# New York State
# Regents Examination in
# Global History and Geography II

## 2019 Technical Report

Prepared for the New York State Education Department
by Pearson

**March 2020**

# Copyright

Developed and published under contract with the New York State Education Department by Pearson.

# 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 will provide New York State (NYS) with documentation of the purposes of the Regents Examination, scoring information, evidence of both reliability and validity of the exam, scaling information, and guidelines for score reporting for the June 2019 administration. The June 2019 administration was the first operational administration of the Regents Examination in Global History and Geography II. 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" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p.123).[1] Please note that a technical report, by design, addresses technical documentation of a testing program; other aspects of a testing program (content standards, scoring guides, guide to test interpretation, etc.) are thoroughly addressed and referenced in supporting documents.

The Regents Examination in Global History and Geography II was given in June 2019 to students enrolled in New York State schools. The examination is based on the NYS K-12 Social Studies Framework, adopted by the NYS Board of Regents in April 2014.

## 1.2 PURPOSES OF THE EXAM

The Regents Examination in Global History and Geography II measures examinee achievement against the NYS K-12 Social Studies Framework. 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 in order 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: http://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf. Results of the Regents Examination in Global History and Geography II may also be used to satisfy various locally established requirements throughout the state.

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

## 1.3 TARGET POPULATION (STANDARD 7.2)

The examinee population for the Regents Examination in Global History and Geography II is composed of students in Grade 10 who have completed a course in Global History and Geography.

Table 1 provides a demographic breakdown of all students who took the June 2019 administration of the Regents Examination in Global History and Geography II. All analyses in this report are based on the population described in Table 1. Annual Regents Examination results 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 exam administered in June 2019 (see http://data.nysed.gov/). Item-level data used for the analyses in this report are reported by districts on a similar timeline, yet through a different collection system; therefore, the n-sizes in this technical report will differ from publicly reported counts of student test-takers.

**Table 1 Total Examinee Population: Regents Examination in Global History and Geography II**

| | June Admin* | |
|---|---|---|
| **Demographics** | **Number** | **Percent** |
| **All Students** | 131,481 | 100.00 |
| **Race/Ethnicity** | | |
| American Indian/Alaska Native | 1,082 | 0.82 |
| Asian/Native Hawaiian/Other Pacific Islander | 15,588 | 11.86 |
| Black/African American | 26,364 | 20.06 |
| Hispanic/Latino | 37,331 | 28.41 |
| Multiracial | 2,220 | 1.69 |
| White | 48,839 | 37.16 |
| **English Language Learner/Multilingual Learner** | | |
| No | 119,797 | 91.11 |
| Yes | 11,684 | 8.89 |
| **Economically Disadvantaged** | | |
| No | 57,038 | 43.38 |
| Yes | 74,443 | 56.62 |
| **Gender** | | |
| Female | 64,873 | 49.36 |
| Male | 66,551 | 50.64 |
| **Student with a Disability** | | |
| No | 111,385 | 84.72 |
| Yes | 20,096 | 15.28 |

*Note: Fifty-seven students were not reported in the Race/Ethnicity and Gender groups; however, they are reflected in "All Students."

# 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).

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

In the mean score formula above, the individual item scores ($x_i$) are summed and then divided by the total number of students ($n$). For 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.[2] For example, if a MC item has a *p*-value of 0.89, it means that 89 percent of the students answered the item correctly. Additionally, this value might 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 zero) to the maximum possible score. 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 examination. Items with very high or very low *p*-values receive added scrutiny during all follow-up analyses, including item response theory 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 2019 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 and Table 3 for multiple-choice and constructed-response items, respectively, range from 0.43 to 0.91, with a mean of 0.71. Table 2 and Table 3 also show a standard deviation (SD) of item score and item mean (Table 3, 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 Regents Examination in Global History and Geography II would be more

[2] For MC items with four response options, random guessing would lead to an expected *p*-value of 0.25.

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 zero), meaning that the item is likely discriminating well between high- and low-performing students. Point-biserial values are computed for each answer option, including correct and incorrect options (commonly referred to as "distractors"). Finally, point-biserial values for each distractor are an important part of the analysis. The point-biserial values on the distractors are typically negative. 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 and Table 3 provide the point-biserial values for the correct response and three distractors (Table 2, only) for the June 2019 administration of the Regents Examination in Global History and Geography II. The point-biserial values for correct answers are 0.27 or higher for all items, indicating that the items are discriminating well between high- and low-performing examinees. Point-biserial values for all distractors are negative, indicating that examinees are responding to the items as expected during item and rubric development.

**Table 2 Multiple-Choice Item Analysis Summary: Regents Examination in Global History and Geography II**

| Item | Number of Students | *p*-Value | SD | Point-Biserial | Point-Biserial Distractor 1 | Point-Biserial Distractor 2 | Point-Biserial Distractor 3 |
|---|---|---|---|---|---|---|---|
| 1 | 131,481 | 0.90 | 0.30 | 0.34 | -0.18 | -0.22 | -0.16 |
| 2 | 131,481 | 0.53 | 0.50 | 0.35 | -0.18 | -0.14 | -0.17 |
| 3 | 131,481 | 0.63 | 0.48 | 0.36 | -0.18 | -0.24 | -0.09 |
| 4 | 131,481 | 0.86 | 0.34 | 0.32 | -0.14 | -0.15 | -0.22 |
| 5 | 131,481 | 0.66 | 0.47 | 0.50 | -0.16 | -0.38 | -0.20 |
| 6 | 131,481 | 0.67 | 0.47 | 0.38 | -0.29 | -0.18 | -0.13 |
| 7 | 131,481 | 0.84 | 0.36 | 0.44 | -0.29 | -0.22 | -0.22 |
| 8 | 131,481 | 0.91 | 0.28 | 0.43 | -0.29 | -0.23 | -0.18 |
| 9 | 131,481 | 0.65 | 0.48 | 0.45 | -0.24 | -0.27 | -0.17 |
| 10 | 131,481 | 0.55 | 0.50 | 0.42 | -0.03 | -0.16 | -0.39 |
| 11 | 131,481 | 0.57 | 0.50 | 0.39 | -0.25 | -0.10 | -0.20 |
| 12 | 131,481 | 0.71 | 0.45 | 0.45 | -0.24 | -0.19 | -0.29 |
| 13 | 131,481 | 0.80 | 0.40 | 0.51 | -0.18 | -0.26 | -0.35 |
| 14 | 131,481 | 0.79 | 0.41 | 0.41 | -0.21 | -0.26 | -0.18 |
| 15 | 131,481 | 0.72 | 0.45 | 0.39 | -0.14 | -0.30 | -0.20 |
| 16 | 131,481 | 0.78 | 0.41 | 0.38 | -0.19 | -0.25 | -0.15 |
| 17 | 131,481 | 0.74 | 0.44 | 0.38 | -0.17 | -0.22 | -0.22 |
| 18 | 131,481 | 0.46 | 0.50 | 0.35 | -0.20 | -0.21 | -0.10 |

| Item | Number of Students | *p*-Value | SD | Point-Biserial | Point-Biserial Distractor 1 | Point-Biserial Distractor 2 | Point-Biserial Distractor 3 |
|---|---|---|---|---|---|---|---|
| 19 | 131,481 | 0.49 | 0.50 | 0.43 | -0.25 | -0.17 | -0.19 |
| 20 | 131,481 | 0.43 | 0.50 | 0.37 | -0.17 | -0.20 | -0.23 |
| 21 | 131,481 | 0.78 | 0.42 | 0.43 | -0.21 | -0.23 | -0.24 |
| 22 | 131,481 | 0.75 | 0.43 | 0.54 | -0.27 | -0.31 | -0.27 |
| 23 | 131,481 | 0.82 | 0.39 | 0.45 | -0.29 | -0.20 | -0.23 |
| 24 | 131,481 | 0.60 | 0.49 | 0.38 | -0.18 | -0.14 | -0.22 |
| 25 | 131,481 | 0.80 | 0.40 | 0.38 | -0.27 | -0.17 | -0.15 |
| 26 | 131,481 | 0.53 | 0.50 | 0.45 | -0.24 | -0.20 | -0.19 |
| 27 | 131,481 | 0.83 | 0.38 | 0.44 | -0.20 | -0.26 | -0.26 |
| 28 | 131,481 | 0.62 | 0.49 | 0.27 | -0.22 | -0.14 | -0.10 |

**Table 3 Constructed-Response Item Analysis Summary: Regents Examination in Global History and Geography II**

| Item | Min. Score | Max. Score | Number of Students | Mean | SD | *p*-Value | Point-Biserial |
|---|---|---|---|---|---|---|---|
| 29 | 0 | 1 | 131,481 | 0.85 | 0.36 | 0.85 | 0.47 |
| 30 | 0 | 1 | 131,481 | 0.86 | 0.35 | 0.86 | 0.45 |
| 31 | 0 | 1 | 131,481 | 0.81 | 0.40 | 0.81 | 0.52 |
| 32 | 0 | 1 | 131,481 | 0.79 | 0.41 | 0.79 | 0.38 |
| 33 | 0 | 1 | 131,481 | 0.84 | 0.37 | 0.84 | 0.43 |
| 34 | 0 | 1 | 131,481 | 0.80 | 0.40 | 0.80 | 0.50 |
| 35 | 0 | 1 | 131,481 | 0.68 | 0.47 | 0.68 | 0.55 |
| 36 | 0 | 5 | 131,481 | 2.59 | 1.10 | 0.52 | 0.80 |

## 2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOT

Figure 1 shows a scatter plot of item discrimination values (*y*-axis) and item difficulty values (*x*-axis). The descriptive statistics of *p*-value and point-biserial values, including mean, minimum, Q1, median, Q3, and maximum, are also presented in Table 4.

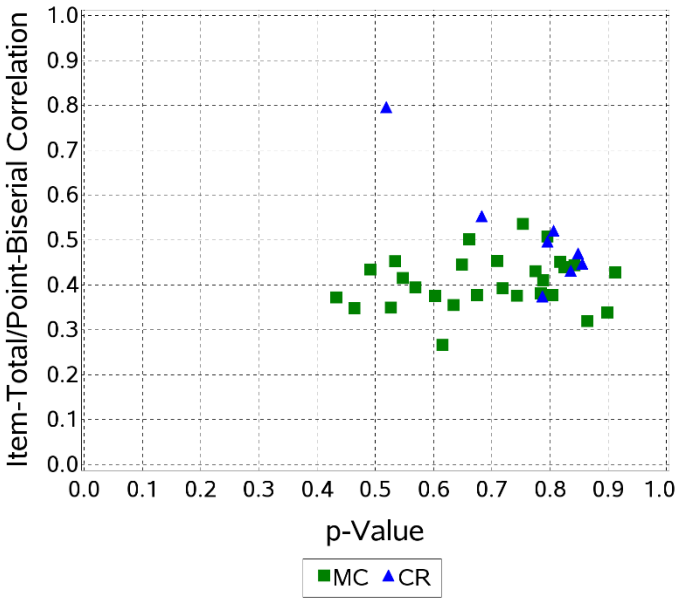


**Figure 1 Scatter Plot: Regents Examination in Global History and Geography II**

**Table 4 Descriptive Statistics in *p*-value and Point-Biserial Correlation: Regents Examination in Global History and Geography II**

| Statistics | N | Mean | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|---|
| *p*-value | 36 | 0.71 | 0.43 | 0.61 | 0.75 | 0.81 | 0.91 |
| Point-Biserial | 36 | 0.43 | 0.27 | 0.38 | 0.43 | 0.45 | 0.80 |

## 2.4 OBSERVATIONS AND INTERPRETATIONS

The *p*-values for the MC items range from about 0.43 to 0.91, while the *p*-values for the CR items (Table 3) range from about 0.52 to 0.86. From the difficulty distributions illustrated in the plot, 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 Global History and Government II 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 it 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, Swaminathan, & Rogers, 1991). This performance value is often referred to as $\theta$. Performance and $\theta$ will be used interchangeably through the remainder of this report.

A fundamental advantage of IRT is that it links examinee performance and item difficulty estimates and places them on the same scale, allowing for an evaluation of examinee 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 examinees across items and test forms.

This chapter outlines the procedures used for calibrating the operational Regents Examination in Global History and Geography II items. 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. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes the Rasch item statistics.

## 3.1 DESCRIPTION OF THE RASCH MODEL

The Rasch model (Rasch, 1960) was used to calibrate multiple-choice items, and the partial credit model, or PCM (Wright & Masters, 1982), was used to calibrate constructed-response 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 by

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

where $\theta_n$ represents examinee 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:

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

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 examinee performance and item difficulty that are theoretically invariant across random samples of the same examinee 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 2. 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 examinee performance at the top and lower performance and easier items at the bottom.

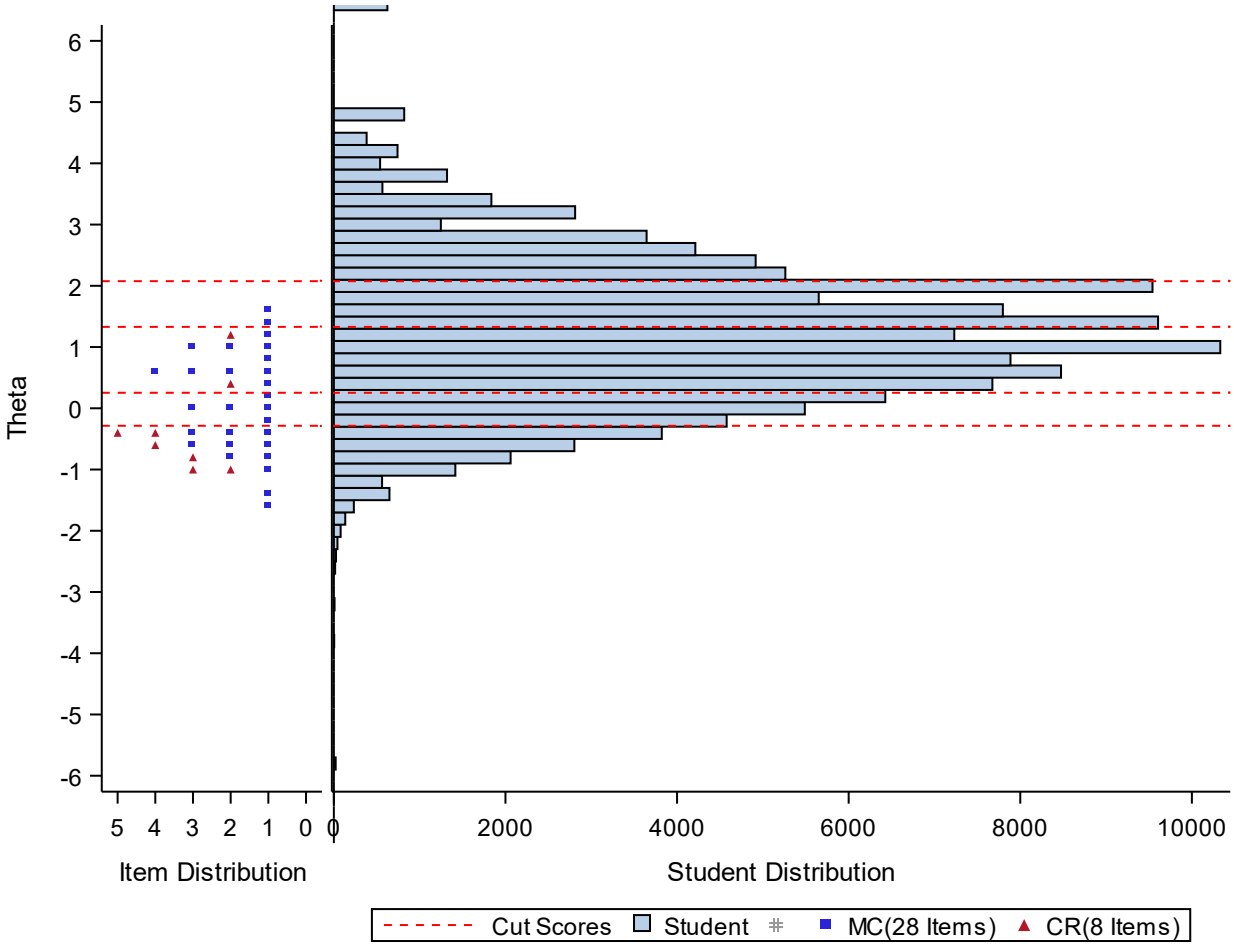


**Figure 2 Student Performance Map: Regents Examination in Global History and Geography II**

## 3.4 CHECKING RASCH ASSUMPTIONS

Since 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. It should be noted that only operational items were analyzed, since they are the basis of student scores.

*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 if 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 decide how many factors exist in principal components. For the parallel analysis, 100 random data sets of sizes equal to the original data were created. For each random data set, a PCA was performed and the resulting eigenvalues stored. Then, for each component, the upper 95th percentile value of the distribution of the 100 eigenvalues from the random data sets was 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 shows the PCA and parallel analysis results for the Regents Examination in Global History and Geography II in June 2019. 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 from 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 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 through 10 are much lower than the first component demonstrates that there is only one dominant component, showing evidence of unidimensionality.

As a rule of thumb, Reckase (1979) proposed that the variance explained by the primary dimension should be greater than 20 percent in order 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 than those from the parallel analysis.

As shown in Figure 3, the primary dimension explained less than 20 percent, at 20.39 percent 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 1.80, and the second value is not significantly different from the parallel analysis. Overall, the PCA suggests that the test is reasonably unidimensional.

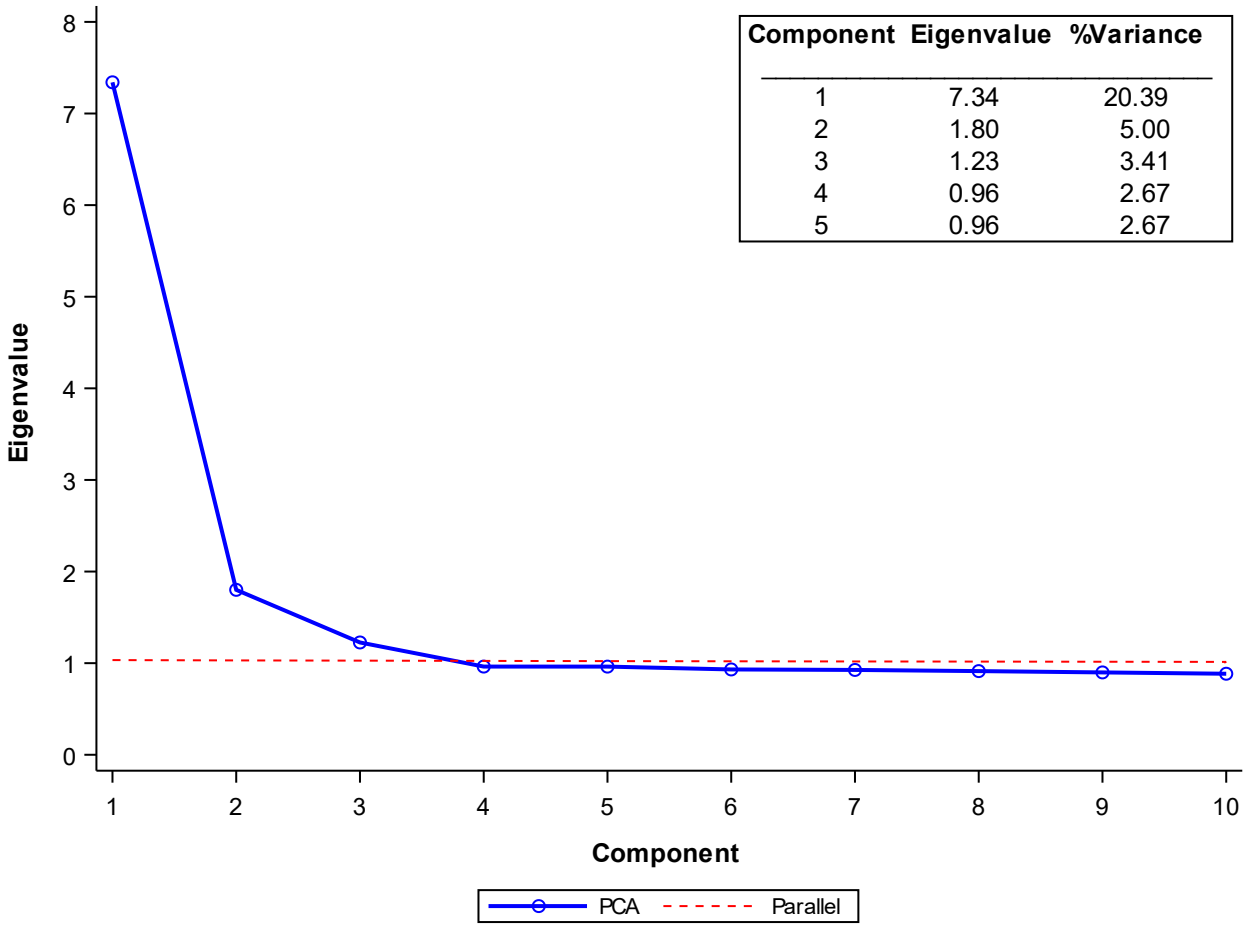


**Figure 3 Scree Plot: Regents Examination in Global History and Geography II**

*Local Independence*

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

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

This formula essentially states that the probability of any pattern of responses across all items (**x**), after conditioning on the examinee's true score ($\theta$) 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 actually 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 the 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:

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

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 ($\theta$) and item parameter estimates. Next, deviations (residuals) between the examinees' 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. It is noted that the raw score residual correlation essentially corresponds to Yen's $Q_3$ index, a popular statistic used to assess local independence. The expected value for the $Q_3$ statistic is approximately $-1/(k-1)$ when no local dependence exists, where $k$ is test length (Yen, 1993). Thus, the expected $Q_3$ values should be approximately −0.01 for the items on the exam. Absolute index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses.

Table 5 shows the summary statistics — mean, standard deviation, minimum, maximum, and several percentiles ($P_{10}$, $P_{25}$, $P_{50}$, $P_{75}$, $P_{90}$) — for all the residual correlations for each test. The total number of item pairs (N) and the number of pairs with the absolute residual correlations greater than 0.20 are also reported in this table. There are three item pairs with an absolute residual correlation greater than 0.20. The mean residual correlation was slightly negative, at −0.02. All residual correlations are small with a maximum absolute value of 0.39, suggesting that local item independence generally holds for the Regents Examination in Global History and Geography II.

**Table 5 Summary of Item Residual Correlations: Regents Examination in Global History and Geography II**

| Statistic Type | Value |
|---|---|
| N | 630 |
| Mean | -0.02 |
| SD | 0.05 |
| Minimum | -0.17 |
| $P_{10}$ | -0.06 |
| $P_{25}$ | -0.04 |
| $P_{50}$ | -0.02 |
| $P_{75}$ | 0.00 |
| $P_{90}$ | 0.01 |
| Maximum | 0.39 |
| >\|0.20\| | 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 ($\theta$) relative to item difficulty.

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 6 presents the summary statistics of INFIT mean square statistics for the Regents Examination in Global History and Geography II, including the number of items, mean, standard deviation, and minimum and maximum values.

The number of items within a targeted range of [0.7, 1.3] is also reported in Table 6. The mean INFIT value is 1.00, with all 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. One of the items is outside of this ideal range. These results indicate that the Rasch model fits the Regents Examination in Global History and Geography II item data well.

**Table 6 Summary of INFIT Mean Square Statistics: Regents Examination in Global History and Geography II**

| | INFIT Mean Square | | | | | |
|---|---|---|---|---|---|---|
| | N | Mean | SD | Min | Max | [0.7, 1.3] |
| Global History and Geography | 36 | 0.99 | 0.10 | 0.86 | 1.32 | [35/36] |

Items for the Regents Examination in Global History and Geography II were field tested in 2017 and 2018, and a separate technical report was produced for each year to document the full test development, scoring, scaling, and data analysis conducted.

## 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 the New York State Education Department 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 map for the June 2019 administration. Note that statistics presented in the test map 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, for the six-point essay item, its parameters can be unstable when estimated across as small a sample as is typically used. Additionally, the six-point essay items are scored by one rater for the field test. On the other hand, for the operational administration, the five-point essay item is scored by two raters and the average score over the two raters is the final score. Thus, it makes the essay item have 11 score points, where the score ranges from 0 to 5 with an increment of 0.5. Therefore, a set of constants is used for these items' parameters on the first operational examination in June 2019.

The Regents Examination in Global History and Geography II has four cut scores, which are set at the scale scores of 55, 65, 79, and 85. Table 7 presents a score range of each performance level and associated performance level descriptors. One of the primary considerations during test construction was to select items so as to 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 0.133. It should be noted that 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 among the items from administration to administration.

**Table 7 Policy Performance Level Descriptors: Regents Examination in Global History and Geography II**

| Performance Level | Scale Score Range | Performance Level Descriptors |
|---|---|---|
| 5 | 85-100 | Level 5: Students performing at Level 5 exceed the expectations of the Framework with distinction for Global History and Geography II. |
| 4 | 79-84 | Level 4: Students performing at Level 4 fully meet the expectations of the Framework for Global History and Geography II. They are likely prepared to succeed in the next level of coursework. |
| 3 | 65-78 | Level 3: Students performing at Level 3 minimally meet the expectations of the Framework for Global History and Geography II. They meet the content area requirements for a Regents diploma but may need additional support to succeed in the next level of coursework. |
| 2 | 55-64 | Level 2: Students performing at Level 2 partially meet the expectations of the Framework for Global History and Geography II. Students with disabilities performing at this level meet the content area requirements for a local diploma but may need additional support to succeed in the next level of coursework |
| 1 | 1-54 | Level 1: Students performing at Level 1 do not demonstrate sufficient knowledge, skills, and practices embodied by the Framework for Global History and Geography II for classification into a performance level. |

The relationship between raw and scale scores is explicated in the scoring tables for each administration. The tables for the June 2019 administration can be found in Appendix B. These tables are the end product of the following scaling procedure.

All Regents Examinations are equated back to a base scale, which is held constant from year to year. Specifically, they are equated to the base scale through the use of 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 examination, the base administration was 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. In addition, a standard setting was held to determine the passing and passing with distinction cut scores in the raw score metric. The standard setting report is available in http://www.p12.nysed.gov/assessment/reports/2019/global-history-2-technical-report-2019.pdf. The scale score points of 55, 65 and 85 were set to correspond to those raw score cuts. A fourth-degree polynomial is required to fit a line exactly to four arbitrary points (e.g., the raw scores corresponding to the four critical scale scores of 0, 55, 65, 85, and 100). The general form of this best-fitting line is:

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

where SS is the scaled score, RS is the raw score, and m0 through m3 are the transformation constants that convert the raw score into the scale score (note that m0 will always be equal to zero in this application, since a raw score of zero corresponds to a scale score of zero). A

subscript for a person on both dependent and independent variables is not included 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 examination. This initial relationship between the raw and scale scores became the base scale.

The minimum (zero) and maximum possible raw scores are assigned scale scores of 0 and 100, respectively. In the event that 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 zero have scale scores that round to zero, their scale scores are instead set equal to one.

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. If no scale score rounds to these critical cuts, then 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). It is 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 examinee score estimates maximize consistency and therefore minimize error or, theoretically speaking, that examinees 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 both influence reliability estimates. Tests with fewer items or with a lack of heterogeneity in scores tend to produce lower reliability estimates. Second, reliability is specifically 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. Of course, systematic error sources also exist.

The remainder of 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
- decision consistency
- 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 ($\rho_X^2$) is defined as the ratio of true score variance ($\sigma_T^2$) to the observed score ($\sigma_X^2$), 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 ($\sigma_E^2$). Put differently, 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 will be 0.0 if none of the test score variances are true. If all test score variances were true, the index would equal 1.0. Such scores

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

would be pure random noise (i.e., all measurement error). 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, it is clear that larger coefficients are more desirable because they indicate that the test scores are less influenced by random error.

*Coefficient Alpha*

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. Note that 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), $\sigma_X^2$ is the variance of the observed total test scores, and $\sigma_{Yi}^2$ 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), discussed further below.

*Traditional Standard Error of Measurement*

The standard error of measurement is defined as the standard deviation of the distribution of observed scores for students with identical true scores. Because the SEM is an index of the random variability in test scores in test score units, it represents important information for test score users.

The SEM formula is provided below.

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

This formula indicates that the value of the SEM depends on both the reliability coefficient (the Coefficient Alpha, as detailed previously) and the standard deviation of test scores. If the reliability were equal to 0.00 (the lowest possible value), the SEM would be equal to the standard deviation of the test scores. If test reliability were equal to 1.00 (the highest possible value), the SEM would be 0.0. In other words, a perfectly reliable test has no measurement error (Harvill, 1991). Additionally, the value of the SEM takes the group variation (i.e., score

standard deviation) into account. Consider that a SEM of 3 on a 10-point test would be very different from a SEM of 3 on a 100-point test.

*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. SEMs help place "reasonable limits" (Gulliksen, 1950) around observed scores through the construction of an approximate score band. Often referred to as confidence intervals, these bands are constructed by taking the observed scores, *X*, and adding and subtracting a multiplicative factor of the SEM. As an example, students with a given true score will have observed scores that fall between ±1 SEM about two-thirds of the time.[4] For ±2 SEM confidence intervals, this increases to approximately 95 percent.

The Coefficient Alpha and associated SEM for the Regents Examination in Global History and Geography II are provided in Table 8.

**Table 8 Reliabilities and Standard Errors of Measurement: Regents Examination in Global History and Geography II**

| Subject | Coefficient Alpha | SEM |
|---|---|---|
| Global History and Geography | 0.88 | 3.11 |

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

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

*Conditional Standard Error of Measurement*

Every time an assessment is administered, the score the student receives 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 true score ($\theta$), 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. Since there is a one-to-one correspondence between the raw score and $\theta$ in the Rasch model, we can apply this concept more generally to all students who obtained a particular raw score and calculate the probability of obtaining each possible raw score, given the students' estimated $\theta$. The standard deviation of this conditional distribution is defined as the conditional standard error of measurement (CSEM). The computer program POLYCSEM (Kolen, 2004) was used to carry out the mechanics of this computation.

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

The relationship between $\theta$ and the scale score is not expressible in a simple mathematical form because it is a blend of the third-degree polynomial relationship between the raw and scale scores and the nonlinear relationship between the expected raw and $\theta$ scores. Additionally, as the exam is equated from year to year, the relationship between the raw and scale scores moves away from the original third-degree polynomial relationship to one that is also no longer expressible in a simple mathematical form. In the absence of a simple mathematical relationship between $\theta$ and the scale scores, the CSEMs that are available for each $\theta$ 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, Zeng, and Hanson (1996) for dichotomously scored items and extended by Wang, Kolen, and Harris (2000) to polytomously scored items. For tests such as the Regents Examination in Global History and Geography that have a one-to-one relationship between raw ($\theta$) 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 $\theta$.

Consider an examinee with a certain performance level. If it were possible to measure this examinee's performance perfectly, without any error, this measure could be called the examinee's "true score," as discussed earlier. This score is equal to the expected raw score. However, whenever an examinee takes a test, the observed test score always includes some level of measurement error. Sometimes, this error is positive, and the examinee achieves a higher score than would be expected, given the examinee's level of $\theta$. Other times, it is negative, and the examinee achieves a lower-than-expected score. If we could give an examinee the same test multiple times and record the observed test scores, the resulting distribution would be the conditional distribution of raw scores for that examinee's level of $\theta$ with a mean value equal to the examinee's expected raw (true) score. The CSEM for that level of $\theta$ 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 $\theta$ is the compound multinomial distribution (Wang et al., 2000). An algorithm to compute this can be found in Hanson (1994) and Thissen, Pommerich, Billeaud, and Williams (1995) and is also implemented in the computer program POLYCSEM (Kolen, 2004). The compound multinomial distribution yields the probabilities that an examinee with a given level of $\theta$ has of achieving 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 $\theta$.

*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 scaled scores through the construction of an approximate score band. The confidence intervals are constructed by adding and subtracting a multiplicative factor of the CSEM.

*Conditional Standard Error of Measurement Characteristics*

The relationship between the scale score CSEM and $\theta$ depends both 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 $\theta$ (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 (2005, p. 357) 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)."

*Results and Observations*

The relationship between raw and scale scores for the Regents Examination 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 first quarter of the scale score range and then are compressed relative to the raw scores for the remaining three-quarters of the range, though there are slight variations. The CSEMs for the Regents Examination can be expected to have inverted-U shaped patterns, with some variations.

Figure 4 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 in Figure 5 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.

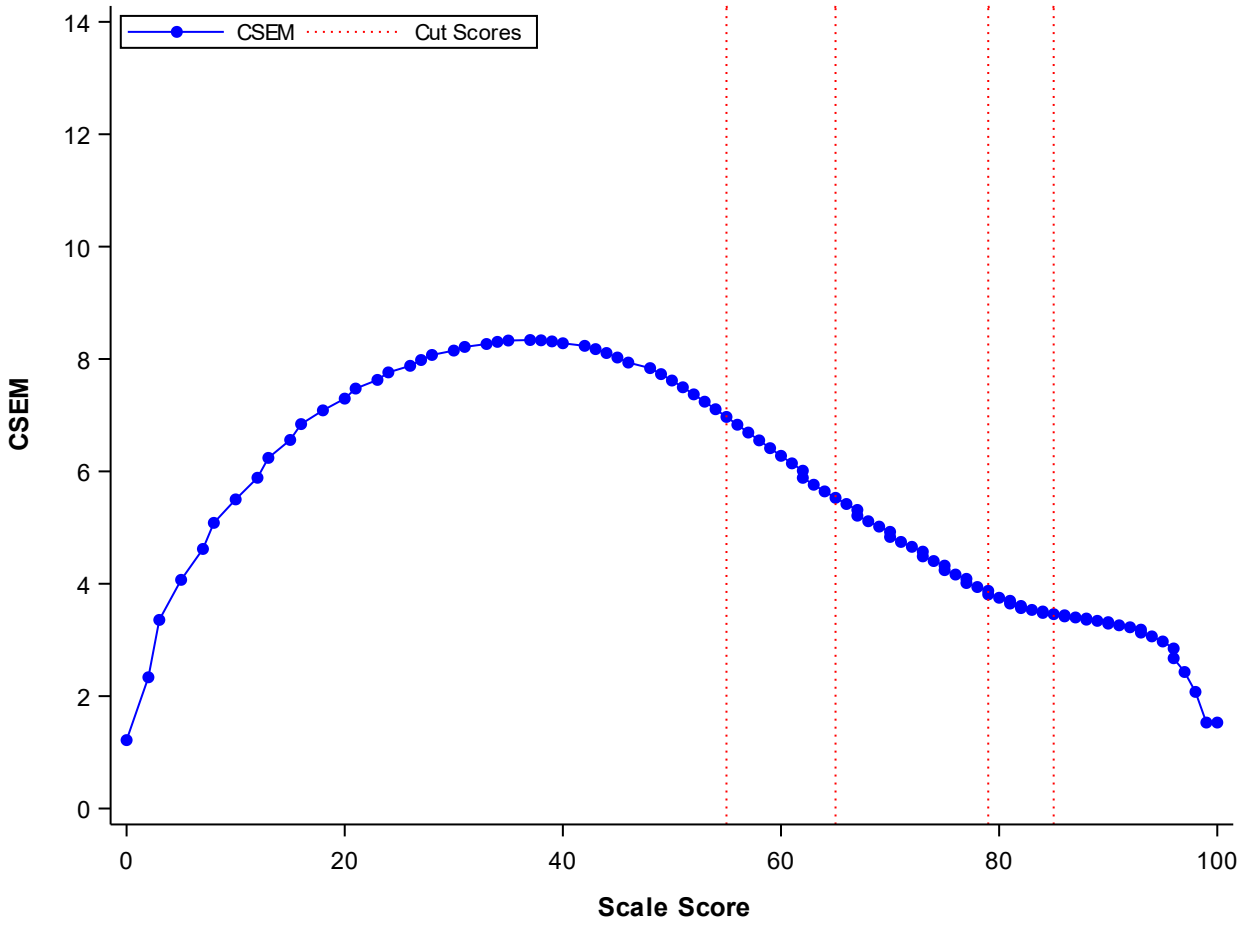


**Figure 4 Conditional Standard Error Plot: Regents Examination in Global History and Geography II**

## 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, which 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 achievement level for each student can be replicated upon retesting using an equivalent form (Huynh, 1976). Decision consistency answers the following question: What is the agreement in classifications between the two non-overlapping, 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 the following tables.

| | | TEST ONE | | |
|---|---|---|---|---|
| | | LEVEL I | LEVEL II | MARGINAL |
| TEST TWO | LEVEL I | φ11 | φ12 | φ1● |
| | LEVEL II | φ21 | φ22 | φ2● |
| | MARGINAL | φ●1 | φ●2 | 1 |

**Figure 5 Pseudo-Decision Table for Two Hypothetical Categories**

| | | TEST ONE | | | | |
|---|---|---|---|---|---|---|
| | | LEVEL I | LEVEL II | LEVEL III | LEVEL IV | MARGINAL |
| TEST TWO | LEVEL I | φ11 | φ12 | φ13 | φ14 | φ1● |
| | LEVEL II | φ21 | φ22 | φ23 | φ24 | φ2● |
| | LEVEL III | φ31 | φ32 | φ33 | φ34 | φ3● |
| | LEVEL IV | φ41 | φ42 | φ43 | φ44 | φ4● |
| | MARGINAL | φ●1 | φ●2 | φ●3 | φ●4 | 1 |

**Figure 6 Pseudo-Decision Table for Four Hypothetical Categories**

If a student is classified as being in one category based on Test One's score, how probable would it be that the student would be reclassified as being in the same category if the student took Test Two (a non-overlapping, 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 two formulas, respectively:

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

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

The sum of the diagonal entries — that is, the proportion of students classified by the two forms into exactly the same achievement level — signifies the overall consistency.

Classification accuracy refers to the agreement of the observed classifications of students with the classifications made on the basis of their true scores. As discussed above, 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 student's true score plus measurement error, or $Observed = True + Error$. Decision accuracy is an index to determine the extent to which measurement error causes a classification different from the one expected from the true score.

Since true scores are unobserved and decision consistency is computed based on a single administration of the Regents Examination in Global History and Geography II, a statistical model using solely data from the available administration is used to estimate the true scores and to project the consistency and accuracy of classifications (Hambleton & Novick, 1973). Although a number of procedures are available, a well-known method developed by Livingston and Lewis (1995) that utilizes 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, since classification and accuracy using 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.

*Results and Observations*

The results for the dichotomies created by the three cut scores are presented in Table 8. For example, the statistics under '2/3' indicate the decision consistency and accuracy when the achievement levels are divided into two categories: one for the second and lower achievement level and the other for the third and higher achievement levels. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. Decision consistency ranged from 0.84 to 0.91, and the decision accuracy ranged from 0.88 to 0.94. Both decision consistency and accuracy values based on individual cut points indicate very good consistency and accuracy of examinee classifications. Refer to Table 9.

**Table 9 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 | 0.84 | 0.85 |
| Accuracy | 0.96 | 0.93 | 0.89 | 0.90 |

Mean scale scores were computed based on reported race/ethnicity, English language learner/multilingual learner status, economically disadvantaged status, gender, and student with a disability status. The results are reported in Table 10.

**Table 10 Group Means: Regents Examination in Global History and Geography II**

| Demographics | Number | Mean Scale Score | SD Scale Score |
|---|---|---|---|
| **All Students*** | 131,481 | 74.65 | 13.82 |
| **Race/Ethnicity** | | | |
| American Indian/Alaska Native | 1,082 | 71.42 | 13.30 |
| Asian/Native Hawaiian/Other Pacific Islander | 15,588 | 80.49 | 11.77 |
| Black/African American | 26,364 | 68.45 | 14.11 |
| Hispanic/Latino | 37,331 | 70.61 | 13.65 |
| Multiracial | 2,220 | 77.50 | 12.40 |
| White | 48,839 | 79.18 | 11.99 |
| **English Language Learner/Multilingual Learner** | | | |
| No | 119,797 | 75.85 | 13.14 |
| Yes | 11,684 | 62.26 | 14.54 |
| **Economically Disadvantaged** | | | |
| No | 57,038 | 79.42 | 12.00 |
| Yes | 74,443 | 70.99 | 14.02 |
| **Gender** | | | |
| Female | 64,873 | 75.36 | 13.27 |
| Male | 66,551 | 73.97 | 14.30 |
| **Student with a Disability** | | | |
| No | 111,385 | 76.67 | 12.67 |
| Yes | 20,096 | 63.46 | 14.60 |

*Note: Fifty-seven students were not reported in the Race/Ethnicity and Gender groups; however, they are reflected in "All Students."

## 4.4 STATE PERCENTILE RANKINGS

State percentile rankings based on scale score distributions are noted in Table 11. The percentiles are based on the distribution of all students taking the Regents Examination in Global History and Geography II. Note that the Regents Examination in Global History and Geography II range from 0 to 100, but 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 11 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 | 55 |
| 2 | 1 | 28 | 1 | 54 | 9 | 80 | 59 |
| 3 | 1 | 29 | 1 | 55 | 10 | 81 | 63 |
| 4 | 1 | 30 | 1 | 56 | 10 | 82 | 67 |
| 5 | 1 | 31 | 1 | 57 | 11 | 83 | 70 |
| 6 | 1 | 32 | 1 | 58 | 12 | 84 | 73 |
| 7 | 1 | 33 | 1 | 59 | 13 | 85 | 77 |
| 8 | 1 | 34 | 1 | 60 | 14 | 86 | 80 |
| 9 | 1 | 35 | 1 | 61 | 15 | 87 | 83 |
| 10 | 1 | 36 | 1 | 62 | 17 | 88 | 85 |
| 11 | 1 | 37 | 2 | 63 | 19 | 89 | 88 |
| 12 | 1 | 38 | 2 | 64 | 20 | 90 | 90 |
| 13 | 1 | 39 | 2 | 65 | 21 | 91 | 92 |
| 14 | 1 | 40 | 2 | 66 | 23 | 92 | 93 |
| 15 | 1 | 41 | 2 | 67 | 25 | 93 | 95 |
| 16 | 1 | 42 | 3 | 68 | 27 | 94 | 96 |
| 17 | 1 | 43 | 3 | 69 | 28 | 95 | 97 |
| 18 | 1 | 44 | 3 | 70 | 31 | 96 | 98 |
| 19 | 1 | 45 | 4 | 71 | 33 | 97 | 99 |
| 20 | 1 | 46 | 4 | 72 | 35 | 98 | 99 |
| 21 | 1 | 47 | 5 | 73 | 38 | 99 | 99 |
| 22 | 1 | 48 | 5 | 74 | 40 | 100 | 99 |
| 23 | 1 | 49 | 5 | 75 | 43 | | |
| 24 | 1 | 50 | 6 | 76 | 46 | | |
| 25 | 1 | 51 | 7 | 77 | 50 | | |

# Chapter 5: Validity (Standard 1)

Restating the purposes and uses of the Regents Examination in Global History and Geography II, this exam measures examinee achievement against the New York State Learning Standards. The exam is prepared by teacher examination committees and New York State Education Department 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 requirements: http://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf. Results of the Regents Examination in Global History and Geography II may also be used to satisfy various locally established requirements throughout the state.

The validity of score interpretations for the Regents Examination in Global History and Geography II is supported by multiple sources of evidence. Chapter 1 of the *Standards for Educational Psychological Testing* (AERA et al., 2014) specifies five sources of validity evidence that are important to gather and document to support validity claims for an assessment:

- test content
- response processes
- internal test structure
- relation to other variables
- consequences of testing

It is important to note that these categories are not mutually exclusive. One source of validity evidence often falls into more than one category, as discussed in more detail in this chapter. Nevertheless, these classifications provide a useful framework within the *Standards* (AERA et al., 2014) for the discussion and documentation of validity evidence, therefore they are used here. 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. The NYS K-12 Social Studies Framework can be found at http://www.p12.nysed.gov/ciai/socst/ssrg.html.

*Content Validity*

Content validity is concerned with the proper definition of the construct and evidence that the test provides an accurate measure of examinee performance within the defined construct. The test blueprint for the Regents Examination in Global History and Geography II is essentially the design document for constructing the exam. It provides an explicit definition of the content domain that is to be represented on the exam. The test development process (discussed in the next section) is in place to ensure, to the extent possible, that the blueprint is met in all operational forms of the exam. Table 12 displays the item types used to assess each standard on the exam.

**Table 7 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% |

*See pages 21–40 for details on item type: http://www.p12.nysed.gov/assessment/ss/hs/framework/ghg2/educator-guide-ghg2-19.pdf

*Item Development Process*

Test development for the Regents Examination in Global History and Geography II 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 New York State 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 New York State-certified educators may participate in this process. The New York State Education Department 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 Examination in Global History and Geography II, 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.

Figure 7 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.

Initial item development is conducted under the criteria and guidance provided by the Department. Both multiple-choice and constructed-response items are included in the Regents Examination in Global History and Geography II to ensure appropriate coverage of the construct domain.

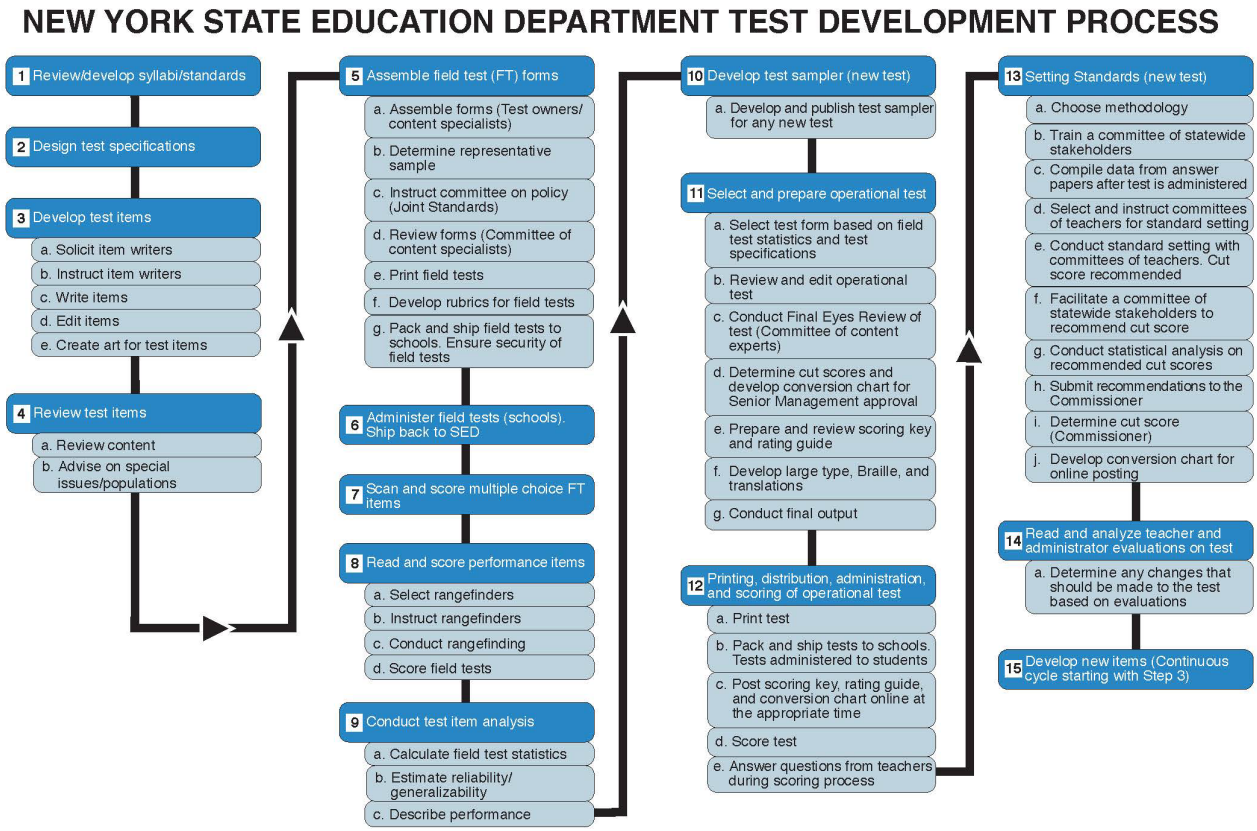


**Figure 7 New York State Education Department Test Development Process**

*Item Review Process*

The item review process helps to ensure 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. This process allows high-quality items to be 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 examinees is fundamental 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 examinee is being asked to demonstrate by responding to each item. Each of these elements of the review process is in place, ultimately, to target fairness for all students by targeting consistency in examinee scores and providing evidence of the validity of their interpretations.

Specifically, the item review process articulates the four major item characteristics that the New York State Education Department 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 (e.g., multiple-choice questions, scaffolding constructed-responses, and 5-point constructed-response questions)

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

The third category of item review, alignment, addresses how each item measures a given standard. This category asks the reviewer to comment on key aspects of how the item addresses and calls for the skills demanded by the standards.

The fourth category addresses the specific demands for different item types and formats. Reviewers evaluate each item to ensure that it conforms to the given requirements. For example, multiple-choice items must have, among other characteristics, one unambiguously correct answer and several plausible, but incorrect, answer choices. Following these reviews, only items that are approved by an assigned educator panel move forward for field testing.

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, the New York State Education Department is committed to ongoing standards review over time and periodically solicits thoughtful, specific responses from stakeholders about individual standards within the NYS P–12 Standards.

## 5.2 EVIDENCE BASED ON RESPONSE PROCESSES

The second source of validity evidence is based on examinee response processes. This standard requires evidence that examinees 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. Accordingly, 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 Examination in Global History and Geography II include the item development process, with attention paid to mitigating the introduction of construct-irrelevant variance. The development process described in the previous sections details the process and attention given to reducing the potential for construct

irrelevance in response processes 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, as well as in the results of statistical analyses, which are covered in the following two sections.

*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, which are contained in the *School Administrator's Manual, Secondary Level Examinations* (http://www.p12.nysed.gov/assessment/manuals/), have been developed and implemented for the New York State Regents testing program. All secondary-level Regents Examinations are administered under these standard conditions to support valid inferences for all students. These standard 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 http://www.p12.nysed.gov/assessment/hsgen/.

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

1) Quality training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, Wong, & Kwong, 2010; Gorman & Rentsch, 2009; Schleicher, Day, Bronston, Mayes, & Riggo, 2002; Woehr & Huffcutt, 1994; Johnson, Penny, & Gordon, 2008; Weigle, 1998)
2) Detection and correction of rating bias (McQueen & Congdon, 1997; Congdon & McQueen, 2000; Myford & Wolfe, 2009; Barkaoui, 2011; Patz, Junker, Johnson, & Mariano, 2002)
3) Consistency or reliability of ratings (Congdon & McQueen, 2000; Harik, Clauser, Grabovsky, Nungester, Swanson, & Nandakumar, 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, Linn, Brennan, & Haertel, 1995; Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000; Smith, 1993; Leacock, Gonzalez, & Conarroe, 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, during which many more items than appear on the operational test are administered to a representative sample of students, a set of "anchor" papers representing student responses across the range of possible responses for constructed-response items is selected. The objective of these "range-finding" efforts is to create a training set for scorer training and execution, the scores from which 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 question. Committee members familiarize themselves with each item type and score a number of responses that are representative of each of the different score points. After the independent scoring is completed, the committee reviews and discusses their results and determines consensus scores for the student responses. During this process, 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 constructed-response items. The final model response sets for the June 2019 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 be aware of and control for sources of variation in scoring. One possible source of variation in constructed-response scores is unintended rater bias associated with items and examinee responses. Because the rater is often unaware of such bias, 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, which 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 effectively controlled by training practices with a strict focus on rubric requirements.

The training process for operational scoring by state educators begins with a review and discussion of actual student work on constructed-response test items. This helps raters understand the range and characteristics typical of examinee responses, as well as the kinds of mistakes that 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 constructed-response 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 fundamentally contributes to the validity of examinee response processes. The rubric specifies what the examinee 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 examinees. To facilitate the development of constructed-response 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 alternative approaches and other legitimate methods.

In support of the goal (valid score interpretations for each examinee) such scoring training procedures are implemented for Regents Examination in Global History and Geography II. Operational raters are selected based on expertise in the exam subject and are assigned a specific set of items to score. No more than approximately one-half of the items on the test are assigned to any one rater. This increases the consistency of scoring across examinee 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 examinees. Additionally, raters are not allowed to score the responses of their own students.

*Statistical Analysis*

One statistic that is useful for evaluating the response processes for multiple-choice 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 examinees are not responding the way in which the item developer intended. As documented in Table 2, the point-biserial statistics for distractors in the multiple-choice items all appear to be very low, indicating that, for the most part, examinees 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. This requires that test developers evaluate the test structure to ensure the test 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 subgroups of examinees 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
- IRT model fit
- test reliability
- classification consistency and accuracy
- test dimensionality

*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 Regents Examination in Global History and Geography II show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values in June 2019 for the Regents Examination in Global History and Geography II range from 0.43 to 0.91, with a mean of 0.71. Furthermore, the point-biserial values (discussed in the following section) for these items indicate that they are discriminating the high performers well. Refer to Chapter 2 of this report for additional details.

*Item Discrimination*

How well the items on a test discriminate between high- and low-performing examinees is an important measure of the structure of a test. Items that do not discriminate well generally provide less reliable information about student performance. Tables 2 and 3 provide point-biserial values on the correct responses, and Table 2 also provides point-biserial values on the three distractors. The values for correct answers are 0.27 or higher for all items; and for all distracters, they are negative or close to zero, indicating that examinees are responding to the items as expected during item and rubric development.

*Differential Item Functioning*

Differential item functioning (DIF) was conducted for gender, race/ethnicity, needs/resource capacity (NRC) categories, and ELL/MLL status based on the data for the June 2019 administration. DIF data is only available after the administration because 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 of these subgroups. The Mantel-Haenszel$\chi^2$ 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 our 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.

*IRT Model Fit*

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the Regents Examination in Global History and Geography II 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 5. The mean INFIT value is 1.00, with 35 of 36 items falling in a targeted range of [0.7, 1.3]. These results indicate that the Rasch model fits the Regents Examination in Global History and Geography II item data well.

*Test Reliability*

As discussed, 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 examinee score estimates maximize consistency and therefore minimize error or, theoretically speaking, that examinees who take a test multiple times would get the same score each time. The reliability estimate for the Transition Regents Examination in Global History and Geography in June 2019 is 0.88, showing high reliability of examinee scores. Refer to Chapter 4 of this report for additional details.

*Classification Consistency and Accuracy*

A decision consistency analysis measures the agreement between the classifications based on two non-overlapping, 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. Decision accuracy is an index to determine 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 and the Regents Examination in Global History and Geography II, both decision consistency and accuracy values are high, indicating very good consistency and accuracy of examinee classifications. The decision consistency ranged from 0.84 to 0.94, and the decision accuracy ranged from 0.89 to 0.96 in June 2019. For the Regents Examination in Global History and Geography II, both decision consistency and accuracy values for all three cut points are high, indicating very good consistency and accuracy of examinee classifications.

*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 assumption might suggest that the test is measuring something other than the intended content and indicate that the quality of the test structure is compromised. A principal components analysis was conducted to test the assumption of unidimensionality, and the results provide strong evidence that a single dimension in the Regents Examination in Global History and Geography II 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 examinee performance.

## 5.4 EVIDENCE BASED ON RELATION 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, these claims could be supported by providing evidence that performance on this 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 examinee score data for other tests measuring the same content as the Regents Examination in Global History and Geography II are ideal, but 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 New York State educators, deeply familiar with both 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 in regard to 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 examinee scores on other tests is not directly supported by either the stated purposes or the development process and research conducted on examinee data. A brief survey of web sites of New York State 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, 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, and to assess their appropriateness for the inferences being made about course placement.

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

Barkaoui, Khaled. (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).* Iowa City, IA: 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) Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995, Summer). Generalizability analysis for educational assessments. Los Angeles, CA: 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). Princeton, NJ: 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*. New York: 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 Park, CA: Sage Publications.

Hanson, B. A. (1994). Extension of Lord-Wingersky algorithm to computing test scores for polytomous items. Retrieved February 17, 2016 from http://www.b-a-h.com/papers/note9401.pdf.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009, Spring). 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*. New York, NY: The Guilford Press.

Kolen, M. J. (2004). POLYCSEM [Computer program]. University of Iowa. Retrieved August 1, 2012, from https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs.

Kolen, M. J., & Brennan, R. L. (2005). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: 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, Claudia, Gonzalez, Erin, Conarroe, Mike. (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]. Chicago: 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*. Hillsdale, NJ: 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), 200-15.

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.

Patz, R. J., Junker, B. W., Johnson, M. S., and 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 Wei, R. R. (2007). Performance assessment for California teachers: Summary of validity and reliability studies for the 2003−04 pilot year. Palo Alto, CA: 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.* New York, NY: 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*. New York, NY: 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 verses 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.* Princeton, NJ: 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 Maps

### Table A.1 Test Map for June 2019 Administration

| Position | Item Type | Max Points | Weight | Standard | Key Idea | Mean | Point-Biserial | Rasch Difficulty | INFIT |
|---|---|---|---|---|---|---|---|---|---|
| 1 | MC | 1 | 1 | 2 | 10.1 | 0.76 | 0.52 | -1.3618 | 0.91 |
| 2 | MC | 1 | 1 | 5 | 10.1 | 0.48 | 0.37 | 0.3159 | 1.17 |
| 3 | MC | 1 | 1 | 5 | 10.2 | 0.54 | 0.41 | -0.0411 | 1.13 |
| 4 | MC | 1 | 1 | 2 | 10.2 | 0.74 | 0.40 | -1.2169 | 1.09 |
| 5 | MC | 1 | 1 | 2 | 10.3 | 0.57 | 0.61 | -0.2081 | 0.85 |
| 6 | MC | 1 | 1 | 2 | 10.3 | 0.58 | 0.43 | -0.2347 | 1.10 |
| 7 | MC | 1 | 1 | 2 | 10.5 | 0.67 | 0.60 | -0.7488 | 0.86 |
| 8 | MC | 1 | 1 | 2 | 10.5 | 0.77 | 0.61 | -1.4195 | 0.82 |
| 9 | MC | 1 | 1 | 2 | 10.5 | 0.48 | 0.46 | 0.2897 | 1.03 |
| 10 | MC | 1 | 1 | 5 | 10.5 | 0.39 | 0.42 | 0.7677 | 1.06 |
| 11 | MC | 1 | 1 | 2 | 10.6 | 0.55 | 0.39 | -0.5455 | 1.14 |
| 12 | MC | 1 | 1 | 3 | 10.6 | 0.67 | 0.53 | -1.2080 | 0.92 |
| 13 | MC | 1 | 1 | 2 | 10.6 | 0.68 | 0.60 | -1.2267 | 0.84 |
| 14 | MC | 1 | 1 | 5 | 10.7 | 0.76 | 0.48 | -1.7596 | 0.94 |
| 15 | MC | 1 | 1 | 2 | 10.7 | 0.70 | 0.52 | -1.3503 | 0.92 |
| 16 | MC | 1 | 1 | 2 | 10.7 | 0.65 | 0.52 | -1.0701 | 0.93 |
| 17 | MC | 1 | 1 | 2 | 10.7 | 0.64 | 0.51 | -1.0340 | 0.96 |
| 18 | MC | 1 | 1 | 2 | 10.8 | 0.42 | 0.32 | 0.1659 | 1.22 |
| 19 | MC | 1 | 1 | 5 | 10.8 | 0.47 | 0.41 | -0.0942 | 1.12 |
| 20 | MC | 1 | 1 | 3 | 10.6 | 0.40 | 0.31 | 0.4210 | 1.24 |
| 21 | MC | 1 | 1 | 2 | 10.9 | 0.70 | 0.52 | -1.2287 | 0.94 |
| 22 | MC | 1 | 1 | 2 | 10.9 | 0.67 | 0.60 | -1.0605 | 0.85 |
| 23 | MC | 1 | 1 | 2 | 10.10 | 0.70 | 0.59 | -1.2575 | 0.84 |
| 24 | MC | 1 | 1 | 2 | 10.10 | 0.51 | 0.39 | -0.1959 | 1.16 |
| 25 | MC | 1 | 1 | 2 | CT | 0.74 | 0.48 | -1.4752 | 0.97 |
| 26 | MC | 1 | 1 | 2 | CT | 0.53 | 0.51 | -0.3137 | 0.98 |
| 27 | MC | 1 | 1 | 5 | 10.10 | 0.61 | 0.58 | -0.7312 | 0.89 |
| 28 | MC | 1 | 1 | 4 | 10.10 | 0.53 | 0.43 | -0.2800 | 1.11 |
| 29 | CRQ | 1 | 1 | 2 | 10.4 | 0.49 | 0.55 | -0.1246 | 0.97 |
| 30 | CRQ | 1 | 1 | 2, 4 | 10.7 | 0.65 | 0.57 | -1.0353 | 0.93 |
| 31 | CRQ | 1 | 1 | 2, 4, 5 | CT | 0.44 | 0.59 | 0.1716 | 0.91 |
| 32 | CRQ | 1 | 1 | 2, 4, 5 | 10.7 | 0.40 | 0.44 | 0.6172 | 1.12 |
| 33 | CRQ | 1 | 1 | 2, 4, 5 | 10.7 | 0.40 | 0.57 | 0.5805 | 0.92 |
| 34a | CRQ | 1 | 1 | 2, 4, 5 | 10.7 | 0.49 | 0.57 | 0.1220 | 0.92 |
| 34b | CRQ | 1 | 1 | 2, 4, 5 | 10.7 | 0.42 | 0.62 | 0.4715 | 0.85 |
| 35 | ESS | 5 | 3 | 2, 3, 4, 5 | CT | 1.09 | 0.73 | 1.3179 | 0.87 |

# Appendix B: Raw-to-Theta-to-Scale Score Conversion Tables

### Table B.1 Score Table for June 2019 Administration

| Raw Score | Ability | Scale Score |
|---|---|---|
| 0.0 | -5.7848 | 0.000 |
| 0.5 | -4.5689 | 1.718 |
| 1.0 | -3.8590 | 3.422 |
| 1.5 | -3.4372 | 5.110 |
| 2.0 | -3.1336 | 6.783 |
| 2.5 | -2.8950 | 8.440 |
| 3.0 | -2.6976 | 10.082 |
| 3.5 | -2.5289 | 11.707 |
| 4.0 | -2.3812 | 13.315 |
| 4.5 | -2.2496 | 14.907 |
| 5.0 | -2.1307 | 16.481 |
| 5.5 | -2.0223 | 18.037 |
| 6.0 | -1.9225 | 19.576 |
| 6.5 | -1.8299 | 21.097 |
| 7.0 | -1.7433 | 22.600 |
| 7.5 | -1.6621 | 24.085 |
| 8.0 | -1.5853 | 25.551 |
| 8.5 | -1.5125 | 26.998 |
| 9.0 | -1.4432 | 28.426 |
| 9.5 | -1.3769 | 29.836 |
| 10.0 | -1.3132 | 31.226 |
| 10.5 | -1.2518 | 32.598 |
| 11.0 | -1.1925 | 33.950 |
| 11.5 | -1.1350 | 35.282 |
| 12.0 | -1.0791 | 36.596 |
| 12.5 | -1.0245 | 37.890 |
| 13.0 | -0.9712 | 39.165 |
| 13.5 | -0.9189 | 40.420 |
| 14.0 | -0.8676 | 41.656 |
| 14.5 | -0.8170 | 42.872 |
| 15.0 | -0.7671 | 44.069 |
| 15.5 | -0.7177 | 45.247 |
| 16.0 | -0.6688 | 46.406 |
| 16.5 | -0.6203 | 47.545 |
| 17.0 | -0.5721 | 48.666 |
| 17.5 | -0.5241 | 49.768 |
| 18.0 | -0.4763 | 50.851 |
| 18.5 | -0.4285 | 51.916 |
| 19.0 | -0.3809 | 52.962 |
| 19.5 | -0.3332 | 53.990 |
| 20.0 | -0.2854 | 55.000 |
| 20.5 | -0.2375 | 55.992 |
| 21.0 | -0.1896 | 56.967 |
| 21.5 | -0.1414 | 57.925 |
| 22.0 | -0.0931 | 58.865 |
| 22.5 | -0.0445 | 59.789 |
| 23.0 | 0.0044 | 60.697 |
| 23.5 | 0.0536 | 61.588 |
| 24.0 | 0.1030 | 62.464 |
| 24.5 | 0.1528 | 63.324 |
| 25.0 | 0.2030 | 64.169 |
| 25.5 | 0.2536 | 65.000 |
| 26.0 | 0.3045 | 65.816 |
| 26.5 | 0.3559 | 66.619 |
| 27.0 | 0.4078 | 67.408 |
| 27.5 | 0.4601 | 68.184 |
| 28.0 | 0.5130 | 68.948 |
| 28.5 | 0.5664 | 69.699 |
| 29.0 | 0.6203 | 70.439 |
| 29.5 | 0.6749 | 71.168 |
| 30.0 | 0.7301 | 71.887 |
| 30.5 | 0.7861 | 72.595 |
| 31.0 | 0.8427 | 73.295 |
| 31.5 | 0.9001 | 73.985 |
| 32.0 | 0.9584 | 74.667 |
| 32.5 | 1.0175 | 75.341 |
| 33.0 | 1.0777 | 76.009 |
| 33.5 | 1.1389 | 76.670 |
| 34.0 | 1.2012 | 77.325 |
| 34.5 | 1.2648 | 77.975 |
| 35.0 | 1.3296 | 78.621 |
| 35.5 | 1.3959 | 79.263 |
| 36.0 | 1.4637 | 79.902 |
| 36.5 | 1.5332 | 80.539 |
| 37.0 | 1.6044 | 81.175 |
| 37.5 | 1.6775 | 81.810 |
| 38.0 | 1.7525 | 82.444 |
| 38.5 | 1.8297 | 83.080 |
| 39.0 | 1.9093 | 83.717 |
| 39.5 | 1.9912 | 84.357 |
| 40.0 | 2.0756 | 85.000 |
| 40.5 | 2.1628 | 85.647 |
| 41.0 | 2.2530 | 86.300 |
| 41.5 | 2.3463 | 86.958 |
| 42.0 | 2.4430 | 87.623 |
| 42.5 | 2.5435 | 88.296 |
| 43.0 | 2.6481 | 88.978 |
| 43.5 | 2.7574 | 89.669 |
| 44.0 | 2.8721 | 90.371 |
| 44.5 | 2.9927 | 91.085 |
| 45.0 | 3.1203 | 91.811 |
| 45.5 | 3.2560 | 92.551 |
| 46.0 | 3.4014 | 93.305 |
| 46.5 | 3.5587 | 94.076 |
| 47.0 | 3.7312 | 94.863 |
| 47.5 | 3.9246 | 95.668 |
| 48.0 | 4.1489 | 96.492 |
| 48.5 | 4.4236 | 97.336 |
| 49.0 | 4.7954 | 98.201 |
| 49.5 | 5.4206 | 99.089 |
| 50.0 | 6.5383 | 100.000 |

# Appendix C: Item Writing Guidelines

### GENERAL RULES 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 Hints 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 HINTS

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

1. Cover up the responses and read just the stem.

2. 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, there may be a problem with the stem.

3. 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.