guide, the ELAGP assessment:

- Includes 20 items for a total of 74 points;
- Consists of two 90-minute units;
- Utilizes one test blueprint embedded with an LAT and short passage set (unit 1), and RST (unit 2);
- Includes items aligned to the grade 10 standards and evidence statements;
- Contains the following item types: Evidence-Based Selected Response (EBSR), Technology-Enhanced Constructed Response (TECR), and Prose Constructed Response (PCR) items;
- Incorporates writing tasks that will be scored using the Research Simulation Task and Literary Analysis Task Scoring Rubric.

ELA assessments contain two Prose Constructed Response tasks. These two writing prompts are each scored for two traits: (1) Reading Comprehension and Written Expression, and (2) Knowledge of Language and Conventions. All traits are initially scored as either 0–4 or 0–3. The Written Expression traits are then multiplied by 3 (or weighted) to increase their contribution to the total score, making subclaim scores of 0, 3, 6, 9, and 12 possible. The maximum points possible for ELAGP PCR items are provided in Table 2.2.2.

Table 2.2.1 Form Composition for ELAGP Prose Constructed Response Items

Unit	Task	Claims/Subclaims	EBSR/TECR Points	PCR Points
1	Literary Analysis Task	Reading: Literary Text	8	4
1	Literary Analysis Task	Reading: Vocabulary	4	0
1	Literary Analysis Task	Writing: Written Expression	0	12
1	Literary Analysis Task	Writing: Knowledge of Language and Conventions	0	3
1	Short Passage Set	Reading: Informational Text	6	—
1	Short Passage Set	Reading: Vocabulary	2	—
2	Research Simulation Task	Reading: Informational Text	12	4
2	Research Simulation Task	Reading: Vocabulary	4	0
2	Research Simulation Task	Writing: Written Expression	0	12
2	Research Simulation Task	Writing: Knowledge of Language and Conventions	0	3

Table 2.2.2 Contribution of Prose Constructed Response Items to ELAGP

	Possible Points	
Score	**Literary Analysis Task Points**	**Research Simulation Task Points**
Reading Comprehension	4	4
Written Expression	12	12
Knowledge of Language and Conventions	3	3
Total	19	19

2.3. Mathematics Claims and Subclaims

As described in the Mathematics Content Guide for NJGPA, the math component consists exclusively of NJSLA-Algebra I and NJSLA-Geometry items. The form composition of MATGP is described in Table 2.3.1.

Table 2.3.1 Mathematics Form Composition for MATGP

	Subclaims	Number of Items	Number of Points
Mathematics			
	Major Content	13–15	16–18
	Additional & Supporting Content	9–11	12–14
	Expressing Mathematical Reasoning	3	10
	Modeling and Applications	3	15
Total		30	55

2.4. Test Development Activities

The test forms used in NJGPA testing are based on the NMC Affiliate program blueprints for ELA grade 10 and Integrated Math II summative assessments. For the MATGP assessment, only Algebra I and Geometry content were included; Algebra II was excluded. Test development activities were conducted by

New Meridian Corporation (NMC) under the guidance of NJDOE content leads, with input from New Jersey educators.

The process of creating forms is undertaken each year. Potential forms that meet the content guidelines and psychometric standards for both test components are identified by the NMC psychometric team and reviewed by NMC content specialists for subject matter appropriateness. Potential forms that meet psychometric and content guidelines are then reviewed by NJDOE staff and by New Jersey educators during test form verification.

Test items used in the ELA and Math components are drawn from banks of items used for NMC Affiliate ELA grade 10, Algebra I, and Geometry. There are no field-test items on the MATGP assessment.

2.4.1. Item Development Process

Items used on the NJGPA were created for the PARCC/NMC Affiliate program. The PARCC/NMC Affiliate program item development process is described in this section.

Test and item development activities were conducted by Pearson under the guidance and oversight of the K–12 state leads, the Higher Education Leadership Team, the Technical Advisory Committee, the Operational Working Group (OWG) members from each of the member states, the Text and Content Item Review Committees, and staff members from New Meridian, the project manager.

Developing high-quality assessment content with authentic stimuli for computer-based tests (CBT) and paper-based tests (PBT) measuring rigorous standards was a complex process involving the services of many experts, including assessment designers, psychometricians, managers, trainers, content providers, content experts, editors, artists, programmers, technicians, human scorers, advisors, and members of the OWGs.

2.4.1.1. Text Selection for ELA

Using the Passage Selection Guidelines, English language arts subject matter experts were trained to search for appropriate passages to support an annual pool of passages for consideration. Guided by the test specifications documents, contracted subject matter experts worked to deliver the number of texts specified in the annual asset development plan. The Passage Selection Guidelines provided a text complexity framework and guidance on selecting a variety of text types and passages that allowed for a range of standards/evidences to be demonstrated to meet the assessment claims. ELA tests are based on authentic texts, including multimedia stimuli. Authentic texts are grade-appropriate texts that are not developed for the assessment or to achieve a particular readability metric but reflect the original language of the authors. Staff content experts reviewed the passages for adherence to the Passage Selection Guidelines to meet the annual asset development plan described above in the number and distribution of genres and topics prior to review and consideration by the Text Review Committee. ELA item development was not conducted until after texts were approved by the Text Review Committee.

2.4.1.2. Item Development

Item writers were recruited, trained, and managed to develop the number of items specified in the annual asset development plan. Prior to committee reviews, staff reviewed the items for content accuracy, alignment to the standards, range of difficulty, adherence to universal design principles (which maximize the participation of the widest possible range of students), bias and sensitivity, and copy editing to enable the accurate measurement of the standards.

2.4.2. Item and Text Review Committees

Members of the OWGs for ELA and mathematics, state-level experts, local educators, post-secondary faculty, and community members conducted rigorous reviews of every item and passage being developed for the summative assessment system to ensure all test items are of the highest quality, aligned to the standards, and fair for all student populations. All reviewers were nominated by their state education agency. The purpose of the educator reviews was to provide feedback on the quality, accuracy, alignment, and appropriateness of the test passages and items developed annually for the summative assessments. The meetings were conducted either in person or virtually and included large group training on the expectations and processes of each meeting, followed by breakout meetings of subject-level working committees where additional training was provided.

2.4.2.1. Text Review

The Text Review Committee meets to review and approve the texts eligible for item development. Participants reviewed and provided feedback about the grade-level appropriateness, content, and potential bias concerns, and reached consensus about which texts would move forward for development. The Text Review Committee was made up of members of both Content Item Review and Bias and Sensitivity Review Committees.

2.4.2.2. Content Item Review

During Content Item Review, committees reviewed and edited test items for adherence to the foundational documents, basic universal design principles, Accessibility Guidelines, associated item metadata, and the Style Guide. Committees accessed the item content within the Pearson Assessment Banking for Building and Interoperability (ABBI) system, which previews how the passages and items will be displayed in an operational online environment. Committees also verified that the appropriate scoring rule had been applied to each item. The Content Item Review Committees were made up of OWG members and educators nominated by participating states.

2.4.2.3. Bias and Sensitivity Review

Educators and community members made up the committee that reviewed items and tasks to confirm that there were no bias or sensitivity issues that would interfere with a student's ability to achieve his or her best performance. The committee reviewed items and tasks to evaluate adherence to the Fairness and Sensitivity Guidelines and to ensure that items and tasks do not unfairly advantage or disadvantage one student or group of students over another. Bias and Sensitivity Committee members made edits and modifications to items and passages to eliminate sources of bias and improve accessibility for all students.

2.4.2.4. Editorial Review

The Editorial Review Committee consisted of editors who reviewed up to 10 percent of the items and tasks. The committee reviewed the items for grammar, punctuation, clarity, and adherence to the Style Guide.

2.4.2.5. Data Review

Following testing, ELA educator content and bias committee members met to evaluate field-tested items and associated performance data. The focus of data review is to review items with questionable performance data. Items with data that suggest good performance are generally accepted into item banks. Items demonstrating poor performance data are excluded from item banks. Items with questionable performance data are reviewed for appropriateness, level of difficulty, and potential gender, ethnicity, or other bias. The data review committees then recommended acceptance or rejection of questionable field-test items for inclusion in item banks. Items that were approved by the committee are eligible for use on operational summative assessments.

2.4.3. Operational Test Construction

In conjunction with NJDOE content specialists, NMC creates one online core form and one paper-based core form for each NJGPA subject component. Core forms are constructed to meet the blueprint and psychometric properties outlined in the test construction specifications. Operationally, each component is assessed via one online operational form for the general student population.

Students requiring paper-based accommodations are assessed by the paper accommodated form (AC2). Other accommodations are applied to the online core form. Pearson tracks the online accommodated form(s) as an operational form distinct from the operational online form for the general student population. This form is referred to as the online accommodated form (AC1). The AC1 form contains the same items as the online operational form, but the items have been modified for the specific accommodation of the form (e.g., closed captioning). For the MATGP assessment only, an additional operational form, the Spanish language form, is created by Pearson. The accommodations available in each content area are outlined in section 2.4.3.3.

2.4.3.1. Test Construction Activities

After the data review meetings and prior to the test construction meetings, psychometricians constructed initial versions of all core forms. Content specialists reviewed the initial core forms based on the support documents and specific processes to achieve fair parallel forms. The following steps were used to construct the operational core forms taken to the Test Construction Committee for review.

1. Online forms were constructed to match the blueprint and test construction specifications.
2. Paper forms were constructed to match the blueprint and test construction specifications.
3. Accommodated and accessibility forms were constructed to match the blueprint, test construction specifications, and Accessibility, Accommodations, and Fairness (AAF) constraints.

The test construction process included iterative steps between content specialists and psychometricians. Custom test construction reports generated by the New Meridian psychometric team provided information on adherence to blueprint and statistical averages/distributions of item difficulty and discrimination, describing the forms and allowing comparison of the forms. These reports facilitated content changes to

better achieve the test construction goals. Linking across administrations for operational forms was accomplished by including prior operational items on the current operational test forms.

NMC assessment specialists identified forms for each NJGPA subject component suitable for use as the accommodated forms. The content of these forms was also reviewed by accessibility specialists, allowing for content changes prior to the Test Construction Committee meetings.

These test construction activities provided the meeting materials necessary to conduct test form verification meetings. These meeting materials included:

- the proposed items for the initial operational core forms and the accommodated forms described above;
- reports describing each form and comparing parallel forms; and
- recommended accommodated forms.

2.4.3.2. Test Form Verification Meeting to Review Test Construction Inputs

Members of the Content Item Review Committees and the AAF experts participated in the building of operational core forms that met the summative assessment requirements. During this process, members met in a virtual meeting to review and make recommendations for changes so that test forms conformed to both the content and psychometric requirements of the assessment.

2.4.3.3. Accommodated Form Review Process

In addition to participating in many of the development activities, including the Text Review and the Bias and Sensitivity Review meetings, the AAF experts reviewed the proposed accommodated forms at the Test Form Verification meeting for accessibility to make sure that the content can be accommodated for students with disabilities and English learners without changing the underlying measured construct.

Forms were identified to support the following accommodations:

Accommodated Base 1 (AC1)

- Closed captioning
- Text-to-speech first form
- Spanish online
- Spanish text-to-speech

Accommodated Base 2 (AC2)

- Spanish paper (also serves Spanish LP, Spanish human reader paper)
- Spanish human reader/human signer online
- Base accommodated paper (serves braille, LP, human reader paper)
- Human reader/human signer online
- Assistive technology screen reader
- Assistive technology non-screen reader
- American Sign Language (ASL)

Accommodated Base 3 (mathematics only)

- Text-to-speech second form

Spanish is used for mathematics only. Closed captioning is used for ELA only.

At the conclusion of the meetings, all test forms were constructed to meet test blueprints and requirements, and if necessary, reflect the operational linking design. Each test form reflected the test blueprint in terms of content, item types, and test length, as well as expected difficulty and performance along the ability continuum. Linking sets were proportionally representative of the operational test blueprint. The operational core forms, linking item set, and field-test item sets were reviewed during the test form verification meeting and approved prior to the test administration.

2.4.3.4. Spanish-Language Assessments for Mathematics

For English learners, the mathematics assessments are offered in Spanish, as well as in Spanish-language large print and text-to-speech versions. Once the operational form was approved, the items were transadapted. Transadaptation differs from translation in that it takes into consideration the grade-level appropriateness of the words, as well as the linguistic and cultural differences that exist between speakers of two different languages. Accounting for these differences allows the item to measure the achievement of Spanish language speakers in the same way that the original version of the item does for native speakers of English. The Spanish Glossary provided guidance to the translator in grade-level and culturally appropriate transadaptation. For the Spanish language text-to-speech form, the alternate text (used for description and/or text in art and graphics) was transadapted from the alternate text for the English language version of the text-to-speech form. Phonetic markup, which guides how the text-to-speech reader pronounces content-specific words and phrases, was also applied in this process.

In addition to the expert review of potential content for all accommodated forms conducted by the AAF experts with assistance from content experts at the test construction meetings, the transadapted forms underwent additional quality checks: a Pearson Spanish copy edit services review and approval, and an AAF expert review and approval.

Chapter 3. Assessment Administration

3.1 Test Security and Administration Policies

The administration of the NJGPA is a secure testing event. Maintaining the security of test materials before, during, and after the test administration is crucial to obtaining valid and reliable results. School Test Coordinators are responsible for ensuring that all personnel with authorized access to secure materials are trained in and subsequently act in accordance with all security requirements.

School Test Coordinators must implement chain-of-custody requirements for specified materials. School Test Coordinators are responsible for distributing materials to Test Administrators, collecting materials from Test Administrators, returning secure test materials, and securely destroying certain specified materials after testing.

The administration of the summative assessment includes both secure and nonsecure materials, and these materials are further delineated by whether they are "scorable" or "nonscorable," depending on whether the assessments were administered via paper/pencil (i.e., paper-based assessments) or online (i.e., computer-based assessments).

3.1.1. Secure vs. Nonsecure Materials

Participating states and agencies define secure materials as those that must be closely monitored and tracked to prevent unauthorized access to or prohibited use or distribution of secure content, such as test items, reading passages, student work, and so on. For paper-based tests (PBTs), secure materials include both used and unused test booklets and used scratch paper, while for computer-based tests (CBTs), secure materials include student testing tickets, secure administration scripts (e.g., mathematics read-aloud), and used scratch paper. Nonsecure materials are defined as any authorized testing materials that do not include secure content (e.g., test items or student work). These include test administration manuals, unused scratch paper, and mathematics reference sheets that have not been written on, and so on.

3.1.2. Scorable vs. Nonscorable Materials

Paper-based assessments have both scorable and nonscorable materials, while computer-based assessments have only nonscorable materials. Scorable materials for paper-based assessments consist of used. Scorable materials must be returned to the vendor to be scored. All other materials for PBTs, such as blank (i.e., unused) test booklets, test administration manuals, scratch paper, mathematics reference sheets, and so on, are deemed nonscorable. For CBTs, there are no scorable materials as student work is submitted electronically for scoring. Thus, there are limited physical materials to return (e.g., secure administration scripts for certain accommodations).

Students taking the CBT may not have access to secure test materials, including printed student testing tickets, prior to testing. Printed mathematics reference sheets (if applicable) and scratch paper must be new and unmarked.

Students taking the PBT may not have access to scorable or nonscorable secure test content before or after testing. Scorable secure materials that are to be provided by Test Administrators to

students include answer documents. Nonscorable secure materials that are distributed by Test Administrators to paper-based testing students include large-print test booklets, braille test booklets, scratch paper (paper used by students to take notes and work through items), and printed mathematics reference sheets.

School Test Coordinators are required to maintain a tracking log to account for the collection and destruction of test materials, including mathematics reference sheets and scratch paper written on by students. As part of the test administration policy, schools are required to maintain the Chain-of-Custody Form or tracking log of secure materials for at least three years unless otherwise directed by state policy. Copies of the Chain-of-Custody Form for paper-based testing are included in each local education agency (LEA) or school's test materials shipment.

Test administrators are not to have extended access to test materials before or after administration (except for certain accessibility or accommodations purposes). Test administrators must document the receipt and return of all secure test materials (used and unused) to the school test coordinator immediately after testing.

All test security and administration policies are found in the *Test Coordinator Manual* and the *Test Administrator Manual*.

3.2. Accessibility Features and Accommodations

3.2.1. Participation Guidelines for Assessments

All students, including students with disabilities (SWDs) and English learners (ELs), are required to participate in statewide assessments and have their assessment results be part of the state's accountability systems, with narrow exceptions for ELs in their first year in a U.S. school, and certain SWDs who have been identified by the Individualized Education Program (IEP) team to take their state's alternate assessment. Federal laws governing student participation in statewide assessments include the Individuals with Disabilities Education Act of 2004 (IDEA), Section 504 of the Rehabilitation Act of 1973 (reauthorized in 2008), and the Elementary and Secondary Education Act (ESEA) of 1965, as amended by the Every Student Succeeds Act (ESSA). All students can receive accessibility features on the summative assessments.

Four distinct groups of students may receive accommodations on the summative assessments:

1. SWDs who have an IEP.
2. Students with a Section 504 plan who have a physical or mental impairment that substantially limits one or more major life activities, have a record of such an impairment, or are regarded as having such an impairment but who do not qualify for special education services.
3. Students who are ELs.
4. Students who are ELs with disabilities and have an IEP or 504 plan.

These students are eligible for accommodations intended for both SWDs and ELs. Testing accommodations for SWDs or students who are ELs must be documented according to the guidelines and requirements outlined in the *Accessibility Features and Accommodations Manual*.

3.2.2. Accessibility System

Through a combination of universal design principles and accessibility features, participating states and agencies designed an inclusive assessment system by considering accessibility from initial design and through item development, field testing, and implementation of the assessments for all students, including SWDs, ELs, and ELs with disabilities. Accommodations may still be needed for some SWDs and ELs to assist in demonstrating what they know and can do. However, the accessibility features available to students should minimize the need for accommodations during testing and ensure the inclusive, accessible, and fair testing of the diverse students being assessed.

3.2.3. What are Accessibility Features?

On computer-based assessments, accessibility features are tools or preferences that are either built into the assessment system or provided externally by Test Administrators and may be used by any student taking the summative assessments (i.e., students with and without disabilities, gifted students, ELs, and ELs with disabilities). Since accessibility features are intended for all students, they are not classified as accommodations. Students should have the opportunity to select and practice using the accessibility features prior to testing to determine which are appropriate for use on the assessment. Consideration should be given to the supports a student finds helpful and consistently uses during instruction. Practice tests that include accessibility features are available for teacher and student use throughout the year.

3.2.4. Accommodations for Students with Disabilities and English Learners

It is important to ensure that performance in the classroom and on assessments is influenced minimally, if at all, by a student's disability or linguistic/cultural characteristics that may be unrelated to the content being assessed. For the summative assessments, accommodations are considered adjustments to the testing conditions, test format, or test administration that provide equitable access during assessments for SWDs and students who are ELs. In general, the administration of the assessment should not be the first occasion on which an accommodation is introduced to the student. To the extent possible, accommodations should do the following:

- Provide equitable access during instruction and assessments.
- Mitigate the effects of a student's disability.
- Not reduce learning or performance expectations.
- Not change the construct being assessed.
- Not compromise the integrity or validity of the assessment.

Accommodations are intended to reduce and/or eliminate the effects of a student's disability and/or English language proficiency level; however, accommodations should never reduce learning expectations by reducing the scope, complexity or rigor of an assessment. Moreover, accommodations provided to a student on the summative assessments must be generally consistent with those provided for classroom instruction and classroom assessments. There are some accommodations that may be used for instruction and for formative assessments that are not allowed for a summative assessment because they impact the validity of the assessment results; for example, allowing a student to use a thesaurus or access the internet during an

assessment. There may be consequences (e.g., excluding a student's test score) for the use of non-allowable accommodations during assessments. It is important for educators to become familiar with NJDOE policies regarding accommodations used for assessments.

To the extent possible, accommodations should adhere to the following principles:

- Accommodations enable students to participate more fully and fairly in instruction and assessments and demonstrate their knowledge and skills.
- Accommodations should be based upon an individual student's needs rather than on the category of a student's disability, level of English language proficiency alone, level of or access to grade-level instruction, amount of time spent in a general classroom, current program setting, or availability of staff.
- Accommodations should be based on a documented need in the instruction/assessment setting and should not be provided for the purpose of giving the student an enhancement that could be viewed as an unfair advantage.
- Accommodations for SWDs must be described and documented in the student's appropriate plan (i.e., either a 504 plan or an approved IEP) and must be provided if they are listed.
- Accommodations for ELs should be described and documented.
- Students who are ELs with disabilities are eligible to receive accommodations for both SWDs and ELs.
- Accommodations should become part of the student's program of daily instruction as soon as possible after completion and approval of the appropriate plan.
- Accommodations should not be introduced for the first time during the testing of a student.
- Accommodations should be monitored for effectiveness.
- Accommodations used for instruction should also be used, if allowable, on local district assessments and state assessments.
- In the following scenarios, the school must follow each state's policies and procedures for notifying the state assessment office:
    - A student was provided a test accommodation that was not listed in his or her IEP/504 plan/documentation for an English learner, or
    - A student was not provided a test accommodation that was listed in his or her IEP/504 plan/documentation for an English learner.

3.2.5. Unique Accommodations

A comprehensive list of accessibility features and accommodations designed to increase access to the summative assessments and that will result in valid, comparable assessment scores was provided in the *Accessibility Features and Accommodations Manual*. However, SWDs or ELs may require additional accommodations that are not already listed. Participating states and agencies individually review requests for unique accommodations in their respective states and provide a determination as to whether the accommodation would result in a valid score for the student, and if so, would approve the request.

3.2.6. Emergency Accommodations

Emergency accommodation may be appropriate for a student who incurs a temporary disabling

condition that interferes with test performance shortly before or during the assessment window. A student, whether or not they already have an IEP or 504 plan, may require accommodation as a result of a recently occurring accident or illness. Cases include a student who has a recently fractured limb (e.g., arm, wrist, or shoulder); a student whose only pair of eyeglasses has broken; or a student returning to school after a serious or prolonged illness or injury. Emergency accommodation should be given only if the accommodation will result in a valid score for the student (i.e., does not change the construct being measured by the test[s]). If the principal (or designee) determines that a student requires an emergency accommodation on the summative assessment, an Emergency Accommodation Form must be completed and maintained in the student's assessment file. The parent must be notified that emergency accommodation was provided. If appropriate, the Emergency Accommodation Form may also be submitted to the District Assessment Coordinator to be retained in the student's central office file. Requests for emergency accommodations will be approved after it is determined that use of the accommodation would result in a valid score for the student.

3.2.7. Student Refusal Form

If a student refuses an accommodation listed in his or her IEP, 504 plan, or (if required by the member state) an EL plan, the school should document in writing that the student refused the accommodation, and the accommodation must be offered and remain available to the student during testing. This form must be completed and placed in the student's file and a copy must be sent to the parent on the day of refusal. Principals (or designee) should work with Test Administrators to determine who, if any others, should be informed when a student refuses an accommodation documented in an IEP, 504, or (if required by the member state) EL plan.

3.3. Testing Irregularities and Security Breaches

Any action that compromises test security or score validity is prohibited. These may be classified as testing irregularities or security breaches. Below are examples of activities that compromise test security or score validity (note that these lists are not exhaustive). It is highly recommended that School Test Coordinators discuss other possible testing irregularities and security breaches with Test Administrators during training.

Examples of test security breaches and irregularities include but are not limited to:

Electronic Devices:

- Using a cell phone or other prohibited handheld electronic device (e.g., smartphone, iPod, smart watch, personal scanner) while secure test materials are distributed, while students are testing, after a student turns in his or her test materials, or during a break. (*Exception:* Test Coordinators, Technology Coordinators, Test Administrators, and Proctors are permitted to use cell phones in the testing environment only in cases of emergencies or when timely administration assistance is needed; LEAs may set additional restrictions on allowable devices as needed.)

Test Supervision:

- Coaching students during testing, including giving students verbal or nonverbal cues, hints, suggestions, or paraphrasing or defining any part of the test.
- Engaging in activities (e.g., grading papers, reading a book, newspaper, or magazine) that prevent proper student supervision at all times while secure test materials are still distributed or while students are testing.
- Leaving students unattended for any period of time while secure test materials are distributed or while students are testing.
- Deviating from testing time procedures.
- Allowing cheating of any kind.
- Providing unauthorized persons with access to secure materials.
- Unlocking a test in PearsonAccess^{next} during non-testing times.
- Failing to provide a student with a documented accommodation or providing a student with an accommodation that is not documented and therefore not appropriate.
- Allowing students to test before or after the state's test administration window.

Test Materials:

- Losing a student test booklet or answer document.
- Losing a student testing ticket.
- Leaving test materials unattended or failing to keep test materials secure at all times.
- Reading or viewing the passages or test items before, during, or after testing (*Exception:* Administration of a human reader/signer accessibility feature for mathematics or accommodation for English language arts, which requires a Test Administrator to access passages or test items).
- Copying or reproducing (e.g., taking a picture of) any part of the passages or test items or any secure test materials or online test forms.
- Revealing or discussing passages or test items with anyone, including students and school staff, through verbal exchange, email, social media, or any other form of communication.
- Removing secure test materials from the school's campus or removing them from locked storage for any purpose other than administering the test.

Testing Environment:

- Allowing unauthorized visitors in the testing environment.
- Failing to follow administration directions exactly as specified in the *Test Administrator Manual*.
- Displaying testing aids in the testing environment (e.g., a bulletin board containing relevant instructional materials) during testing.

All instances of security breaches and testing irregularities must be reported to the School Test Coordinator immediately. The Form to Report a Testing Irregularity or Security Breach must be completed within two school days of the incident.

If any situation occurred that could cause any part of the test administration to be compromised, schools should refer to the *Test Coordinator Manual* for each state's policy and immediately follow those steps. Instructions for the School Test Coordinator or LEA Test Coordinator to report a testing irregularity or security breach are available in the *Test Coordinator Manual*.

3.4. Data Forensics Analyses

Maintaining the validity of test scores is essential in any high-stakes assessment program, and misconduct represents a serious threat to test score validity. When used appropriately, data forensic analyses can serve as integral components of a wider test security protocol. The results of these data forensic analyses may be instrumental in identifying potential cases of misconduct for further follow-up and investigation.

The following data forensics analyses were conducted on the operational assessments:

- Response Change Analysis.
- Aberrant Response Analysis
- Plagiarism Analysis
- Longitudinal Performance Modeling
- Internet and Social Media Monitoring
- Off-Hours Testing Monitoring

An overview of each data forensics analysis method is provided next.

3.4.1. Response Change Analysis

Response change analysis looks at how often student answers are changed, focusing specifically on an excessive number of wrong answers changed to right answers. In traditional paper-based, multiple-choice testing programs, this is sometimes referred to as "erasure analysis." The rationale for erasure analysis is that a teacher or administrator who is intent on improving classroom performance might be motivated to change student responses after the answer sheets are collected. A clustered number of student answer documents from the same school or classroom with unusually high numbers of answers changed from wrong to right might provide evidence to support follow-up investigation. The response change analysis extended the traditional erasure method to account for issues specific to computer-based testing as well as the variety of item types on the summative assessments, such as partial-credit, multi-part, and multiple-select items.

3.4.2. Aberrant Response Analysis

Aberrant response pattern detection analysis looks at the unusualness of student responses compared with what would be expected. Most simply, this can be thought of as quantifying the extent to which higher-scoring students miss easy questions and lower-scoring students answer difficult questions correctly. While it would be difficult to draw a definitive inference about a single student flagged as having an aberrant response pattern, a cluster of students with aberrant response patterns within a classroom or school might warrant further investigation.

3.4.3. Plagiarism Analysis

Plagiarism analysis compares the responses given for a group of written composition items, looking for high degrees of similarity. For the summative assessments, the primary item type of interest was the prose constructed-response tasks in the English language arts content area. This analysis was conducted for prose constructed-response tasks administered online using some of the same artificial intelligence techniques that are applied in automated essay scoring. Specifically, this method was based on latent semantic analysis (LSA) technology to detect possible plagiarism. Using LSA, the content of each constructed response was compared against the content of every other constructed response, and a measure that indicated the degree of similarity was generated for each pair of response comparison. Because LSA provided a semantic representation of language, rather than a syntactic or word-based representation, it allowed the detection of potential copying behaviors, even when students or administrators substituted synonymous words or phrases.

3.4.4. Longitudinal Performance Monitoring

Longitudinal performance modeling evaluates the performance on the summative assessments across test administrations and identifies unusual performance gains in the unit of interest (e.g., school or district). A weighted least squares (WLS) regression methodology was evaluated and recommended by the Technical Advisory Committee for implementation starting in spring 2017. The WLS identified unusual changes in test performance across two consecutive administrations of the assessment. In the WLS regression approach, mean current year scale scores are regressed on mean prior year scale scores, weighting by unit sample size.

Standardized residuals are calculated by dividing raw residuals by their respective standard deviations. Units with a standardized residual exceeding 3.0 are flagged for unexpected performance.

3.4.5. Internet and Social Media Monitoring

Internet and social media monitoring were conducted by Caveon LLC. Caveon's team monitored English-language websites and searchable forums that were publicly available for suspected proxy testing solicitations and website postings that contain, or appear to contain, infringements of protected operational test content.

The internet and social media outlets monitored included popular websites (such as Facebook and

Twitter), blogs, discussion forums, video archives, document archives, brain dumps, auction sites, media outlets, peer-to-peer servers, and so on. Caveon's process generated regular updates that categorize identified threats by level of actual or potential risk based upon the representations made on the websites, or actual analysis of the proffered content. For example, categorizations typically ranged from "cleared" (lowest risk but bookmarked for continued monitoring) to "severe" (highest risk). Note that this process only considered potential breaches of secure item content, not violations of testing administration policies. Potential breaches were reported directly to the state(s) implicated for further action. Summary reports describing the threats were provided through notification emails.

3.4.6. Off-Hours Testing Monitoring

Off-hours testing monitoring is a process that checks for suspicious activity occurring outside computer-based testing windows. Participating states and agencies have set start and end times for administering computer-based assessments. Authorized users in the state role were allowed to override these start and end times. The off-hours testing monitoring tracked such overrides and logged them in an operational report. States could use this report to follow up with the organizations.

3.5. Quality Control of Test Administration

Pearson provided high-quality materials in a timely and efficient manner to meet the needs of the test administration. Since the majority of printing work was done in-house, it was possible to fully control the production environment, press schedule, and quality process for print materials.

Additionally, strict security requirements were employed to protect secure materials production. Chapter 3 provides details on the secure handling of test materials. Materials were produced according to the style guide and to the detailed specifications supplied in the materials list.

Pearson Print Service operates within the sanctions of an ISO 9001:2008 Quality Management System, and practices process improvement through lean principles and employee involvement.

Raw materials (paper and ink) used for scannable forms production were manufactured exclusively for Pearson Print Service using specifications created by Pearson Print Service. Samples of ink and paper were tested by Pearson prior to use in production. Project specialists were the point of contact for incoming production.

Purchase orders and other order information were assessed against manufacturing capabilities and assigned to the optimal production methodology. Expectations, quality requirements, and cost considerations were foremost in these decisions. Prior to release for manufacture, order information was checked against specifications, technical requirements, and other communication that includes expected outcomes. Records of these checks were maintained.

Files for image creation flow through one of two file preparation functions: digital pre-press for digital print methodology, or plateroom for offset print methodology. Both the digital prepress and plateroom functions verify content, file naming, imposition, pagination, numbering stream, registration of technical components, color mapping, workflow, and file integrity. Records of these checks are created and saved.

Offset production requires printing that uses a lithographic process. Offline finishing activities are required to create books and package offset output. Digital output may flow through an inkjet digital production line or a sheet-fed toner application process in the Xpress Center. A battery of quality checks was performed in these areas. The checks included color match, correct file selection, content match to proof, litho-code to serial number synchronization, registration of technical components, ink density controlled by densitometry, inspection for print flaws, perforations, punching, pagination, scanning requirements, and any unique features specified for the order. Records of these checks and samples pulled from planned production points were maintained. Offline finishing included cutting, shrink-wrapping, folding, and collating. The collation process has three robust inline detection systems that inspected each book for:

- Caliper validation that detects too few or too many pages. This detector will stop the collator if an incorrect caliper reading is registered.
- An optical reader that will only accept one sheet. Two or zero sheets will result in a collator stoppage.
- The correct bar code for the signature being assembled. An incorrect or upside-down signature will be rejected by the bar code scanner and will result in a collator stoppage.

Pearson's Quality Assurance department personnel inspected print output prior to collation and shipment. Quality Assurance also supported process improvement, work area documentation, audited process adherence, and established training programs for employees.

Chapter 4. Item Scoring

4.1. Machine-scored Items

4.1.1. Key-based Items

Pearson performed a key review prior to the test administration to verify that the scoring (answer) keys were correct for each item. Once the forms were constructed and approved for publication, an independent key review was performed by an experienced third-party vendor. The vendor reviewed each item and confirmed that the key was correct. If discrepancies were identified, a Pearson senior content specialist or content manager reviewed the flagged item(s) and worked with the New Meridian content staff to resolve the issue.

4.1.2. Rule-based Items

Rule-based scoring refers to item types that use various scoring models. Participating states and agencies use Question and Test Interoperability item type implementation based on scoring model rules. Examples of these item types include choice interaction, which presents a set of choices where one or more choices can be selected; text entry, where the response is entered in a text box; hot spot or text interaction, where an area in a graph or text in a paragraph (for example) can be highlighted; or match interaction, where an association can be made between pairs of choices in a set. These items include the scoring rules and correct responses as part of their item XML (markup language) coding.

During the initial stages of item development, Pearson staff worked closely with participating states and agencies to first delineate the rules for the scoring rubrics and then adjust those rules based on student responses. During item studies in spring 2015, Pearson content staff received input from the staff of participating states and agencies to develop a thorough rule-based scoring process that met their needs.

Pearson worked with the item developers to review initial scoring rules created during item development. Once the rule-based scoring process was approved, and prior to test construction, Pearson content staff worked closely with the item developers to finalize scoring rubrics for items to be scored via the rule-based scoring method. The proposed scoring rubrics were sent for review, and if any additional changes were needed or new rules added, Pearson documented and applied the requested edits.

During test construction, Pearson monitored and evaluated the scoring and updated the scoring keys/scoring rules in the item bank. After the tryout items were scored, Pearson prepared a frequency distribution of student responses for each scored item or task using a rule-based approach and compared this to the expected response based on correct answers to ensure that scoring keys and rules were appropriately applied. The content team analyzed the student response data to determine if scoring was acceptable using the item metadata and the student

response file in conjunction with any potential item issues as flagged by psychometrics. These frequency distributions included an indication of right/wrong and other identifying information defined by participating states and agencies, and those items that showed a statistical anomaly, whereby the frequency distribution was outside of the expected range, were sent to content experts to verify that the items were coded with the correct key.

Following the Rule-Based Scoring Educator Committee's review, which occurred prior to year one test construction, Pearson analyzed the feedback from the committees and made recommendations about adjustments to the scoring rubrics based on the results of the reviews. Upon submission of the results, Pearson worked with the staff of participating states and agencies to discuss these findings and determine next steps prior to the completion of scoring.

Following the initial development of the rule-based scoring rubrics, Pearson has continued to monitor and evaluate new item development to ensure the scoring rules established are maintained within all item types as approved.

Pearson continues to use several avenues to monitor scoring each year. Prior to testing, a third-party key review checks operational and field-test items for correct keys. Any disputed items go to a second review with Pearson content experts, and anything still in question is taken before the task force for review and possible key change. During testing, Pearson creates early testing files for frequency distribution analysis, whereby items for which an incorrect key receives a high distribution of responses are further evaluated for accuracy. After testing, all responses are again evaluated for the distribution of responses and potential scoring abnormalities during psychometric analysis. These processes are the same for both paper and online modes of testing.

4.2. Human or Hand-scored Items

Constructed-response items were scored by human scorers in a process referred to as hand scoring. Online training units were used to train all scorers. The online training units included prompts (items), passages, rubrics, training sets and qualification sets. Scorers who successfully completed the training and demonstrated they could correctly score student responses based on the guidelines in the online training units were permitted to score student responses using the ePEN2 (Electronic Performance Evaluation Network, second generation) scoring platform. All online and paper responses were scored within the ePEN2 system. Pearson monitored quality throughout scoring.

Pearson staff roles and responsibilities were as follows:

- Scorers applied scores to student responses.
- Scoring supervisors monitored the work of a team of scorers through review of scorer statistics and backreading, which is a review of responses scored by each scorer. When backreading, a supervisor sees the scores applied by scorers, which helps the supervisor provide additional coaching or instruction to the scorer being backread.
- Scoring directors managed the scoring quality of a subset of items and monitored the work of supervisors and scorers for their assigned items. Directors backread responses scored by supervisors and scorers as part of their quality-monitoring duties.

- English language arts (ELA) and mathematics content specialists managed the scoring quality and monitored the work of the scoring directors.
- The project manager documented the procedures, identified risks, and managed day-to-day administrative matters.
- A portfolio manager provided oversight for the entire scoring process.

All Pearson employees involved in the scoring or the supervision of scoring possessed at least a four-year college degree.

4.2.1. Scorer Training

Key steps in the development of scorer training materials were range-finding and rangefinder review meetings, where educators and administrators from states met to interpret the scoring rubrics and determine consensus scores for student responses. Range-finding meetings were held prior to scoring field-test items, and range-finding review meetings were held prior to scoring operational items.

At range-finding meetings, educators and administrators from states reviewed student responses and used scoring rubrics to determine consensus scores. Those responses scored in rangefinding were used to create field-test scorer training sets. After items were selected for operational testing, educators and administrators attended rangefinder review meetings to review and approve proposed operational scorer training sets.

When developing scorer training materials, Pearson scoring directors carefully reviewed detailed notes and records from range-finding and rangefinder review committee meetings. Training sets were developed using the responses scored by the committees and additional suitable student response samples (as needed). All scorer training sets were reviewed and approved prior to scorer training.

During training, scorers reviewed training sets of scored student responses with annotations that explained the rationale for the score assigned. The anchor set was the primary reference for scorers as they internalized the rubric during training. Each anchor set consisted of responses that were clear examples of student performance at each score point. The responses selected were representative of typical approaches to the task and arranged to reflect a continuum of performance. All scorers had access to the anchor set when they were training and scoring and were directed to refer to it regularly during scoring.

Practice sets were used in training to help trainees practice applying the scoring guidelines. Scorers reviewed the anchor sets, scored the practice sets, and then were able to compare their assigned scores for the practice sets to the actual assigned scores to help them learn.

Qualification sets were used to confirm that scorers understood how to score student responses accurately. Qualification sets were composed of responses that were clear examples of score points. Scorers were required to meet specified agreement percentages on qualification sets in order to score student responses.

Pearson has developed two types of training sets to train scorers: prototype and abbreviated sets. Prototype training sets were complete training sets consisting of anchor, practice, and qualification sets (refer to 4.2.2 for information on the qualification process). In ELA, there was one prototype training set per task type (i.e., Research Simulation, Literary Analysis, Narrative Writing). In mathematics, a prototype training set was built for a grouping of similar items for a total of approximately three to four prototype sets per course.

The prototype training approach promoted consistency in scoring, as each subsequent abbreviated training set for the ELA task type or mathematics item grouping was based on the prototype. Once a prototype was chosen, full training materials were developed for that item. Then, scorers at each grade level were trained to score a particular item type using the prototype training materials for that type.

Abbreviated training sets were prepared for all items not selected for prototype training sets. The abbreviated training sets included an anchor set and two practice sets so scorers could internalize the scoring standards for these new items, which were similar to prototype items they had previously scored.

Anchor and practice sets for both prototype and abbreviated items included annotations for each response. Annotations are formal written explanations of the score for each student response.

Table 4.2.1 details the composition of the anchor sets, practice sets, and qualification sets.

Table 4.2.1 Training Materials Used During Scoring

Training Set Development	
Description	**Specification**
Anchor Set	
The anchor set is the primary reference for scorers as they internalize the rubric during training. All scorers have access to the anchor set when they are training and scoring and are directed to refer to it regularly.	The anchor set for mathematics prototype items comprises three annotated responses per score point. The anchor set for subsequent abbreviated items for mathematics comprises one to three annotated responses per score point.

The anchor set comprises clear examples of student performance at each score point. The responses selected may be representative of typical approaches to the task or arranged to reflect a continuum of performance.	The anchor sets for ELA prototype items comprise three annotated responses per score point. Anchor sets for prototype items include separate complete anchor sets for each applicable scoring trait (Reading Comprehension and Written Expression, and Conventions for Research Simulation and Literary Analysis tasks, Written Expression for Narrative Writing tasks, and Knowledge of Language and Conventions for all task types).
Practice Sets	
Practice sets are used to help trainees develop experience in independently applying the scoring guide (the rubric) to student responses. Some of these responses clearly reinforce the scoring guidelines presented in the anchor set. Other responses are selected because they are more difficult to evaluate, fall near the boundary between two score categories, or represent unusual approaches to the task. The practice sets provide guidance and practice for trainees in defining the line between score categories, as well as applying the scoring criteria to a wider range of types of responses.	The practice sets for the mathematics prototype and abbreviated items include two to three sets of ten annotated responses. ELA practice sets for prototype items include two sets of five annotated responses and two sets of 10 annotated responses. The subsequent ELA practice sets for abbreviated items include two sets of ten annotated responses.
Training Set Development	
Description	**Specification**
Qualification Sets	

Qualification sets are used to confirm that scorer trainees understand the scoring criteria and are able to assign scores to student responses accurately. The responses in these sets are selected to reinforce the application of the scoring criteria illustrated in the anchor set. Scorer trainees must demonstrate acceptable performance on these sets by meeting a predetermined standard for accuracy in order to qualify to score. Pearson scoring staff defined and documented qualifying standards in conjunction with participating states and agencies prior to scoring.	The qualification sets for mathematics prototype items include three sets of 10 responses each (not annotated). The subsequent mathematics abbreviated items for mathematics do not include qualification sets. The qualification sets for ELA prototype items include three sets of 10 responses each (not annotated). The subsequent ELA abbreviated items do not include qualification sets.

4.2.2. Scorer Qualification

To score items, scorers were required to show that they were able to apply scoring methodology accurately through a qualification process. Scorers were asked to apply scores to three qualification sets consisting of 10 responses each. ELA scorers applied a score for each trait on each response in the qualification sets. Literary Analysis and Research Simulation tasks each had two traits: the Reading Comprehension and Written Expression trait and the Knowledge of Language and Conventions trait. The Narrative Writing Task had two traits: Written Expression and Knowledge of Language and Conventions. Mathematics scorers applied a score for each part of an item that was a constructed response. The number of constructed-response parts for each mathematics item ranged from one to four. Scorers were required to match the approved score at a percentage agreed to by participating states and agencies to qualify.

For ELA qualification, scorers were required to meet the following three conditions:

1. On at least one of the three qualifying sets, at least 70 percent of the ratings on each of the two scoring traits (considered separately) must agree exactly with the approved scores.
2. On at least two of the three qualifying sets, at least 70 percent of the ratings (combined across the three scoring traits) must agree exactly with the approved scores.
3. Combining over the three qualifying sets and across the two scoring traits, at least 96 percent of the ratings must be within one point of the approved scores.

For mathematics qualification, the requirements were based on the item type and score point range. Because mathematics items can have one or more scoring traits, a scorer needed to achieve the requirements as set forth in Table 4.2.2 separately for each scoring trait (when applicable to the item).

Table 4.2.2 Mathematics Qualification Requirements

Category	Score Point Range	Perfect Agreement	Within One Point
2	0–1	90%	100%
3	0–2	80%	96%
4	0–3	70%	96%
5	0–4	70%	95%
6	0–5	70%	95%
7	0–6	70%	95%

On at least two of the three qualifying sets, a scorer was required to meet the "perfect agreement" percentage indicated in the table above for each category. "Perfect agreement" was achieved when the scores applied exactly matched the approved scores. Over the three qualifying sets, a scorer was required to meet the "within one point" percentage indicated in the table above for each category. The average is exclusive to each trait, so an item with multiple scoring traits would have multiple-trait rating averages within one point of the approved score.

4.2.3. Managing Scoring

Pearson created a hand-scoring specifications document that detailed the hand-scoring schedule, customer requirements, range-finding plans, quality management plans, item information, and staffing plans for each scoring administration.

4.2.4. Monitoring Scoring

4.2.4.1. Second Scoring

During scoring, Pearson's ePEN2 scoring system automatically and randomly distributed a minimum of 10 percent of student responses for second scoring; scorers had no indication whether a response had been scored previously. Humans applied the second score for all mathematics items. The second scoring for ELA was performed either by human scorers or by Pearson's Intelligent Essay Assessor. If the first and second scores applied were nonadjacent, a third and occasionally a fourth score were assigned to resolve scorer disagreements. When a resolution score (i.e., third score) was nonadjacent to one or both of the first and second scores, the content specialist or scoring director would apply an adjudication score (fourth score). If a response was scored more than once, the rules in Table 4.2.3 were applied to determine the final score.

Table 4.2.3 Scoring Hierarchy Rules

Score Type	Rank	Final Score Calculation
Adjudication	1	If an adjudication score is assigned, this is the final score.
Resolution	2	If no adjudication score is assigned, this is the final score.

Score Type	Rank	Final Score Calculation
Backread	3	If no adjudication or resolution score is assigned, the latest backreading score is the final score.
Human First Score	4	If no adjudication, resolution, or backreading score is assigned, this is the final score.
Human Second Score	5	If no adjudication, resolution, backreading, or human first score is assigned, this is the final score.
Intelligent Essay Assessor Score	6	If no human score is assigned, this is the final score.

4.2.4.2. Backreading

Backreading was one of the major responsibilities of Pearson Scoring Supervisors and a primary tool for proactively guarding against scorer drift, where scorers score responses in comparison to one another instead of in comparison to the training responses. Scoring supervisory staff used the ePEN2 backreading tool to review scores assigned to individual student responses by any given scorer in order to confirm that the scores were correctly assigned and to give feedback and remediation to individual scorers. Pearson backread approximately 5 percent of the hand-scored responses. Backreading scores did not override the original score but were used to monitor scorer performance.

4.2.4.3. Validity

Validity responses are prescored responses strategically interspersed in the pool of live responses. These responses were not distinguishable from any other responses so that scorers were not aware they were scoring validity responses rather than live responses. The use of validity responses provided an objective measure that helped ensure that scorers were applying the same standards throughout the project. In addition, validity was at times shared with scorers in a process known as “validity as review.” Validity as review provided scorers automated, immediate feedback: a chance to review responses they mis-scored, with reference to the correct score and a brief explanation of that score. One validity response was sent to scorers for every 25 “live” responses scored.

Validity agreement requirements for scorers are listed in Table 4.2.4. Scorers had to meet the required validity agreement percentages to continue working on the project. Scorers who did not maintain expected agreement statistics were given a series of interventions culminating in a targeted calibration set: a test of scorer knowledge. Scorers who did not pass targeted calibration were removed from scoring the item, and all the scores they assigned were deleted.

Table 4.2.4 Scoring Validity Agreement Requirements

Subject	Score Point Range	Perfect Agreement	Within One Point
Mathematics	0–1	90%	96%
Mathematics	0–2	80%	96%

Subject	Score Point Range	Perfect Agreement	Within One Point
Mathematics	0–3	70%	96%
Mathematics	0–4	65%	95%
Mathematics	0–5	65%	95%
Mathematics	0–6	65%	95%
ELA	Multi-trait	65%	96%

*A numerical score compared to a blank or condition code score will have a disagreement greater than 1 point.

4.2.4.4. Calibration Sets

Calibration sets are special sets created during scoring to help train scorers on particular areas of concern or focus. Scoring directors used calibration sets to reinforce range-finding standards, introduce scoring decisions, or address scoring issues and trends. Calibration was used either to correct a scoring issue or trend or to continue scorer training by introducing a scoring decision. Calibration was administered regularly throughout scoring.

4.2.4.5. Inter-rater Agreement

Inter-rater agreement is the agreement between the first and second scores assigned to student responses and is the measure of how often scorers agree with each other. Pearson scoring staff used inter-rater agreement statistics as one factor in determining the need for continuing training and intervention on both individual and group levels. Inter-rater agreement expectations are shown in Table 4.2.5.

Table 4.2.5 Inter-Rater Agreement Expectations and Results

Subject	Score Point Range	Perfect Agreement Expectation	Perfect Agreement Result	Within One Point Expectation*	Within One Point Result
Mathematics	0–1	90%	98%	96%	100%
Mathematics	0–2	80%	97%	96%	100%
Mathematics	0–3	70%	96%	96%	99%
Mathematics	0–4	65%	94%	95%	99%
Mathematics	0–5	65%	91%	95%	98%
Mathematics	0–6	65%	95%	95%	98%
ELA	Multi-trait	65%	83%	96%	100%

*A numerical score compared to a blank or condition code score will have a disagreement greater than 1 point.

Pearson's ePEN2 scoring system included comprehensive inter-rater agreement reports that allowed supervisory personnel to monitor both individual and group performance. Based on reviews of these reports, scoring experts targeted individuals for increased backreading and feedback, and if necessary, retraining.

The perfect agreement rate for mathematics responses scored by two scorers ranged from 76 to 100 percent, and the within-one-point rate ranged from 96 to 100 percent. For all ELA responses scored by two scorers, the perfect agreement rate ranged from 69 percent to 100 percent, and the

within-one-point rate ranged from 97 percent to 100 percent.

The results for the ELA PCR are provided in Chapter 4.3.7 “Inter-rater Agreement for Prose Constructed-Response.”

4.3. Automated Scoring for PCRs

Automated scoring performed by Pearson’s Intelligent Essay Assessor (IEA) was the default option for scoring the summative assessment’s online prose constructed-response (PCR) tasks. Under the default option, it was assumed that operational scores for approximately 90 percent of the online PCR responses would be assigned by IEA for the spring administration. The operational scores for the remaining online responses were assigned by human scorers. Human scoring was applied to responses that were scored while IEA was being trained as well as to additional responses routed to human scoring when there was uncertainty about the automated scores.

For 10 percent of responses, a second “reliability” score was assigned. The purpose of the reliability score was to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. When IEA provided the first score of record, the second reliability score was a human score.

4.3.1. Concepts Related to Automated Scoring

The sections below describe concepts related to automated scoring.

4.3.1.1. Continuous Flow

Continuous flow scoring results in an integrated connection between human scoring and automated scoring. It refers to a system of scoring where either an automated score, a human score, or both can be assigned based on a predetermined asynchronous operational flow.

4.3.1.2. Training of IEA Using Operational Data

Continuous flow scoring facilitates the training of IEA using human scores assigned to operational online data collected early in the administration. Once IEA obtains sufficient data to train, it can be “turned on” and becomes the primary source of scoring (although human scoring continues for the 10 percent reliability sample and other responses that may be routed accordingly).

4.3.1.3. Smart Routing

Smart routing refers to the practice of using automated scoring results to detect responses that are likely to be challenging to score and applying automated routing rules to obtain one or more additional human scores. Smart routing can be applied prompt by prompt to the extent needed to meet scoring quality criteria for automated scoring.

4.3.1.4. Quality Criteria for Evaluating Automated Scoring

The state leads approved specific quality criteria for evaluating automated scoring. The primary evaluation criteria for IEA were based on responses to validity papers with “known” scores assigned by experts. For each prompt scored, a set of validity papers is used to monitor the

human-scoring process over time. Validity papers are seeded into human scoring throughout the administration. The expectation is that IEA can score validity papers at least as accurately as humans can.

Additional measures of inter-rater agreement for evaluating automated scoring were proposed based on the research literature (Williamson et al., 2012). These measures were previously utilized in Pearson's automated scoring research and include Pearson correlation, kappa, quadratic-weighted kappa, exact agreement, and standardized mean difference. These measures are computed between pairs of human scores, as well as between IEA and humans, to evaluate how performance was the same or different. Criteria for evaluating the training of IEA given these measures include the following:

- Pearson correlation between IEA-human should be within 0.1 of human-human.
- Kappa between IEA-human should be within 0.1 of human-human.
- Quadratic-weighted kappa between IEA-human should be within 0.1 of human-human.
- Exact agreement between IEA-human should be within 5.25 percent of human-human.
- Standardized mean difference between IEA-human should be less than 0.15.

The specific criteria for evaluating IEA included both primary and secondary criteria and are noted below:

- Primary Criteria — Based on responses to validity papers: With smart routing applied as needed, IEA agreement is as good as or better than human agreement for each trait score.
- Contingent Primary Criteria — Based on the training responses if validity responses are not available: With smart routing applied as needed, IEA-human exact agreement is within 5.25 percent of human-human exact agreement for each trait score.
- Secondary Criteria — Based on the training responses: With smart routing applied as needed, IEA-human differences on statistical measures for each trait score are within the Williamson et al. (2012) tolerances for subgroups with at least 50 responses.

4.3.1.5. Hierarchy of Assigned Scores for Reporting

When multiple scores are assigned for a given response, the following hierarchy determines which score was reported operationally:

- The IEA score is reported if it is the only score assigned.
- If an IEA score and a human score are assigned, the human score is reported.
- If a first human score and a second human score are assigned, the first human score is reported.
- If a backread score and human and/or IEA scores are assigned, the backread score is reported if there is no resolution or adjudication score assigned.
- If a resolution score is assigned and an adjudicated score is not assigned, the resolution score is reported (note that if nonadjacent scores are encountered, responses are automatically routed to resolution).
- If an adjudicated score is assigned, it is reported (note that if a resolution score is nonadjacent to the other scores assigned, responses are automatically routed to adjudication).

4.3.2. Sampling Responses Used for Training IEA

For prompts trained using 2024 operational data, the early performance of human scoring was closely monitored to verify that an appropriate set of data would be available for training IEA. In particular, several characteristics of the human scoring data were monitored, including:

- Exact agreement between human scorers (the goal was for this to be at least 65 percent for each trait).
- Exact agreement between human scores is conditioned on score point (the goal was for this to be at least 50 percent for each trait).
- The number of responses at each score point (the goal was to have at least 40 responses at the highest score points in the training samples used by IEA).
- The number of responses with two human scores assigned (note that IEA "ordered" additional scoring of responses during the sampling period as needed).

Although the desired characteristics of the training data were easily achieved for some prompts, they were more challenging to achieve for others. For some prompts, a subset of scores was reset, and clarifying directions were provided to scorers to improve human-human agreement. For other prompts, special sampling approaches were used to increase the number of responses that received top scores. In addition, a healthy percentage of responses were backread during the sampling period, and these scores, as well as double human scores, were all part of the data used to train IEA.

4.3.3. Primary Criteria for Evaluating IEA Performance

The primary criteria for evaluating IEA performance are based on evaluating validity papers and are stated as follows: With smart routing applied as needed, IEA agreement is as good as or better than human agreement for each trait score.

To operationalize the primary criteria for a given prompt, the following general steps are undertaken:

1. Determine agreement of the human scores with the validity papers for each trait.
2. Calculate agreement of the IEA scores with the validity papers for each trait.
3. Compare the IEA validity agreement with the human agreement.
4. If the IEA validity agreement is greater than or equal to the human agreement for each trait, IEA can be deployed operationally.

In addition to looking at the overall validity agreement, conditional agreement was also examined. In general, it was desirable for IEA to exceed 65 percent agreement at every score point as well as be close to or exceed the human validity agreement at each score point.

4.3.4. Contingent Primary Criteria for Evaluating IEA Performance

For many of the prompts trained in 2024, it was not possible to utilize human-scored validity responses in evaluating IEA performance. In these cases, IEA was evaluated based on IEA-human exact agreement for each trait score and compared to agreement based on responses that were double-scored by humans. A portion of the data was held out for evaluating IEA-human exact

agreement according to the following steps:

1. Determine exact agreement of the two human scores with each other for each trait.
2. Calculate agreement of the IEA scores with the human scores for each trait.
3. Compare the IEA-human agreement with the human-human agreement.
4. If the IEA-human agreement is within 5.25 percent of the human-human agreement, IEA can be deployed operationally.

In addition to the overall comparison, the following performance thresholds were targeted in the test data set: (1) at least 65 percent overall IEA-human agreement; and (2) 50 percent IEA-human agreement by score point (i.e., conditioned on the human score). These targets went beyond the contingent primary criteria approved by the state leads.

4.3.5. Applying Smart Routing

With smart routing, the quality of automated scoring can be increased by routing responses that are more likely to disagree with a human score to receive an additional human score.

When human scorers read a paper, they typically apply integer scores based on a scoring rubric. When there is strong agreement between two independent human readers, the readers might both assign a score of 3 such that the average score over both raters is also a 3 (i.e., (3+3)/2 = 3). IEA simulates this behavior, but because its scores come from an artificial intelligence algorithm, it generates continuous (i.e., decimalized) scores. In this case, the IEA score might be a 2.9 or 3.1. When human readers disagree on the score for a paper, say one reader gives the paper a score of 3 and another reader gives the paper a score of 4, the average of the two scores would be 3.5 (i.e., 3+4=7/2=3.5). For this paper, IEA would likely provide a score between 3 and 4, say 3.4 or 3.6. Because this continuous score needs to be rounded to an integer score for reporting, it might be reported as a 3 or a 4, depending on the rounding rules. Smart routing involves routing those responses with “in between” IEA scores to additional human scoring because the nature of the responses suggests there may be less confidence in the IEA score. Since these “in between” IEA scores are based on modeling human scores, it follows that human scores may be less certain as well, and thus such responses tend to be the ones that it makes sense to have double-scored and possibly to resolve if the IEA and human scores are nonadjacent.

Smart routing was utilized as needed to help IEA achieve targeted quality metrics (e.g., validity agreement or agreement with human scorers). Smart routing involved the application of the following four steps:

1. The continuous IEA score for each of the two trait scores was rounded to the nearest score interval of 0.2, starting from zero. For example, IEA scores between 0 and 0.1 were rounded to an interval score of 0, scores between 0.1 and 0.3 were rounded to an interval score of 0.2, scores between 0.3 and 0.5 were rounded to an interval score of 0.4, and so on.
2. Within each of these intervals, the percentage of exact agreement between IEA integer scores and the human scores was calculated for each trait.
3. For each prompt, agreement rates were evaluated by rounding interval. Those intervals for which the agreement rates were below a designated threshold for either trait were

identified.

4. Once IEA scoring was implemented, responses within intervals for which IEA-human agreement was below the designated threshold were routed for additional human scoring.

In training IEA, the scoring models without smart routing were evaluated first by applying either the primary validity criteria or the contingent criteria as described in Chapter 4.3.4. For those prompts that did not meet these criteria, increasing smart routing thresholds were applied in an iterative fashion to filter scores and evaluate the remaining scores against the criteria. That is, in any one iteration a particular smart routing threshold was applied such that only scores falling in intervals for which exact agreement exceeded the threshold were included in evaluating the criteria. If the primary or contingent criteria were not met with this level of smart routing, an increased smart routing threshold was applied iteratively until the primary or contingent criteria were met, or the maximum threshold reached. If the criteria were still not met after a maximum threshold was applied, different models were investigated and/or additional human scoring data utilized until an IEA scoring model was found that met the criteria.

4.3.6. Evaluation of Secondary Criteria for Evaluating IEA Performance

The secondary criteria for evaluating IEA performance involved comparing agreement indices for IEA-human scoring for various demographic subgroups. Because of the importance of protecting personally identifiable information, student demographic data is stored and managed separately from the performance scoring data. For this reason, it was not possible to evaluate subgroup performance in real time as IEA was being trained.

For those prompts trained on early operational data, attempts were made to prioritize the data being returned from the field to include data from states or districts where more diverse populations of students were anticipated. In addition, requests for additional human scores were made to increase the likelihood that there would be sufficient numbers of responses with two human scores for most of the demographic subgroups of interest.

Once IEA was trained and deployed, scoring sets used in training were matched to demographic information so that agreement between IEA and human scorers could be evaluated across subgroups. The analysis was conducted for the eleven comparison groups outlined in Table 4.3.1.

Table 4.3.1 Comparison Groups

Group Type	Comparison Groups
Sex	Female Male
Ethnicity	American Indian/Alaska Native Asian Black/African American Hispanic/Latino Native Hawaiian or Other Pacific Islander Two or More Races White
	English Language Learners (ELL)

Group Type	Comparison Groups
Special Instructional Needs	Students with Disabilities (SWD) Economically Disadvantaged

IEA-human agreement indices were calculated for all cases with an IEA score and at least one human score. Human-human agreement was calculated for all cases with two human scores.

To evaluate the training of IEA for subgroups, the following criteria approved by the state leads for subgroups with at least 50 IEA-human scores and at least 50 human-human scores were applied:

- Pearson correlation between IEA-human should be within 0.1 of human-human.
- Kappa between IEA-human should be within 0.1 of human-human.
- Quadratic-weighted kappa between IEA-human should be within 0.1 of human-human.
- Exact agreement between IEA-human should be within 5.25 percent of human-human.
- Standardized mean difference between IEA-human should be less than ±0.15 (this criterion was applied to subgroups with at least 50 IEA-human scores).

Although it was not expected that these criteria would be met for all subgroups for all prompts, if results of the evaluation between IEA and human scoring for subgroups for any prompt indicated that IEA performance persistently failed on the criteria listed above, consideration would be given to resetting the responses scored by IEA and reverting to human scoring until such time that an alternate IEA model could be established with improved subgroup performance.

In addition to the secondary criteria approved by the State Leads, the performance of IEA was compared to the following targets on the various measures for subgroups with at least 50 responses:

- Pearson correlation between IEA-human should be 0.70 or above.
- Kappa between IEA-human should be 0.40 or above.
- Quadratic-weighted kappa between IEA-human should be 0.70 or above.
- Exact agreement between IEA-human should be 65 percent or above.

These targets were not intended to be directly applied in decisions about whether to deploy IEA operationally or not. Such targets may or may not be met by human scoring for any particular prompt and/or subgroup, and if they are not met by human scoring, they are unlikely to be met by IEA scoring. Nevertheless, comparisons to these targets provided additional information about IEA performance (and human scoring) in an absolute sense.

4.3.7. Inter-rater Agreement for Prose Constructed-Response

This section presents the inter-rater agreement for operational results for the online PCR tasks by trait and grade level. PCR items are scored on two traits: (1) Reading Comprehension and Written Expression and (2) Knowledge of Language and Conventions for Research Simulation for Literary Analysis tasks and (1) Written Expression and (2) Knowledge of Language and Conventions for the Narrative task.

For 10 percent of responses, a second "reliability" score was assigned. The purpose of the reliability score is to provide data for evaluating the consistency of scoring, which is done by evaluating scoring agreement. Inter-rater agreement is the agreement between the first and second scores assigned to student responses and is the measure of how often scorers agree with each other. Pearson scoring staff used inter-rater agreement indices as one factor in determining the needs for continuing training and intervention on both individual and group levels. Inter-rater agreement expectations are provided in Table 4.2.5 in Chapter 4.2.4. For ELA PCR traits, the expectation for agreement is an inter-rater agreement of 65 percent or higher between two scorers. When IEA provided the first score of record, the second reliability score was a human score. For a subset of responses, the first and second score were both human scores.

Table 4.3.2 presents the average agreement across the PCRs for each grade level by trait. The number of prompts included in the analyses is listed for each grade level. The agreement indices (exact agreement, kappa, quadratic-weighted kappa, and Pearson correlation) were calculated separately by PCR for each trait (Reading Comprehension and Written Expression or Written Expression and Conventions). For each grade level, the agreement indices were averaged across the PCRs. Table 4.3.2 presents the average count and the average for the agreement indices.

The exact agreement for the PCR traits is above the criteria of a 65 percent agreement rate for all PCRs. The strength of agreement between raters is moderate to substantial agreement as defined by Landis and Koch (1977) for all PCRs. The quadratic-weighted kappa (QW Kappa) distinguishes between differences in ratings that are close to each other versus larger differences. The weighted kappa is substantial to almost perfect agreement for all grades. The Pearson correlations (*r*) ranged from 0.75 to 0.95.

Table 4.3.2 Prose Constructed-Response Average Agreement Indices by Test

			Written Expression				Conventions			
Test	**N-PCRs**	**Count**	**Exact**	**Kappa**	**QW Kappa**	***r***	**Exact**	**Kappa**	**QW Kappa**	***r***
ELAGP	4	27,923	74.53	0.57	0.74	0.75	75.3	0.6	0.78	0.78

4.4. Quality Control of Scoring

Quality control of answer document processing and scoring involves all aspects of the scoring procedures, including key-based and rule-based machine scoring and hand scoring for constructed-response items and performance tasks.

For the 2015 operational administration, Pearson's validation team prepared test plans used throughout the scoring process. Test plan preparation was organized around detailed specifications that continued to the 2024 operational administration.

Based on lessons learned from previous administrations, the following quality steps were implemented:

- Raw score validation (e.g., score key validation; evidence statement; field-test non-score; double-grid combinations; possible correct combination, if applicable; out-of-range/negative test cases).
- Matching (e.g., validation of high-confidence criteria, low-confidence criteria, cross document, external or forced matching by customer; prior to and after data updates; extract file of matched and unmatched documents).
- Demographic update tests (e.g., verification of data extract against corresponding layout; valid values for updatable fields; invalid values for updatable/non-updatable fields; negative test for nonexistent record or empty file).

The following components were added to the quality control process specifically for the program. These additional steps were introduced to address issues with item-level scoring that were identified in the 2014 field-test administration:

- XML Validation: A combination of automated validation against 100 percent of item XMLs and human inspection of XML from selected difficult item types or composite items.
- Administration/End-to-End Data Validation: An automated generation of response data from approved test maps that have known conditions against the operational scoring systems and data generation systems to verify scoring accuracy.
- Psychometric Validation: Verification of data integrity using criteria typically used in psychometric processes (e.g., statistical key checks) and categorization of identified issues to help inform investigation by other groups.
- Content Validation: An examination, by subject matter experts, of all items using a combination of automated tools to generate response and scoring data.

In addition to the steps described above, the following quality control process for answer keys and scoring that was implemented for the first operational administration was as follows:

- Pearson's psychometrics team conducted empirical analyses based on preliminary data files and flagged items based on statistical criteria.
- Pearson's content team reviewed the flagged items and provided feedback on the accuracy of content, answer keys, and scoring.
- Items potentially requiring changes were added to the product validation log for further investigation by other Pearson teams.
- Staff was notified of items for which keys or scoring changes were recommended.
- Participating states and agencies approved or rejected scoring changes.
- All approved scoring changes were implemented and validated prior to the generation of the data files used for psychometric processing.

Chapter 5. Performance Standards and Standards Validation

The score scales and performance standards for ELAGP and MATGP are based on score scales created and maintained for the PARCC/New Meridian Affiliate program. The ELAGP score scale derives from the grade 10 Affiliate summative assessment scale and the MATGP from the Affiliate Integrated Math II scale. The performance standard level of 725 scale score relates to the performance levels established for the Affiliate program. For further documentation on the setting of these scales, see the most recent New Meridian Affiliate Technical Report Alternate Blueprint.

5.1. Performance Standards

Student performance on both components of the NJGPA is compared with performance standards to determine if each student meets criteria to be considered Graduation Ready. Those students not meeting the criteria are Not Yet Graduation Ready. The cut scores that delineate Graduation Ready from Not Yet Graduation Ready were validated by New Jersey subject matter experts (SMEs) as described in the next section of this chapter. The suggested cut scores were approved by the New Jersey State Board of Education in May 2023.

5.2. Standard Setting Process

This section discusses two distinct standard-setting processes that apply to the NJGPA: The first subsection will explain the standard-setting processes of the PARCC/NMC Affiliate program, including the initial calibration of performance levels in high school ELA and mathematics assessments, which contributed to NJGPA development. The final subsection will outline a standards validation meeting held in February 2022 that validated the recommended 725 scale score used for the NJGPA Graduation Ready cut score.

The extensive standard-setting activities described in the following section regarding the PARCC/NMC Affiliate program allowed for a streamlined, standard validation process for NJGPA cut scores. PARCC/NMC Affiliate assessments report five performance levels. PARCC/NMC Affiliate cut scores corresponding to level three are defined as approaching academic performance expectations. Level three is anchored to a scale score of 725. The level three scale score of 725 was chosen as the target scale score for the NJGPA standards validation meetings.

5.2.1. Performance Level Setting (PLS) Process for the PARCC/NMC Affiliate System

NJGPA score scales were derived from score scales previously created for the PARCC/NMC Affiliate ELA grade 10 and Integrated Math II summative assessments. One of the main objectives of the PARCC/NMC Affiliate system was to provide information to students, parents, educators, and administrators as to whether students are on track in their learning for success after high school, defined as college- and career-readiness. To set performance levels associated with this objective, participating states and agencies used the evidence-based standard setting (EBSS) method (Beimers et al., 2012) for the PLS

process. The EBSS method is a systematic method for combining various considerations into the process for setting performance levels, including policy considerations, content standards, educator judgment about what students should know and be able to demonstrate, and research to support policy goals related to college- and career-readiness. A defined multistep process was used to allow a diverse set of stakeholders to consider the interaction of these elements in recommending performance-level threshold scores for each assessment.

The seven steps of the EBSS process that were followed to establish performance standards for the summative assessments are:

- Step 1: Define outcomes of interest and policy goals.
- Step 2: Develop research, data collection, and analysis plans.
- Step 3: Synthesize the research results.
- Step 4: Conduct pre-policy meeting.
- Step 5: Conduct PLS meetings with panels.
- Step 6: Conduct reasonableness review with post-policy panel.
- Step 7: Continue to gather evidence in support of standards.

A summary of key components within these steps is provided below. Additional details about each step in the PLS process are provided in the *Performance Level Setting Technical Report*.

5.2.1.1. Research Studies

Participating states and agencies conducted two research studies in support of their policy goals—the benchmarking study and the postsecondary educators' judgment (PEJ) study. The benchmarking study included a review of the literature relative to college- and career-readiness as well as consideration of the percentage of students obtaining a level equivalent to college- and career-readiness on a set of external assessments (e.g., ACT, SAT, NAEP). The PEJ study involved a group of nearly 200 college faculty reviewing items on the Algebra II and ELA grade 11 assessments and making judgments about the level of performance needed on each item to be academically ready for an entry-level college-credit-bearing course in mathematics or ELA. Additional details about the benchmarking study can be found in the *Performance Level Setting Technical Report* as well as in the *PARCC Benchmarking Study Report*. Additional details about the PEJ study can be found in the *Performance Level Setting Technical Report* as well as in the *Postsecondary Educators' Judgment Study Final Report*.

5.2.1.2. Pre-Policy Meeting

Prior to the PLS meetings, a pre-policy meeting was convened to determine reasonable ranges that would be shown to panelists during the high school PLS meetings. Pre-policy meeting participants included representatives from both K–12 and higher education who served in roles such as commissioner/superintendent, deputy/assistant commissioner, state board member, director of assessment, director of academic affairs, senior policy associate, and so on. The reasonable ranges recommended by the pre-policy meeting defined the minimum and maximum percentage of students that would be expected to be classified as college- and career-ready. The pre-policy meeting participants reviewed the test purpose, how the performance standards will be used, and the results of the research studies to provide recommendations for the reasonable ranges without viewing any student performance data.

5.2.1.3. Performance Level Setting Meetings

The task of the PLS committee was to recommend four threshold scores that would define the five performance levels for each assessment. Participating states and agencies solicited nominations from all Affiliate states that had administered the assessments in 2014–2015 for panelists to serve on the PLS committees. Nominations were solicited both from state departments of public education (K–12) and higher education (primarily for participation on the high school panels). When selecting panelists, an emphasis was placed on those educators who had content knowledge as well as experience with a variety of student groups, and attempted to balance the panels in terms of state representation.

Participating states and agencies used an extended modified Angoff (Yes/No) method to collect educator judgments on the items. This method asked panelists to review each item on a reference form of the assessment and to make the following judgment: How many points would a borderline student at each performance level likely earn if they answered the question?

This extension to the Angoff standard setting method (Plake et al., 2005) allowed for the incorporation of the multipoint items by asking educators to evaluate (Yes or No) whether a borderline student would earn the maximum number of points on an item, a lesser number of points on an item, or no points on the item. In the case of a single-point or multiple-choice item, this task simplifies to the standard Yes/No method.

After receiving training on the PLS procedure, panelists participated in three rounds of judgments for each assessment. Within each round, panelists were asked to consider the items in the test form, starting with the performance-based assessment (PBA) component and then the end-of-year (EOY) component. Each panelist made a judgment for the Level 2 performance level, followed by judgments for the Level 3 performance level, the Level 4 performance level, and the Level 5 performance level, in this order. The panelists entered their item judgments for each round by completing an online item judgment survey. Educator judgments were summed across items to create an estimated total score on the reference form for each performance level threshold. Feedback data relative to panelist agreement, student performance on the items, and student performance on the test as a whole were provided in between each of the three rounds of judgment. Panelists were shown the pre-policy reasonable ranges prior to making their Round 1 judgments and again as feedback data following each round of judgment.

A dry run of the PLS meeting process was held for grade 11 ELA and Algebra II in order to evaluate the implementation of the PLS method with the innovative characteristics of the summative assessments. The results of the dry run PLS meeting were used to implement improvements in the process for the operational PLS meetings. Additional information about the methods and results of the dry run PLS meeting is available in the full report in the *Performance Level Setting Dry-Run Meeting Report*.

Additional information about the methods and results of the PLS meetings is available in the *Performance Level Setting Technical Report*.

5.2.1.4. Post-Policy Reasonableness Review

Performance standards for all summative assessments were recommended by PLS committees and reviewed by the Governing Board and (for the Algebra II, Integrated Mathematics III, and ELA grade 11 assessments) the Advisory Committee on College Readiness as part of a post-policy reasonableness review. This group reviewed both the median threshold score recommendations from each committee and

the variability in the threshold scores as represented by the standard error of judgment (SEJ) of the committee. Adjustments to the median threshold scores that were within 2 SEJ were considered to be consistent with the PLS panels' recommendation.

In addition to voting to adopt the performance standards based on the committee's recommendations, this group also voted to conduct a shift in the performance levels to better meet the intended inferences about student performance. Holding the college- and career-ready (or on-track) expectations (i.e., the current level 4) constant, performance levels above this expectation were combined, and performance levels below this expectation were expanded to create the final system of performance levels with three below and two above the college- and career-ready (or on-track) expectation. The shift in performance levels was accomplished using a scale anchoring process that involved two primary steps. In the first step, the top two performance levels, above college- and career-ready (or on track), were combined into a single performance level, and an additional performance level below college- and career-ready (or on track) was created by empirically determining the midpoint between the existing two levels. In the second step, the performance level descriptors (PLDs) were updated using items that discriminated student performance well at this level to create a PLD aligned with the new empirically determined performance level. At the same time, PLDs for all performance levels were reviewed for consistency and continuity. Members of the original PLS committees were recruited to participate in this process. Additional information about this process can be found in the *Performance Level Setting Technical Report*.

5.2.2. NJGPA Standards Validation

The New Jersey Department of Education (NJDOE) conducted virtual standards validation meetings on February 1–2, 2022, facilitated by NMC staff. Educators from throughout the state of New Jersey participated in these two-day meetings. The main goal of the meetings was to have SMEs recommend psychometric cut scores which delineate two performance levels: "Graduation Ready" and "Not Yet Graduation Ready" for NJGPA ELA and mathematics. Two additional aims of the meetings designed to prepare and assist SMEs in the identification of cut scores were:

- To allow workshop participants (i.e., subject matter experts [SMEs]) to gain an understanding of the assessment content and performance level descriptors (PLDs).
- To have SMEs learn a standard-setting methodology based on the Bookmark standard-setting procedure.

NMC facilitated the Bookmark standard-setting procedures to validate the psychometric cut scores for the ELAGP and MATGP assessments set at the 725 scale score level. The Bookmark standard-setting procedure (Lewis, Mitzel & Green, 1996) was developed in 1996 and fully documented in *Setting Performance Standards: A Guide to Establishing and Evaluating Performance Standards on Tests* (Cizek & Bunch, 2007). The procedure is an iterative set of activities intended to produce suggested cut scores based on the judgments of SMEs. The Bookmark standard-setting procedure was developed to provide a less complex method for judges to determine a cut score (Mitzel et al., 2001) and is one of the most used procedures for establishing cut scores for large-scale K–12 assessments (Baldwin et al., 2019).

For the NJGPA standards validation meeting, the standard Bookmark procedure was modified. The SMEs were provided with the cut score levels, and the discussion centered on how well the scores defined the knowledge, skills, and abilities that would distinguish students who are ready for graduation from those

who are not yet ready. The SMEs placed their bookmarks in the locations they felt best described the graduation-ready level of performance. The calculated cut scores were then compared to the established cut score levels.

The key activities for the NJGPA standards validation included:

- Group discussion of PLDs and students at the threshold.
- Taking the NJGPA.
- SME examination of the test items (no item statistics provided).
- Group discussion of each item.
- Individual SME bookmark placement (Round 1 rating).
- Group discussion of Round 1 results.
- Individual SME bookmark placement (Final rating).

An Ordered Item Booklet (OIB) is integral to the Bookmark standard-setting procedure. Each NJGPA OIB contained items and metadata from the item bank. Each OIB consisted of a number of items which appeared on the Spring 2022 forms and some other items chosen to represent the range of topics and difficulty defining the overall item bank available for forms construction. NMC psychometricians created OIBs by ordering items from easiest to most difficult according to an item response theory (IRT) measure. The item statistics of record from prior administrations of the items were used to order the OIB.

During standard setting, SMEs proceeded through the OIB one page at a time and asked themselves whether they believed the "just barely meeting expectations," or "threshold," student would have at least a 67% probability of answering the item correctly (or obtaining the given score point). If the answer was "Yes," then the SME proceeded to the next page. The page on which the SME answered "No" became the bookmark between the adjacent performance levels of Not Yet Graduation Ready and Graduation Ready. A key element to the procedure was the reconciliation of the PLDs and a threshold student's demonstrated knowledge, skills, and abilities (KSAs).

Following each of the bookmark placement rounds, NMC collected the SMEs' bookmark placements (provided as a page number) and computed the median, median absolute difference, minimum, and maximum placement by group (i.e., breakout room) and by committee. NMC psychometricians then presented these results to the SMEs. Following the first round of bookmark placement, SMEs had the opportunity to discuss the subset of items within the individually bookmarked pages, considering the PLDs and the KSAs required by a threshold student. Following the first group discussions, SMEs had the opportunity to update their bookmark location based on the shared discussions. At the culmination of the final round, the committee's median value was used to calculate the associated psychometric cut.

The cut scores suggested by the SMEs were calculated as IRT theta scores, which allows their application across various forms of the assessment. The cut score theta for the ELAGP assessment is -0.31. The cut score theta for the MATGP assessment is -0.049. The cut score for both subjects corresponds to a scale score of 725.

At the conclusion of the workshop, the SMEs were asked to complete a process evaluation survey. The survey collected information about the SMEs' experiences in recommending a psychometric cut score for the NJGPA.

Chapter 6. Item Analysis

6.1. Overview

This chapter describes the results of the classical item analyses conducted for operational items on the spring operational forms. For the NJGPA, all forms were pre-equated, meaning that the scoring was based on item parameters estimated using data from prior assessment administrations. By contrast, item analysis results in this section are derived from the item statistics from post-administration analysis. Post-administration analyses are presented as evidence of actual item performance on the spring 2024 test.

An operational item may appear on multiple test forms as described in section 2.4.3. The tables below list unique item counts for the assessments, and the reported item statistics may be based on student responses across multiple occurrences of an item.

Spoiled or "do not score" items, if any, are excluded from the total test score in item analysis. These items are removed from scoring because of item performance, technical scoring issues, content concerns, or multiple/no correct answers. There were no spoiled items in the 2024 NJGPA testing. Thus, there were no exclusions.

For each item, analyses were conducted at the item level for each form. These analyses included difficulty (p-value and pseudo p-value) and discrimination (item-total correlation).

6.2. Data Screening Criteria

Analyses were performed on an Incomplete Data Matrix (IDM) that was generated from the results file. These analyses were done by form. Student records were removed prior to running the analyses if the records met any of the criteria indicated in the test calibration specs.

Items may not be scored due to a student omitting the item or to the student not yet reaching an item within the test. "Omitted" (i.e., skipped) items are items for which a student did not provide a response when items coming before and after have student responses. Item response scores for "omits" are re-coded as '0' in the classical test theory (CTT) analyses and IRT IDM files whereas "Not reached" and "Not administered" items are considered missing and therefore do not contribute to the statistics.

6.3. Classical Item Analysis

6.3.1. Item Difficulty

The p-value for dichotomous items—that is, one-point items scored as either correct or incorrect—is the mean item score, which is calculated as the proportion of examinees who answer the item correctly among the total number of examinees having scored data for the item. The formula to calculate the *p*-value for dichotomous items is:

Equation 6.1

$$p\text{-}value = \bar{x_i} = \frac{1}{n} \cdot \sum_{1}^{n} x_i$$

where x_i are the individual student item scores on item *i* and *n* is the total number of students having scored data for the item.

For polytomous items, the pseudo *p*-value is calculated by dividing the average score on the item by the maximum obtainable points possible for the item.

Equation 6.2

$$pseudo\ p\text{-}value = \frac{\bar{x}}{T}$$

where $\bar{x}$ is the mean item score and *T* is the maximum obtainable points possible for the item.

P-value and pseudo p-value item statistics are bounded between 0 and 1. For these statistics, higher values indicate easier items while lower values indicate more difficult items. Frequently, the p-value and pseudo p-value are reported as percentages that are calculated by multiplying the proportions by 100. For instance, a p-value of 0.67 means that 67 percent of the students answered the dichotomous item correctly. On the other hand, a pseudo p-value of 0.67 means the average score obtained for the item among all students with scored data is 67 percent of the total maximum obtainable points possible for the polytomous item.

6.3.2. Response Option or Score Point Proportions

A dichotomous item's alternate response options (i.e., commonly known as being distractors) are plausible but incorrect options that are included to test common misconceptions or miscalculations. Ideally, all response options should garner a proportion of student responses. For a given response option, the proportion is calculated by the simple formula

Equation 6.3

$$proportion = \frac{N_O}{N_T}$$

where N_O is the number of students selecting the specific option and N_T is the total number of students having scored data for the item.

In the case of polytomous items, the proportion is calculated for each score point as the number of students obtaining the specific score point (N_{SP}) divided by the total number of students having scored data for the item (N_T).

Equation 6.4

$$proportion = \frac{N_{SP}}{N_T}.$$

6.3.3. Item-Total Correlations

The item-total correlation is the relationship between students' performances on a specific item and students' performances on the assessment overall. Possible values for the item-total correlation are bounded between -1 and +1. The correlation will be positive when the mean total test score of the students answering the item correctly is greater than the mean total test score of the students having an incorrect or omitted response for the item. A negative item-total correlation indicates that students with lower mean total test scores were more likely to answer the question correctly than students with higher total test scores. A negative item-total correlation may indicate that an item has multiple correct answers or an incorrect answer key.

The point-biserial correlation (Crocker & Algina, 1986) is one possible item-total correlation for dichotomously scored items. However, the correlation will be spuriously high because the item of interest is also included in the calculation of the total test score (i.e., correlating with itself; Henrysson, 1963). Therefore, a correction is made by calculating the means after excluding the item from the calculation of the total test score (i.e., the total operational test score not including the item of interest for the calculation)

Equation 6.5

$$r_{\text{pbis}} = \frac{(\bar{M}'_{+} - \overline{M'})}{S'}\sqrt{p/(1-p)}$$

where $\bar{M}'_{+}$ is the mean score with the item excluded for students who answered the item correctly, $\overline{M'}$ is the mean score with the item excluded for all students who either answered incorrectly or have an omitted response, S' is the standard deviation of the distribution of students' scores (with the item excluded for all students), and p is the item p-value (difficulty).

The Pearson correlation (polyserial), calculated after excluding the item of interest, is typically computed for polytomous items by this equation:

Equation 6.6

$$r = \frac{\sum(x_i - \bar{x})(y_i' - \bar{y}\prime)}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i' - \bar{y}\prime)^2}};$$

where x_i is the student score point on the item, $\bar{x}$ is the mean score for the item, y_i' is the total score with the item excluded for the student, and $\bar{y}'$ is the mean total score with the item excluded for all students (Lemke & Wiersma, 1976).

6.3.4. Results of Classical Item Analysis

Item analysis results are reported in the below are computed as the weighted average across test forms. Each form is weighted according to its contribution to the total population. Therefore, operational core forms, which are administered to larger numbers of students, contribute more to the computations than accommodated forms, which are administered to smaller numbers of students.

Table 6.3.1 Summary of Post-Administration P-values for NJGPA Operational Items

Test	N Items	Mean P-Value	SD P-Value	Min. P-Value	Max. P-Value	Median P-Value
ELAGP	20	0.49	0.14	0.24	0.75	0.48
MATGP	30	0.40	0.21	0.13	0.75	0.40

Note: SD = standard deviation

Table 6.3.1 presents post-administration summary statistics for item p-values for the operational items for NJGPA. The weighted mean p-values range from 0.40 in MATGP to 0.49 for ELA, indicating that most items were of moderate difficulty. The standard deviations range from 0.14 in ELAGP to 0.21 in MATGP demonstrating that the forms contained items assessing a range of difficulties with MATGP showing slightly more variability than ELAGP. Mean difficulty tended to be relatively consistent across grades.

Table 6.3.2 Summary of Post-Administration ITC for NJGPA Operational Items

Test	N Items	Mean ITC	SD ITC	Min. ITC	Max. ITC	Median ITC
ELAGP	20	0.52	0.17	0.23	0.93	0.50
MATGP	30	0.71	0.11	0.45	0.85	0.75

Note. SD = standard deviation.

Table 6.3.2 presents post-administration summary statistics for item-total correlations for operational items for NJGPA. The weighted mean correlations ranges from 0.52 in ELAGP to 0.71 in MATGP and standard deviations range from 0.11 to 0.17. Correlations tended to be robust, indicating the items discriminated well between students that performed better overall versus students that performed worse overall.

Chapter 7. Item Response Theory Analysis, Calibration and Scaling

7.1. Overview

The ELA and mathematics core forms are linked to their respective base reporting scales via pre-equating. This section of the technical report describes the item response theory (IRT) models used for pre-equating, item calibration and calculation of students' scale scores for ELAGP and MATGP. Descriptive statistics of the distributions of item parameter estimates for each assessment component are also included in this chapter.

7.2. IRT Models

The items on mathematics operational core forms were calibrated using the IRT two-parameter logistic (2PL) model (Birnbaum, 1968) for dichotomously scored items and the Generalized Partial Credit Model (GPCM) (Muraki, 1997) for polytomously scored items. ELA items were calibrated using the GPCM. To be concise, the two models are expressed using a single formula since the 2PL model can be considered algebraically nested within the GPCM, which is denoted as:

Equation 7.1

$$P_{im}(\theta_j) = \frac{exp\left[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})\right]}{\sum_{v=0}^{M_i-1} exp\left[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{ik})\right]}$$

where $a_i(\theta_j - b_i + d_{ik}) \equiv 0$; $P_{im}(\theta_j)$ is the probability of the j^{th} student with θ_j getting score m on item i; D is the IRT scale constant (1.7); a_i is the discrimination parameter of item i; b_i is the item difficulty parameter of item i; d_{ik} is the k^{th} step deviation parameter for item i; M_i is the number of score categories of item i with possible item scores as consecutive integers from zero to $M_i - 1$; and v indexes the response categories and is iterated from zero to $M_i - 1$. For items with just two response categories, with one category being scored as a correct response and the other category scored as an incorrect response, the d_{ik} parameter becomes zero since no step parameters are needed, making the 2PL model a special case of the GPCM.

7.3. Summary Statistics and Distributions from IRT Analyses

Table 7.3.1 present summary statistics for the pre-equated IRT (*a*- and *b*-) parameter estimates, the standard errors (SEs) of the parameter estimates, and the IRT model fit values (chi-square and adjusted fit) for ELA and mathematics components. The summary statistics for IRT parameter estimates include all the items administered in the spring administration. In IRT, the *a*-parameter refers to the item's ability to discriminate the performance of test takers. The *b*-parameter represents the item's level of difficulty, with larger, positive values reflecting a harder item. The items summarized in Table 7.3.1 include all unique

items on online general forms, paper forms and online accommodated forms. ELAGP PCR traits are included in the item counts as IRT parameters are calculated for each trait. Thus, each PCR item is counted twice: once for each scored trait.

Table 7.3.1 IRT Parameter Estimates Summary for All Items

			***b* Estimates Summary**				***a* Estimates Summary**			
Test	**No. of Score Points**	**No. of Items**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
ELAGP	44	100	0.52	0.36	0.13	1.60	0.70	0.90	-1.18	4.03
MATGP	58	102	0.68	0.27	0.16	1.23	0.83	1.14	-1.5	2.78

Note. SD = standard deviation.

Items on spring 2024 forms used pre-equated IRT parameters that were calculated during item calibration in previous years. Table 7.3.2 shows the source years for the IRT item parameters for the 2024 ELA assessments.

Table 7.3.2 IRT Parameter Distribution by Year for All Items for ELA Assessments

Test	**No. of Items**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2022**	**2023**
ELAGP	44	0	0	0	3	6	0	35	
MATGP	56	0	0	0	0	0	0	54	4

7.4. Scale Scores

Student NJGPA results in ELA and mathematics are reported using scales that designate student performance into one of two performance levels that delineate the knowledge, skills, and practices students are able to demonstrate. Students meeting the graduation performance standards described in Chapter 5 of this report are described as Graduation Ready. Students not meeting these standards are Not Yet Graduation Ready.

The ELA and mathematics components are designed to measure and report results in categories called master claims and subclaims. Major claims (or simply “claims”) are at a higher level than subclaims, with content representing multiple subclaims contributing to each claim outcome. A summative scale score is reported for the mathematics assessment. A summative scale score and claim scores for Reading and Writing are reported for the ELA assessment.

7.4.1. Establishing the Reporting Scales

There are 201 defined summative scale score points for both the ELA and mathematics components, with the scale ranging from 650 (lowest obtainable scale score) to 850 (highest obtainable scale score). Scale scores are calculated from the IRT theta score calculated for each student from their test responses. Scale score calculations are based on a linear transformation between two sets of score values: IRT theta scores and scale scores. Not all possible scale scores may be realized in a scoring table. The relationship between raw and scale scores is expressed through scaling constants that define the mathematical relationship

between the two score scales. The threshold scores and scaling constants for the overall ELAGP and MATGP scales are reported in Table 7.4.1.

To facilitate the linear transformation between the IRT-derived theta scale and the reporting scale score range, anchor points were specified. Anchors were selected at the 725 and 750 scale score levels. The 725 scale score level was chosen as an anchor due to the cut score being set at this level. Psychometric analysis was conducted to identify other possible anchor points. The 700 and 750 levels were considered. These levels are among the defined performance levels in the grade 10 ELA and Integrated Math II NMC assessments but would only serve as 2nd anchor points for the ELAGP and MATGP scales. Pearson and NMC psychometricians calculated the 750 scale score anchor points for ELAGP and MATGP scales at approximately the same theta score differences from the 725 threshold scores, as would define the 750 scale score level in similar NMC Affiliate program test scales.

Table 7.4.1 Threshold Scores and Scaling Constants for NJGPA

Test	Theta	Scale Score	A	B
ELAGP	-0.310	725	43.2227	738.399
MATGP	-0.049	725	29.80448	726.4604

7.4.2. ELA Reading and Writing Claim Scales

As with the summative scale scores, scale scores for Reading and Writing were defined for each test as a linear transformation of the IRT theta scale. There are 81 defined scale score points possible for Reading, ranging from 10 to 90 and 51 defined scale score points possible for Writing, ranging from 10 to 60. Not all possible scale scores may be realized in a scoring table. The same IRT theta scale was used for Reading and Writing as was used for the ELA summative scores. The scaling constants for the Reading and Writing scales are reported in Table 7.4.2.

Table 7.4.2 Scaling Constants for Reading and Writing ELAGP Claims

	Reading		Writing	
Assessment	AR	BR	AW	BW
ELAGP	17.28907	45.35961	8.644537	32.67981

Table 7.4.3 reports the threshold scores for ELA and mathematics subclaim scales. These values are set at 1.5 times greater than and less than the standard error (+/- 1.5 SE) calculated for each threshold score. For ELA 1.5 SE was calculated to be 0.470 theta and mathematics was calculated to be 0.4526 theta.

Table 7.4.3 NJGPA Subclaim Theta Cut Scores

Assessment	Theta
ELAGP	-0.7804
	0.1605
MATGP	-0.5164
	0.4148

7.4.3. Creating Conversion Tables

A conversion table relates the number of points earned by a student for the ELA summative score, the mathematics summative score, the Reading claim score, or the Writing claim score to the corresponding scale score for the test form administered to that student. An IRT inverse test characteristic curve (TCC) approach is used to develop the relationship between point scores and IRT ability parameters or theta scores. In carrying out the calculations, estimates of item parameters and thetas are substituted for parameters in the formulas in each of the following steps:

Step 1: The expected item score (i.e., estimated item true score) is calculated for every theta in the selected range (between -15 and +15, in 0.0001 increments) based on the GPCM for both dichotomous and polytomous items

Equation 7.2

$$s_i(\theta_j) = \sum_{m=0}^{M_i-1} mP_{im}(\theta_j)$$

Equation 7.3

$$P_{im}(\theta_j) = \frac{exp[\sum_{k=0}^{m} Da_i(\theta_j - b_i + d_{ik})]}{\sum_{v=0}^{M_i-1} exp[\sum_{k=0}^{v} Da_i(\theta_j - b_i + d_{iv})]}$$

where $a_i(\theta_j - b_i + d_{i0}) \equiv 0$; $s_i(\theta_j)$ is the expected item score for item i on theta, θ_j; $P_{im}(\theta_j)$ is the probability of a student, j, with θ_j getting score m on item i; m_i is the number of score categories of item i; with possible item scores as consecutive integers from 0 to $M_i - 1$; D is the IRT scale constant (1.7); a_i is the item slope parameter; b_i is the item location parameter reflecting overall item difficulty; d_{ik} is a location parameter incrementing the overall item difficulty to reflect the difficulty of earning score category k; v is the number of score categories. Since the 2PL model can be considered a special case of the GPCM, the latter can be used to calibrate dichotomously scored items, resulting in 2PL model item parameters.

Step 2: The expected (weighted) test score for every theta in the selected range is calculated as

Equation 7.4

$$T_j = \sum_{i=1}^{I} w_i \, s_i(\theta_j)$$

Where T_j is the expected (weighted) test score on theta,θ_j; w_i is the item weight for item i I is the total number of items in the test form.

Step 3: The estimated conditional standard error of measurement (CSEM) is calculated for each theta in the selected range as

Equation 7.5

$$CSEM_j = \sqrt{\frac{1}{\sum_{i=1}^{I} L_i(\theta_j)}}$$

Equation 7.6

$$L_i(\theta_j) = (Da_i)^2[s_{i2}(\theta_j) - s_i^2(\theta_j)]$$

Equation 7.7

$$s_{i2}(\theta_j) = \sum_{m=0}^{M_i-1} m^2 P_{im}(\theta_j)$$

where $L_i(\theta_j)$ is the estimated item information function for item i on theta, θ_j.

Step 4: Every raw score is matched with a theta, where $\theta_j$is the theta for a raw score r_h , if $T_j - r_h$ is minimum across all T_j.

Step 5: The reported scale score is calculated. Using the A and B scaling constants in Table 7.4.1, each theta value is converted to a scale score and each theta CSEM to a scale score CSEM:

Equation 7.8

$$ScaleScore = A \times \theta + B$$

Equation 7.9

$$CSEM = CSEM_\theta \times A$$

The scale scores are rounded to the nearest whole number, and CSEMs are rounded to the tenths place. Furthermore, the scale scores are truncated with the lowest obtainable scale score (LOSS) of 650 and the highest obtainable scale score (HOSS) of 850.

Figure 7.4.1 and Figure 7.4.2 contains TCC, CSEM, and estimated test information function (TIF) curves for all operational NJGPA forms. Within each figure, the curve is reported on the theta scale, and the vertical dotted line indicates the performance level threshold score on the theta scale.

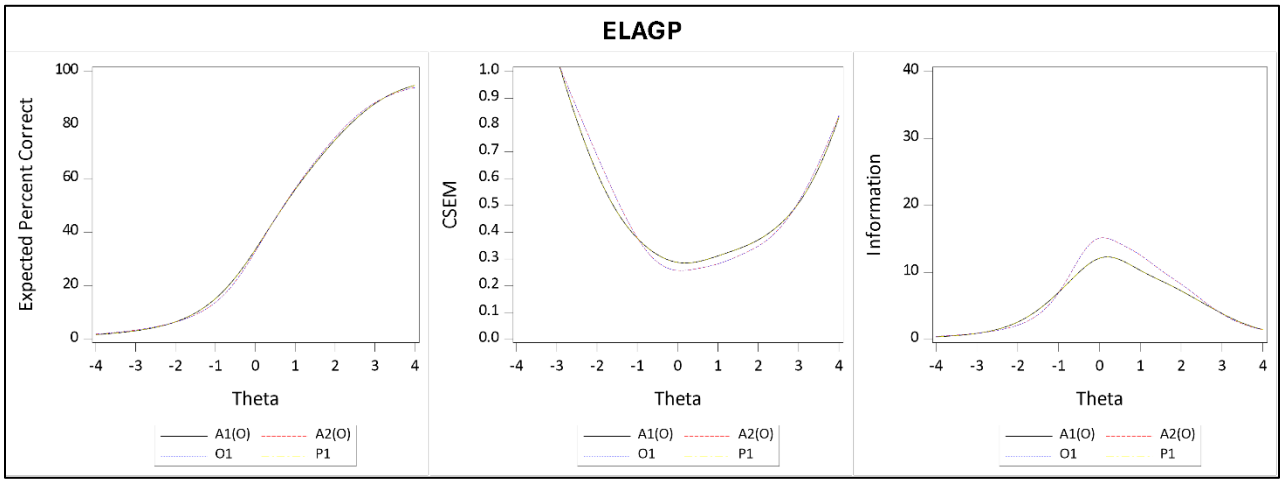

Figure 7.4.1 ELAGP Test Characteristic Curves, CSEM Curves, and Information Curves

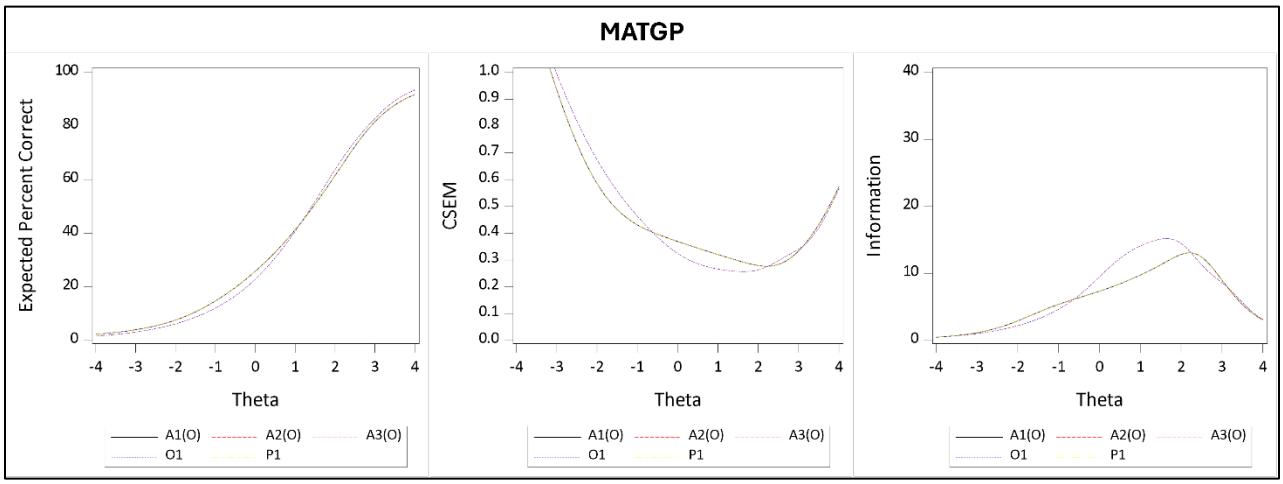

Figure 7.4.2 MATGP Test Characteristic Curves, CSEM Curves, and Information Curves

7.5. Scale Score Distributions

7.5.1. ELA Score Distributions

The overall performance of students taking the 2024 ELAGP core forms is reported in the cumulative score distributions given in Table 7.5.1. Score distribution information is also illustrated in Figure 7.5.1. The vertical axis of each graph represents the proportion of students earning the scale score point indicated along the horizontal axis. For the ELAGP score distribution, the *y*-axis ranges from 0 to 0.08 and the *x*-axis from 650 to 850.

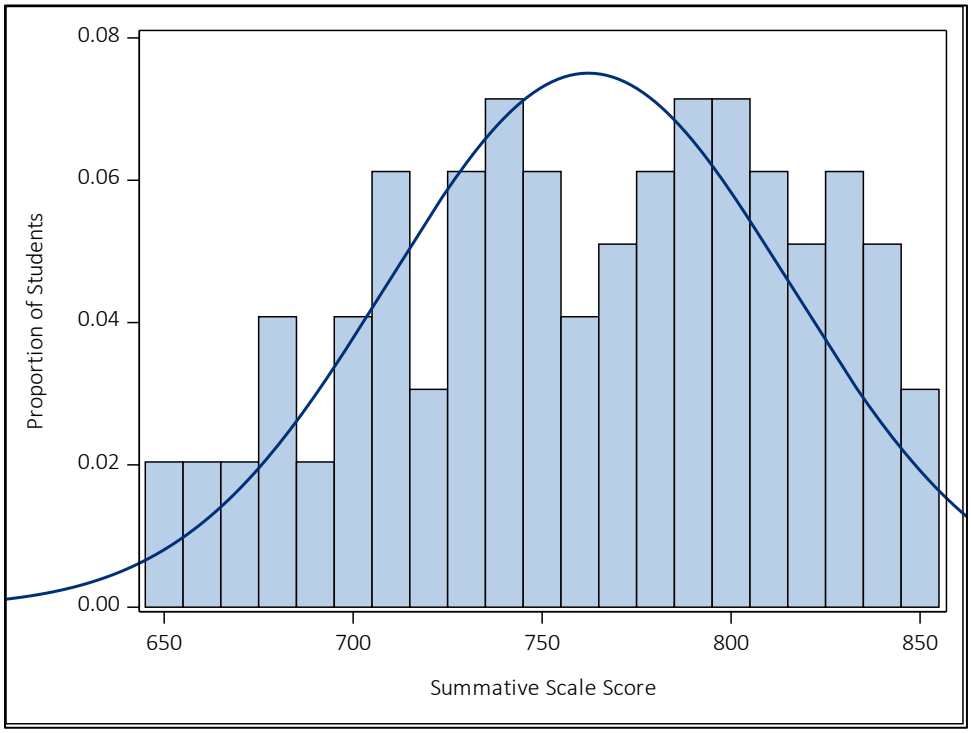

Figure 7.5.1 ELAGP Score Distribution

The performance of students taking 2024 ELAGP is reported in the cumulative score distributions in Table 7.5.1. Note that since this table reports actual student performance, not all scale score bands may be realized.

Table 7.5.1 ELAGP Scale Score Cumulative Frequencies

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	2,511	2.42	2,511	2.42
655-659	1,297	1.25	3,808	3.67
660-664	8	0.01	3,816	3.68
665-669	1,348	1.30	5,164	4.98
670-674	9	0.01	5,173	4.99
675-679	1,363	1.31	6,536	6.30
680-684	1,277	1.23	7,813	7.53
685-689	1,246	1.20	9,059	8.73
690-694	1,190	1.15	10,249	9.88
695-699	1,123	1.08	11,372	10.96
700-704	1,049	1.01	12,421	11.97

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
705-709	1,023	0.99	13,444	12.96
710-714	2,089	2.01	15,533	14.97
715-719	1,077	1.04	16,610	16.01
720-724	2,147	2.07	18,757	18.08
725-729	2,280	2.20	21,037	20.28
730-734	2,569	2.48	23,606	22.76
735-739	2,728	2.63	26,334	25.39
740-744	3,011	2.90	29,345	28.29
745-749	3,335	3.22	32,680	31.51
750-754	5,551	5.35	38,231	36.86
755-759	4,063	3.92	42,294	40.78
760-764	4,354	4.20	46,648	44.98
765-769	4,611	4.45	51,259	49.43
770-774	4,728	4.56	55,987	53.99
775-779	2,392	2.31	58,379	56.30
780-784	4,946	4.77	63,325	61.07
785-789	4,915	4.74	68,240	65.81
790-794	4,894	4.72	73,134	70.53
795-799	2,367	2.28	75,501	72.81
800-804	4,477	4.32	79,978	77.13
805-809	4,203	4.05	84,181	81.18
810-814	1,996	1.92	86,177	83.10
815-819	3,695	3.56	89,872	86.66
820-824	1,695	1.63	91,567	88.29
825-829	3,013	2.91	94,580	91.20
830-834	1,319	1.27	95,899	92.47
835-839	1,244	1.20	97,143	93.67
840-844	1,135	1.09	98,278	94.76
845-849	1,030	0.99	99,308	95.75
850	4,387	4.23	103,695	100.00

7.5.2. Mathematics Score Distributions

The overall performance of students taking the 2024 ELAGP core forms is reported in the cumulative score distributions given in Table 7.5.2. Score distribution information is also illustrated in Figure 7.5.2. The vertical axis of each graph represents the proportion of students earning the scale score point indicated along the horizontal axis. For the MATGP score distribution, the *y*-axis ranges from 0 to 0.08 and the *x*-axis from 650 to 850.

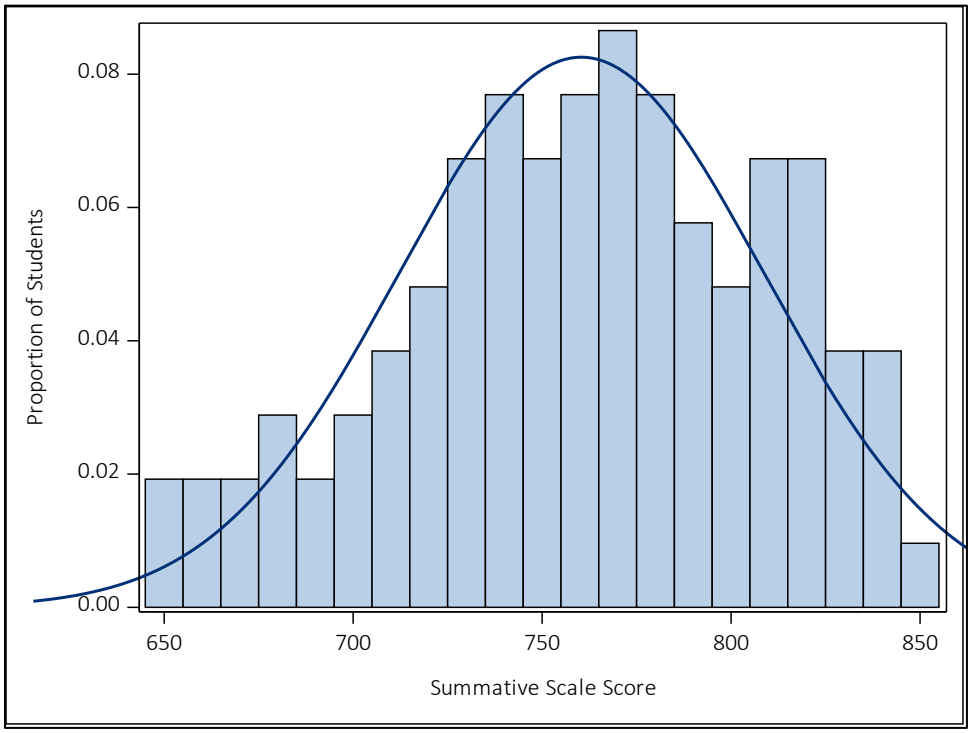

Figure 7.5.2 MATGP Score Distribution

The performance of students taking 2024 MATGP is reported by form in the cumulative score distributions given in Table 7.5.2. Note that since this table reports actual student performance, not all scale score bands may be realized.

Table 7.5.2 MATGP Scale Score Cumulative Frequencies

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
650-654	823	0.79	823	0.79
655-659	—	—	823	0.79
660-664	1,546	1.48	2,369	2.27
665-669	126	0.12	2,495	2.39
670-674	3,018	2.89	5,513	5.28

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
675-679	233	0.22	5,746	5.50
680-684	4,735	4.54	10,481	10.04
685-689	—	—	10,481	10.04
690-694	5,749	5.51	16,230	15.55
695-699	6,089	5.83	22,319	21.38
700-704	387	0.37	22,706	21.75
705-709	5,911	5.66	28,617	27.41
710-714	5,407	5.18	34,024	32.59
715-719	4,780	4.58	38,804	37.17
720-724	7,956	7.62	46,760	44.79
725-729	4,014	3.85	50,774	48.64
730-734	3,657	3.50	54,431	52.14
735-739	5,981	5.73	60,412	57.87
740-744	5,298	5.08	65,710	62.95
745-749	2,481	2.38	68,191	65.33
750-754	4,381	4.20	72,572	69.53
755-759	3,943	3.78	76,515	73.31
760-764	3,540	3.39	80,055	76.70
765-769	3,221	3.09	83,276	79.79
770-774	4,155	3.98	87,431	83.77
775-779	2,506	2.40	89,937	86.17
780-784	2,222	2.13	92,159	88.30
785-789	2,137	2.05	94,296	90.35
790-794	1,953	1.87	96,249	92.22
795-799	1,704	1.63	97,953	93.85
800-804	1,584	1.52	99,537	95.37
805-809	765	0.73	100,302	96.10
810-814	1,267	1.21	101,569	97.31
815-819	556	0.53	102,125	97.84
820-824	503	0.48	102,628	98.32
825-829	447	0.43	103,075	98.75
830-834	376	0.36	103,451	99.11

Score Band	Count	Percent	Cumulative Count	Cumulative Percent
835-839	281	0.27	103,732	99.38
840-844	244	0.23	103,976	99.61
845-849	—	—	103,976	99.61
850	380	0.36	104,356	100.00

7.5.3. ELA Major Claims Score Distributions

Score distributions are also presented for the Reading and Writing sub scores, in Figure 7.5.3 and Figure 7.5.4, respectively. For the Reading distribution, the *y*-axis ranges from 0 to 0.10 and the *x*-axis from 10 to 90. For the Writing distribution, the *y*-axis ranges from 0 to 0.08 and the *x*-axis from 10 to 60.

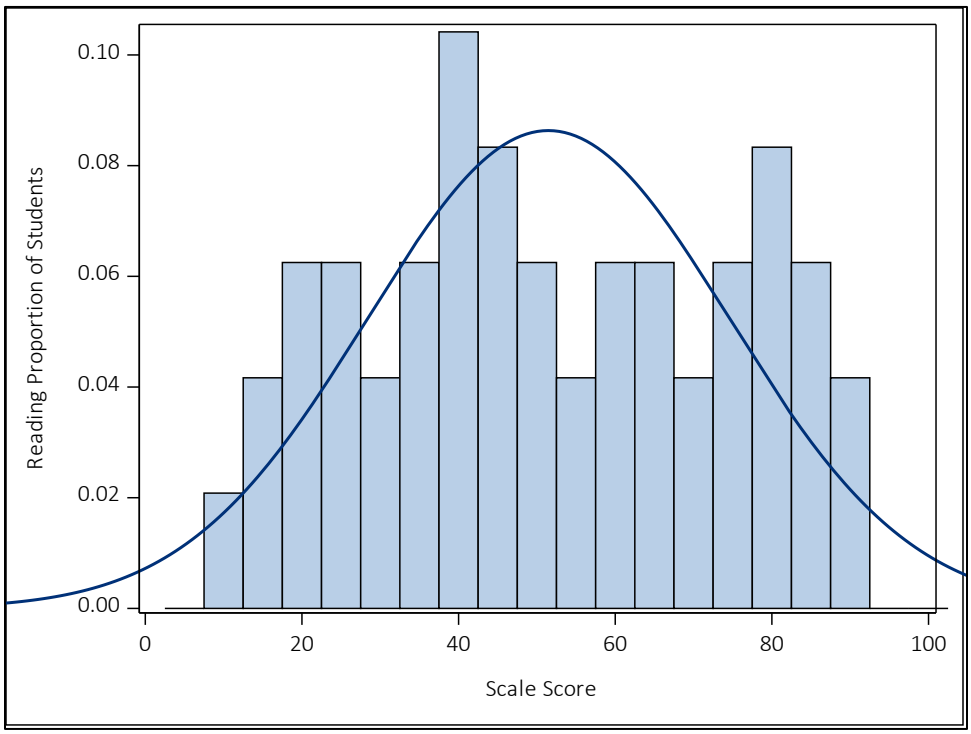

Figure 7.5.3 ELAGP Reading Score Distribution

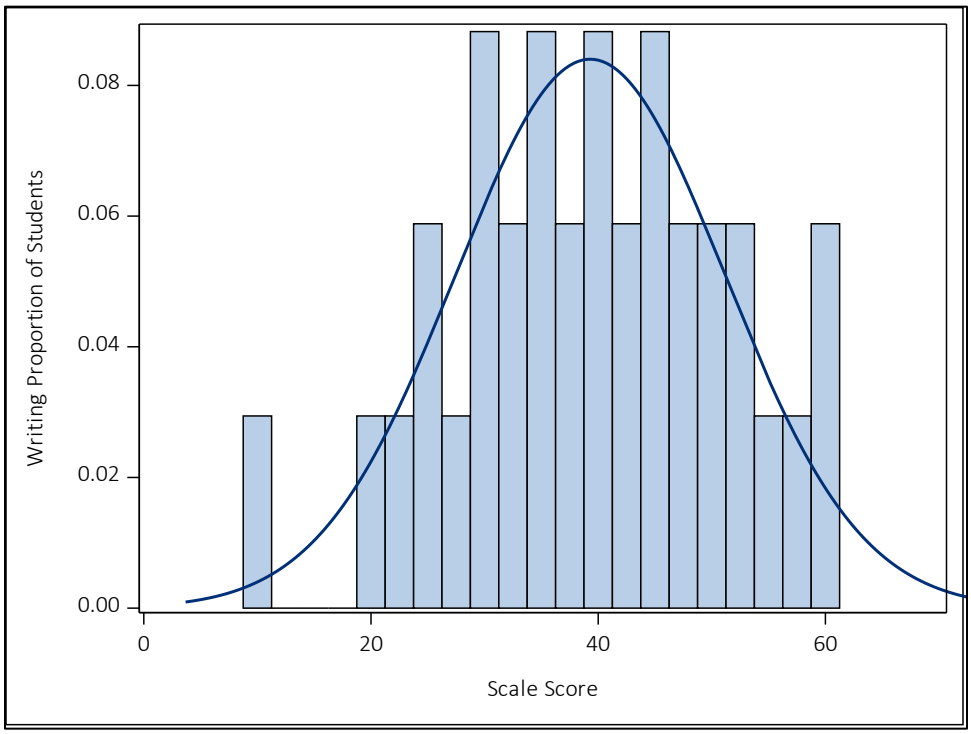

Figure 7.5.4 ELAGP Writing Score Distribution

7.5.4. Scale Score Distributions for Student Demographic Groups of Interest

The performance of demographic groups of students is summarized in Table 7.5.3 and Table 7.5.4 for ELA and mathematics, respectively. Each table reports the number of students in ethnic groups, and other demographic variables including gender, economic and English Learner status, and students with disabilities. For each demographic grouping the mean scale score and standard deviation is presented along with the minimum and maximum scale scores for reference. Table 7.5.3 also summarizes group performance on the Reading and Writing scores.

Table 7.5.3 ELAGP Subgroup Performance for Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		**103,695**	**766.85**	**48.63**	**650**	**850**
Gender	Female	50,526	774.49	46.68	650	850
	Male	53,005	759.51	49.33	650	850
Ethnicity	American Indian/Alaska Native	176	754.77	48.26	650	850
	Asian	10,818	800.12	39.23	650	850
	Black or African American	14,844	749.76	46.35	650	850

Group Type	Subgroup	N	Mean	SD	Min.	Max.
	Hispanic/Latino	34,090	749.81	50.42	650	850
	Native Hawaiian or Pacific Islander	230	780.35	48.78	650	850
	Two or more races	2,415	775.19	43.87	650	850
	White	41,074	777.92	42.12	650	850
Economic Status*	Not Economically Disadvantaged	69,956	775.50	46.12	650	850
	Economically Disadvantaged	33,739	748.91	48.81	650	850
English Learner Status	Non-English Learner	96,336	772.30	44.91	650	850
	English Learner	7,359	695.44	37.84	650	841
Disabilities	Students without Disabilities	82,311	773.78	46.36	650	850
	Student with Disability (SWD)	21,384	740.17	47.95	650	850
Reading Summative Score		**103,695**	**55.80**	**19.64**	**10**	**90**
Gender	Female	50,526	57.91	19.07	10	90
	Male	53,005	53.76	19.96	10	90
Ethnicity	American Indian/Alaska Native	176	50.56	19.18	10	90
	Asian	10,818	69.56	16.69	10	90
	Black or African American	14,844	49.20	18.46	10	90
	Hispanic/Latino	34,090	48.42	19.79	10	90
	Native Hawaiian or Pacific Islander	230	60.66	18.69	10	90
	Two or more races	2,415	59.95	17.76	10	90
	White	41,074	60.45	17.19	10	90
Economic Status*	Not Economically Disadvantaged	69,956	59.48	18.77	10	90
	Economically Disadvantaged	33,739	48.16	19.19	10	90
English Learner Status	Non-English Learner	96,336	57.94	18.30	10	90
	English Learner	7,359	27.78	14.27	10	90
Disabilities	Students without Disabilities	82,311	58.39	18.88	10	90
	Student with Disability (SWD)	21,384	45.83	19.30	10	90
Writing Summative Score		**103,695**	**38.09**	**13.01**	**10**	**60**
Gender	Female	50,526	40.52	12.15	10	60
	Male	53,005	35.77	13.38	10	60

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Ethnicity	American Indian/Alaska Native	176	35.29	13.45	10	60
	Asian	10,818	45.59	9.78	10	60
	Black or African American	14,844	33.87	13.14	10	60
	Hispanic/Latino	34,090	34.32	14.04	10	60
	Native Hawaiian or Pacific Islander	230	41.16	13.89	10	60
	Two or more races	2,415	39.64	11.74	10	60
	White	41,074	40.70	11.21	10	60
Economic Status	Not Economically Disadvantaged	69,956	40.04	12.16	10	60
	Economically Disadvantaged	33,739	34.05	13.75	10	60
English Learner Status	Non-English Learner	96,336	39.49	11.98	10	60
	English Learner	7,359	19.84	12.06	10	60
Disabilities	Students without Disabilities	82,311	39.89	12.21	10	60
	Student with Disability (SWD)	21,384	31.17	13.66	10	60

Note. SD = standard deviation. Economic status was based on participation in the National School Lunch Program (NSLP): receipt of free or reduced-price lunch (FRL).

Table 7.5.4 MATGP Subgroup Performance for Scale Scores

Group Type	Subgroup	N	Mean	SD	Min.	Max.
Full Summative Score		**104,356**	**734.94**	**38.85**	**650**	**850**
Gender	Female	50,943	735.01	37.33	650	850
	Male	53,251	734.83	40.25	650	850
Ethnicity	American Indian/Alaska Native	176	725.75	39.86	650	850
	Asian	10,839	774.86	38.59	650	850
	Black or African American	14,952	715.01	30.32	650	850
	Hispanic/Latino	34,606	718.72	32.01	650	850
	Native Hawaiian or Pacific Islander	229	746.38	40.37	650	850
	Two or more races	2,413	741.41	38.34	650	850
	White	41,080	744.95	35.34	650	850
Economic Status*	Not Economically Disadvantaged	70,209	743.35	39.22	650	850
	Economically Disadvantaged	34,147	717.64	31.71	650	850

Group Type	Subgroup	N	Mean	SD	Min.	Max.
English Learner Status	Non-English Learner	96,615	737.81	38.40	650	850
	English Learner	7,741	699.01	23.41	650	817
Disabilities	Students without Disabilities	82,979	740.03	38.36	650	850
	Student with Disability (SWD)	21,377	715.15	34.12	650	850
Language Form	Spanish	2,864	695.90	20.19	650	771

Chapter 8. Student Demographics and Differential Item Functioning (DIF)

8.1. Overview of Test-Taking Population

More than 100,000 New Jersey high school juniors attempted the NJGPA in spring 2024. Chapter 8 reports important demographic descriptions of the testing population in each grade and subject as well as results of Differential Item Functioning (DIF) analysis. DIF analyses are utilized to detect systematic differences in performance on the test items between demographic groups.

8.2. Rules for Inclusion of Students in Analyses

Criteria for inclusion of students were implemented prior to all operational analyses. These rules were established by NMC psychometrics in consultation with NJDOE to determine which, if any, student records should be removed from analyses. This data screening process resulted in higher-quality, albeit slightly smaller, data sets.

Student response data were included in analyses when:

1. Valid form numbers were observed for each unit for online assessments.
2. Student records were not flagged as "void" (i.e., do not score).
3. The student attempted at least 25 percent of the items in each unit or form.

Additionally, in cases where students had more than one valid record, the record with the higher raw score was chosen. Records for students with administration issues or anomalies were excluded from analyses.

8.3. Time to Attempt Assessment Items

It is important to understand how long students may take to answer each item on the assessment. For each operational form, Table 8.3.1 reports the minimum, maximum, median, and mean time to answer the items on the form and the items in each unit that make up the test form. Also reported are the standard deviation of the mean and the time for 80% of respondents to answer (P80).

Table 8.3.1 Time in Seconds for All Test Items and Items by Unit

Test	N	Min	Median	P80	Max	Mean	SD	Unit 1 Mean	Unit 2 Mean
ELAGP	103,695	0	200	475	17076	384.68	550.62	358.33	416.87
MATGP	104,356	0	145	351	10,579	235.45	267.09	213.56	264.07

Note. SD=Standard Deviation.

8.4. Demographics

Table 8.4.1 presents the number and percentage of students who took the ELAGP and MATGP assessments in each mode, either computer-based test (CBT) or paper-based test (PBT). Table 8.4.1 also provides this information for students who took the mathematics component in Spanish.

Markedly more students tested online than on paper for both content areas. For ELA, the percentage of online students by grade level was greater than 99 percent. For all mathematics students, the percentage of students testing online was greater than 99 percent. The percentage of students taking Spanish-language mathematics online forms was greater than or equal to 99 percent.

Table 8.4.1 ELA Test Takers

Test	% of All Students	N Students	N CBT	% CBT	N PBT	% PBT
ELAGP	100.0	103,695	103,617	99.9	78	0.1
MATGP	100.0	104,356	104,237	99.9	119	0.1
MATGP Spanish	100.0	2,864	2,862	99.9	2	0.1

Note. n/r = not reported due to n<20. CBT=Computer-based test. PBT=Paper-based test.

Table 8.4.2 summarizes demographic information for students with valid ELA scores, and Table 8.4.3 presents demographics for students with valid mathematics scores. Percentages are not reported in instances where fewer than 20 students were tested.

Table 8.4.2 ELAGP Test Taker Demographic Information

Demographic	N	Percent
Economically Disadvantaged	35,737	39.0
Student with Disability (SWD)	19,610	21.4
English Learner	11,807	12.9
Male	46,286	50.5
Female	45,404	49.5
American Indian/ Alaska Native	216	0.2
Asian	9,974	10.9
Black or African American	12,447	13.6
Hispanic/Latino	30,565	33.3
White/Caucasian	34,822	38.0
Native Hawaiian or Pacific Islander	162	0.2
Two or more races	3,482	3.8
Unknown Ethnicity	25	0.0

Table 8.4.3 MATGP Test Taker Demographic Information

Demographic	N	Percent
Economically Disadvantaged	36,910	39.4
Student with Disability (SWD)	19,636	20.9
English Learner	13,893	14.8
Male	47,340	50.5
Female	46,413	49.5
American Indian/ Alaska Native	216	0.2
Asian	10,187	10.9
Black or African American	12,566	13.4
Hispanic/Latino	32,082	34.2
White/Caucasian	35,026	37.4
Native Hawaiian or Pacific Islander	166	0.2
Two or more races	3,487	3.7
Unknown Ethnicity	26	0.0

8.5. Differential Item Functioning

Differential item functioning (DIF) is a procedure that matches students based on total test scores to compare the performance of similarly able students across subgroups. The procedure identifies two contrasting groups called focal and reference for which differences in item performances are computed. Table 8.5.1 indicates the focal and comparison groups used in DIF comparisons. At least 100 students in each group (focal and reference) and 300 total students across the two groups are required for DIF procedures to be conducted. For the procedures described next, positive DIF values indicate that, for students of similar ability, the focal group has a higher mean item score than the reference group. Negative DIF values indicate that, for students of similar ability, the focal group has a lower mean item score than the reference group.

Table 8.5.1 DIF Comparison Groups

Comparison Type	Focal Group (N≥100)	Reference Group (N≥100)
Gender	Female	Male
Ethnicity	African American	White
	Asian	White
	American Indian/Alaska Native	White
	Hispanic	White

Comparison Type	Focal Group (N≥100)	Reference Group (N≥100)
	Pacific Islander	White
	Multiple	White
Economic Status	Economically Disadvantaged	Not Economically Disadvantaged
English Learners	English Learner	English Proficient (including former English learners)
Students with an IEP	IEP	No IEP

8.5.1. Dichotomous Items: Mantel-Haenszel

The Mantel-Haenszel (MH) chi-square approach (Mantel & Haenszel, 1959) is used to detect DIF in dichotomously scored, one-point items. The range of total scores is divided into 10 stratifications (*S*) based on raw score performance, and those strata are used to match samples from each group. Contingency tables (such as in Table 8.5.2) for each stratum are constructed for the responses to the item in which *S* represents the strata, *Wrs* and *Wfs* represent the number of students (in the reference and focal groups, respectively) who answer the item incorrectly, *Rrs* and *Rfs* represent the number of students (in the reference and focal groups, respectively) who answer the item correctly, and *Nts* represents the total number of students (*Wrs + Rrs + Wfs + Rfs)*.

Table 8.5.2 Mantel-Haenszel Contingency Table

Score Stratum (S)	Incorrect/Wrong (O)	Correct/Right (1)	Total
Reference	*Wrs*	*Rrs*	*Wrs + Rrs*
Focal	*Wfs*	*Rfs*	*Wfs + Rfs*
Total	*Wrs + Wfs*	*Rrs + Rfs*	*Nts*

A common odds ratio is computed across all intervals of matched groups using the following formula (Dorans & Holland, 1993):

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S} R_{rs} W_{fs} \Big/ N_{ts}}{\sum_{s=1}^{S} R_{fs} W_{rs} \Big/ N_{ts}}.$$

Furthermore, the Mantel-Haenszel delta statistic (MHD-DIF) (Holland & Thayer, 1988) is computed to measure the degree and magnitude of DIF using the formula

$$MH_{D\text{-}DIF} = -2.35 \ln(\hat{\alpha}_{MH}).$$

8.5.2. Polytomous Items: Standardized Mean Difference

For polytomous items, the MHD-DIF is not calculated. Rather, a standardized mean difference (SMD) is calculated using a contingency table that extends the possible items scores beyond 1 point using this formula,

$$SMD = \sum_{s} w_{Fs} m_{Fs} - \sum_{s} w_{Fs} m_{Rs}$$

where $w_{Fs} = n_{F+s}/n_{F++}$ is the focal group proportion at the *sth* stratification variable; $m_{Fs} = (1/n_{F+s})F_s$ is the focal group's mean item score in the sth stratum; and $m_{Rs} = (1/n_{R+s})R_s$ is the reference group's mean item score in the *sth* stratum. Because the focal group proportion is used in both terms of the equation, the reference group's item mean is weighted, whereas the focal group's item mean is unweighted.

The effect size (ES) is then computed by dividing by the total group standard deviation (SD) using this equation:

$$ES = \frac{SMD}{SD}.$$

By using Mantel's chi-square statistic (1963), the magnitude of the ES is interpreted using Golia's (2012) rules.

8.5.3. DIF Classification

Based on the DIF statistics and significance tests, items are classified into three categories: A, B, or C (as in Table 8.5.3). Category A items contain negligible difference in performance and Category B items exhibit slight to moderate difference in performance, while Category C items exhibit moderate to large difference in performance. Items flagged with C-DIF during the preliminary analyses were provided to both the ELA and mathematics test development managers and the accessibility, accommodations, and fairness (AAF) specialist as part of the preliminary analysis communication plan.

Table 8.5.3 DIF Classifications

Analysis	Criteria
Differential Item Functioning (DIF)	+ Favors the focal group – Favors the reference group
Mantel-Haenszel	A. Negligible – MH is not significantly different from 0 OR (MH is significantly different from 0 AND has a delta absolute value < 1). B. Slight to Moderate – MH is significantly different from 0 AND the absolute value of delta is < 1.5 AND has a delta absolute value greater than or equal to 1. C. Moderate to Large – MH is significantly different from 1 AND delta has an absolute value greater than or equal to 1.5.
Standardized Mean Difference	A. Negligible – is not significantly different from 0 OR has an absolute value ≤ 0.17. B. Slight to Moderate – is significantly different from 0 AND 0.17 < |ES| ≤ 0.25. C. Moderate to Large – is significantly different from 0 AND has an absolute value > 0.25.

8.5.4. Differential Item Functioning Results

Table 8.5.4 presents DIF results for ELAGP and Table 8.5.5 presents DIF results for MATGP, offering insights into potential disparities in item performance across different demographic groups. DIF analysis helps identify whether certain test items exhibit differential difficulty or functioning for distinct subgroups.

Table 8.5.4 ELAGP Post-Administration Differential Item Functioning

		C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
DIF Comparison	**N Items**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Female vs. Male	20			1	5	19	95				
White vs. Black/ African American	20					20	100				
White vs. Hispanic/ Latino	20	2	10	2	10	16	80				
White vs. Asian	20					18	90	2	10		
White vs. American Indian/ Alaska Native	20	2	10			18	90				
White vs. Native Hawaiian or Pacific Islander	20	1	5			19	95				
White vs. Two or more races	20					20	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	20					20	100				
Non-English Learner vs. English Learner	20	3	15	5	25	12	60				
Student without Disability vs. Student with Disability	20					20	100				

For the ELAGP assessment, the following comparison groups exhibited some degree of DIF. The Female vs Male and White vs. Native Hawaiian or Pacific Islander comparisons, one item each exhibited B- (B minus) DIF. For the White vs Hispanic/Latino comparison, two items exhibited C- (C minus) DIF, and two items exhibited B- DIF. For the White vs American Indian/Alaska Native comparison, two items exhibited C- DIF. For the Non-English Learner vs English Learner comparison, three items exhibited C- DIF and five items exhibited B- DIF. For the Asian vs White comparison, two items exhibited B+ (B plus) DIF.

Table 8.5.5 MATGP Post-Administration Differential Item Functioning

		C- DIF		B- DIF		A DIF		B+ DIF		C+ DIF	
DIF Comparison	**N Items**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Female vs. Male	30	1	3	1	3	28	93				
White vs. Black/ African American	30			1	3	29	97				
White vs. Hispanic/ Latino	30					30	100				
White vs. Asian	30					26	87	3	10	1	3
White vs. American Indian/ Alaska Native	30			1	3	29	97				
White vs. Native Hawaiian or Pacific Islander	30			1	3	29	97				
White vs. Two or more races	30					30	100				
Not Economically Disadvantaged vs. Economically Disadvantaged	30					30	100				
Non-English Learner vs. English Learner	30			1	3	28	93	1	3		
Student without Disability vs. Student with Disability	30					30	100				

For the MATGP assessment the following comparison groups exhibited some degree of DIF. For the Male vs Female comparison, one item exhibited C- (C minus) DIF and one item exhibited B- (B minus) DIF. For the White vs Black/African American, White vs American Indian/Alaska Native, White vs Native Hawaiian or Pacific Islander, and Non-English Learner vs English Learner comparisons, one item each exhibited B- DIF. For the Asian vs White comparison, three items exhibited B+ (B plus) DIF and one item exhibited C+ (C plus) DIF.

Chapter 9. Reliability

9.1. Overview

Reliability focuses on the extent to which differences in scores reflect true differences in the level of knowledge, skills, and abilities being assessed rather than chance fluctuations. Thus, reliability measures the level of consistency of the scores that would result if the assessment were to be repeatedly administered under the same conditions. Any degree of inconsistency is assumed due to random fluctuations that occur during administration. The sources of random fluctuations can be internal or external for the students, internal for the assessment, and/or from other phenomena that randomly occur during administration. For example, random fluctuation can be due to the use of multiple forms of the assessment administered to the students, or the assignment of raters assigned to score students' responses to constructed-response item prompts. In statistical terms, the variance in the distribution of scores, essentially the observed differences among students, is partly due to real differences in the levels of knowledge, skills, and abilities being assessed (true variance) and partly due to differences caused by random errors that customarily occur in the measurement process (error variance). Reliability is the proportion of the total variance that is true variance. Psychometricians use statistical formulas to estimate the level of reliability of students' scores.

There are several different ways to estimate reliability. The type of raw score reliability estimate reported here is an internal consistency measure, which is derived from an analysis of the consistency of the performances of students among items within an assessment. An internal consistency reliability estimate is used because it serves as a good estimate of reliability when using multiple items within a test form, but it does not consider form-to-form variation due to lack of test form parallelism, nor can it provide information regarding score reliability across repeated administrations due to day-to-day variations, for example, day-to-day changes of students' states of health or of the administration environment.

Reliability coefficients range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely students are to obtain very similar scores upon repeated administrations if the students do not change in their level of the knowledge or skills measured by the assessment. Acceptable ranges of reliability tend to exceed 0.6, with values over 0.8 considered good to excellent (Cortina, 1993; Schmitt, 1996). Estimates lower than 0.5 may indicate a lack of internal consistency.

In classical test theory, standard errors of measurement (SEM) quantify the amount of error in the scores. SEM is the extent to which students' scores tend to differ from the scores they would receive if the test were perfectly reliable. As the SEM increases, the amount of measurement error increases, and the variability of students' observed scores is likely to increase across hypothetical repeated administrations. Observed scores with large SEMs pose a challenge to the valid interpretation of a single score. Classical test theory reliability and SEM estimates were calculated for each NJSLA test form, and the weighted average score is reported here.

9.2. Reliability and SEM Estimation

Coefficient alpha (Cronbach, 1951) is a reliability measure for use when there are dichotomously or polytomously scored items (Brennan, 2001). The coefficient is calculated by using both the items' variances and the observed variance of the total raw scores in the following formula:

Equation 9.1

$$\alpha_x = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_x^2}\right)$$

in which n is the number of items, σ_i^2 is the variance of scores on each item, and σ_x^2 is the variance of the total raw scores. Coefficient alpha is a lower bound estimate of the reliability of the distribution of total raw scores. For example, if coefficient alpha has a value of .90 for the estimated level of reliability, the level of reliability (as a theoretical quantity) would be even higher in value. When other administration conditions and contexts are held constant, the more items the test form includes, the greater the value of coefficient alpha, and in turn, the greater the reliability of scores. Additionally, when sample sizes become smaller and more homogeneous, lower reliability estimates are obtained.

The formula for calculating the classical test theory SEM is

Equation 9.2

$$SEM = \sigma_x\sqrt{1-\alpha_x}$$

where σ_x is the standard deviation of the total raw scores and α_x is the value of coefficient alpha as computed above.

9.3 Scale Score Reliability Estimation

Like the level of classical test theory reliability, the level of scale score reliability can range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain similar scores upon repeated testing occasions. Because the scale scores are computed via a procedure that differs from the calculation of the total raw scores, coefficient alpha cannot be computed for scale scores. Instead, Kolen et al.'s (1996) method for scale score reliability is calculated. For details of scale score calculation, please review Chapter 7.4.

The general formula for the reliability coefficient,

Equation 9.3

$$\rho = 1 - \frac{\sigma^2(E)}{\sigma^2(X)},$$

involves the error variance $\sigma^2(E)$, and the total observed score variance $\sigma^2(X)$. Using Kolen et al.'s (1996) method, conditional raw score distributions are estimated using Lord and Wingersky's (1984) recursion

formula. The conditional raw score distributions are transformed into conditional scale score distributions. Denoting X as the raw total score ranging from 0 to X, and s as a resulting scale score after transformation, the conditional distribution of scale scores is written as $P(X = x|\theta)$. The mean and variance $\sigma^2[s(X)]$ of this distribution can be computed using these scores and their associated probabilities.

The average error variance of the scale scores is computed as

Equation 9.4

$$\sigma^2(Error_{scale}) = \int_\theta \sigma^2(s(X)|\theta)\, g(\theta) d\theta$$

where $g(\theta)$ is the ability distribution. The square root of the error variance is the conditional standard error of measurement of the scale scores.

Just as the reliability of raw scores is one minus the ratio of error variance to total variance, the reliability of scale scores is one minus the ratio of the average variance of measurement error for scale scores to the total variance of scale scores.

Equation 9.5

$$\rho_{scale} = 1 - \frac{\sigma^2(Error_{scale})}{\sigma^2[s(X)]}$$

The program POLYCSEM (Kolen, 2004) was used to estimate scale score error variance and reliability.

9.3. Reliability Results

9.3.1. Raw Score Reliability Results

Table 9.3.1 summarize test reliability estimates for the total testing group for ELAGP and MAGP. Please note that during operational test form construction described in section 2.4.3 of this report, multiple parallel operational forms of the accommodated core form may be constructed to facilitate delivery and scoring. The tables below report the weighted average statistics for all operational forms created for the ELA and mathematics assessments. Estimates were calculated for the total group, and for subgroups of 100 or more students, who were administered a specific test form. Since reliabilities were averaged across test forms, the minimum and maximum reliabilities are also provided. Minimum reliability reports the reliability for the test form with the lowest value and maximum reliability reports the reliability for the test form with the largest value. Typically, test forms administered to larger numbers of students (operational online forms) tend to have higher reliability compared to test forms administered to smaller numbers of students (accommodated forms).

Mean reliabilities tended to be robust for each NJGPA subject, with mean reliabilities ranging from 0.87 for ELAGP to 0.84 for MATGP. The lowest minimum reliability for a test form was 0.85 in ELAGP.

Table 9.3.1 Summary of Raw Score Test Reliability Estimates for Total Group

Test	Number of Forms	Avg. Max Possible Score	Avg. Raw Score SEM	Average Reliability	Minimum Reliability		Maximum Reliability	
					N	Alpha	N	Alpha
ELAGP	3	74	5.49	0.87	3881	0.85	99411	0.88
MATGP	4	56	3.04	0.91	443	0.86	90402	0.93

9.3.2. Scale Score Reliability Results

Table 9.3.2 reports the scale score reliability and SEM estimates for ELAGP and MATGP for spring 2024.

Table 9.3.2 Summary of Scale Score Test Reliability Estimates for Total Group

Test	Number of Forms	Avg. Scale Score SEM	Avg. Scale Score Reliability	Min. Scale Score Reliability	Max. Scale Score Reliability
ELAGP	3	13.5	0.95	0.93	0.96
MATGP	4	12.71	0.9	0.88	0.92

9.4. Reliability Results for Demographic Groups of Interest

Raw score reliability statistics were also calculated for student demographic groups. The results for these gender, ethnic and student needs groups, along with similar statistics calculated for the students taking various accommodated forms, are reported in Table 9.4.1 for the ELAGP assessment and Table 9.4.2 for the MATGP assessment. All forms had a maximum score of 74. Reliability estimates are provided for gender, ethnicity, special instruction needs, and accommodated forms. For MATGP, reliability is also provided for the Spanish-language forms. For some demographic and accommodation categories, the tables note "n/r" (not reported) for SEM and Alpha reliability coefficient, indicating that the reliability estimates are not available for those specific accommodation types due to insufficient sample sizes.

Table 9.4.1 ELAGP Summary of Test Reliability Estimates for Subgroups

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Total Group	74	5.49	0.87	3,881	0.85	99,411	0.88
Gender							
Male	74	5.39	0.87	2,496	0.85	225	0.88
Female	74	5.54	0.86	1,384	0.84	48,965	0.87

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Ethnicity							
Black/African American	74	5.15	0.85	677	0.83	14,099	0.87
Asian/Pacific Islander	74	5.26	0.84	127	0.83	10,902	0.85
Hispanic/Latino	74	5.49	0.87	1,335	0.84	129	0.88
American Indian/Alaska Native	74	5.76	0.88	170	0.88	170	0.88
Two or more races	74	5.51	0.86	2,338	0.86	2,338	0.86
White	74	5.44	0.85	180	0.83	1661	0.85
Special Instruction Needs							
Economically Disadvantaged	74	5.33	0.86	1,534	0.83	171	0.88
Not Economically Disadvantaged	74	5.53	0.86	232	0.85	67,377	0.87
English Learner	74	4.3	0.83	7,258	0.83	7,258	0.83
Non-English Learner	74	5.44	0.86	3,791	0.85	392	0.87
Students with Disabilities (SWD)	74	5.33	0.87	3,881	0.85	182	0.88
Students without Disabilities	74	5.63	0.87	221	0.86	82,090	0.87
Students Taking Accommodated Forms							
ASL	74	n/r	n/r	n/r	n/r	n/r	n/r
Closed Caption	74	5.62	0.88	103	0.88	103	0.88
Human Reader	74	5.55	0.88	264	0.88	264	0.88
Non-Screen Reader	74	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	74	n/r	n/r	n/r	n/r	n/r	n/r

Note: n/r = not reported.

Table 9.4.2 MATGP Summary of Test Reliability Estimates for Subgroups

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Total Group	56	3.04	0.91	443	0.86	90,402	0.93
Gender							
Male	56	3.05	0.91	238	0.87	45,500	0.93
Female	56	3.03	0.9	205	0.84	44,752	0.92
Ethnicity							
Black/African American	56	2.61	0.85	982	0.82	12,959	0.88
Asian/Pacific Islander	56	3.75	0.93	424	0.92	462	0.94
Hispanic/Latino	56	2.7	0.86	4,679	0.82	27,858	0.9
American Indian/Alaska Native	56	3.03	0.93	145	0.93	145	0.93
Two or more races	56	3.14	0.91	112	0.88	2,186	0.92
White	56	3.2	0.9	187	0.85	1910	0.92
Special Instruction Needs							
Economically Disadvantaged	56	2.69	0.86	3,922	0.83	27,832	0.9
Not Economically Disadvantaged	56	3.22	0.91	260	0.85	4,221	0.93
English Learner	56	2.33	0.72	3,130	0.62	4,316	0.81
Non-English Learner	56	3.13	0.91	424	0.86	5,013	0.93
Students with Disabilities (SWD)	56	2.7	0.87	2,756	0.8	15,643	0.92
Students without Disabilities	56	3.17	0.91	281	0.85	5,327	0.93
Students Taking Accommodated Forms							
American Sign Language	55	n/r	n/r	n/r	n/r	n/r	n/r
Human Reader	55	3.12	0.87	284	0.87	284	0.87
Non-Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r

	Avg. Max. Raw Score	Avg. SEM	Avg. Reliability	Minimum Reliability		Maximum Reliability	
				N	Alpha	N	Alpha
Screen Reader	55	n/r	n/r	n/r	n/r	n/r	n/r
Text-to-Speech	55	3.07	0.91	10,649	0.91	10,649	0.91
Students Taking Translated Forms							
Spanish Language	55	2.18	0.57	443	0.54	2,419	0.58

Note: n/r = not reported.

9.5. Reliability Estimates of Subclaim Scores

Table 9.5.1 presents average reliability estimates for various ELAGP subclaim scores across different forms, providing insights into the consistency and dependability of these scores. The table includes subclaim scores for Reading (RD), Reading: Literature (RL), Reading: Information (RI), Reading: Vocabulary (RV), Writing (WR), Written Expression (WE), and Knowledge of Language and Conventions (WKL).

Table 9.5.1 Average ELAGP Reliability Estimates for Subscores

	Reading: Total		Reading: Literature		Reading: Information		Reading: Vocabulary		Writing: Total		Writing Expression		Writing: Knowledge Language and Conventions	
Test	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability
ELAGP	44-44	0.82	12-12	0.59	20-20	0.69	12-12	0.48	30-30	0.84	24-24	0.87	06-06	0.87

Table 9.5.2 presents reliability estimates for the various subscores across mathematics forms. The subscores include Major Content (MC), Additional & Supporting Content (ASC), Expressing Mathematical Reasoning (MR), and Modeling & Applications (M&A). Each form is associated with corresponding score points and alpha reliability coefficients for the mentioned subscores.

Table 9.5.2 Average Math Reliability Estimates for Subscores

	Major Content		Additional & Supporting Content		Mathematics Reasoning		Modeling Practice	
Test	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability	Range of Raw Score	Average Reliability
MATGP	16-16	0.81	14-14	0.69	10-10	0.70	15-15	0.72

9.6. Reliability of Classification

Classification accuracy is defined as the extent to which the actual classifications of test takers (on the basis of their test scores) agree with those that would be made on the basis of their true scores—if their true scores could somehow be known. The term consistency refers to the agreement between classifications based on two nonoverlapping, equally difficult forms of the test (parallel forms) (Livingston & Lewis, 1995).

We used Livingston and Lewis's (1995) approach, which is intended to handle situations where items are not equally weighted and/or some or all the items are polytomously scored. This method is formulated as:

Equation 9.6

$$\tilde{n} = \frac{(\mu_x - X_{min})(X_{max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)},$$

where X_{min} is the lowest score for X, X_{max} is the highest score, μ_x is the mean, σ_x^2 is the variance, and r is the reliability. This method models the distribution of the true scores and of scores on a parallel form by using a four-parameter beta distribution.

As seen in the above formula, classification accuracy and consistency indices rely on the interaction between several different factors related to test design and standard-setting decisions. These factors include the number of score cuts, test reliability, measurement accuracy at the cut score, distance between adjacent cuts, location of the cut scores on the ability scale, and percentage of students around a cut score (Ercikan & Julian, 2002; Lee et al., 2002).

Classification accuracy indices quantify the percentage of students who are accurately placed below and/or above a given cut score. For example, a classification accuracy index of 0.88 means that were students to be classified twice, once according to their observed score and once according to their true score, 88% of those students would be classified in the same category both times. Similarly, classification consistency indices give the percentage of students classified consistently below and/or above a given cut score. For example, a classification index of 0.84 means that were two parallel forms to be administered to students, 84% of those students would be classified in the same way for both forms.

Table 9.6.1 reports the classification accuracy and classification consistency for classifying each student into one of the two performance levels for NJGPA.

Table 9.6.1 Classification Accuracy Indices at Cut Score Level for NJGPA

Test	Classification Accuracy: Proportion Accurately Classified	Classification Consistency: Proportion Consistently Classified
ELAGP	0.96	0.94

Test	Classification Accuracy: Proportion Accurately Classified	Classification Consistency: Proportion Consistently Classified
MATGP	0.92	0.86

Chapter 10. Validity

10.1. Overview

The *Standards for Educational and Psychological Testing*, issued jointly by the American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014), states the following:

> 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 its validity (or evidence of lack of validity), including design, content specifications, item development, and psychometric characteristics. The spring 2022–2023 NJGPA administration provided an opportunity to gather evidence of validity based on both test content and on the internal structure of the tests.

NMC applies the principles of universal design, as articulated in materials developed by the National Center for Educational Outcomes (NCEO) at the University of Minnesota (Thompson et al., 2002).

10.2. Evidence Based on Test Content

This section describes evidence of validity from the creation of items used on NJGPA forms which come from PARCC/NMC Affiliate ELA Grade 10, Algebra I and Geometry banks.

Evidence based on content of achievement tests is supported by the strength and clarity of the relationship between test items and content standards. The degree to which the test measures what it claims to measure is known as construct validity. The NJGPA adheres to the principles of evidence-centered design, in which the standards to be measured are identified and the performance a student needs to achieve to meet those standards is delineated in the evidence statements. Test items are reviewed for adherence to universal design principles, which maximize the participation of the widest possible range of students.

During the creation of ELA grade 10, Algebra I and Geometry items, Pearson built spreadsheets at the evidence statement level that incorporate the probability statements from the test blueprints and attrition rates at committee review and data review. The basis of our entire item development is driven by the use of these item development target spreadsheets. Before item development begins, these target spreadsheets are used to develop an internal item writing plan to correlate with the expectations of the test design. These were reviewed and approved by state leads and NMC.

In addition to the evidence statements, content was aligned through the articulation of performance in the performance level descriptors. At the policy level, the performance level descriptors include policy claims about the educational achievement of students who attain a particular performance level. They also include a broad description of the grade-level knowledge, skills, and practices that students performing at a particular achievement level are able to demonstrate. Those policy-level descriptors were the foundation for the subject- and grade-specific performance level descriptors, which, along with the evidence frameworks, guided the development of the items and tasks.

The college- and career-ready determinations (CCRD) in English language arts (ELA) and mathematics describe the academic knowledge, skills, and practices students must demonstrate to show readiness for success in entry-level, credit-bearing college courses and relevant technical courses. To validate the determinations, a postsecondary educator judgment study and a benchmark study were conducted of the SAT, ACT, National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), Programme of International Student Assessment (PISA), and Progress in International Reading Literacy Study (PIRLS) tests (McClarty et al., 2015).

Gathering construct validity evidence for the assessments was embedded in the process by which the assessment content was developed. At each step in the assessment development process, participating state educators, assessment experts, and bias and sensitivity experts reviewed texts, items, and tasks for accuracy, appropriateness, and to eliminate bias.

Please see Chapter 2 for an overview of the operational forms and the form development process. For more information on the origins of items, forms, and test blueprints, please consult chapter 10 of the most recent New Meridian Affiliate Technical Report Alternate Blueprint. The items and tasks were field tested prior to their use on an assessment. Items demonstrating questionable statistical performance, either within the testing population as a whole or within demographic groups, are reviewed at data review meetings.

10.3. Evidence Based on Internal Structure

Analyses of the internal structure of a test typically involve studies of the relationships among test items and/or test components (i.e., subclaims) in the interest of establishing the degree to which the items or components appear to reflect the construct on which a test score interpretation is based (AERA et al., 2014, p. 16). The term construct is used here to refer to the characteristics that a test is intended to measure; in the case of the NJGPA, the characteristics of interest are the knowledge and skills defined by the test blueprint for the ELA and mathematics components.

NJGPA scoring provides the overall ELA and mathematics test scores, the Reading claim score, and the Writing claim score, as well as ELA subclaim and mathematics subclaim scores. The goal of reporting at

this level is to provide criterion-referenced data to assess the strengths and weaknesses of a student's achievement in specific topics for each test component.

10.3.1. Intercorrelations

The ELAGP assessment comprises two claim scores—Reading (RD) and Writing (WR)— and five subclaim scores—Reading Literature (RL), Reading Information (RI), Reading Vocabulary (RV), Written Expression (WE), and Knowledge of Language and Conventions (WKL). The RD claim score is a composite of RL, RI, and RV. The WR claim score is a composite of WE and WKL and comprises only PCR items, with the same PCR items in each subclaim. The ELA operational test analyses were performed by evaluating the separate trait scores of WE and WKL, and for some PCR items, also RL or RI; therefore, the trait scores were used for the intercorrelations.

The MATGP assessment has four subclaim scores—Major Content (MC), Mathematical Reasoning (MR), Modeling Practice (MP), and Additional and Supporting Content (ASC).

High total group internal consistencies as well as similar reliabilities across subgroups provide additional evidence of validity. High reliability of test scores implies that the test items within a domain are measuring a single construct, which is a necessary condition for validity when the intention is to measure a single construct. Refer to Chapter 9 for reliability estimates for the overall population, subgroups of interest, as well as for claims and subclaims for ELA and subclaims for mathematics.

Another way to assess the internal structure of a test is through the evaluation of correlations among scores. These analyses were conducted between the ELA Reading and Writing claim scores and the ELA subclaims (RL, RI, RV, WE, and WKL) and between the mathematics subclaims. If these components within a content area are strongly related to each other, this is evidence of unidimensionality.

A series of tables is provided to summarize the results for the spring 2024 administration. *Table 10.3.1* presents the correlations observed between the ELA Reading and Writing subclaim scores for ELA. The intercorrelations for mathematics are provided in Table 10.3.2.

Table 10.3.1 ELAGP Average Intercorrelations between Subclaims

Subclaims	N Students	RD	RL	RI	RV	WR	WE	WKL
RD	103,695	1						
RL	103,695	0.83	1					
RI	103,695	0.92	0.66	1				
RV	103,695	0.79	0.52	0.58	1			
WR	103,695	0.77	0.73	0.74	0.50	1		
WE	103,695	0.77	0.73	0.73	0.49	1.00	1	
WKL	103,695	0.76	0.71	0.72	0.50	0.97	0.95	1

Note. RD=Reading, RL=Reading: Literature, RI=Reading: Information, RV=Reading: Vocabulary, WR=Writing, WE=Written Expression, WKL=Knowledge of Language and Conventions.

ELAGP intercorrelations range from moderate to high. The intercorrelations of reading subclaim scores range from 0.49 to 0.92 and likely indicate measurement of a single reading construct, with Reading Vocabulary a possibly related construct. The intercorrelations of writing subscores range from 0.95 to 1.0, indicating there is likely only one writing dimension being measured. The intercorrelations between reading and writing subscores, with the exception of Reading Vocabulary, likely indicate a unidimensional measurement factor. The WR, WE, and WKL scores tended to be highly correlated; this is expected given

that these three intercorrelations are based on the trait scores from the same Writing items. RL, RI, and RV, all subclaims of Reading, are moderately to highly correlated. Additionally, the WR claim and the WE and WKL subclaims are moderately correlated with RD subclaims (of RL, RI, and RV). These moderate to high ELA intercorrelations amongst the subclaims are sufficiently high to provide evidence that the ELA tests are unidimensional. The moderate intercorrelations among the subclaims and claims suggest the claims may be sufficient for individual student reporting.

Table 10.3.2 MATGP Average Intercorrelations between Subclaims

Subclaims	N Students	MC	ASC	EMR	M&A
MC	93,968	1			
ASC	93,968	0.72	1		
MR	93,968	0.71	0.67	1	
MP	93,968	0.81	0.68	0.68	1

Note. MC=Major Content, ASC=Additional & Supporting Content, MR=Mathematics Reasoning, MP=Modeling Practice.

MATGP intercorrelations are moderate. The main observable pattern in the mathematics intercorrelations is that the MC subclaim generally has slightly higher correlations with the ASC, EMR, and M&A subclaims; the intercorrelations amongst the ASC, EMR, and M&A subclaims are usually slightly lower. The mathematics intercorrelations are sufficiently high to suggest that mathematics tests are likely to be unidimensional with some minor secondary dimensions.

All observed mathematics intercorrelations between subclaims are moderate. The main observable pattern in the mathematics intercorrelations is that the Major Content subclaim generally has slightly higher correlations with the Additional and Supporting Content, Mathematical Reasoning, and Modeling Practice subclaims; the intercorrelations among the Additional and Supporting Content, Mathematical Reasoning, and Modeling Practice subclaims are usually slightly lower. The mathematics intercorrelations are sufficiently high to suggest that the mathematics tests are likely to be unidimensional with some minor secondary dimensions.

10.3.2. Reliability

The reliability analyses presented in Chapter 9 of this technical report provide information about the internal consistency of the summative assessments. Internal consistency is typically measured via correlations amongst the items on an assessment and provides an indication of how much the items measure the same general construct. As do the subclaim intercorrelations, the reliability estimates indicate that the items within each assessment are measuring the same construct and provide further evidence of unidimensionality.

10.3.3. Local Item Dependence

Local item independence was evaluated for ELA and mathematics. Local independence is one of the primary assumptions of item response theory (IRT) that states the probability of success on one item is not influenced by performance on other items when controlling for ability level. This implies that ability or theta accounts for the associations among the observed items. Local item dependence (LID), when present, essentially overstates the amount of information predicted by the IRT model. It can exert other

undesirable psychometric effects and represents a threat to validity since other factors besides the construct of interest are present. Classical statistics are also affected when LID is present because estimates of test reliability, like IRT information, can be inflated (Zenisky et al., 2003).

The LID issue affects the choice of item scoring in IRT calibrations. Specifically, if evidence suggests these items indeed have local dependence, then it might be preferable to sum the item scores into clusters or testlets as a method of minimizing LID. However, if these items do not appear to have strong local item dependence, then retaining the scores as individual item scores in an IRT calibration is preferred since more information concerning item properties is retained. During the initial operational administration of the summative assessments in spring 2015, a study that included two methods of investigating the presence of LID was conducted. A description of the study methods and findings is available in Chapter 10 of the most recent New Meridian Affiliate Technical Report Alternate Blueprint.

10.4. Evidence from Special Studies

During the creation and evolution of the testing programs that inform ELA grade 10, Algebra I, and Geometry, several research studies provided additional validity evidence for the goals of assessing more rigorous academic expectations, helping to prepare students for college and careers, and providing information back to teachers and parents about their students' progress toward college and career readiness. Details of these studies can be found in the most recent version of the New Meridian affiliate technical report.

References

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

Baldwin, P., Margolis, M.J., Clauser, B.E., Mee, J., Winward, M. (2019). The choice of response probability in bookmark standard setting: An experimental study. *Educational Measurement: Issues and Practice, 39*(1), 37–44.

Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012). Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments. *Bulletin*, Issue 21. Pearson Education, Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical Theories of Mental Test Scores* (pp 397–472). Reading, MA: Addison Wesley Publishing.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38(*4), 295–317.

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(*3), 297–334.

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.

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

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

Ercikan, K, & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269–294.

Henrysson, S. (1963). Correction of Item-Total Correlations in item analysis. *Psychometrika, 28*(2), 211–218.

Holland, P. W. & Thayer, D. T. (1988). Differential item performances and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Kolen, M. J. (2004). POLYCSEM windows console version [Computer software]. Iowa City, IA: The Center for Advanced Studies in Measurement and Assessment (CASMA), The University of Iowa.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33(2), 129–140.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412–432.

Lemke, E., & Wiersma, W. (1976). *Principles of psychological measurement*. Chicago, Ill: McNally.

Lewis, D., Mitzel, H., & Green, D. (1996, June). Standard setting: A bookmark approach. In D. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 453–461.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 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.

McClarty, K. L., Korbin, J. L., Moyer, E., Griffin, S., Huth, K., Carey, S., & Medberry, S. (2015). *PARCC benchmarking study*. Pearson Educational Measurement, Iowa City, IA: Pearson.

Mitzel, H., Lewis, D., Patz, R., & Green, D. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.

Muraki. E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement 16*(2), 159–176.

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

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments*. Synthesis Report.

Zenisky, A. L., Hambleton, R. K., & Sireci, S.C. (2003). *Effects of local dependence on the validity of IRT item test, and ability statistics*. (Technical Report). American College Admissions Test.

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.

(Appendices are numbered in reference to the chapter in the body of the text for which they expand or amplify the information reported.)

Appendix 6

A.6.1. Classical Item Analysis Statistics

Table A.6.1.1 ELAGP Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.53	0.52
Item 2	0.46	0.36
Item 3	0.38	0.56
Item 4	0.25	0.23
Item 5	0.24	0.42
Item 6	0.51	0.66
Item 7	0.47	0.93
Item 8	0.48	0.35
Item 9	0.62	0.58
Item 10	0.54	0.58
Item 11	0.45	0.47
Item 12	0.38	0.44
Item 13	0.73	0.39
Item 14	0.40	0.58
Item 15	0.65	0.56
Item 16	0.54	0.92
Item 17	0.38	0.37
Item 18	0.61	0.47
Item 19	0.75	0.56
Item 20	0.34	0.45

Table A.6.1.2 MATGP Item Analysis Statistics

Item	P-value	Polyserial
Item 1	0.74	0.59
Item 2	0.73	0.75
Item 3	0.15	0.78
Item 4	0.17	0.73
Item 5	0.38	0.77
Item 6	0.25	0.69
Item 7	0.50	0.82
Item 8	0.24	0.64
Item 9	0.30	0.66
Item 10	0.20	0.82
Item 11	0.43	0.52
Item 12	0.15	0.78
Item 13	0.49	0.77
Item 14	0.13	0.53
Item 15	0.13	0.77
Item 16	0.73	0.70
Item 17	0.34	0.84
Item 18	0.14	0.85
Item 19	0.69	0.72
Item 20	0.75	0.72
Item 21	0.53	0.80
Item 22	0.47	0.55
Item 23	0.68	0.45
Item 24	0.45	0.83
Item 25	0.44	0.79
Item 26	0.31	0.82
Item 27	0.35	0.68
Item 28	0.22	0.75
Item 29	0.52	0.52
Item 30	0.55	0.76

Appendix 7

A.7.1. IRT Threshold Scores and Scaling Constants

Table A.7.1.1 ELAGP IRT Parameters by Item

Item	Max Points	a	b	d1	d2	d3	d4	d5
Item 1	2	0.28	0.69	0.00	-2.98	2.98		
Item 2	2	0.27	0.66	0.00	-0.50	0.50		
Item 3	2	0.48	1.17	0.00	0.35	-0.35		
Item 4	2	0.13	4.03	0.00	-2.10	2.10		
Item 5	2	0.31	2.33	0.00	0.11	-0.11		
Item 6	2	0.55	0.56	0.00	-0.52	0.52		
Item 7	4	1.14	1.16	0.00	1.44	0.66	-0.58	-1.51
Item 8	3	1.19	0.62	0.00	0.90	0.13	-1.04	
Item 9	2	0.43	-0.33	0.00	-2.81	2.81		
Item 10	2	0.37	1.43	0.00	0.25	-0.25		
Item 11	2	0.22	0.17	0.00	0.72	-0.72		
Item 12	2	0.27	0.35	0.00	-4.69	4.69		
Item 13	2	0.16	1.83	0.00	-1.51	1.51		
Item 14	2	0.35	0.19	0.00	-2.38	2.38		
Item 15	2	0.24	1.81	0.00	-2.98	2.98		
Item 16	2	0.29	0.45	0.00	-0.95	0.95		
Item 17	4	0.93	1.24	0.00	1.41	0.56	-0.40	-1.57
Item 18	3	0.99	0.54	0.00	0.82	0.10	-0.92	
Item 19	2	0.18	0.62	0.00	-3.47	3.47		
Item 20	2	0.31	0.12	0.00	-4.92	4.92		
Item 21	2	0.46	0.43	0.00	-0.21	0.21		
Item 22	2	0.41	0.84	0.00	1.19	-1.19		
Item 23	2	0.34	1.60	0.00	0.94	-0.94		
Item 24	2	0.29	-1.18	0.00	0.53	-0.53		
Item 25	2	0.59	1.13	0.00	0.69	-0.69		
Item 26	2	0.39	0.10	0.00	-0.63	0.63		
Item 27	4	1.39	0.68	0.00	1.10	0.59	-0.30	-1.39
Item 28	3	1.60	0.20	0.00	0.70	0.10	-0.80	
Item 29	2	0.22	0.13	0.00	-1.82	1.82		

Item	Max Points	a	b	d1	d2	d3	d4	d5
Item 30	2	0.25	0.12	0.00	-2.76	2.76		
Item 31	2	0.29	1.22	0.00	-0.12	0.12		
Item 32	2	0.40	-0.09	0.00	-0.98	0.98		
Item 33	2	0.45	0.77	0.00	0.61	-0.61		
Item 34	2	0.58	0.12	0.00	0.08	-0.08		
Item 35	4	1.10	1.17	0.00	1.53	0.74	-0.65	-1.62
Item 36	3	1.18	0.59	0.00	0.92	0.18	-1.10	
Item 37	2	0.27	1.52	0.00	-0.10	0.10		
Item 38	2	0.47	-0.04	0.00	1.87	-1.87		
Item 39	2	0.50	-0.58	0.00	-0.56	0.56		
Item 40	2	0.26	1.65	0.00	-1.60	1.60		
Item 41	2	0.55	0.58	0.00	0.20	-0.20		
Item 42	2	0.29	1.22	0.00	-0.94	0.94		
Item 43	2	0.64	-1.08	0.00	-1.07	1.07		
Item 44	2	0.66	-0.06	0.00	-0.38	0.38		

Table A.7.1.2 MATGP IRT Parameters by Item

Item	Max Points	a	b	d1	d2	d3	d4	d5	d6	d7
Item 1	2	0.45	-1.29	0.00	1.25	-1.25				
Item 2	6	0.40	1.63	0.00	-1.74	3.06	-1.30	2.26	-2.67	0.39
Item 3	2	0.16	1.40	0.00	2.60	-2.60				
Item 4	2	0.52	-0.80	0.00	0.20	-0.20				
Item 5	6	0.56	2.52	0.00	1.16	1.41	-2.09	0.12	-0.83	0.23
Item 6	6	0.42	2.17	0.00	1.08	-0.11	0.28	0.09	-0.59	-0.76
Item 7	2	0.66	0.80	0.00	0.03	-0.03				
Item 8	2	0.71	-0.39	0.00	0.39	-0.39				
Item 9	3	0.50	1.68	0.00	0.35	0.12	-0.47			
Item 10	1	0.64	1.49							
Item 11	1	1.23	0.37							
Item 12	1	0.35	0.87							
Item 13	1	0.72	1.59							
Item 14	1	0.94	0.79							
Item 15	2	0.43	0.70	0.00	1.01	-1.01				
Item 16	1	0.97	0.05							
Item 17	1	0.71	1.35							
Item 18	1	1.18	1.65							
Item 19	2	0.33	0.93	0.00	1.07	-1.07				
Item 20	4	0.71	2.54	0.00	0.96	0.83	0.51	-2.30		
Item 21	1	0.68	0.70							
Item 22	1	0.36	-1.08							
Item 23	1	0.44	0.65							
Item 24	2	0.35	0.39	0.00	1.34	-1.34				
Item 25	1	0.56	2.78							
Item 26	3	0.61	2.12	0.00	-0.17	0.39	-0.23			
Item 27	1	0.98	2.39							
Item 28	1	1.05	-1.50							
Item 29	1	0.59	-1.03							

Item	Max Points	a	b	d1	d2	d3	d4	d5	d6	d7
Item 30	1	0.32	1.96							
Item 31	1	1.14	1.11							
Item 32	3	0.80	2.18	0.00	-0.70	0.88	-0.18			
Item 33	1	0.79	-0.38							
Item 34	1	0.64	-0.60							
Item 35	1	1.05	-0.62							
Item 36	3	1.03	2.50	0.00	0.61	-0.09	-0.52			
Item 37	1	0.82	-0.84							
Item 38	4	0.60	2.22	0.00	1.03	0.08	-0.86	-0.25		
Item 39	1	1.07	0.26							
Item 40	1	0.56	0.95							
Item 41	3	0.97	2.23	0.00	0.47	-0.28	-0.18			
Item 42	1	0.72	2.48							
Item 43	1	0.40	0.65							
Item 44	1	0.34	-0.95							
Item 45	1	0.84	0.60							
Item 46	1	1.09	0.48							
Item 47	2	0.35	1.71	0.00	1.53	-1.53				
Item 48	1	1.00	1.57							
Item 49	1	1.02	-0.95							
Item 50	1	0.84	0.58							
Item 51	2	1.09	1.26	0.00	0.52	-0.52				
Item 52	2	0.59	1.18	0.00	0.56	-0.56				
Item 53	1	0.45	0.24							
Item 54	3	0.54	1.69	0.00	-1.07	1.84	-0.77			
Item 55	1	0.82	0.06							
Item 56	1	0.37	0.37							
Item 57	1	0.31	0.54							
Item 58	1	0.73	0.02							