New York State Regents Examination in Global History and Geography II

2022 Technical Report

Prepared for the New York State Education Department
by Pearson

June 2024

Copyright

Secure Materials.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. INTRODUCTION

This technical report for the Regents Examination in Global History and Geography II, based on the New York State (NYS) Learning Standards, provides New York State with documentation of the purposes of the Regents Examination, scoring information, evidence of both the reliability and validity of the exams, scaling information, and guidelines for score reporting for the June 2022 administration. As the *Standards for Education and Psychological Testing* discusses in Standard 7, "The objective of the documentation is to provide test users with the information needed to help them assess the nature and quality of the test, the resulting scores, and the interpretations based on the test scores" (AERA et al., 2014, p. 123).[1] **Please note:** A technical report, by design, addresses the technical documentation of a testing program. Other aspects of a testing program (e.g., content standards, scoring guides, the guide to test interpretation) are thoroughly addressed and referenced in supporting documents. All analyses in this report were conducted using the operational items only.

Typically, the Regents Examinations are administered each year in August, January, and June to students enrolled in New York State schools. However, testing in 2020 and 2021 was disrupted by the COVID-19 pandemic, starting with the cancellation of the June 2020 administration. In 2021, the New York State Education Department (NYSED) cancelled all but four of the Regents Examinations scheduled to be administered in June 2021, and all the Regents Examinations scheduled to be administered in August 2021. The four examinations administered in June 2021 were Algebra I, Earth Science, English Language Arts and Living Environment, as required to comply with the federal Every Student Succeeds Act (ESSA). NYSED also cancelled the January 2022 administration of the Regents Examination Program in response to the ongoing impact of COVID-19, as described online at https://www.nysed.gov/news/2022/january-2022-regents-examinations-cancelled-due-ongoing-pandemic. Regular testing activities resumed in June 2022.

1.2. PURPOSES OF THE EXAM

The Regents Examination in Global History and Geography II measures student achievement against the NYS Learning Standards. The exam is prepared by teacher examination committees and New York State Education Department (NYSED) subject matter and testing specialists. Further, it provides teachers and students with important information about student learning and performance against the established curriculum standards. Results of this exam may be used to identify student strengths and needs to guide classroom teaching and learning. The exam also provides students, parents, counselors, administrators, and college admissions officers with objective and easily understood achievement information that may be used to inform empirically based educational and vocational decisions about students.

As a state-provided objective benchmark, the Regents Examination in Global History and Geography II is intended for use in satisfying state testing requirements for students who have finished a course in Global History and Geography II. A passing score on the exam counts toward requirements for a high school diploma, as described in the New York State diploma

[1] References to specific *Standards* will be placed in parentheses throughout the technical report to provide further context for each section.

requirements.[2] Results of the Regents Examinations may also be used to satisfy various locally established requirements throughout the state.

1.3. TARGET POPULATION (STANDARD 7.2)

The student population for the Regents Examination in Global History and Geography II is composed of students who have completed a course in Global History and Geography II. Table 1.1 presents a demographic breakdown of all students who took the June 2022 Regents Examination in Global History and Geography II. All analyses in this report are based on the population described in this table. Annual Regents Examination results reported in the New York State Report Cards are those reported in the Student Information Repository System (SIRS) as of the reporting deadline. The results include the exams administered in June 2022 (see https://data.nysed.gov/). If a student takes the same exam multiple times in the year, only the highest score is included in these results. However, as previously mentioned, this exam was only administered in June 2022 during the 2021–2022 school year. Item-level data used for the analyses in this report are reported by districts on a similar timeline but through a different collection system. These data include all student results for each administration. Therefore, the n-counts in this technical report will differ from publicly reported counts of student test takers.

Table 1.1. Total Student Population in June 2022: Regents Examination in Global History and Geography II

Demographics	N	%
All Students	187,860	100.00
Race/Ethnicity		
American Indian/Alaska Native	1,193	0.64
Asian/Native Hawaiian/Other Pacific Islander	19,396	10.33
Black/African American	28,845	15.36
Hispanic/Latino	48,849	26.00
Multiracial	5,118	2.72
White	84,444	44.95
English Language Learner		
No	176,031	93.70
Yes	11,829	6.30
Economically Disadvantaged		
No	97,224	51.75
Yes	90,636	48.25
Gender		
Female	93,026	49.52
Male	94,770	50.45
Nonbinary	49	0.03
Student with a Disability		
No	159,715	85.02
Yes	28,145	14.98

Note. 15 students were not reported in the race/ethnicity and gender groups, but they are reflected in "All Students."

[2] The New York State diploma requirements are located online at https://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf.

Chapter 2: Classical Item Statistics (Standard 4.10)

This chapter provides an overview of the two most familiar item-level statistics obtained from classical item analysis: item difficulty and item discrimination. The following results pertain to the operational Regents Examination in Global History and Geography II items.

2.1. ITEM DIFFICULTY

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

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

The individual item scores (x_i) are summed and then divided by the total number of students (n). For multiple-choice (MC) items, student scores are represented by 0s and 1s (0 = wrong answer, 1 = correct answer). With 0–1 scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. Therefore, this is also the proportion correct for the item, or the *p*-value. In theory, *p*-values can range from 0.00 to 1.00 on the proportion-correct scale. For example, if an MC item has a *p*-value of 0.89, it means 89% of the students answered the item correctly. This value might also suggest that the item was relatively easy and/or that the students who attempted the item were relatively high achievers. For constructed-response (CR) items, mean scores can range from the minimum possible score (usually 0) to the maximum possible score (e.g., 6 points for some mathematics items). To facilitate average score comparability across MC and CR items, mean item performance for CR items is divided by the maximum score possible so that the *p*-values for all items are reported as a ratio from 0.0 to 1.0.

Although the *p*-value statistic does not consider individual student ability in its computation, it provides a useful view of overall item difficulty and can provide an early and simple indication of items that are too difficult for the population of students taking the exam. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory (IRT) analyses that factor student ability into estimates of item difficulty. Such items may be removed from the item pool during the test development process, as field testing typically reveals that they add very little measurement information.

Items for the June 2022 Regents Examination in Global History and Geography II show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values, presented in Table 2.1 and Table 2.2 for MC and CR items, respectively, range from 0.38 to 0.93, with a mean of 0.70. The tables also show a standard deviation (SD) of item score and item mean (Table 2.2 only).

2.2. ITEM DISCRIMINATION

At the most general level, estimates of item discrimination indicate an item's ability to differentiate between high and low performance on an exam. It is expected that students who perform well on the exam would be more likely to answer any given item correctly, while low-performing students (i.e., those who perform poorly on the exam overall) would be more likely to answer the same item incorrectly. Pearson's product-moment correlation coefficient (also commonly referred to as a point-biserial correlation) between item scores and test scores is used to indicate discrimination (Pearson, 1896). The correlation coefficient can range from −1.0 to +1.0. If high-scoring students

tend to get the item correct while low-scoring students do not, the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., above 0), indicating that the item is likely discriminating well between high- and low-performing students. Point-biserial correlations are computed for each answer option, including correct and incorrect options (commonly referred to as distractors). Point-biserial values for each distractor are typically negative and an important part of the analysis. Positive values can indicate that higher-performing students are selecting an incorrect answer or that the item key for the correct answer should be checked.

Table 2.1 and Table 2.2 present the point-biserial values on the correct response and three distractors (Table 2.1 only) for the June 2022 Regents Examination in Global History and Geography II. The values for correct answers are 0.19 or higher for all items, indicating that the items are generally discriminating well between high- and low-performing students. Point-biserial values for all distractors are negative, indicating that students are generally responding to the items as expected during item and rubric development.

Table 2.1. MC Item Analysis Summary: Regents Examination in Global History and Geography II

Item	#Students	*p*-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
01	187,860	0.84	0.37	0.45	-0.33	-0.24	-0.14
02	187,860	0.78	0.41	0.43	-0.26	-0.11	-0.29
03	187,860	0.48	0.50	0.49	-0.18	-0.22	-0.25
04	187,860	0.82	0.39	0.42	-0.25	-0.19	-0.23
05	187,860	0.62	0.48	0.38	-0.14	-0.22	-0.22
06	187,860	0.91	0.29	0.19	-0.12	-0.08	-0.10
07	187,860	0.86	0.34	0.43	-0.19	-0.20	-0.31
08	187,860	0.42	0.49	0.33	-0.17	-0.14	-0.16
09	187,860	0.55	0.50	0.39	-0.17	-0.27	-0.19
10	187,860	0.61	0.49	0.41	-0.15	-0.21	-0.26
11	187,860	0.83	0.37	0.32	-0.19	-0.16	-0.18
12	187,860	0.78	0.41	0.38	-0.26	-0.14	-0.18
13	187,860	0.57	0.49	0.35	-0.15	-0.18	-0.18
14	187,860	0.50	0.50	0.44	-0.26	-0.16	-0.18
15	187,860	0.64	0.48	0.39	-0.27	-0.18	-0.15
16	187,860	0.77	0.42	0.46	-0.25	-0.26	-0.24
17	187,860	0.50	0.50	0.25	-0.02	-0.06	-0.24
18	187,860	0.40	0.49	0.32	-0.01	-0.09	-0.28
19	187,860	0.66	0.47	0.56	-0.23	-0.35	-0.25
20	187,860	0.66	0.47	0.45	-0.25	-0.21	-0.24
21	187,860	0.80	0.40	0.42	-0.17	-0.24	-0.25
22	187,860	0.58	0.49	0.47	-0.28	-0.27	-0.16
23	187,860	0.78	0.41	0.38	-0.29	-0.16	-0.21
24	187,860	0.81	0.39	0.39	-0.22	-0.21	-0.23
25	187,860	0.87	0.33	0.45	-0.24	-0.27	-0.23
26	187,860	0.78	0.41	0.47	-0.17	-0.28	-0.28
27	187,860	0.38	0.49	0.41	-0.11	-0.26	-0.10
28	187,860	0.59	0.49	0.52	-0.16	-0.29	-0.28

Table 2.2. CR Item Analysis Summary: Regents Examination in Global History and Geography II

Item	#Students	Min. Score	Max. Score	Mean	SD	*p*-Value	Point-Biserial
29	187,860	0	1	0.76	0.43	0.76	0.49
30	187,860	0	1	0.89	0.32	0.89	0.39
31	187,860	0	1	0.80	0.40	0.80	0.51
32	187,860	0	1	0.77	0.42	0.77	0.54
33	187,860	0	1	0.93	0.26	0.93	0.36
34	187,860	0	1	0.87	0.33	0.87	0.48
35	187,860	0	1	0.75	0.43	0.75	0.54
36	187,860	0	5	2.63	1.06	0.53	0.77

2.3. DISCRIMINATION ON DIFFICULTY SCATTERPLOT

Figure 2.1 presents a scatterplot of the item discrimination values (*y*-axis) and item difficulty values (*x*-axis). Table 2.3 presents the descriptive statistics of *p*-value and point-biserial values, including mean, minimum, Q1, median, Q3, and maximum.

Figure 2.1. Scatterplot: Regents Examination in Global History and Geography II

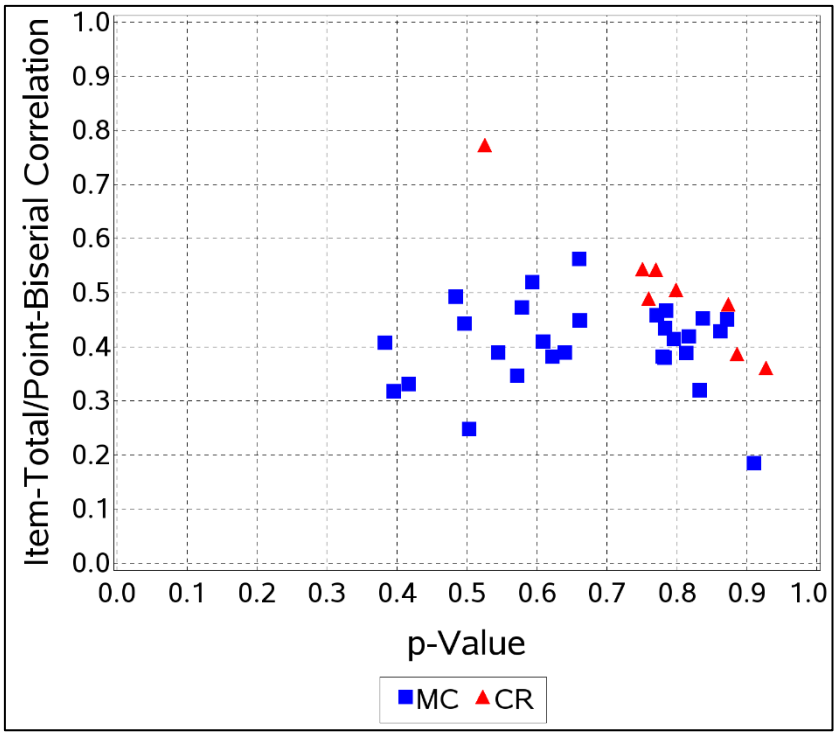


Table 2.3. Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Global History and Geography II

Statistic	N	Mean	Min.	Q1	Median	Q3	Max.
p-value	36	0.70	0.38	0.58	0.77	0.82	0.93
Point-Biserial	36	0.43	0.19	0.38	0.43	0.48	0.77

2.4. OBSERVATIONS AND INTERPRETATIONS

The *p*-values for the MC items ranged from about 0.38 to 0.91, while the *p*-values for the CR items ranged from about 0.53 to 0.93. The difficulty distributions illustrated in the plot indicate that a wide range of item difficulties appeared on each exam, which was one test development goal.

Chapter 3: IRT Calibrations, Equating, and Scaling (Standards 2 and 4.10)

The item response theory (IRT) model used for the Regents Examinations is based on the work of Georg Rasch (Rasch, 1960). The Rasch model has a long-standing presence in applied testing programs. IRT has several advantages over classical test theory and has become the standard procedure for analyzing item response data in large-scale assessments. According to van der Linden and Hambleton (1997), "The central feature of IRT is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying ability." Ability, in this sense, can be thought of as performance on the test and is defined as "the expected value of observed performance on the test of interest" (Hambleton et al., 1991). This performance value is often referred to as θ. Performance and θ will be used interchangeably throughout the remainder of this report.

A fundamental advantage of IRT is that it links student performance and item difficulty estimates and places them on the same scale, allowing for an evaluation of student performance that considers the difficulty of the test. This is particularly valuable for final test construction and test form equating, as it facilitates a fundamental attention to fairness for all students across items and test forms.

This chapter outlines the procedures used for calibrating the operational items, including an introduction to the Rasch model, the results from evaluations of the adequacy of the Rasch assumptions, and the Rasch item statistics. Generally, item calibration is the process of assigning a difficulty, or item "location," estimate to each item on an assessment so that all items are placed onto a common scale.

3.1. DESCRIPTION OF THE RASCH MODEL

The Rasch model (Rasch, 1960) was used to calibrate the MC items, and the partial credit model (PCM; Wright & Masters, 1982) was used to calibrate the CR items. The PCM extends the Rasch model for dichotomous (0, 1) items so that it accommodates the polytomous CR item data. Under the PCM model, for a given item *i* with m_i score categories, the probability of person *n* scoring *x* (x = 0, 1, 2, ... m_i) is given as follows:

$$P_{ni}(X = x) = \frac{\exp \sum_{j=0}^{x} (\theta_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\theta_n - D_{ij})}$$

where θ_n represents student ability, and D_{ij} is the step difficulty of the j^{th} step on item *i*. D_{ij} can be expressed as $D_{ij} = D_i - F_{ij}$, where D_i is the difficulty for item *i* and F_{ij} is a step deviation value for the j^{th} step. For dichotomous MC items, the PCM reduces to the standard Rasch model, and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person *n* getting item *i* correct, as follows:

$$P_{ni}(X = 1) = \frac{\exp(\theta_n - D_i)}{1 + \exp(\theta_n - D_i)}$$

The Rasch model places both performance and item difficulty (estimated in terms of log-odds or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of student performance and item difficulty that are theoretically invariant across random samples of the same student population.

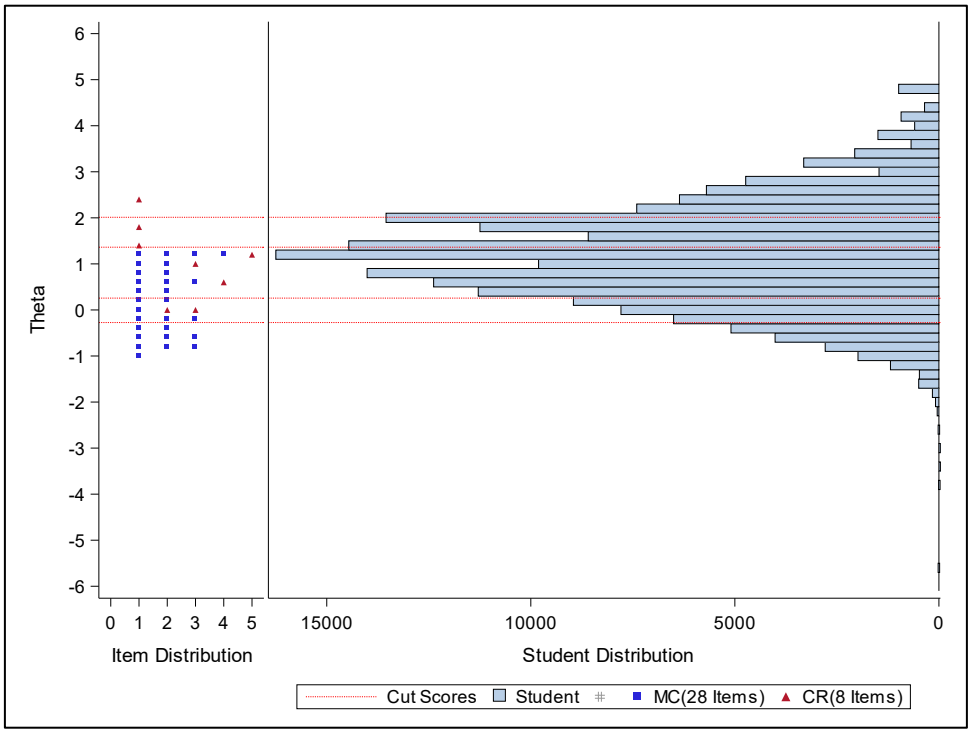
3.2. SOFTWARE AND ESTIMATION ALGORITHM

Item calibration was implemented via the WINSTEPS 3.60 computer program (Linacre, 2005), which employs unconditional (UCON) joint maximum likelihood estimation (JMLE).

3.3. ITEM DIFFICULTY–STUDENT PERFORMANCE MAP

The distributions of the Rasch item logits (item difficulty estimates) and student performance are shown on the item difficulty–student performance map presented in Figure 3.1. This graphic illustrates the location of student performance and item difficulty on the same scale, along with their respective distributions and cut scores (indicated by the horizontal dotted lines). The figure shows more difficult items and higher student performance at the top and lower performance and easier items at the bottom.

Figure 3.1. Student Performance Map: Regents Examination in Global History and Geography II



3.4. CHECKING RASCH ASSUMPTIONS

Because the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the Regents Examination in Global History and Geography II, the validity of the inferences from these results depends on the degree to which the assumptions of the model were met and how well the model fits the test data. Therefore, it is important to check these assumptions. This section evaluates the dimensionality of the data, local item independence, and item fit. Only operational items were analyzed as they are the basis of student scores.

3.4.1. Unidimensionality

Rasch models assume that one dominant dimension determines the differences in students' performances. Principal components analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant components exist among the items. If any other dimensions are found, the unidimensionality assumption would be violated.

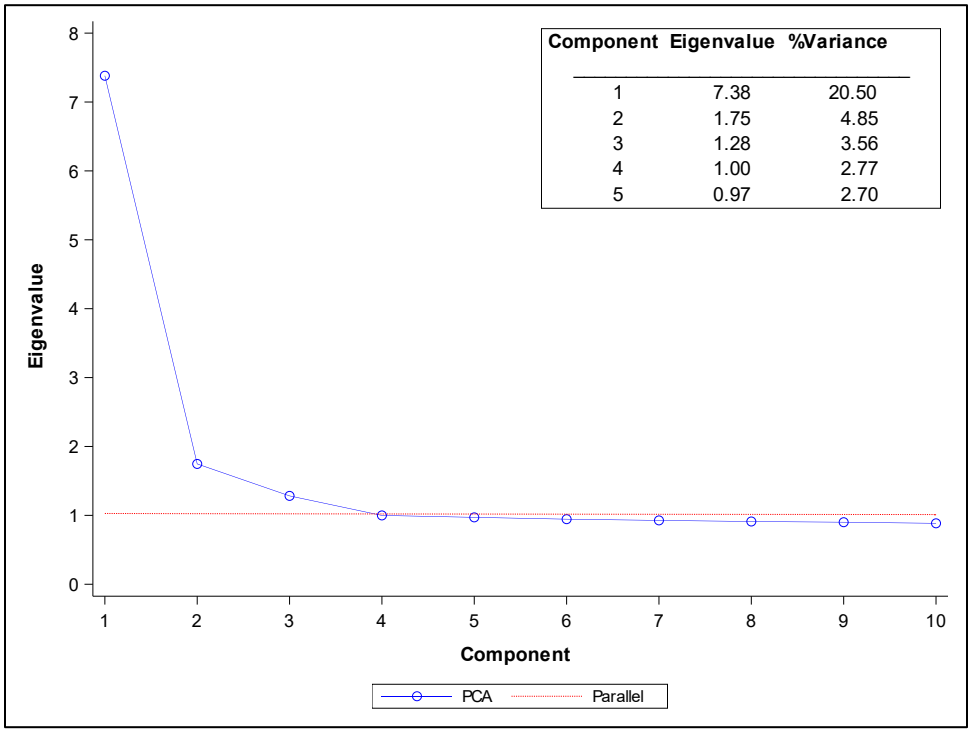
A parallel analysis (Horn, 1965) was conducted to help distinguish components that are real from components that are random. Parallel analysis is a technique used to determine how many factors exist in principal components. For the parallel analysis, 100 random datasets of sizes equal to the original data were created. For each random dataset, a PCA was performed and the resulting eigenvalues stored. For each component, the upper 95th percentile value of the distribution of the 100 eigenvalues from the random data sets was then plotted. Given the size of the data generated for the parallel analysis, the reference line is essentially equivalent to plotting a reference line for an eigenvalue of 1.

Figure 3.2 presents the PCA and parallel analysis results for the Regents Examination in Global History and Geography II. The results include the eigenvalues and the percentage of variance explained for the first five components, as well as the scree plots. The scree plots show the eigenvalues plotted by component number and the results of a parallel analysis. Although the total number of components in the PCA is the same as the total number of items in a test, Figure 3.2 shows only the first 10 components. This view is sufficient for interpretation because components are listed in descending eigenvalue order. The fact that the eigenvalues for components 2–10 are much lower than the first component demonstrates that there is one dominant component, showing evidence of unidimensionality.

Reckase (1979) proposed that the variance explained by the primary dimension should be greater than 20% to indicate unidimensionality. However, as this rule is not absolute, it is helpful to consider three additional characteristics of the PCA and parallel analysis results: (1) whether the ratio of the first to the second eigenvalue is greater than 3, (2) whether the second value is not much larger than the third value, and (3) whether the second value is not significantly different from those from the parallel analysis.

As shown in Figure 3.2, the primary dimension explained 20.50% of the total variance for the Regents Examination in Global History and Geography II. The eigenvalue of the second dimension is less than one-third of the first at 1.75, and the second value is not significantly different from the parallel analysis. Overall, the PCA suggests that the test is reasonably unidimensional.

Figure 3.2. Scree Plot: Regents Examination in Global History and Geography II



3.4.2. Local Independence

Local independence is a fundamental assumption of IRT. This means that, for statistical purposes, a student's response to any one item should not depend on the student's response to any other item on the test. In formal statistical terms, Test *X*, which comprises items X_1, X_2, …X_n, is locally independent with respect to the latent variable θ if, for all $x = (x_1, x_2, \ldots x_n)$ and θ:

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{I} P(X_i = x_i \mid \theta)$$

This formula essentially states that the probability of any pattern of responses across all items (**x**), after conditioning on the student's true score (θ) as measured by the test, should be equal to the product of the conditional probabilities across each item (i.e., the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of local independence. A weak form of local independence (WLI) is proposed by McDonald (1979). The distinction is important because many indicators of local dependency are framed by WLI. For WLI, the conditional covariances of all pairs of item responses, conditioned on the abilities, are assumed to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on abilities, is the product of the probabilities of responses to these two items, as shown below. Based on the WLI, the following expression can be derived:

$$P(X_i = x_i, X_j = x_j \mid \theta) = P(X_i = x_i \mid \theta) P(X_j = x_j \mid \theta)$$

Marais and Andrich (2008) point out that local item dependence in the Rasch model can occur in two ways that may be difficult to distinguish. The first way occurs when the assumption of unidimensionality is violated. Here, other nuisance dimensions besides a dominant dimension determine student performance (this can be called "trait dependence"). The second way occurs when responses to an item depend on responses to another item. This is a violation of statistical independence and can be called "response dependence." By distinguishing the two sources of local dependence, one can see that, while local independence can be related to unidimensionality, the two are different assumptions and therefore require different tests.

Residual item correlations, provided in WINSTEPS for each item pair, were used to assess the local dependence between the Regents Examination in Global History and Geography II items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using θ and item parameter estimates. Next, deviations (residuals) between the students' expected and observed performance are determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. The raw score residual correlation essentially corresponds to Yen's Q3 index, a popular statistic used to assess local independence. The expected value for the Q3 statistic is approximately -1/(k − 1) when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected Q3 values should be approximately -0.03 for the items on the exam. Index values greater than 0.20 indicate a degree of local dependence that should be examined by test developers (Chen & Thissen, 1997).

Because the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Table 3.1 presents the summary statistics (i.e., the mean, standard deviation, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, P_{90})) for all the residual correlations. The table also presents the total number of item pairs (N) and the number of pairs with residual correlations greater than 0.20. There are no item pairs with residual correlations greater than 0.20. The mean residual correlations are slightly negative, and the values are close to -0.02. Most of the correlations are very small, with one pair exceeding 0.20 with a value of 0.28, suggesting that local item independence generally holds for the Regents Examination in Global History and Geography II.

Table 3.1. Summary of Item Residual Correlations: Regents Examination in Global History and Geography II

Statistic	Value
N	630
Mean	-0.02
SD	0.04
Minimum	-0.20
P_{10}	-0.06
P_{25}	-0.04
P_{50}	-0.02
P_{75}	-0.01
P_{90}	0.01
Maximum	0.28
>|0.20|	1.00

3.4.3. Item Fit

An important assumption of the Rasch model is that the data for each item fit the model. WINSTEPS provides two item fit statistics (INFIT and OUTFIT) for evaluating the degree to which the Rasch model predicts the observed item responses for a given set of test items. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (Zstd with mean = 0 and variance = 1). MnSq values are more oriented toward practical significance, while Zstd values are more oriented toward statistical significance. INFIT MnSq values are the average of standardized residual variance (the difference between the observed score and the Rasch-estimated score divided by the square root of the Rasch-model variance). The INFIT statistic is weighted by the θ relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., informative, on-target responses).

The expected MnSq value is 1.0 and can range from 0.0 to infinity. Deviation in excess of the expected value can be interpreted as noise or lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (too unpredictable, too much noise). Rules of thumb regarding "practically significant" MnSq values vary.

Table 3.2 presents the summary statistics of INFIT mean square statistics for the Regents Examination in Global History and Geography II, including the mean, standard deviation, and minimum and maximum values. The table also presents the number of items within a targeted range of [0.7, 1.3]. The mean INFIT value is close to 0.99, with all but one of the 36 items falling in the targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as a guide for ideal fit, fit values outside of the range are considered individually. Overall, these results indicate that, for most items, the Rasch model fits the Regents Examination in Global History and Geography II item data well.

Table 3.2. Summary of INFIT Mean Square Statistics: Regents Examination in Global History and Geography II

N	Mean	SD	Min.	Max.	[0.7, 1.3]
36	0.99	0.11	0.85	1.41	[35/36]

3.5. SCALING OF OPERATIONAL TEST FORMS

Operational test items were selected based on content coverage, content accuracy, and statistical quality. The sets of items on each operational test conformed to the coverage determined by content experts working from the learning standards established by NYSED and explicated in the test blueprint. Each item's classical and Rasch statistics were used to assess item quality. Items were selected to vary in difficulty to accurately measure students' abilities across the ability continuum. Appendix A: contains the operational test maps for the June 2022 administration. The statistics presented in the test maps were generated based on the field test data.

All Regents Examinations are pre-equated, meaning that the parameters used to derive the relationship between the raw and scale scores are estimated prior to the construction and administration of the operational form. These field tests are administered to as small a sample of students as possible to minimize the effect on student instructional time throughout the state. The small n-counts associated with such administrations are sufficient for reasonably accurate estimation of most items' parameters. However, the parameters for the 10-point essay item can be unstable when estimated across as small a sample as is typically used. Therefore, a set of constants is used for these items' parameters on operational examinations. These constants were set by NYSED and are based on the values in the bank for all essay items. The Regents Examination in Global History and Geography II only has one 10-point item with fixed constants as follows: $D = 1.22$, $F_0 = 0.00$, $F_1 = -0.63$, $F_2 = -3.74$, $F_3 = -0.97$, $F_4 = -2.07$, $F_5 = 0.07$, $F_6 = -0.40$, $F_7 = 1.56$, $F_8 = 1.02$, $F_9 = 2.84$, and $F_{10} = 2.32$.

The Regents Examination in Global History and Geography II has four cut scores, which are set at the scale scores of 55, 65, 79 (floating), and 85. The third cut point at 85 was determined at the most recent Global History and Geography II standard setting session during which a raw score associated with assessment performance at this performance level was determined. The scale score associated with this raw score was then determined and used for the performance level cut moving forward.

A primary consideration during test construction was to select items that would minimize changes in the raw scores corresponding to these scale scores. Maintaining a consistent mean Rasch difficulty level from administration to administration facilitates this. For this assessment, the target value for the mean Rasch difficulty was set at -0.285. The raw scores corresponding to the scale score cut scores may still fluctuate, even if the mean Rasch difficulty level is maintained at the target value, due to differences in the distributions of the Rasch difficulty values between the items from administration to administration.

The relationship between raw and scale scores is explicated in the scoring table for the June 2022 administration, presented in Appendix B. This table is the end product of the following scaling procedure.

All Regents Examinations are equated back to a base scale held constant from year to year. Specifically, they are equated to the base scale using a calibrated item pool. The Rasch difficulties from the items' initial administration in a previous year's field test are used to equate the scale for the current administration to the base administration. For this exam, the base administration was the June 2019 administration. Scale scores from the June 2022 administration are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2019 administration.

When the base administration was concluded, the initial raw-score-to-scale-score relationship was established. Three raw scores were fixed at specific scale scores. Scale scores of 0 and 100 were fixed to correspond to the minimum and maximum possible raw scores. A standard setting had been held to determine the "passing" and "passing with distinction" cut scores in the raw score metric. The scale score points of 55, 65, 79, and 85 were set to correspond to those raw score cuts. A fourth-degree polynomial is required to fit a line exactly to six arbitrary points (e.g., the raw scores corresponding to the six critical scale scores of 0, 55, 65, 79, 85, and 100). The general form of this best-fitting line is as follows:

$$SS = m4 * RS^4 + m3 * RS^3 + m2 * RS^2 + m1 * RS^1 + m0$$

where *SS* is the scale score, *RS* is the raw score, and *m0* through *m4* are the transformation constants that convert the raw score into the scale score (*m0* will always be equal to 0 in this application, as a raw score of 0 corresponds to a scale score of 0). A subscript for a person on both dependent and independent variables is not present for simplicity. The above relationship and the values of *m1* to *m4* specific to this subject were then used to determine the scale scores corresponding to the remainder of the raw scores on the exam. This initial relationship between the raw and scale scores became the base scale.

The Rasch difficulty parameters for the items on the base form were then used to derive a raw score to Rasch student ability (theta score) relationship. This allowed the relationship between the Rasch theta score and the scale score to be known, mediated through their common relationship with the raw scores. In succeeding years, each test form was selected from the pool of items that had been tested in previous years' field tests, each of which had known Rasch item difficulty parameter(s). These known parameters were then used to construct the relationship between the raw and Rasch theta scores for that particular form. Because the Rasch difficulty parameters are all on a common scale, the Rasch theta scores were also on a common scale with previously administered forms. The remaining step in the scaling process was to find the scale score equivalent for the Rasch theta score corresponding to each raw score point on the new form using the theta-to-scale score relationship established in the base year. This was done via linear interpolation.

This process results in a relationship between the raw scores on the form and the overall scale scores. The scale scores corresponding to each raw score are then rounded to the nearest integer for reporting on the conversion chart (posted at the close of each administration). The only exceptions are for the minimum and maximum raw scores and the raw scores that correspond to the scaled cut scores of 55, 65, 79, and 85. The minimum (0) and maximum possible raw scores are assigned scale scores of 0 and 100, respectively. If there are raw scores less than the maximum with scale scores that round to 100, their scale scores are set equal to 99. A similar process is followed with the minimum score; if any raw scores other than 0 have scale scores that round to 0, their scale scores are instead set equal to 1.

With regard to the cuts, if two or more scale scores round to 55, 65, or 85, the lowest raw score's scale score is set equal to 55, 65, or 85, and the scale scores corresponding to the higher raw scores are set to 56, 66, or 86 as appropriate. This rule does not apply for the third floating cut score. If no scale score rounds to these critical cuts, the raw score with the largest scale score that is less than the cut is set equal to the cut. The overarching principle, when two raw scores both round to either scale score cut, is that the lower of the raw scores is always assigned to be equal to the cut so that students are never penalized for this ambiguity.

Chapter 4: Reliability (Standard 2)

Test reliability is a measure of the internal consistency of a test (Cronbach, 1951), or a measure of the extent to which the items on a test provide consistent information about student mastery of a domain. Reliability should ultimately demonstrate that student score estimates maximize consistency and therefore minimize error or, theoretically speaking, that students who take a test multiple times would get the same score each time.

According to the *Standards for Educational and Psychological Testing*, "A number of factors can have significant effects on reliability/precision, and in some cases, these factors can lead to misinterpretations of test scores, if not taken into account" (AERA et al., 2014, p. 38). First, test length and the variability of observed scores can influence reliability estimates. Tests with fewer items or with a lack of heterogeneity in scores tend to produce lower reliability estimates. Second, reliability is concerned with random sources of error. Accordingly, the degree of inconsistency due to random error sources is what determines reliability: less consistency is associated with lower reliability, and more consistency is associated with higher reliability. Systematic error sources also exist.

This chapter discusses reliability results for the Regents Examination in Global History and Geography II and three additional statistical measures to address the multiple factors affecting an interpretation of the exam's reliability: standard errors of measurement (SEMs), decision consistency, and group means.

4.1. RELIABILITY INDICES (STANDARD 2.20)

Classical test theory describes reliability as a measure of the internal consistency of test scores. The reliability (ρ_X^2) is defined as the ratio of true score variance (σ_T^2) to the observed score variance (σ_X^2), as presented in the equation below. The total variance contains two components: (1) the variance in true scores and (2) the variance due to the imperfections in the measurement process (σ_E^2). In other words, total variance equals true score variance plus error variance.[3]

$$\rho_X^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

Reliability coefficients indicate the degree to which differences in test scores reflect true differences in the attribute being tested rather than random fluctuations. Total test score variance (i.e., individual differences) is partly due to real differences in the construct (true variance) and partly due to random error in the measurement process (error variance).

Reliability coefficients range from 0.0 to 1.0. The index is 0.0 if none of the test score variances are true. Such scores would be pure random noise (i.e., all measurement error). If all test score variances were true, the index would equal 1.0. If the index achieved a value of 1.0, scores would be perfectly consistent (i.e., contain no measurement error). Although values of 1.0 are never achieved in practice, larger coefficients are more desirable because they indicate that the test scores are less influenced by random error.

[3] A covariance term is not required, as true scores and error are assumed to be uncorrelated in classical test theory.

Reliability is most often estimated using the formula for coefficient alpha, which provides a practical internal consistency index. Coefficient alpha can be conceptualized as the extent to which an exchangeable set of items from the same domain would result in a similar rank ordering of students. Relative error is reflected in this index. Excessive variation in student performance from one sample of items to the next should be of particular concern for any achievement test user. A general computational formula for coefficient alpha is as follows:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma_{Yi}^2}{\sigma_X^2}\right)$$

where N is the number of parts (items), σ_X^2 is the variance of the observed total test scores, and σ_{Yi}^2 is the variance of part *i*.

4.2. STANDARD ERROR OF MEASUREMENT (STANDARDS 2.13, 2.14, 2.15)

Reliability coefficients best reflect the extent to which measurement inconsistencies may be present or absent. The standard error of measurement (SEM) is another indicator of test score precision that is better suited for determining the effect of measurement inconsistencies for the scores obtained by individual examinees. This is particularly so for conditional SEMs (CSEMs).

4.2.1. Traditional Standard Error of Measurement

The SEM is defined as the standard deviation of the distribution of observed scores for students with identical true scores, as shown below. The SEM is an index of the random variability in test scores in test score units and therefore represents important information for test score users.

$$SEM = SD\sqrt{1 - \alpha}$$

This formula indicates that the value of the SEM depends on both the reliability coefficient (the coefficient alpha) and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value), the SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value), the SEM would be 0.0. Therefore, a perfectly reliable test has no measurement error (Harvill, 1991). The value of the SEM also takes the group variation (i.e., score standard deviation) into account. For example, an SEM of 3 on a 10-point test would be very different from an SEM of 3 on a 100-point test.

4.2.2. Traditional Standard Error of Measurement Confidence Intervals

The SEM is an index of the random variability in test scores reported in actual score units, which is why it has such great utility for test score users. SEMs allow statements regarding the precision of individual test scores and help place “reasonable limits” (Gulliksen, 1950) around observed scores through construction of an approximate score band. Often referred to as confidence intervals, these bands are constructed by taking the observed scores, X, and adding and subtracting a multiplicative factor of the SEM. For example, students with a given true score will have observed scores that fall between ±1 SEM about two-thirds of the time.[4] For ±2 SEM confidence intervals, this increases to about 95%.

[4] Some prefer the following interpretation: If a student were tested an infinite number of times, the ±1 SEM confidence intervals constructed for each score would capture the student’s true score 68% of the time.

Table 4.1 presents the coefficient alpha and associated SEM for the Regents Examination in Global History and Geography II.

Table 4.1. Coefficient Alpha and SEM: Regents Examination in Global History and Geography II

Coefficient Alpha	SEM
0.88	3.05

Assuming normally distributed scores, one would expect about two-thirds of the observations to be within one standard deviation of the mean. An estimate of the standard deviation of the true scores can be computed as follows:

$$\hat{\sigma}_T = \sqrt{\hat{\sigma}_x^2 - \hat{\sigma}_x^2(1-\hat{\rho}_{xx})}$$

4.2.3. Conditional Standard Error of Measurement

Every time an assessment is administered, a student's score contains some error. If the same exam were administered an infinite number of times to the same student, the mean of the distribution of the student's raw scores would be equal to the student's true score (θ) (i.e., the score obtained with no error), and the standard deviation of the distribution of the student's raw scores would be the conditional standard error. Because there is a one-to-one correspondence between the raw score and θ in the Rasch model, this concept can be applied more generally to all students who obtained a particular raw score, and the probability of obtaining each possible raw score can be calculated given the students' estimated θ. The standard deviation of this conditional distribution is defined as the CSEM. The computer program POLYCSEM (Kolen, 2004) was used to carry out the mechanics of this computation.

The relationship between θ and the scale score is not expressible in a simple mathematical form because it is a blend of the fourth-degree polynomial relationship between the raw and scale scores and the nonlinear relationship between the expected raw and θ scores. In addition, as the exam is equated from year to year, the relationship between the raw and scale scores moves away from the original fourth-degree polynomial relationship to one that is also no longer expressible in simple mathematical form. In the absence of a simple mathematical relationship between θ and the scale scores, the CSEMs available for each θ score via Rasch IRT cannot be converted directly to the scale score metric.

The use of Rasch IRT to scale and equate the Regents Examinations does, however, make it possible to calculate CSEMs by using the procedures described by Kolen et al. (1996) for dichotomously scored items and extended by Wang et al. (2000) to polytomously scored items. For tests such as the Regents Examinations that do not have a one-to-one relationship between raw (θ) and scale scores, the CSEM for each achievable scale score can be calculated by using the compound multinomial distribution to represent the conditional distribution of raw scores for each level of θ.

Consider a student with a certain performance level. If it were possible to measure this student's performance perfectly, without any error, this measure could be called the student's true score, which is equal to the expected raw score. However, whenever a student takes a test, the observed test score always includes some level of measurement error. Sometimes this error is positive, and the student achieves a higher score than would be expected given the student's level of θ. Other times it is negative, and the student achieves a lower-than-expected score. If a student could be given the same test multiple times and their observed test scores recorded, the resulting distribution would be the conditional distribution of raw scores for that student's level of θ with a mean value equal to the student's expected raw (true) score. The CSEM for that level of θ in the raw score metric is the square root of the variance of this conditional distribution.

The conditional distribution of raw scores for any level of θ is the compound multinomial distribution (Wang et al., 2000). An algorithm to compute this can be found in Hanson (1994) and in Thissen et al. (1995) and is also implemented in the computer program POLYCSEM (Kolen, 2004). The compound multinomial distribution yields the probabilities that a student with a given level of θ has of attaining each achievable raw (and accompanying scale) score. The point values associated with each achievable raw or scale score point can be used to calculate the mean and variance of this distribution in the raw or scale score metric, respectively. The square root of the variance is the CSEM of the raw or scale score point associated with the current level of θ.

4.2.4. Conditional Standard Error of Measurement Confidence Intervals

CSEMs allow statements regarding the precision of individual test scores. Like SEMs, they help place reasonable limits around observed scale scores through the construction of an approximate score band. The confidence intervals are constructed by adding and subtracting a multiplicative factor of the CSEM.

4.2.5. Conditional Standard Error of Measurement Characteristics

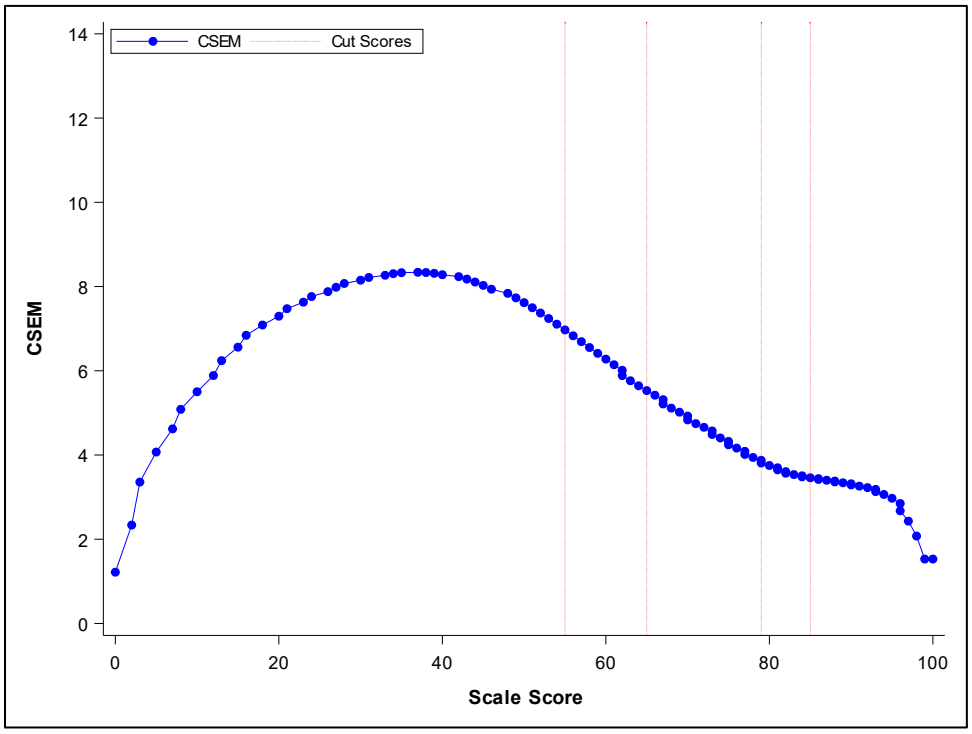
The relationship between the scale score CSEM and θ depends on the nature of the raw-to-scale score transformation (Kolen & Brennan, 2005; Kolen & Lee, 2011) and on whether the CSEM is derived from the raw scores or from θ (Lord, 1980). The pattern of CSEMs for raw scores and linear transformations of the raw score tend to have a characteristic "inverted-U" shape, with smaller CSEMs at the ends of the score continuum and larger CSEMs toward the middle of the distribution. Achievable raw score points for these distributions are spaced equally across the score range. Kolen and Brennan state, "When, relative to raw scores, the transformation compresses the scale in the middle and stretches it at the ends, the pattern of the conditional standard errors of measurement will be concave up (U-shaped), even though the pattern for the raw scores was concave down (inverted-U shape)" (2005, p. 357).

4.2.6. Results and Observations

The relationship between raw and scale scores for the Regents Examinations tends to be roughly linear from scale scores of 0 to 65 and then concave down from about 65 to 100. In other words, the scale scores track linearly with the raw scores for the lower two-thirds of the scale score range and are then compressed relative to the raw scores for the remaining one-third of the range, though there are variations. The CSEMs for the Regents Examinations can be expected to have inverted-U shaped patterns, with some variations.

Figure 4.1 shows this type of CSEM variation for the Regents Examination in Global History and Geography II, where the compression of raw score to scale scores between the cut scores of 65 and 85 slightly changes the shape of the curve. This type of expansion and compression can be seen by looking at the changing density of raw score points along the scale score range on the horizontal axis. Specifically, the raw scores are expanded up to a scale score of about 65 followed by very noticeable compression through a scale score of about 95.

Figure 4.1. Conditional Standard Error Plot: Regents Examination in Global History and Geography II



4.3. DECISION CONSISTENCY AND ACCURACY (STANDARD 2.16)

In a standards-based testing program, there is interest in knowing how accurately students are classified into performance categories. In contrast to the coefficient alpha that is concerned with the relative rank-ordering of students, it is the absolute values of student scores that are important in decision consistency and accuracy.

Classification consistency refers to the degree to which the performance level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). Decision consistency answers the question, what is the agreement in classifications between the two nonoverlapping, equally difficult forms of the test? If two parallel forms of the test were given to the same students, the consistency of the measure would be reflected by the extent to which the classification decisions based on the first set of test scores matched the decisions based on the second set of test scores.

Consider Table 4.2 and Table 4.3. If a student is classified as being in one category based on the Test 1 score, how probable would it be that the student would be reclassified as being in the same category if the student took Test 2 (a nonoverlapping, equally difficult form of the test)? This proportion is a measure of decision consistency. The proportions of correct decisions, φ, for two and four categories are computed by the following formulas, respectively:

$$\varphi = \varphi_{11} + \varphi_{22}$$
$$\varphi = \varphi_{11} + \varphi_{22} + \varphi_{33} + \varphi_{44}$$

The sum of the diagonal entries (i.e., the proportion of students classified by the two forms into the same performance level) signifies the overall consistency.

Table 4.2. Pseudo-Decision Table for Two Hypothetical Categories

		Test 1		
		Level 1	Level 2	Marginal
Test 2	Level 1	φ11	φ12	φ1●
	Level 2	φ21	φ22	φ2●
	Marginal	φ●1	φ●2	1

Table 4.3. Pseudo-Decision Table for Four Hypothetical Categories

		Test 1				
		Level 1	Level 2	Level 3	Level 4	Marginal
Test 2	Level 1	φ11	φ12	φ13	φ14	φ1●
	Level 2	φ21	φ22	φ23	φ24	φ2●
	Level 3	φ31	φ32	φ33	φ34	φ3●
	Level 4	φ41	φ42	φ43	φ44	φ4●
	Marginal	φ●1	φ●2	φ●3	φ●4	1

Classification accuracy refers to the agreement of the observed classifications of students with the classifications made based on their true scores. An observed score contains measurement error, while a true score is theoretically free of measurement error. A student's observed score can be formulated by the sum of the true score plus measurement error ($Observed = True + Error$). Decision accuracy is an index to determine the extent to which measurement error causes a classification different than the one expected from the true score. Because true scores are unobserved and decision consistency is computed based on a single administration of the Regents Examination, a statistical model using data only from the available administration is used to estimate the true scores and to project the consistency and accuracy of classifications (Hambleton & Novick, 1973). Although several procedures are available, a well-known method developed by Livingston and Lewis (1995) that implements a specific true score model is used.

Several factors might affect decision consistency and accuracy. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications and less measurement error. Another factor is the location of the cut score in the score distribution. More consistent and accurate classifications are observed when the cut scores are located away from the mass of the score distribution. The number of performance levels is also a consideration. Consistency and accuracy indices based on four performance levels should be lower than those based on two performance levels. This is not surprising, as classification and accuracy based on four performance levels would allow more opportunity to change performance levels. Hence, there would be more classification errors and less accuracy with four performance levels, resulting in lower consistency indices.

Table 4.4 presents the results for the dichotomies created by the four corresponding cut scores. For example, the statistics under "2/3" indicate the decision consistency and accuracy when the performance levels are divided into two categories: one for the second and lower performance level, and the other for the third and higher performance levels. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method (1995). The decision consistency ranged from 0.85 to 0.94, and the decision accuracy ranged from 0.89 to 0.96. Both decision consistency and accuracy values are high, indicating very good consistency and accuracy of student classifications.

Table 4.4. Decision Consistency and Accuracy Results: Regents Examination in Global History and Geography II

Statistic	1/2	2/3	3/4	4/5
Consistency	0.94	0.90	0.85	0.86
Accuracy	0.96	0.93	0.89	0.90

4.4. GROUP MEANS (STANDARD 2.17)

The student population for the Regents Examination in Global History and Geography II is composed of students who have completed a course in Global History and Geography II. Table 4.5 presents the overall mean scale score that was computed based on all students who took the exam.

Table 4.5. Group Means: Regents Examination in Global History and Geography II

Demographics	#Students	Mean Scale Score	SD Scale Score
All Students	187,860	74.64	13.25
Race/Ethnicity			
American Indian/Alaska Native	1,193	70.70	13.48
Asian/Native Hawaiian/Other Pacific Islander	19,396	80.54	11.06
Black/African American	28,845	68.41	13.96
Hispanic/Latino	48,849	70.69	13.16
Multiracial	5,118	76.12	12.93
White	84,444	77.66	11.98
English Language Learner			
No	176,031	75.42	12.83
Yes	11,829	63.06	13.99
Economically Disadvantaged			
No	97,224	78.55	11.52
Yes	90,636	70.44	13.70
Gender			
Female	93,026	75.21	12.61
Male	94,770	74.07	13.83
Nonbinary	49	79.14	11.49
Student with a Disability			
No	159,715	76.61	11.98
Yes	28,145	63.47	14.51

Note. 15 students were not reported in the race/ethnicity and gender groups, but they are reflected in "All Students."

4.5. STATE PERCENTILE RANKINGS

Table 4.6 presents the state percentile rankings based on scale score distributions of all students taking the Regents Examination in Global History and Geography II for the June 2022 administration. The scale scores for range from 0 to 100, and some scale scores may not be obtainable depending on the raw score-to-scale score relationship for a specific administration. The percentile ranks are computed in the following manner:

- A student's assigned "state percentile rank" will be the cumulative percentage of students scoring at the immediate lower score plus half of the percentage of students obtaining the given score.
- Students who obtain the highest possible score will receive a percentile rank of 99.

Table 4.6. State Percentile Ranking for Scale Score: Regents Examination in Global History and Geography II

Scale Score	Percentile Rank	Scale Score	Percentile Rank	Scale Score	Percentile Rank	Scale Score	Percentile Rank
0	1	26	1	52	7	78	53
1	1	27	1	53	8	79	57
2	1	28	1	54	8	80	60
3	1	29	1	55	9	81	65
4	1	30	1	56	10	82	68
5	1	31	1	57	11	83	71
6	1	32	1	58	12	84	75
7	1	33	1	59	12	85	77
8	1	34	1	60	13	86	81
9	1	35	1	61	14	87	85
10	1	36	1	62	16	88	87
11	1	37	1	63	17	89	90
12	1	38	2	64	19	90	92
13	1	39	2	65	20	91	94
14	1	40	2	66	22	92	95
15	1	41	2	67	23	93	96
16	1	42	2	68	26	94	97
17	1	43	3	69	28	95	98
18	1	44	3	70	30	96	98
19	1	45	3	71	33	97	99
20	1	46	4	72	36	98	99
21	1	47	4	73	38	99	99
22	1	48	4	74	40	100	99
23	1	49	5	75	44	–	–
24	1	50	6	76	47	–	–
25	1	51	6	77	50	–	–

Chapter 5: Validity (Standard 1)

To restate the purposes and uses of the Regents Examination in Global History and Geography II, this exam measures student achievement against the NYS Learning Standards and was prepared by teacher examination committees and NYSED subject matter and testing specialists. Further, it provides teachers and students with important information about student learning and performance against the established curriculum standards. Results of this exam may be used to identify student strengths and needs, which may be used to guide classroom teaching and learning. The exam also provides students, parents, counselors, administrators, and college admissions officers with objective and easily understood achievement information that may be used to inform empirically based educational and vocational decisions about students. As a state-provided objective benchmark, the Regents Examination in Global History and Geography II is intended for use in satisfying state testing requirements for students who have finished a course in Global History and Geography II. A passing score on the exam counts toward requirements for a high school diploma, as described in the New York State diploma requirements (https://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf). Results of the exam may also be used to satisfy various locally established requirements throughout the state.

The validity of score interpretations for this exam is supported by the following sources of evidence, which are important to gather and document to support validity claims for an assessment as provided by the *Standards for Educational Psychological Testing* (AERA et al., 2014):

- Test content
- Response processes
- Internal test structure
- Relation to other variables
- Consequences of testing

While these categories are not mutually exclusive, as one source of validity evidence often falls into more than one category, they provide a useful framework within the *Standards* for the discussion and documentation of validity evidence. The process of gathering evidence of the validity of score interpretations is best characterized as ongoing throughout test development, administration, scoring, reporting, and beyond.

5.1. EVIDENCE BASED ON TEST CONTENT

The validity of test content is fundamental to arguments that test scores are valid for their intended purpose. It demands that a test developer provide evidence that test content is well aligned within the framework and standards used in curriculum and instruction. Accordingly, detailed attention was given to this correspondence between standards and test content during test design and construction. The Regents Examination in Global History and Geography II measure student achievement on the NYS K–12 Social Studies Framework, located at https://www.p12.nysed.gov/ciai/socst/ssrg.html.

5.1.1. Content Validity

Content validity is concerned with the proper definition of the construct and evidence that the test provides an accurate measure of student performance within the defined construct. The test blueprint is essentially the design document for test construction that provides an explicit definition of the construct domain that is to be represented on the exam. The test development process is in place to ensure, to the extent possible, that the blueprint is met in all operational forms of the exam. Table 5.1 presents a summary of the Global History and Geography II blueprint.

Table 5.1. Test Blueprint: Regents Examination in Global History and Geography II

Item Type	Approximate Weighting
Stimulus-based multiple-choice	54%
Short-answer constructed-response	17%
Enduring issues essay	29%

5.1.2. Item Development Process

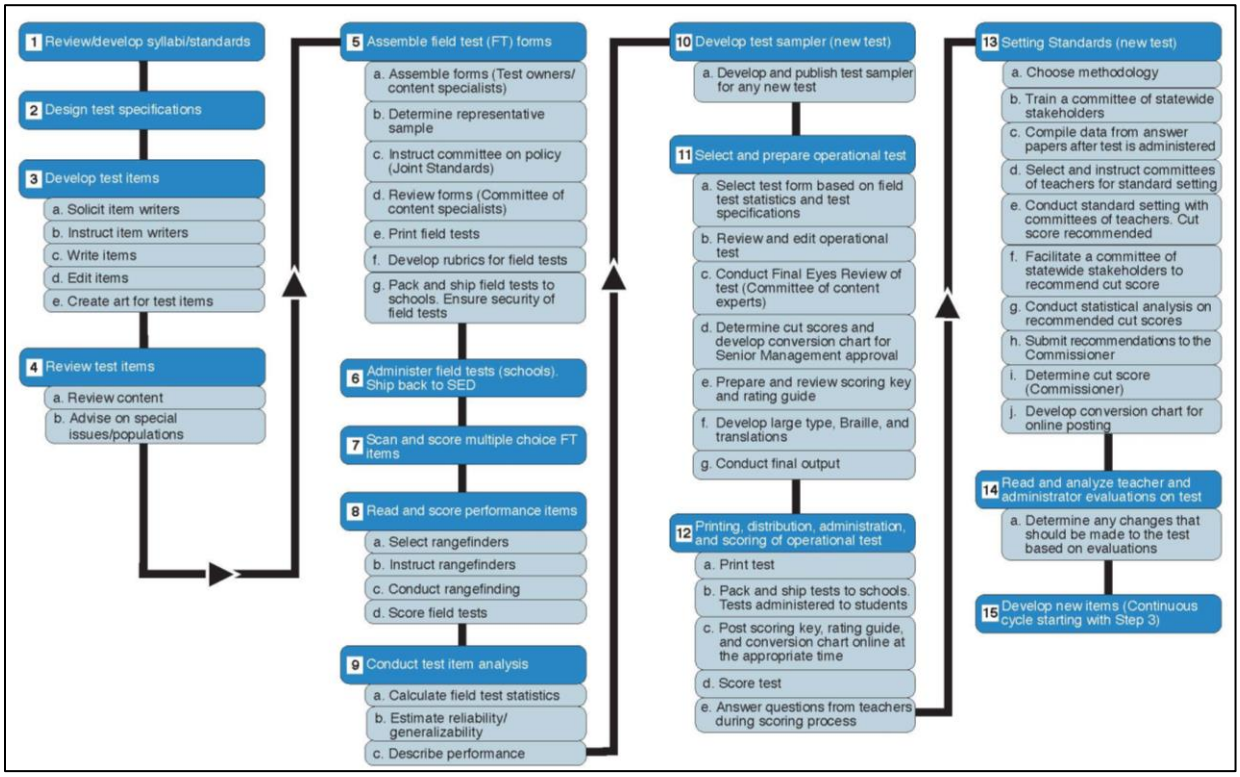
Test development is a detailed, step-by-step process of development and review cycles. An important element of this process is that all test items are developed by NYS educators in a process facilitated by state subject matter and testing experts. Bringing experienced classroom teachers into this central item development role serves to draw a strong connection between classroom and test content.

Only NYS-certified educators may participate in this process. NYSED asks for nominations from districts, and all recruiting is done with diversity of participants in mind, including diversity in gender, ethnicity, geographic region, and teaching experience. Educators with item writing skills from throughout the state are retained to write all items for the Regents Examinations, under strict guidelines that leverage best practices (see Appendix C). State educators also conduct all item quality and bias reviews to ensure that item content is appropriate to the construct being measured and fair for all students. Finally, educators use the defined standards, test blueprint targets, and statistical information generated during field testing to select the highest-quality items for use in the operational test.

Initial item development is conducted under the criteria and guidance provided by multiple documents, including the blueprint, item writing criteria, and a content verification checklist. Both MC and CR items are included on the exam to ensure appropriate coverage of the construct domain. The item writing guidelines in Appendix C provide detailed information about how the items are developed.

Figure 5.1 summarizes the full test development process, with Steps 3 and 4 addressing initial item development and review. This figure also demonstrates the ongoing nature of ensuring the content validity of items through field test trials and final item selection for operational testing.

Figure 5.1. NYSED Test Development Process



5.1.3. Item Review Process

The item review process assists in the consistent application of rigorous item reviews intended to assess the quality of the items developed and identify items that require edits or removal from the pool of items to be field-tested. The following criteria help to ensure that high-quality items are continually developed in a manner that is consistent with the test blueprint.

All reviewers participate in rigorous training designed to assist in a consistent interpretation of the standards throughout the item review process. This is a critical step in item development because consistency between the standards and what the items are asking students is a fundamental form of evidence of the validity of the intended score interpretations. Another integral component of this item review process is to review the scoring rules, or rubrics, for their clarity and consistency in what the student is being asked to demonstrate by responding to each item. Each of these elements is in place to target fairness for all students by targeting consistency in student scores and providing evidence of the validity of their interpretations. Following these reviews, only items that are approved by an assigned educator panel move forward for field testing.

Specifically, the item review process articulates four major item characteristics that NYSED looks for when developing quality items:

1. Language and graphical appropriateness
2. Sensitivity/bias
3. Alignment of measurement to standards
4. Conformity to the expectations for the specific item types and formats

Each section of the criteria includes pertinent questions that help reviewers determine whether an item is of sufficient quality. Within the first two categories, criteria for language appropriateness are used to help ensure that students understand what is asked in each item and that the language in the item does not adversely affect a student's ability to perform the required task. Similarly, the sensitivity/bias criteria are used to evaluate whether items are unbiased, non-offensive, and not disadvantageous to any given subgroup(s). The mathematics art criteria assess the appropriateness and clarity of any graphics that are used within items.

The third category of the item review (alignment) addresses how each item measures a given mathematics standard. This criterion asks the reviewer to comment on key aspects of how the item addresses and calls for the skills demanded by the standards. These criteria also prompt reviewers to comment on how more than one standard is addressed by a given item. The fourth category (criteria framework) addresses the demands for different item types and formats. Reviewers evaluate each item to ensure that it conforms to the given requirements. For example, MC items must have one unambiguously correct answer and several plausible, but incorrect, answer choices.

Ongoing attention is also given to the relevance of the standards used to guide curriculum and assessment. Consistent with a desire to assess this relevance, NYSED is committed to ongoing standards review over time and periodically solicits thoughtful, specific responses from stakeholders about individual standards within the NYS P–12 Learning Standards.

5.2. EVIDENCE BASED ON RESPONSE PROCESSES

The second source of validity evidence is based on student response processes and requires evidence that students are responding in the manner intended by the test items and rubrics and that raters are scoring those responses in a manner that is consistent with the rubrics. It is important to control and monitor whether construct-irrelevant variance in response patterns has been introduced at any point in the test development, administration, or scoring processes.

The controls and monitoring in place for the Regents Examinations include the item development process, with attention paid to mitigating the introduction of construct-irrelevant variance. The test development process details the methods used and attention given to reducing the potential for construct irrelevance in response processes. This is accomplished by attending to the quality and alignment of test content to the test blueprint and to the item development guidelines (Appendix C). Further evidence is documented in the test administration and scoring procedures and in the results of statistical analyses.

5.2.1. Administration and Scoring

Adherence to standardized administration procedures is fundamental to the validity of test scores and their interpretation, as such procedures allow for adequate and consistently applied conditions for scoring the work of every student who takes the examination. For this reason, guidelines, as contained in the *School Administrator's Manual* (https://www.nysed.gov/state-assessment/test-manuals), have been developed and implemented. All secondary-level Regents Examinations are administered under these standard conditions to support valid inferences for all students. These procedures also cover testing students with disabilities who are provided testing accommodations consistent with their Individualized Education Programs (IEPs) or Section 504 Accommodation Plans (504 Plans). Full test administration procedures are available at https://www.nysed.gov/state-assessment/high-school-regents-examinations.

The implementation of rigorous scoring procedures supports the validity of the scores. Regents test scoring practices therefore focus on producing high-quality scores. MC items are scored via local scanning at testing centers, and trained educators score CR items. Many studies focus on various elements of producing valid and reliable scores for CR items, but generally, attention to the following all contribute to valid and reliable scores for CR items:

1. Quality training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang et al., 2010; Gorman & Rentsch, 2009; Schleicher et al., 2002; Woehr & Huffcutt, 1994; Johnson et al., 2008; Weigle, 1998)
2. Detection and correction of rating bias (McQueen & Congdon, 1997; Congdon & McQueen, 2000; Myford & Wolfe, 2009; Barkaoui, 2011; Patz et al., 2002)
3. Consistency or reliability of ratings (Congdon & McQueen, 2000; Harik et al., 2009; McQueen & Congdon, 1997; Myford & Wolfe, 2009; Mero & Motowidlo, 1995; Weinrott & Jones, 1984)
4. Rubric designs that facilitate consistency of ratings (Pecheone & Chung, 2007; Wolfe & Gitomer, 2000; Cronbach et al., 1995; Cook & Beckman, 2009; Penny et al., 2000; Smith, 1993; Leacock et al., 2014).

The distinct steps for operational test scoring include close attention to each of these elements and begin before the operational test is selected. After the field test process, a set of anchor papers representing student responses across the range of possible responses for CR items is selected. The objective of these range-finding efforts is to create a training set for scorer training and execution. The scores calculated for this training set are used to generate important statistical information about the item. Training scorers to produce reliable and valid scores is the basis for creating rating guides and scoring ancillaries to be used during operational scoring.

To review and select these anchor papers, NYS educators serve as table leaders during the range-finding session. In the range-finding process, committees of educators receive a set of student papers for each field-tested item. Committee members familiarize themselves with each item type and score several responses that are representative of each score point. After the independent scoring is completed, the committee reviews and discusses their results and determines consensus scores for the student responses. Atypical responses are important to identify and annotate for use in training and live scoring. The range-finding results are then used to build training materials for the vendor's scorers, who then score the rest of the field test responses to CR items. The Rating Guide for the June 2022 administration of the Regents Examination in Global History and Geography II are located at https://www.nysedregents.org/ghg2/home.html.

During the range-finding and field test scoring processes, it is important to control for sources of variation in scoring. One possible source of variation in CR scores is unintended rater bias associated with items and student responses. The rater is often unaware of such bias, so this type of variation may be the most challenging source of variation in scoring to control and measure. Rater biases can appear as severity or leniency in applying the scoring rubric. Bias also includes phenomena such as the halo effect that occurs when good or poor performance on one element of the rubric encourages inaccurate scoring of other elements. These types of rater bias can be controlled by training practices with a strict focus on rubric requirements.

The training process for operational scoring by NYS educators begins with a review and discussion of actual student work on CR items. This helps raters understand the range and characteristics typical of student responses and the kinds of mistakes students commonly make. This information is used to train raters on how to consistently apply key elements of the scoring rubric across the domain of student responses.

Raters then receive training consistent with the guidelines and ancillaries produced after field testing and are allowed to practice scoring prior to the start of live scoring. Throughout the scoring process, there are important procedures for correcting inconsistent scoring or the misapplication of scoring rubrics for CR items. When monitoring and correction do not occur during scoring, construct-irrelevant variation may be introduced. Accordingly, a scoring lead may be assigned to review the consistency of scoring for the lead's assigned staff against model responses and to be available for consultation throughout the scoring process.

Attention to the rubric design also contributes to the validity of student response processes. The rubric specifies what the student needs to provide as evidence of learning based on the question asked. The more explicit the rubric (and the item), the clearer the response expectations are for students. To facilitate the development of CR scoring rubrics, NYSED training for writing items includes specific attention to rubric development, as follows:

- The rubric should clearly specify the criteria for awarding each credit.
- The rubric should be aligned to what is asked for in the item and correspond to the knowledge or skill being assessed.
- Whenever possible, the rubric should be written to allow for alternate approaches and other legitimate methods.

In support of the goal of valid score interpretations for each student, such scoring training procedures are implemented for the Regents Examinations. Operational raters are selected based on expertise in the exam subject and are assigned a specific set of items to score. No more than one-third of the items on the test are assigned to any one rater. This increases the consistency of scoring across student responses by allowing each rater to focus on a subset of items. It also ensures that no one rater is allowed to score the entire test for any one student. This practice reduces the effect of any potential bias of a single rater on individual students. Raters are also not allowed to score the responses of their own students.

5.2.2. Statistical Analysis

A useful statistic for evaluating the response processes for MC items is an item's point-biserial correlation on the distractors. A high point-biserial on a distractor may indicate that students are not able to identify the correct response for a reason other than the difficulty of the item. A finding of poor model fit for an item may also support a finding that students are not responding in the way in which the item developer intended. As shown in Table 2.1, the point-biserial correlations for distractors in the MC items all appear to be negative or close to zero, indicating that, for the most part, students are not being drawn to an unintended construct.

5.3. EVIDENCE BASED ON INTERNAL STRUCTURE

The third source of validity evidence comes from the internal structure of the test and requires that test developers evaluate the test structure to ensure that it is functioning as intended. Such an evaluation may include attention to item interactions, tests of dimensionality, or indications of test bias for or against one or more student subgroups detected by differential item functioning (DIF) analysis. Evaluation of internal test structure also includes a review of the results of classical item analyses, test reliability, and the IRT scaling and equating. The following analyses were conducted for the Regents Examination in Global History and Geography II:

- Item difficulty
- Item discrimination
- Differential item functioning (DIF)
- IRT model fit
- Test reliability
- Classification consistency
- Test dimensionality

5.3.1. Item Difficulty

Multiple analyses allow for an evaluation of item difficulty. For this exam, *p*-values and Rasch difficulty (item location) estimates were computed for MC and CR items. Items for the June 2022 Regents Examination in Global History and Geography II show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values range from 0.38 to 0.93, with a mean of 0.70. The difficulty distribution illustrated in Figure 2.1 shows a wide range of item difficulties on the exam. This is consistent with general test development practice, which seeks to measure student ability along a full range of difficulty. Refer to Chapter 2 of this report for additional details.

5.3.2. Item Discrimination

How well the items on a test discriminate between high- and low-performing students is an important measure of the test structure. Items that do not discriminate well generally provide less reliable information about student performance. Table 2.1 and Table 2.2 provide point-biserial values on the correct responses; Table 2.1 also provides point-biserial values on the three distractors. The values for correct answers are 0.19 or higher for all items, indicating that most items are discriminating well between high- and low-performing students. Point-biserial values for all distractors are negative, indicating that students are responding to the items as expected during item development. Refer to Chapter 2 of this report for additional details.

5.3.3. Differential Item Functioning

Differential Item Functioning (DIF) was conducted for gender, race/ethnicity, needs/resource capacity (NRC) categories, and ELL status based on the data for the June 2022 administration. DIF data are only available after the administration as all Regents exams are pre-equated, meaning that the parameters used to derive the relationship between the raw and scale scores are estimated prior to the construction and administration of the operational form.

The Mantel-Haenszel (Mantel & Haenszel, 1959) and standardized mean difference (Dorans & Schmitt, 1991) methods were used to detect items that may function differently for any subgroup. The Mantel-Haenszelχ^2 is a conditional mean comparison of the ordered response categories for reference and focal groups combined over values of the matching variable score. "Ordered" means that a response earning a score of 1 on an item is better than a response earning a score of 0, a 2 is better than 1, and so on. "Conditional," on the other hand, refers to the comparison of members from the two groups who received the same score on the matching variable (the total test score in this analysis). The results of these analyses were examined by NYSED content specialists to identify potential systematic issues that could be addressed in future item writing.

5.3.4. IRT Model Fit

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the exam provide important evidence that the internal structure of the test is of high technical quality. The number of items within a targeted range of [0.7, 1.3] is reported in Table 3.2. The mean INFIT value is 0.99, with all but one of the items falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as a guide for ideal fit, fit values outside of the range are considered individually. These results indicate that, for most items, the Rasch model fits the item data well.

5.3.5. Test Reliability

Test reliability is a measure of the internal consistency of a test (Cronbach, 1951). It is a measure of the extent to which the items on a test provide consistent information about student mastery of the domain. Reliability should ultimately demonstrate that student score estimates maximize consistency and therefore minimize error or, theoretically, that students who take a test multiple times would get the same score each time. The reliability estimate for the Regents Examination in Global History and Geography II is 0.88, showing high reliability of student scores. Refer to Chapter 4 of this report for additional details.

5.3.6. Classification Consistency and Accuracy

Decision consistency measures the agreement between the classifications based on two nonoverlapping, equally difficult forms of the test. If two parallel forms were given to the same students, the consistency of the measure would be reflected by the extent to which the classification decisions based on the first set of test scores matched the decisions based on the second set of scores. Decision accuracy determines the extent to which measurement error causes a classification different from that expected from the true score. High decision consistency and accuracy provide strong evidence that the internal structure of a test is sound.

For the Regents Examination in Global History and Geography II, both decision consistency and accuracy values are high, indicating very good consistency and accuracy of student classifications. The results for the overall consistency across all five performance levels, as well as for the dichotomies created by the four corresponding cut scores, are presented in Table 4.4. The tabled values are derived with the program BB-Class (Brennan, 2004) using the Livingston and Lewis method. The decision consistency ranged from 0.85 to 0.94, and the decision accuracy ranged from 0.89 to 0.96.

5.3.7. Test Dimensionality

In addition to model fit, a strong assumption of the Rasch model is that the construct measured by a test is unidimensional. Violation of this might suggest that the test is measuring something other than the intended content, indicating that the test structure quality is compromised. The results of a PCA conducted to test the assumption of unidimensionality provide strong evidence that a single dimension in the exam is explaining a large portion of the variance in student response data. This analysis does not characterize or explain the dimension, but a reasonable assumption can be made that the test is largely unidimensional and that the dimension most present is the targeted construct. Refer to Chapter 3 for details of this analysis.

Considering this collection of detailed analyses of the internal structure of the Regents Examination in Global History and Geography II, strong evidence exists that the exam is functioning as intended and is providing valid and reliable information about student performance.

5.4. EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES

Another source of validity evidence is based on the relation of the test to other variables. This source commonly encompasses two validity categories prevalent in the literature and practice: concurrent and predictive validity. To make claims about the validity of a test that is to be used for high-stakes purposes, such as the Regents Examination in Global History and Geography II, these claims could be supported by providing evidence that performance on the Global History and Geography II test correlates well with other tests that measure the same or similar constructs. Although not absolute in its ability to offer evidence that concurrent test score validity exists, such correlations can be helpful for supporting a claim of concurrent validity if the correlation is high. To conduct such studies, matched student score data for other tests measuring the same content as the Regents Examination in Global History and Geography II is ideal. However, the systematic acquisition of such data is complex and costly.

Importantly, a strong connection between classroom curriculum and test content may be inferred by the fact that NYS educators, deeply familiar with the curriculum standards and their enactment in the classroom, develop all content for the Regents Examination in Global History and Geography II. In terms of predictive validity, time is a fundamental constraint on gathering evidence. The gold standard for supporting the validity of predictive statements about test scores requires empirical evidence of the relationship between test scores and future performance on a defined characteristic. To the extent that the objective of the standards is to prepare students for meeting graduation requirements, it will be important to gather evidence of this empirical relationship over time.

5.5. EVIDENCE BASED ON TESTING CONSEQUENCES

There are two general approaches in the literature for evaluating consequential validity. Messick (1995) points out that adverse social consequences invalidate test use mainly if they are due to flaws in the test. In this sense, the sources of evidence documented in this report (based on the construct, internal test structure, response processes, and relation to other variables) serve as a consequential validity argument as well. This evidence supports conclusions based on test scores that social consequences are not likely to be traced to characteristics or qualities of the test itself.

Cronbach (1988), on the other hand, argues that negative consequences could invalidate test use. From this perspective, the test user is obligated to make the case for test use and to ensure appropriate and supported uses. Regardless of perspective on the nature of consequential validity, it is important to caution against uses that are not supported by the validity claims documented for this test. For example, use of this test to predict student scores on other tests is not directly supported by either the stated purposes or by the development process and research conducted on student data. A brief survey of websites for NYS universities and colleges finds that, beyond the explicitly defined use as a testing requirement toward graduation for students who have completed a course in Global History and Geography II, the exam is most commonly used to inform admissions and course placement decisions. Such uses can be considered reasonable, assuming that the competencies demonstrated in the Regents Examination in Global History and Geography II are consistent with those required in the courses for which a student is seeking enrollment or placement. Educational institutions using the exam for placement purposes are advised to examine the scoring rules for the Regents Examination in Global History and Geography II and to assess their appropriateness for the inferences being made about course placement.

5.6. SUMMARY

As stated, the nature of validity arguments is not absolute. Rather, it is supported through ongoing processes and studies designed to accumulate support for validity claims. The evidence provided in this report documents the evidence to date that supports the use of the Regents Examination in Global History and Geography II scores for the purposes described.

References

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

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3).

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* [Computer software] (Version 1.0). University of Iowa.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163–178.

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for mini-CEX*. Advances in Health Sciences Education, 14,* 655–684.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). *Generalizability analysis for educational assessments*. University of California, Center for the Study of Evaluation and The National Center for Research on Evaluation, Standards, and Student Testing.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS Research Report No. 91-49). Educational Testing Service.

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*(5), 1336–1344.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Hambleton, R. K., & Novak, M. R. (1973). Toward an integration of theory and methods for criterion-referenced tests. *Journal of Educational Measurement*, *10*, 159–170.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Item response theory*. Newbury Sage Publications.

Hanson, B. A. (1994). *Extension of Lord-Wingersky algorithm to computing test scores for polytomous items*. http://www.openirt.com/b-a-h/papers/note9401.pdf

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46(*1), 43–58.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practices, 10*(2), 33–41.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 32,* 179–185.

Hoyt, W. T., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4)*,* 403–424.

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika*, *41*, 65–78.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance.* The Guilford Press.

Kolen, M. J., (2004). POLYCSEM [Computer program]. University of Iowa. https://education.uiowa.edu/casma/computer-programs

Kolen, M. J., & Brennan, R. L. (2005). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer-Verlag.

Kolen, M. J. & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice 30*(2), 15–24.

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.

Leacock, C., Gonzalez, E., & Conarroe, M. (2014). *Developing effective scoring rubrics for AI short answer scoring*. McGraw-Hill Education CTB Innovative Research and Development Grant.

Linacre, J. M. (2005). *WINSTEPS Rasch measurement computer program* (Version 3.60) [Computer software]. Winsteps.com.

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. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54–71.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research, 14*, 21–38.

McQueen, J., & Congdon, P. J. (1997, March). *Rater severity in large-scale assessment: Is it invariant?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80*(4), 517–524.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46*(4), 371–389.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2014). *PARCC model content frameworks: Mathematics grades 3–11*.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341–384.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. —III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London, 187*, 253–318.

Pecheone, R. L., & Chung, R. (2007). *Performance assessment for California teachers: Summary of validity and reliability studies for the 2003−04 pilot year*. Stanford University PACT Consortium.

Penny, J., Johnson, R. L., & Gordon, B. (2000). Using rating augmentation to expand the scale of an analytic rubric. *The Journal of Experimental Education, 68*(3), 269–287.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford. Nielsen & Lydiche.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230.

Schleicher, D. J., Day, D. V., Bronston, T., Mayes, B. T., & Riggo, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735–746.

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. A. Hout (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations.* Hampton Press.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer-Verlag.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement*, *37*(2), 141–162.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology, 95*(3), 546–561.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287.

Weinrott, L., & Jones, B. (1984). Overt versus covert assessment of observer reliability. *Child Development, 55,* 1125–1137.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3)*,* 189–205.

Wolfe, E. W., & Gitomer, D. H. (2000). *The influence of changes in assessment design on the psychometric quality of scores*. Educational Testing Service.

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

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

Appendix A: Operational Test Map

Table A.1. June 2022 Test Map

Position	Item Type	Max. Points	Weight	Mean	Point-Biserial	Rasch Difficulty	INFIT
01	MC	1	1	0.79	0.53	-0.9357	0.86
02	MC	1	1	0.69	0.53	-0.3363	0.92
03	MC	1	1	0.40	0.58	1.2218	0.86
04	MC	1	1	0.73	0.53	-0.5282	0.90
05	MC	1	1	0.49	0.50	0.7853	0.98
06	MC	1	1	0.76	0.44	-0.7745	0.97
07	MC	1	1	0.66	0.58	-0.1282	0.86
08	MC	1	1	0.42	0.45	1.1406	1.00
09	MC	1	1	0.52	0.34	0.6163	1.17
10	MC	1	1	0.49	0.49	0.7862	0.98
11	MC	1	1	0.77	0.45	-0.7956	0.97
12	MC	1	1	0.61	0.57	0.1337	0.88
13	MC	1	1	0.54	0.38	0.4432	1.08
14	MC	1	1	0.42	0.41	1.0495	1.02
15	MC	1	1	0.52	0.51	0.6078	0.96
16	MC	1	1	0.54	0.61	0.5228	0.83
17	MC	1	1	0.44	0.37	1.0170	1.14
18	MC	1	1	0.42	0.34	1.1296	1.16
19	MC	1	1	0.54	0.61	0.2827	0.86
20	MC	1	1	0.58	0.49	0.0575	1.03
21	MC	1	1	0.70	0.53	-0.6227	0.92
22	MC	1	1	0.62	0.48	-0.1712	1.01
23	MC	1	1	0.51	0.44	0.4444	1.10
24	MC	1	1	0.73	0.48	-0.8085	0.97
25	MC	1	1	0.74	0.58	-0.5995	0.83
26	MC	1	1	0.70	0.54	-0.3563	0.89
27	MC	1	1	0.42	0.47	1.1383	1.00
28	MC	1	1	0.68	0.57	-0.2597	0.87
29	CR	1	1	0.25	0.49	2.4070	0.99
30	CR	1	1	0.56	0.58	0.5143	0.94
31	CR	1	1	0.33	0.57	1.8301	0.90
32	CR	1	1	0.43	0.57	1.3523	0.92
33	CR	1	1	0.66	0.53	0.0343	0.99
34a	CR	1	1	0.66	0.63	0.0424	0.84
34b	CR	1	1	0.51	0.59	0.9031	0.92
35	CR	5	3	1.19	0.73	1.5498	0.95

Appendix B: Raw-to-Theta-to-Scale Score Conversion Table

Table B.1. June 2022 Score Table

Raw Score	Ability	Scale Score
0	-5.65151	0.000
0.5	-4.43829	2.031
1	-3.73232	3.929
1.5	-3.31437	5.787
2	-3.01460	7.610
2.5	-2.77971	9.399
3	-2.58604	11.156
3.5	-2.42088	12.883
4	-2.27666	14.579
4.5	-2.14843	16.246
5	-2.03287	17.886
5.5	-1.92753	19.499
6	-1.83064	21.085
6.5	-1.74084	22.645
7	-1.65694	24.183
7.5	-1.57813	25.693
8	-1.50374	27.178
8.5	-1.43312	28.641
9	-1.36576	30.079
9.5	-1.30126	31.493
10	-1.23928	32.883
10.5	-1.17948	34.251
11	-1.12160	35.597
11.5	-1.06542	36.920
12	-1.01065	38.221
12.5	-0.95721	39.500
13	-0.90482	40.759
13.5	-0.85345	41.996
14	-0.80283	43.212
14.5	-0.75298	44.406
15	-0.70368	45.579
15.5	-0.65493	46.732
16	-0.60655	47.865
16.5	-0.55853	48.978
17	-0.51077	50.070
17.5	-0.46322	51.142
18	-0.41580	52.195
18.5	-0.36846	53.230
19	-0.32113	54.245
19.5	-0.27382	55.240
20	-0.22645	56.217
20.5	-0.17899	57.178
21	-0.13142	58.119
21.5	-0.08371	59.044
22	-0.03581	59.951
22.5	0.01229	60.840
23	0.06065	61.713
23.5	0.10925	62.572
24	0.15811	63.414
24.5	0.20730	64.240
25	0.25676	65.051
25.5	0.30661	65.849
26	0.35683	66.633
26.5	0.40741	67.402
27	0.45846	68.160
27.5	0.50991	68.903
28	0.56182	69.635
28.5	0.61424	70.356
29	0.66725	71.066
29.5	0.72079	71.766
30	0.77494	72.454
30.5	0.82972	73.134
31	0.88527	73.806
31.5	0.94154	74.470
32	0.99862	75.126
32.5	1.05663	75.775
33	1.11557	76.418
33.5	1.17553	77.055
34	1.23663	77.687
34.5	1.29891	78.315
35	1.36254	78.940
35.5	1.42750	79.561
36	1.49403	80.180
36.5	1.56212	80.797
37	1.63201	81.414
37.5	1.70373	82.031
38	1.77739	82.649
38.5	1.85324	83.268
39	1.93135	83.889
39.5	2.01180	84.514
40	2.09486	85.143
40.5	2.18070	85.777
41	2.26943	86.416
41.5	2.36134	87.061
42	2.45677	87.715
42.5	2.55600	88.377
43	2.65946	89.050
43.5	2.76761	89.731
44	2.88119	90.425
44.5	3.00081	91.131
45	3.12749	91.850
45.5	3.26240	92.584
46	3.40699	93.333
46.5	3.56358	94.098
47	3.73557	94.881
47.5	3.92845	95.682
48	4.15219	96.502
48.5	4.42656	97.343
49	4.79820	98.205
49.5	5.42331	99.091
50	6.54129	100.000

Appendix C: Item Writing Guidelines

GUIDELINES FOR WRITING MULTIPLE-CHOICE ITEMS

1. The item should focus on a single issue, problem, or topic stated clearly and concisely in the stem.
2. The item should be written in clear and simple language, with vocabulary and sentence structure kept as simple as possible.
3. The stem should be written as a direct question or an incomplete statement.
4. The stem should not contain irrelevant or unnecessary detail.
5. The stem should be stated positively. Avoid using negatively stated stems.
6. The phrase which of the following should not be used to refer to the alternatives. Instead use which followed by a noun.
7. The stem should include any words that must otherwise be repeated in each alternative.
8. The item should have one and only one correct answer (key).
9. The distractors should be plausible and attractive to students who lack the knowledge, understanding, or ability assessed by the item.
10. The alternatives should be grammatically consistent with the stem.
11. The alternatives should be parallel with one another in form.
12. The alternatives should be arranged in logical order, when possible.
13. The alternatives should be independent and mutually exclusive.
14. The item should not contain extraneous clues to the correct answer.
15. Items should be written in the third person. Use generic terms instead of proper nouns, such as first names and brand names.

CHECKLIST OF TEST CONSTRUCTION PRINCIPLES
(Multiple-Choice Items)

	YES	NO
1. Is the item significant?		
2. Does the item have curricular validity?		
3. Is the item presented in clear and simple language, with vocabulary kept as simple as possible?		
4. Does the item have one and only one correct answer?		
5. Does the item state one single central problem completely in the stem? (See Helpful Hint below.)		
6. Does the stem include any extraneous material ("window dressing")?		
7. Are all responses grammatically consistent with the stem and parallel with one another in form?		
8. Are all responses plausible (attractive to students who lack the information tested by the item)?		
9. Are all responses independent and mutually exclusive?		
10. Are there any extraneous clues due to grammatical inconsistencies, verbal associations, length of response, etc.?		
11. Were the principles of Universal Design used in constructing the item?		

HELPFUL HINT

To determine if the stem is complete (meaningful all by itself):

- Cover up the responses and read just the stem.
- Try to turn the stem into a short-answer question by drawing a line after the last word. If it is not a good, short-answer item, then there may be a problem with the stem.
- The stem must consist of a statement that contains a verb.

GUIDELINES FOR WRITING CONSTRUCTED-RESPONSE ITEMS

1. The item should focus on a single issue, problem, or topic stated clearly and concisely.
2. The item should be written with terminology, vocabulary and sentence structure kept as simple as possible. The item should be free of irrelevant or unnecessary detail.
3. The item should be written in the third person. Use generic terms instead of proper nouns such as first names and brand names.
4. The item should not contain extraneous clues to the correct answer.
5. The item should assess student understanding of the material by requiring responses that show evidence of knowledge, comprehension, application, analysis, synthesis, and/or evaluation.
6. When a stimulus is used, an introduction is required.
7. The item should clearly specify what the student is expected to do to provide an acceptable response.
8. A group of constructed-response items should be arranged in logical sequence, and each item should test different knowledge, understandings, and/or skills.
9. The stimulus should provide information/data that is scientifically accurate.
10. The source of each stimulus must be clearly identified for all material that is not original.
11. The introduction, stimulus (when used), item, student answer space, and rating guide must correspond.
12. The rating guide must provide examples of correct responses.
13. The rating guide and items should clearly specify if credit is allowed for labeling units. If no credit is allowed for units, the unit should be provided within the student answer space.
14. The rating guide should specify the acceptable range for numerical responses.