New York State Regents Examination in United States History

2015 Technical Report

Prepared for the New York State Department of Education

by

Data Recognition Corporation

April 2016

Developed and published under contract with the New York State Education Department by Data Recognition Corporation.
Copyright © 2016 by the New York State Education Department.

Contents

CHAPTER 1: INTRODUCTION **1**

1.1 Introduction 1
1.2 Purposes of the Exam 1
1.3 Target Population (Standard 7.2) 1

CHAPTER 2: CLASSICAL ITEM STATISTICS (STANDARD 4.10) **3**

2.1 Item Difficulty 3
2.2 Item Discrimination 3
2.3 Discrimination on Difficulty Scatterplots 6
2.4 Observations and Interpretations 7

CHAPTER 3: IRT CALIBRATIONS, EQUATING, AND SCALING (STANDARDS 2, AND 4.10) **8**

3.1 Description of the Rasch Model 8
3.2 Software and Estimation Algorithm 9
3.3 Characteristics of the Testing Population 9
3.4. Item Difficulty-Student Performance Maps 9
3.5 Checking Rasch Assumptions 9
Unidimensionality *10*
Local Independence *11*
Item Fit *13*

CHAPTER 4: RELIABILITY (STANDARD 2) **14**

4.1 Reliability Indices (Standard 2.20) 14
Coefficient Alpha *15*
4.2 Standard Error of Measurement (Standards 2.13, 2.14, 2.15) 15
Traditional Standard Error of Measurement *15*
Conditional Standard Error of Measurement *16*
Results and Observations *18*
4.3 Decision Consistency and Accuracy (Standard 2.16) 18
4.4 Group Means (Standard 2.17) 20
4.5 State Percentile Rankings 21

CHAPTER 5: VALIDITY (STANDARD 1) **23**

5.1 Evidence Based on Test Content *23*
5.2 Evidence Based on Response Processes *26*
5.3 Evidence Based on Internal Structure *29*
5.4 Evidence Based on Relations to Other Variables *31*
5.5 Evidence Based on Testing Consequences *31*

REFERENCES **33**

APPENDIX A – ITEM WRITING GUIDELINES **37**

APPENDIX B – TABLES AND FIGURES FOR AUGUST 2014 ADMINISTRATION **40**

APPENDIX C – TABLES AND FIGURES FOR JANUARY 2015 ADMINISTRATION **46**

List of Tables

Table 1 Total Examinee Population: Regents Examination in United States History 2
Table 2 Multiple-Choice Item Analysis Summary: Regents Examination in United States History 4
Table 3 Constructed-Response Item Analysis Summary: Regents Examination in United States History 6
Table 4 Summary of Item Residual Correlations: Regents Examination in United States History 12
Table 5 Summary of Infit Mean Square Statistics: Regents Examination in United States Histroy 13
Table 6 Reliabilities and Standard Errors of Measurement: Regents Examination in United States History 16
Table 7 Decision Consistency and Accuracy Results: Regents Examination in United States History 20
Table 8 Group Means: Regents Examination in United States History 21
Table 9 State Percentile Ranking for Raw Score – Regents Examination in United States History 22

List of Figures

Figure 1 Scatterplot: Regents Examination in United States History 7
Figure 2 Student Performance Map: Regents Examination in United States History 9
Figure 3 Scree Plots: Regents Examination in United States History 11
Figure 4 Conditional Standard Error Plots: Regents Examination in United States History 18
Figure 5 Pseudo-Decision Table for Two Hypothetical Categories 19
Figure 6 Pseudo-Decision Table for Four Hypothetical Categories 19
Figure 7 New York State Education Department Test Development Process 25

Chapter 1: Introduction

1.1 Introduction

This technical report for the Regents Examination in United States History will provide the state of New York with documentation on the purpose of the Regents Examination, scoring information, evidence of both reliability and validity of the exams, scaling information, and guidelines and reporting information for the August 2014, January 2015, and June 2015 administrations. Chapters 1-5 detail results for the June administration. Results for the January and August administrations are provided in Appendices B and C. 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, equating, etc.) are thoroughly addressed and referenced in supporting documents.

The Regents Examination in United States History is given to students enrolled in New York State schools in June, August, and January. The examination is based on the United States History Core Curriculum which is based on standards 1 2, and 4 in the New York State Learning Standards for Social Studies.

1.2 Purposes of the Exam

The Regents Examination in United States History measures examinee achievement against New York State's (NYS) learning standards. The exam is prepared by teacher examination committees and New York State Department of Education subject and testing specialists and 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 exams also provide 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 United States History is intended for use in satisfying State testing requirements for students who have finished a course of instruction in United States History. 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.p12.nysed.gov/ciai/gradreq/2015GradReq11-15.pdf. Results of the Regents Examination in United States History may also be used to satisfy various locally established requirements throughout the State.

1.3 Target Population (Standard 7.2)

The examinee population for the Regents Examination in United States History is composed of students who have completed a course in United States History.

Table 1 provides a demographic breakdown of all students who took the August 2014, January 2015, and June 2016 Regents Examination in United States History. All analyses in this report are based on

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

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 those exams administered August (2014), January, and June of the reporting year (see http://data.nysed.gov/). If a student takes the same exam multiple times in the year, the highest score only is included in these results. Item-level data used for the analyses in this report are reported by districts on a similar timeline, but through a different collection system. These data include all student results for each administration. Therefore, the n-sizes in this technical report will differ from publically reported counts of student test-takers. Tables 2 – 10 and Figures 1- 7 detail analysis results for the June administration. Corresponding tables and figures for the January and August administrations are included in Appendices B and C.

Table 1 Total Examinee Population: Regents Examination in United States History

	August Admin*		January Admin**		June Admin***	
Demographics	**Number**	**Percent**	**Number**	**Percent**	**Number**	**Percent**
All Students	16429	100	25268	100	176825	100
Race/Ethnicity						
American Indian or Alaska Native	123	0.75	173	0.68	973	0.55
Asian/Native Hawaiian/Other Pacific Islander	1040	6.33	1851	7.33	17548	9.92
Black or African American	5684	34.60	8362	33.09	33631	19.02
Hispanic or Latino	5829	35.48	8913	35.27	40034	22.64
Multiracial	88	0.54	177	0.70	1943	1.10
White	3659	22.27	5788	22.91	82692	46.76
English Proficiency						
No	14417	87.75	21995	87.05	167920	94.96
Yes	2012	12.25	3273	12.95	8905	5.04
Economically Disadvantaged						
No	6077	36.99	7960	31.50	94192	53.27
Yes	10352	63.01	17308	68.50	82633	46.73
Gender						
Female	8580	52.24	12430	49.20	88884	50.27
Male	7843	47.76	12834	50.80	87937	49.73
Student with Disabilities						
No	12747	77.59	19160	75.83	154079	87.14
Yes	3682	22.41	6108	24.17	22746	12.86

*Note: 6 students were not reported in the Ethnicity and Gender group but they are reflected in "All Students".
**Note: 4 students were not reported in the Ethnicity and Gender group but they are reflected in "All Students".
***Note: 4 students were not reported in the Ethnicity and Gender group but 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 only to the operational Regents Examination in United States History 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, 1 = right). 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 multiple-choice item has a *p*-value of 0.89, it means that 89 percent of the students answered the item correctly. Additionally, this value might also suggest that the item was relatively easy and/or 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 Regents Examination in United States History show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values range from .30 to .89, with a mean of .61.

Refer to Tables 2 and 3 for item-by-item *p*-values for multiple-choice and constructed-response items respectively.

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 item. It is expected that students who perform well on the Regents Examination in United States History would be more likely to answer any given item correctly, while low-performing students (i.e., those who perform poorly on the exam overall) would be more likely to answer the same item incorrectly. Pearson's product-moment correlation coefficient (also commonly referred to as a point biserial correlation) between item scores and test scores is used

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

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 right 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 the item is likely discriminating well between high- and low-performing students. Point biserials 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.

Refer to Tables 2 and 3 for point biserial values on the correct response and three distractors (Table 2 only). The values for correct answers are .12 or higher for all items, indicating that the items are discriminating well between high- and low-performing examinees. Point biserials for all item distractors are negative, zero, or very close to zero 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 United States History

Item	Number of Students	*p* Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
1	176825	.89	.31	.25	-.13	-.20	-.12
2	176825	.69	.46	.43	-.14	-.23	-.29
3	176825	.52	.50	.43	-.22	-.21	-.22
4	176825	.79	.40	.37	-.30	-.17	-.14
5	176825	.80	.40	.40	-.27	-.13	-.27
6	176825	.64	.48	.47	-.32	-.20	-.16
7	176825	.88	.33	.31	-.19	-.15	-.17
8	176825	.72	.45	.26	-.13	-.29	.00
9	176825	.60	.49	.47	-.25	-.23	-.20
10	176825	.70	.46	.38	-.12	-.26	-.18
11	176825	.79	.41	.41	-.22	-.22	-.24
12	176825	.79	.41	.18	-.25	.01	-.12
13	176825	.75	.43	.40	-.26	-.16	-.21
14	176825	.82	.39	.30	-.18	-.15	-.18
15	176825	.70	.46	.53	-.28	-.31	-.24
16	176825	.59	.49	.51	-.35	-.28	-.08
17	176825	.60	.49	.45	-.13	-.16	-.34
18	176825	.67	.47	.34	-.05	-.22	-.21
19	176825	.83	.38	.49	-.25	-.27	-.30
20	176825	.76	.43	.37	-.12	-.28	-.24
21	176825	.83	.38	.43	-.18	-.27	-.25
22	176825	.61	.49	.12	-.13	.11	-.24

Item	Number of Students	*p* Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
23	176825	.79	.41	.42	-.19	-.22	-.26
24	176825	.72	.45	.33	-.22	-.14	-.21
25	176825	.69	.46	.36	-.29	-.19	-.05
26	176825	.83	.38	.43	-.22	-.17	-.31
27	176825	.62	.48	.29	-.13	-.14	-.16
28	176825	.82	.38	.49	-.33	-.22	-.25
29	176825	.65	.48	.48	-.20	-.26	-.29
30	176825	.88	.32	.46	-.27	-.24	-.24
31	176825	.80	.40	.46	-.28	-.26	-.18
32	176825	.81	.39	.43	-.15	-.25	-.28
33	176825	.61	.49	.31	-.10	-.26	-.10
34	176825	.63	.48	.30	-.25	-.06	-.15
35	176825	.82	.38	.47	-.17	-.27	-.30
36	176825	.80	.40	.43	-.20	-.19	-.29
37	176825	.87	.34	.43	-.18	-.23	-.30
38	176825	.76	.43	.42	-.19	-.15	-.29
39	176825	.73	.44	.46	-.23	-.25	-.23
40	176825	.66	.47	.37	-.19	-.22	-.14
41	176825	.69	.46	.49	-.20	-.27	-.26
42	176825	.76	.43	.52	-.28	-.26	-.27
43	176825	.90	.30	.38	-.25	-.22	-.15
44	176825	.70	.46	.36	-.17	-.19	-.20
45	176825	.71	.45	.39	-.24	-.20	-.20
46	176825	.61	.49	.45	-.30	-.16	-.19
47	176825	.81	.39	.50	-.30	-.29	-.21
48	176825	.80	.40	.37	-.23	-.19	-.20
49	176825	.77	.42	.43	-.27	-.19	-.21
50	176825	.64	.48	.44	-.17	-.22	-.27

Table 3 Constructed-Response Item Analysis Summary: Regents Examination in United States History

Item	Min. score	Max. score	Number of Students	Mean	SD	*p* Value	Point Biserial
51	0	5	176825	2.38	1.21	.48	.60
52	0	1	176825	.93	.26	.93	.23
53	0	2	176825	1.91	.33	.96	.29
54	0	2	176825	1.85	.41	.93	.30
55	0	1	176825	.92	.27	.92	.26
56	0	1	176825	.95	.23	.95	.25
57	0	1	176825	.96	.21	.96	.28
58	0	1	176825	.93	.25	.93	.29
59	0	1	176825	.96	.19	.96	.22
60	0	2	176825	1.91	.35	.95	.28
61	0	1	176825	.92	.26	.92	.29
62	0	5	176825	2.71	1.04	.54	.59

2.3 Discrimination on Difficulty Scatterplots

Figure 1 shows a scatterplot of item discrimination values (*y*-axis) and item difficulty values (*x*-axis). The distributions of *p*-value and point biserials are also included in the graphic to illustrate the mean, median, total range, and quartile ranges for each.

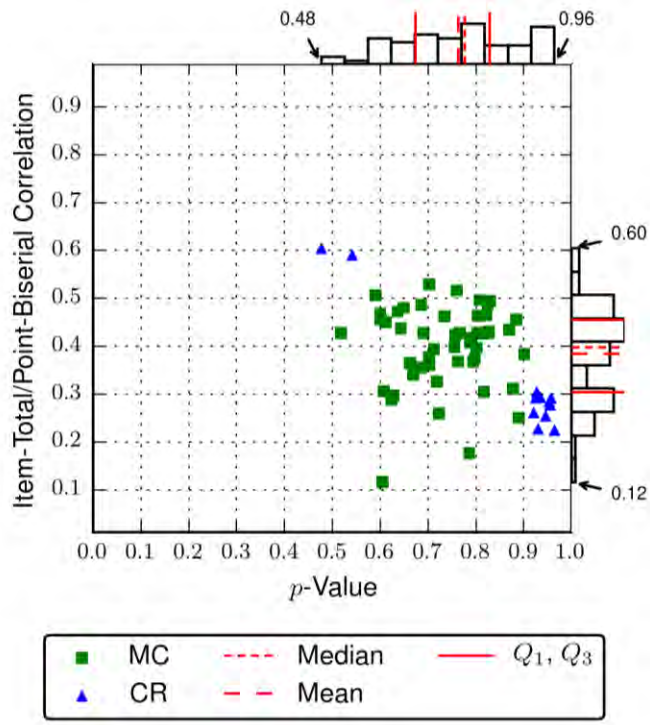

Figure 1 Scatterplot: Regents Examination in United States History

2.4 Observations and Interpretations

The mean *p*-values for the MC items ranged from about 0.52 to 0.90, while the mean proportion-correct values for the CR items (Table 3) ranged from about 0.48 to 0.96. 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 Regents Examination in United States History is based on the work of Georg Rasch (Rasch, 1960). The Rasch model has a long-standing presence in applied testing programs. IRT has several advantages over classical test theory and has become the standard procedure for analyzing item response data in large-scale assessments. According to Van der Linden and Hambleton (1997), "The central feature of IRT is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying ability." Ability in this sense can be thought of as performance on the test and is defined as "the expected value of observed performance on the test of interest" (Hambleton, Swaminathan, and Roger, 1991). This performance value is often referred to as θ. Performance and θ 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 United States History 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 and 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 θ_n represents examinee ability, and D_{ij} is the step difficulty of the jth step on item i. For dichotomous MC items, the RPCM 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:

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

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 2015 computer program (Wright and Linacre, 2015), which employs unconditional (UCON), joint maximum likelihood estimation (JMLE).

3.3 Characteristics of the Testing Population

The data analyses reported here are based on all students who took the Regents Examination in United States History in June 2014. The characteristics of this population are provided in Table 1.

3.4. Item Difficulty-Student Performance Maps

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.

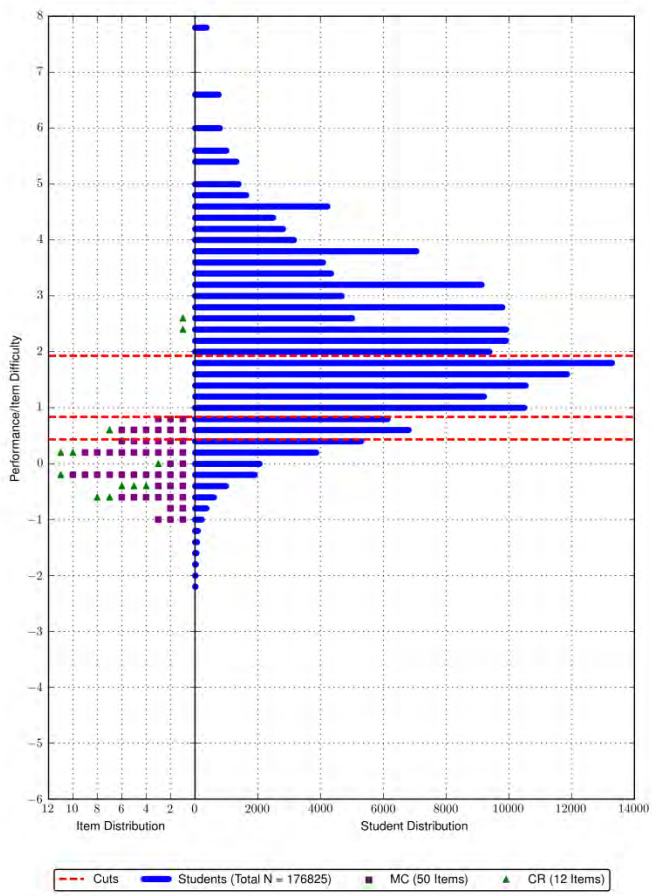

Figure 2 Student Performance Map: Regents Examination in United States History

3.5 Checking Rasch Assumptions

Since the Rasch model was the basis of all calibration, scoring, and scaling analyses associated with the Regents Examination in United States History , 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 student performance. Principal Components Analysis (PCA) can be used to assess the unidimensionality assumption. The purpose of the analysis is to verify whether any other dominant components exist among the items. If any other dimensions are found, the unidimensionality assumption would be violated.

A parallel analysis (Horn, 1965) can be further helpful to help distinguish components that are real from components that are random. Parallel analysis is a technique 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.

Figur 3 shows the PCA and parallel analysis results for the Regents Examination in English Language Arts (Common Core). 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 PCA is same as the total number of items in a test, Figure 3 shows only 10 components. This view is sufficient for interpretation because components are listed in descending eigenvalue order. The lower eigenvalues from components 2 through 10 demonstrates that components beyond 1 are not individually contributing to the explanation of variance in the data.

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

As shown in Figure 3, the primary dimension explained less than 20 percent, but only slightly so at 18.3 percent of the total variance for the Regents Examination in English Language Arts (Common Core). The eigenvalue of the second dimension is less than one third of the first at 2.5, and the second value is not significantly different from the parallel analysis. Overall, the PCA suggests that the test is reasonably unidimensional.

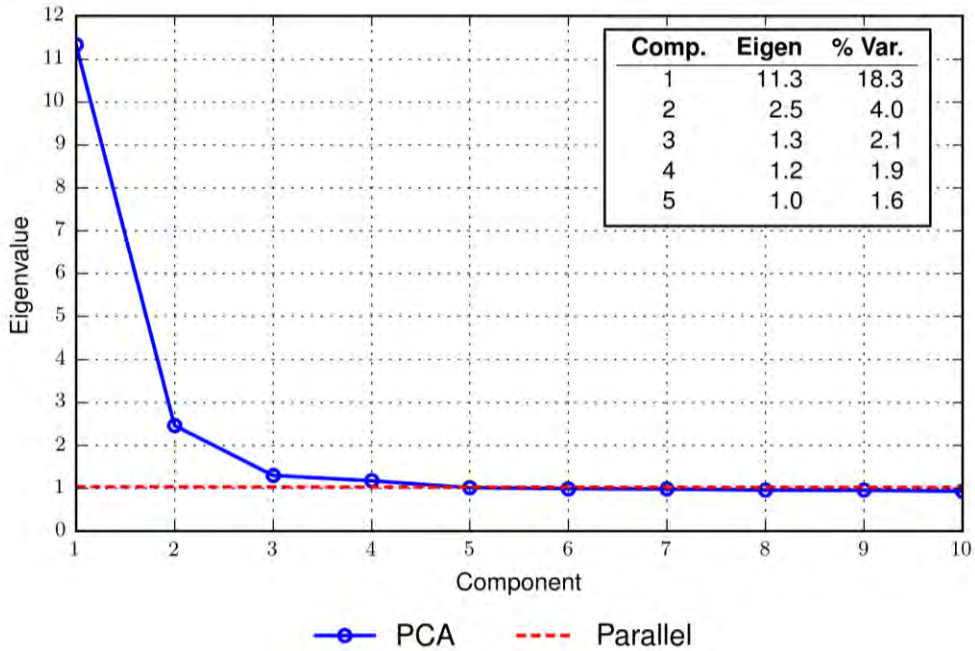

Figure 3 Scree Plots: Regents Examination in United States History

Local Independence

Local independence (LI) is a fundamental assumption of IRT. This means simply, 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, a test X that is comprised of items $X_1, X_2, \ldots X_n$ is locally independent with respect to the latent variable θ if, for all $x = (x_1, x_2, \ldots x_n)$ and θ,

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

This formula essentially states that the probability of any pattern of responses across all items ($\mathbf{x}$), after conditioning on the examinee's true score (θ) as measured by the test, should be equal to the product of the conditional probabilities across each item (cf. 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 abilities, is the product of the probabilities of responses to these two items, as shown below. Based on the WLI, the following expression can be derived:

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

Marais and Andrich (2008) point out that local item dependence in the Rasch model can occur in two ways that may be difficult to distinguish. The first way occurs when the assumption of

unidimensionality is violated. Here, other nuisance dimensions besides a dominant dimension determine student performance (this can be called "trait dependence"). The second violation 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 among the Regents Examination in United States History items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using (θ) and item parameter estimates. Next, deviations (residuals) between the examinees' expected and observed performance is 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 −0.02 for the items on the exam. 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 4 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 residual correlations greater than 0.20 are also reported in this table. There were 9 item pairs with residual correlations greater than 0.20 a maximum of 0.31, suggesting that there may be some local dependence for these items on the Regents Examination in United States History.

Table 4 Summary of Item Residual Correlations: Regents Examination in United States History

Statistic Type	Value
N	1891
Mean	-0.01
SD	0.04
Minimum	-0.15
P_{10}	-0.05
P_{25}	-0.03
P_{50}	-0.01
P_{75}	0.00
P_{90}	0.02
Maximum	0.31
>|0.20|	9

Item Fit

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

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 5 presents the summary statistics of infit mean square statistics for the Regents Examination in United States History , including the 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 5. The mean infit value is 1.00, with 33 of 62 items falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as guide for ideal fit, fit values outside of the range are considered individually. In this case, the maximum value for the items falling outside of the ideal range has an infit of 3.40. The overall high performance of examinees on a test may have the effect of reducing the variance of scores for many items which can in turn, impact model fit negatively.

Table 5 Summary of Infit Mean Square Statistics: Regents Examination in United States Histroy

	Infit Mean Square				
	Mean	**SD**	**Min**	**Max**	**[0.7, 1.3]**
United States History	1.20	0.53	0.46	3.40	33/62

Items for the Regents Examination in United States History were field tested in 2013 and 2014, and a separate technical reports were produced to document the full test development, scoring, scaling, and data analysis conducted. Please refer to http://www.p12.nysed.gov/assessment/reports for details.

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 Regents Examination in United States History 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. It is defined as the proportion of true score variance contained in the observed scores. The total variance contains two components: 1) the variance in true scores and 2) the variance due to the imperfections in the measurement process. 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 is true. If all of the test score variances were true, the index would equal 1.0. Such scores 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

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

practice, it is clear that larger coefficients are more desirable because they indicate that 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. It 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 Alpha is as follows:

$$\frac{\quad}{\quad}\left(\frac{\alpha}{\alpha} - \alpha\frac{\sum\alpha \quad {}^{2}_{Yi}}{2\alpha}\right),$$

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

4.2 Standard Error of Measurement (Standards 2.13, 2.14, 2.15)

Reliability coefficients best reflect the extent to which measurement inconsistencies may be present or absent. The standard error of measurement 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 (SEM) 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.

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

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 an SEM of 3 on a 10-point test would be very different than an 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 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 about 95 percent.

The coefficient alpha and associated SEM for the Regents Examination in United States History are provided in Table 6.

Table 6 Reliabilities and Standard Errors of Measurement: Regents Examination in United States History

Subject	Coefficient Alpha	*SEM*
United States History	0.92	4.21

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 that 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 their true score (θ), the score obtained with no error), and the standard deviation of the distribution of their raw scores would be the conditional standard error. Since there is a one-to-one correspondence between the raw score and θ 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 student's estimated θ. 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.

The relationship between θ 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 along with the nonlinear relationship between the expected raw and θ scores. In addition, as the exam is equated from year to year, the relationship between the raw and scale scores moves away from the original 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 θ and the scale scores, the CSEMs that are available for each θ score via Rasch IRT cannot be converted directly to the scale score metric.

The use of Rasch IRT to scale and equate the Regents Exams does, however, make it possible to calculate CSEMs using the procedures described by Kolen, Zeng, and Hanson (1996) for

[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.

dichotomously scored items and extended by Wang, Kolen, and Harris (2000) to polytomously scored items. For tests such as the Regents Examination in United States History that have a one-to-one relationship between raw (θ) and scale scores, the CSEM for each achievable scale score can be calculated using the compound multinomial distribution to represent the conditional distribution of raw scores for each level of θ.

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, their 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 their level of θ; 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 their observed test scores, the resulting distribution would be the conditional distribution of raw scores for that examinee's level of θ with a mean value equal to the examinee's expected raw (true) score. The CSEM for that level of θ in the raw score metric is the square root of the variance of this conditional distribution.

The conditional distribution of raw scores for any level of θ is the compound multinomial distribution (Wang et al., 2000). An algorithm to compute this can be found in Hanson (1994) and 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 θ 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 θ.

Conditional Standard Error of Measurement Confidence Intervals

CSEMs allow statements regarding the precision of individual tests scores. Like SEMs, they help place reasonable limits around observed scaled scores through 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 θ depends both on the nature of the raw to scale score transformation (Kolen and Brennan, 2005; Kolen and Lee, 2011) and on whether the CSEM is derived from the raw scores or from θ (Lord, 1980). The pattern of CSEMs for raw scores and linear transformations of the raw score tend to have a characteristic "inverted-U" shape, with smaller CSEMs at the ends of the score continuum and larger CSEMs towards 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 Exams tends to be roughly linear from scale scores of 0 to 20 and then concave down from about 20 to 100. In other words, the scale scores track linearly with the raw scores for the 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 Exams 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 United States History in which the compression of raw score to scale scores around the cut score of 65 changes the shape of the curve slightly. This type of expansion and compression can be seen in Figure 4 by looking at the changing density of raw score points along the scale score range on the horizontal axis. Specifically, there is noticeable compression in the middle of the score scale between about 20 and 90 on the score scale.

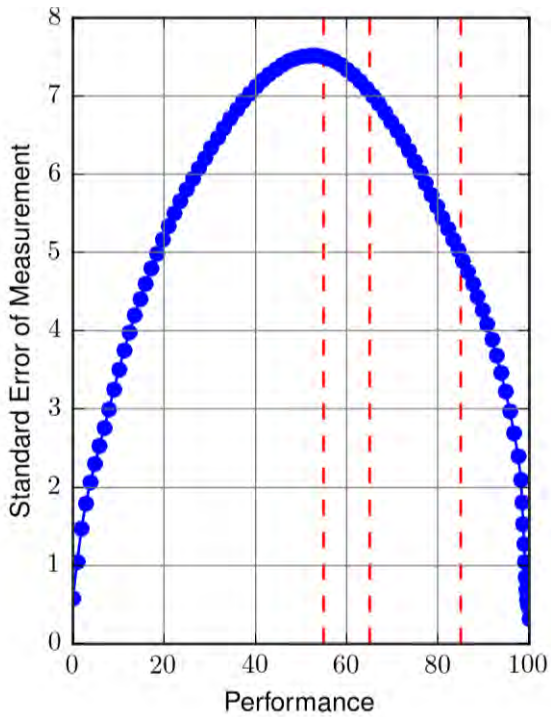

Figure 4 Conditional Standard Error Plots: Regents Examination in United States History

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 made from the first set of test scores matched the decisions based on the second set of test scores. Consider the tables below.

		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 he or she 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 his or her 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 United States History , 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 for four performance levels should be lower than those based on two categories. This is not surprising, since classification and accuracy using four levels would allow more opportunity to change achievement levels. Hence, there would be more classification errors and less accuracy with four achievement levels, resulting in lower consistency indices.

Results and Observations The results for the dichotomies created by the three cut scores, are presented in table 7. The tabled values are derived with the program *BB-Class* (Brennan, 2004) using the Livingston and Lewis method. Decision consistency ranged from 0.87 to 0.95, and the decision accuracy ranged from 0.91 to 0.97. Both decision consistency and accuracy values based on individual cut points indicate very good consistency and accuracy of examinee classifications. Refer to Table 7.

Table 7 Decision Consistency and Accuracy Results: Regents Examination in United States History

Statistic	1/2	2/3	3/4
Consistency	0.95	0.93	0.87
Accuracy	0.97	0.95	0.91

4.4 Group Means (Standard 2.17)

Mean scale scores were computed based on reported gender, race/ethnicity, English Language Learner status, economically disadvantaged status, and student with disability status. The results are reported in Table 8.

Table 8 Group Means: Regents Examination in United States History

Demographics	Number	Mean Scale score	Standard error of group
All Students*	176825	81.03	16.80
Ethnicity			
American Indian/Alaska Native	973	76.26	17.34
Asian/Native Hawaiian/Other Pacific Islander	17548	85.73	15.09
Black/African American	33631	72.53	17.90
Hispanic/Latino	40034	74.30	17.77
Multiracial	1943	84.45	14.42
White	82692	86.73	13.21
English Language Learner			
No	167920	82.03	16.07
Yes	8905	62.17	19.03
Economically Disadvantaged			
No	94192	85.99	14.19
Yes	82633	75.38	17.73
Gender			
Female	88884	81.60	16.26
Male	87937	80.45	17.31
Student with Disabilities			
No	154079	83.33	15.09
Yes	22746	65.44	19.31

*Note: 4 students were not reported in the Ethnicity and Gender group but they are reflected in "All Students".

4.5 State Percentile Rankings

State percentile rankings based on raw score distributions are noted in Table 9. The percentiles are based on the distribution of all students taking the Regents Examination in Earth. 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.
- Students who obtain the lowest possible score (0) will not receive a percentile rank.

Table 9 State Percentile Ranking for Raw Score – Regents Examination in United States History

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

Chapter 5: Validity (Standard 1)

Restating the purpose and uses of the Regents Examination in United States History , this exam measures examinee achievement against New York State's learning standards. The exam is prepared by teacher examination committees and New York State Department of Education subject and testing specialists, and 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 exams also provide 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 United States History is intended for use in satisfying State testing requirements for students who have finished a course of instruction in United States History. 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.p12.nysed.gov/ciai/gradreq/2015GradReq11-15.pdf. Results of the Regents Examination in United States History may also be used to satisfy various locally established requirements throughout the State.

The validity of score interpretations for the Regents Examination in United States History 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, so they are used here. The process of gathering evidence of the validity of score interpretations is best characterized as ongoing throughout the 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 with 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 United States History measures student achievement based on the United States History Core Curriculum which is based on standards 1 2, and 4 in the New York State Learning Standards for Social Studies. The United States History standards can be found at: http://www.p12.nysed.gov/ciai/socst/ssrg.html.

Content Validity

Content validity is necessarily 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 United States History is essentially the design document for constructing the exam. It provides 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 10 displays the targeted item types for each content standard on the exam.

Table 10 Test Blueprint, Regents Examination in United States History

Item Type	Standards
Multiple Choice	See Specifications Grid*
Thematic essay	Standard 1 – United States History
	Standard 2 - Government
Document-based question	Standard 1 – United States History
	Standard 4- Economics

* See pages 114-116 for item level details: www.p12.nysed.gov/ciai/socst/pub/4samus.pdf

Item Development Process

Test development for the Regents Examination in United States History 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 Department of Education 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 around the State are retained to write all items for the Regents Examination in United States History under strict guidelines that leverage best practices (see Appendix A). 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 United States History to ensure appropriate coverage of the construct domain.

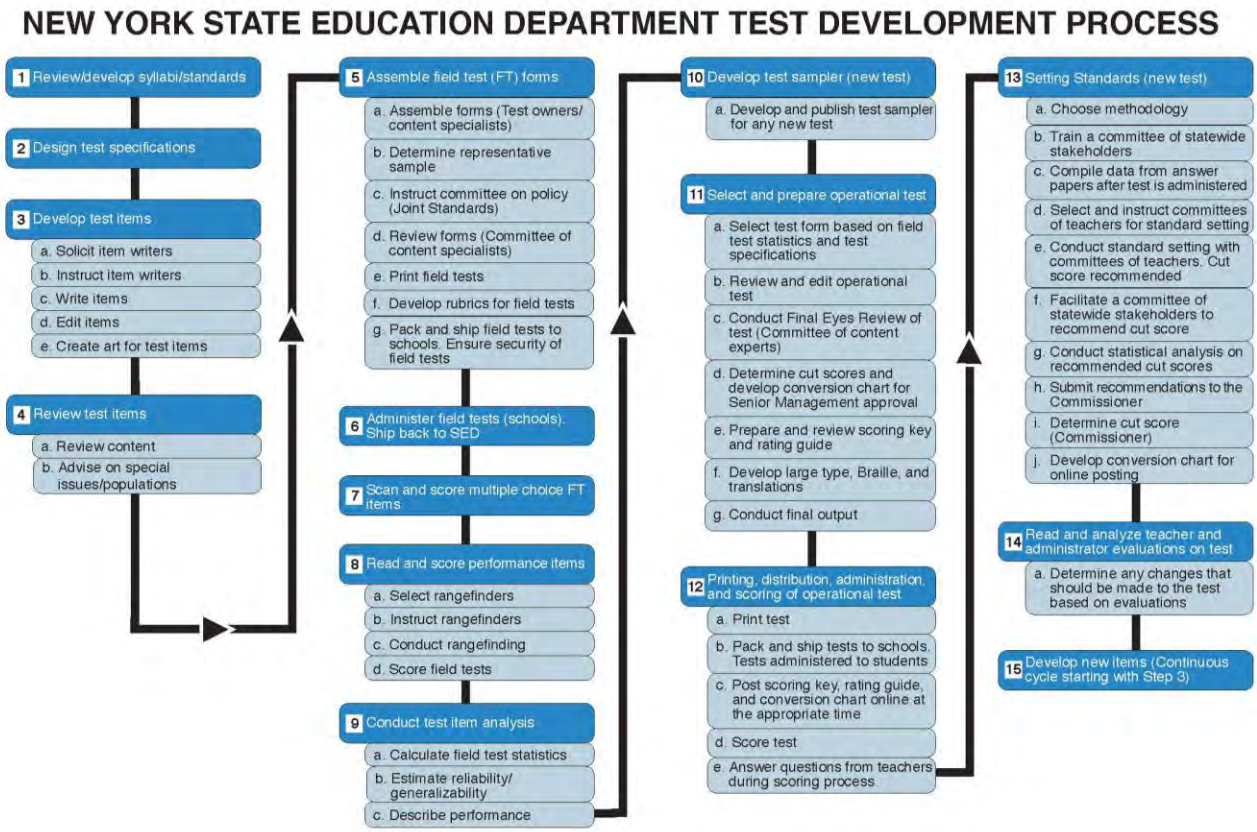

Figure 7 New York State Education Department Test Development Process

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 a fundamental form of evidence of the validity of the intended score interpretations. Another integral component of this item review process is to review the scoring rules, or "rubrics," for their clarity and consistency in what the 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 in developing quality items:

1. language and graphical appropriateness
2. sensitivity/bias
3. fidelity of measurement to standards
4. conformity to the expectations for the specific item types and formats

Each section of the criteria includes pertinent questions that help reviewers determine whether or not an item is of sufficient quality. Within the first two categories, criteria for language appropriateness are used to help ensure that students understand what is asked in each question and that the language in the question does not adversely affect a student's ability to perform the required task. Likewise, 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 Department of Education is committed to ongoing standards review over time and periodically solicits thoughtful, specific responses from stakeholders about individual standards within the NY 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 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 United States History 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 A). Further evidence is documented in the test administration and scoring procedures, as well as 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 titled *School Administrator's Manual, Secondary Level Examinations* (http://www.p12.nysed.gov/assessment/sam/secondary/hssam-update.html) 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, Wang, and Kwong, 2010; Gorman & Rentsch, 2009; Schleicher, Day, Bronston, Mayes, and Riggo, 2002; Woehr & Huffcutt, 1994; Johnson, Penny, and 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, and 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, 2006; Wolfe & Gitomer, 2000; Cronbach, Linn, Brennan, & Haertel, 1995; Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000; Smith, 1993; Leacock, Gonzalez, and Conarro, 2014)

The distinct steps for operational test scoring include close attention to each of these elements and begin before the operational test is even 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 are 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. A consensus on a training for each score point of each item 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 2014 administrations of the Regents Examination in United States History are are located at http://www.p12.nysed.gov/assessment/samplers.

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 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 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 their assigned staff against model responses and is also 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 more clear the response expectations are for examinees. To facilitate the development of constructed-response scoring rubrics, the 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 of valid score interpretations for each examinee, then, such scoring training procedures are implemented for the Regents Examination in United States History. 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 has the effect of increasing the consistency of scoring across examinee responses by allowing each rater to focus on a subset of items. It also assures 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, no rater is allowed to score the responses of his or her 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 the item developer intended them to. 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 that 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 United States History :

- item difficulty
- item discrimination
- differential item functioning
- IRT model fit
- test reliability
- classification consistency
- test dimensionality

Item Difficulty

Multiple analyses allow an evaluation of item difficulty. For this exam, *p*-values and Rasch difficulty (item location) estimates were computed for MC and CR items.[5] Items for the Regents Examination in United States History show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values range from 0.52 to 0.90.

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 0.12 or higher for all items and negative, zero, or very 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) for gender was conducted following field testing of the items in 2011-2014. Sample sizes for subgroups based on ethnicity and English language learner status were unfortunately too small to reliably compute DIF statistics, so only gender DIF analyses were conducted. The Mantel-Haenszel $^{2\alpha}$ and standardized mean difference were used to detect items that may function differently for any of these subgroups. The Mantel $^{2\alpha}$ 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

[5] Refer to the field test report for details: http://www.p12.nysed.gov/assessment/reports.

than a response earning a score of "0," and "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. No items showed DIF for the operational 2015 Physics exam.

Full differential item functioning results are reported in Appendix C of the 2011 and 2012 technical report, and Appendix E of the 2013 and 2014. All reports except 2014 are located at: http://www.p12.nysed.gov/assessment/reports/2014/earthsci-tr14.pdf.

IRT Model Fit

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the Regents Examination in United States History 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 33 of 62 items falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as guide for ideal fit, fit values outside of the range are considered individually. In this case, the maximum value for the items falling outside of the ideal range has an infit of 3.40. The overall high performance of examinees on a test may have the effect of reducing the variance of scores for many items which can in turn, impact model fit negatively.

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 Regents Examination in United States History is 0.92, showing high reliability of examinee scores. Refer to section 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 that the classification decisions made from 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 than expected from the true score. High decision consistency and accuracy provides strong evidence that the internal structure of a test is sound.

For the Regents Examination in United States History , both decision consistency and accuracy values are high, indicating very good consistency and accuracy of examinee classifications. Decision consistency ranged from 0.87 to 0.95, and the decision accuracy ranged from 0.91 to 0.97. Both decision consistency and accuracy values based on individual cut points indicate very good consistency and accuracy of examinee classifications. Refer to Table 7.

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 United States History 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 section 3 for details of this analysis.

Considering this collection of detailed analyses on the internal structure of the Regents Examination in United States History , 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 Relations to Other Variables

Another source of validity evidence is based on the relation of the test to other variables. This source commonly encompasses two validity categories prevalent in the literature and practice—concurrent and predictive validity. To make claims about the validity of a test that is to be used for high stakes purposes, such as the Regents Examination in United States History , 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 United States History 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 United States History.

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

In the literature on validity, there are two general approaches 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, however, 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 by the development process and research conducted on examinee data. A

brief survey of websites for 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 United States History , the exam is most commonly used to inform admissions and course placement decisions. Such uses can be considered reasonable, assuming the competencies demonstrated in the Regents Examination in United States History 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 United States History and to assess their appropriateness for the inferences being made about course placement.

As stated, the nature of validity arguments is not absolute, but 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 United States History 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.

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., Linn, R. L., Brennan, R. T., & Haertel, E. (1995, Summer). Generalizability analysis for educational assessments. Los Angeles, CA: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing. Retrieved February 17, 2016, from www.cse.ucla.edu/products/evaluation/cresst_ec1995_3.pdf.

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. Braun (Eds.), *Test Validity* (pp. 3–17) Hillsdale, NJ: Lawrence Erlbaum.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.

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*, 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 http://www.education.uiowa.edu/casma/computer_programs.htm.

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. Monterey: McGraw-Hill Education CTB.

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

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–72.

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

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

McDonald, R.P. (1979). The structural analysis of multivariate data: A sketch of 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.

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: 341.

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.

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.

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.

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, 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*, 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.

Appendix A – Item Writing Guidelines

General Rules for Writing Multiple-Choice Items

1. ***Use either a direct question or an incomplete statement as the item stem, whichever seems more appropriate to effective presentation of the item.***

   Some item ideas can be expressed more simply and clearly in the incomplete statement style of question. On the other hand, some items seem to require direct question stems for the most effective expression. Teachers should use the item style that seems most appropriate.

2. ***Items should be written in clear and simple language, with vocabulary kept as simple as possible.***

   Like any other item, the multiple-choice item should be perfectly clear. Difficult and technical vocabulary should be avoided unless essential for the purpose of the question. The important elements should generally appear early in the statement of the item, with qualifications and explanations following.

3. ***Each item should have one and only one correct answer.***

   While this requirement is obvious, it is not always fulfilled. Sometimes writers produce items involving issues so controversial and debatable that even experts are unable to agree on one correct answer. More often the trouble is failure to consider the full implications of each response.

4. ***Base each item on a single central problem.***

   A multiple-choice item functions most effectively when the student is required to compare directly the relative merits of a number of specific responses to a definite problem. An item consisting merely of a series of unrelated true-false statements, all of which happen to begin with the same phrase, is unacceptable.

5. ***State the central problem of the item clearly and completely in the stem. (See Helpful Hint #2,476.)***

   The stem should be meaningful by itself. It should be clear and should convey the central problem of the item. It should not be necessary for the student to read and reread all the responses before he/she can understand the basis upon which he/she is to make a choice.

6. ***In general, include in the stem any words that must otherwise be repeated in each response.***

   The stem should contain everything the answers have in common or as much as possible of their common content. This practice serves to make the item shorter, so that it can be read and grasped more quickly.

7. ***Avoid negative statements.***

   Negative statements in multiple-choice items lead to unnecessary difficulties and confusion. Special care must be exercised against the double negative.

8. ***Avoid excessive "window dressing."***

The item should contain only material relevant to its solution, unless selection of what is relevant is part of the problem.

9. ***Make the responses grammatically consistent with the stem and parallel with one another in form.***

10. ***Make all responses plausible and attractive to students who lack the information or ability tested by the item.***

The incorrect responses should be plausible answers. So far as possible, each response should be designed specifically to attract students who have certain misconceptions or who tend to make certain common errors.

11. ***Arrange the responses in logical order, if one exists.***

Where the responses consist of numbers or letters, they should ordinarily be arranged in ascending order. Events should be listed in the order in which they occurred, from earliest to most recent, except when this order would clue the answer. This practice helps insure the student will mark the answer correctly.

12. ***Make the responses independent and mutually exclusive.***

Responses should not be interrelated in meaning. Responses that are not mutually-exclusive, aid the student in eliminating wrong answers and reduce the reliability of the item by decreasing the number of effective, functioning responses.

13. ***Avoid extraneous clues.***

Since the student is required to associate one of several alternative responses with the stem, any aspect of the question that provides an extraneous basis for correctly associating the right answer or for eliminating a wrong response constitutes an undesirable clue.

14. ***Avoid using "all of the above" and "none of the above" as alternatives.***

15. ***Avoid using the phrase "of the following" in the stem.***

CHECKLIST OF TEST CONSTRUCTION PRINCIPLES
(Multiple Choice Items)

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

HELPFUL HINT

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

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 would not be a good-short answer item you may have a problem with the stem.)**

3. **The stem must consist of a statement that contains a verb.**

Appendix B – Tables and Figures for August 2014 Administration

Table B 1 Multiple-Choice Item Analysis Summary: Regents Examination in United States History

Item	Number of Students	*p* Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
1	16429	.44	.50	.20	-.10	-.14	-.03
2	16429	.73	.44	.34	-.20	-.19	-.13
3	16429	.36	.48	.22	-.03	-.14	-.09
4	16429	.54	.50	.31	-.17	-.12	-.13
5	16429	.55	.50	.36	-.14	-.15	-.22
6	16429	.53	.50	.24	-.16	-.12	-.07
7	16429	.55	.50	.10	.04	-.11	-.10
8	16429	.58	.49	.29	-.14	-.14	-.14
9	16429	.65	.48	.34	-.13	-.20	-.18
10	16429	.66	.47	.17	-.16	-.02	-.11
11	16429	.43	.50	.11	.00	-.07	-.12
12	16429	.65	.48	.28	-.20	-.12	-.10
13	16429	.44	.50	.18	-.04	-.10	-.10
14	16429	.72	.45	.24	-.12	-.12	-.13
15	16429	.66	.48	.33	-.23	-.09	-.14
16	16429	.22	.41	.20	-.04	-.13	.00
17	16429	.44	.50	.21	-.05	-.08	-.12
18	16429	.40	.49	.24	-.09	-.13	-.05
19	16429	.51	.50	.18	-.05	-.17	-.01
20	16429	.54	.50	.21	-.04	-.13	-.11
21	16429	.60	.49	.31	-.13	-.15	-.15
22	16429	.81	.39	.40	-.20	-.22	-.21
23	16429	.34	.47	.28	-.04	-.20	-.08
24	16429	.38	.49	.21	-.08	-.07	-.11
25	16429	.29	.45	.26	-.19	-.11	-.03
26	16429	.28	.45	.11	-.01	-.03	-.08
27	16429	.65	.48	.30	-.13	-.16	-.14
28	16429	.54	.50	.29	-.16	-.18	-.05
29	16429	.56	.50	.33	-.18	-.13	-.14
30	16429	.42	.49	.30	-.14	-.15	-.09
31	16429	.64	.48	.33	-.17	-.14	-.18
32	16429	.54	.50	.23	-.14	-.05	-.14
33	16429	.58	.49	.25	-.16	-.12	-.06
34	16429	.47	.50	.24	-.07	-.11	-.13

Item	Number of Students	*p* Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
35	16429	.55	.50	.24	-.11	-.09	-.12
36	16429	.43	.50	.17	-.14	-.10	.01
37	16429	.71	.45	.32	-.18	-.17	-.12
38	16429	.48	.50	.18	-.13	-.11	.00
39	16429	.59	.49	.34	-.15	-.18	-.16
40	16429	.43	.50	.19	-.04	-.09	-.09
41	16429	.56	.50	.22	-.04	-.09	-.17
42	16429	.60	.49	.29	-.19	-.13	-.10
43	16429	.57	.50	.24	-.09	-.11	-.14
44	16429	.56	.50	.19	-.08	-.10	-.08
45	16429	.63	.48	.41	-.17	-.18	-.22
46	16429	.37	.48	.20	-.06	-.08	-.07
47	16429	.64	.48	.29	-.08	-.19	-.15
48	16429	.60	.49	.31	-.16	-.17	-.09
49	16429	.46	.50	.22	-.20	-.02	-.10
50	16429	.45	.50	.24	-.12	-.08	-.11

Table B 2 Constructed-Response Item Analysis Summary: Regents Examination in United States History

Item	Min. score	Max. score	Number of Students	Mean	SD	*p* Value	Point Biserial
51	0	5	16429	2.06	1.00	.41	.40
52	0	1	16429	.92	.28	.92	.25
53	0	1	16429	.87	.33	.87	.25
54	0	2	16429	1.75	.51	.87	.26
55	0	1	16429	.90	.31	.90	.24
56	0	1	16429	.94	.24	.94	.26
57	0	2	16429	1.72	.57	.86	.30
58	0	2	16429	1.64	.63	.82	.30
59	0	2	16429	1.65	.63	.82	.31
60	0	1	16429	.76	.43	.76	.24
61	0	1	16429	.91	.29	.91	.27
62	0	5	16429	1.98	.98	.40	.39

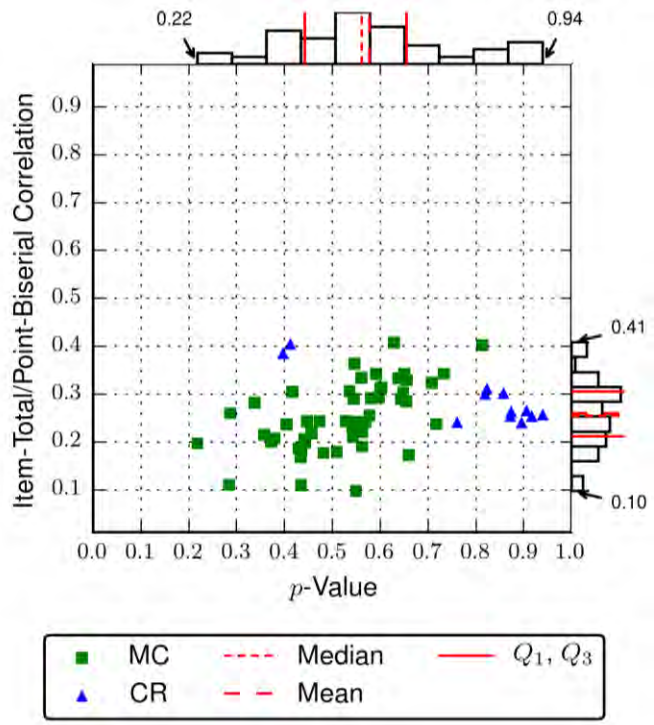

Figure B 1 Scatterplot: Regents Examination in United States History

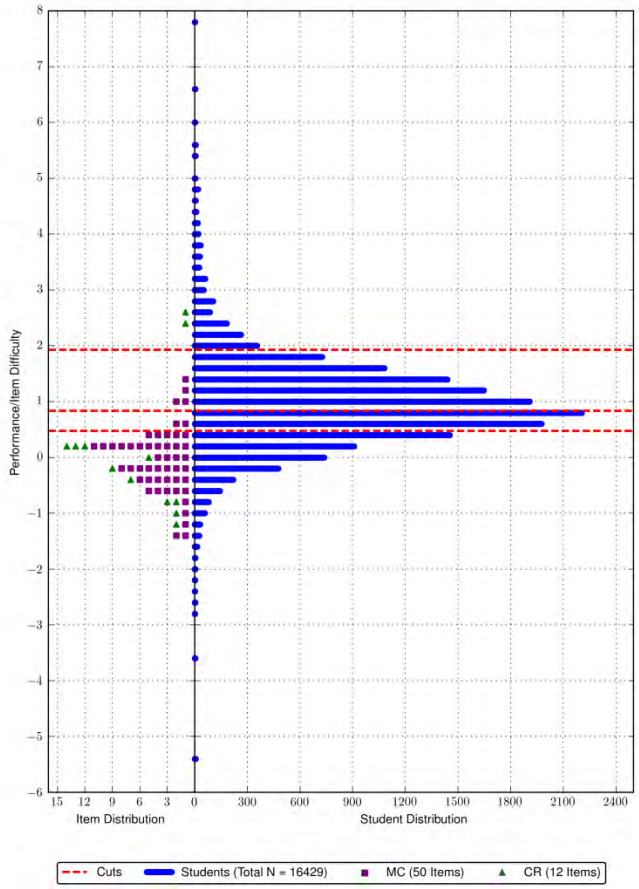

Figure B 2 Student Performance Map: Regents Examination in United States History

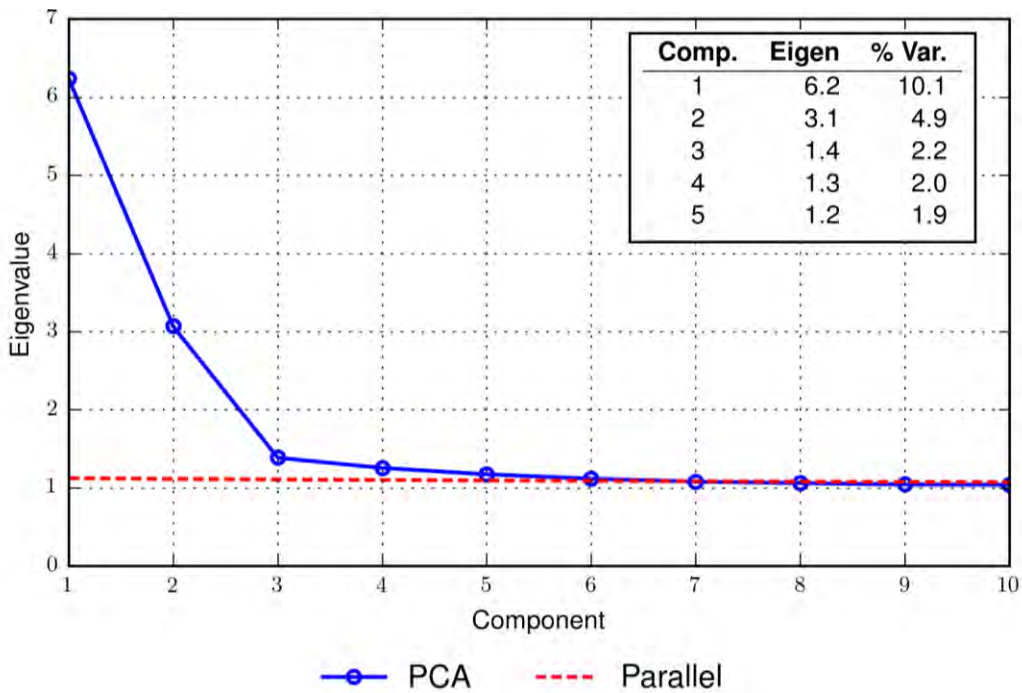

Figure B 3 Scree Plots: Regents Examination in United States History

Table B 3 Summary of Item Residual Correlations: Regents Examination in United States History

Statistic Type	Value
N	1891
Mean	-0.02
SD	0.04
Minimum	-0.14
P_{10}	-0.06
P_{25}	-0.04
P_{50}	-0.02
P_{75}	0.00
P_{90}	0.02
Maximum	0.33
>|0.20|	12

Table B 4 Summary of Infit Mean Square Statistics: Regents Examination in United States History

	Infit Mean Square				
	Mean	**SD**	**Min**	**Max**	**[0.7, 1.3]**
United States History	1.19	0.52	0.48	3.36	46/62

Table B 5 Reliabilities and Standard Errors of Measurement: Regents Examination in United States History

Subject	**Coefficient Alpha**	***SEM***
United States History	0.84	4.72

Table B 6 Decision Consistency and Accuracy Results: Regents Examination in United States History

Statistic	**1/2**	**2/3**	**3/4**
Consistency	0.85	0.82	0.94
Accuracy	0.89	0.87	0.96

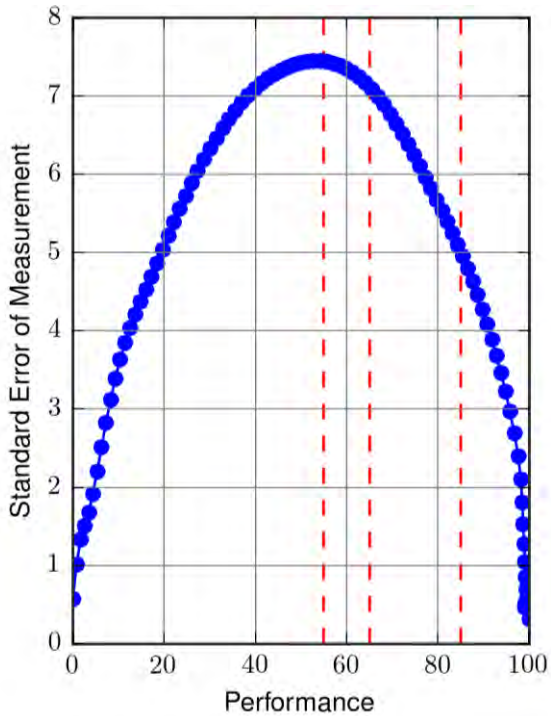

Figure B 4 Conditional Standard Error Plots: Regents Examination in United States History

Table B 7 Group Means: Regents Examination in United States History

Demographics	Number	Mean Scale score	Standard error of group
All Students*	16429	63.39	15.72
Ethnicity			
American Indian/Alaska Native	123	61.77	14.53
Asian/Native Hawaiian/Other Pacific Islander	1040	65.54	16.52
Black/African American	5684	61.70	15.37
Hispanic/Latino	5829	62.19	15.92
Multiracial	88	65.53	12.68
White	3659	67.33	15.05
English Language Learner			
No	14417	64.28	15.43
Yes	2012	57.01	16.33
Economically Disadvantaged			
No	6077	65.43	15.76
Yes	10352	62.20	15.58
Gender			
Female	8580	64.35	14.75
Male	7843	62.35	16.66
Student with Disabilities			
No	12747	65.37	15.14
Yes	3682	56.55	15.80

*Note: 6 students were not reported in the Ethnicity and Gender group but they are reflected in “All Students”.

Appendix C – Tables and Figures for January 2015 Administration

Table C 1 Multiple-Choice Item Analysis Summary: Regents Examination in United States History

Item	Number of Students	*p* Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
1	25268	.48	.50	.14	-.06	-.04	-.08
2	25268	.72	.45	.30	-.12	-.14	-.20
3	25268	.55	.50	.33	-.09	-.20	-.16
4	25268	.70	.46	.27	-.09	-.18	-.14
5	25268	.66	.47	.21	-.11	.00	-.15
6	25268	.46	.50	.35	-.12	-.18	-.14
7	25268	.62	.48	.31	-.11	-.19	-.14
8	25268	.63	.48	.31	-.19	-.12	-.14
9	25268	.71	.45	.26	-.16	-.14	-.09
10	25268	.41	.49	.18	-.06	-.08	-.07
11	25268	.48	.50	.28	-.20	-.12	-.03
12	25268	.60	.49	.27	-.17	-.15	-.05
13	25268	.64	.48	.34	-.23	-.14	-.14
14	25268	.65	.48	.37	-.12	-.17	-.24
15	25268	.51	.50	.40	-.18	-.14	-.21
16	25268	.69	.46	.33	-.14	-.18	-.17
17	25268	.35	.48	.26	-.14	-.11	-.08
18	25268	.56	.50	.36	-.18	-.19	-.13
19	25268	.46	.50	.40	-.14	-.20	-.17
20	25268	.70	.46	.38	-.17	-.20	-.19
21	25268	.54	.50	.27	-.06	-.17	-.13
22	25268	.73	.44	.36	-.19	-.18	-.17
23	25268	.61	.49	.31	-.13	-.14	-.18
24	25268	.46	.50	.22	-.05	-.10	-.15
25	25268	.49	.50	.33	-.10	-.19	-.15
26	25268	.52	.50	.45	-.19	-.17	-.23
27	25268	.43	.49	.17	-.12	-.20	.03
28	25268	.49	.50	.18	-.07	-.04	-.12
29	25268	.46	.50	.41	-.17	-.19	-.16
30	25268	.39	.49	.29	-.14	-.05	-.15
31	25268	.39	.49	.32	-.08	-.14	-.18
32	25268	.42	.49	.38	-.21	-.14	-.12
33	25268	.36	.48	.26	-.11	-.02	-.17
34	25268	.70	.46	.34	-.15	-.16	-.20

Item	Number of Students	*p* Value	SD	Point Biserial	Point Biserial Distractor 1	Point Biserial Distractor 2	Point Biserial Distractor 3
35	25268	.54	.50	.36	-.25	-.10	-.13
36	25268	.40	.49	.28	-.17	-.16	.02
37	25268	.76	.43	.35	-.18	-.21	-.14
38	25268	.47	.50	.34	-.21	-.17	-.06
39	25268	.67	.47	.37	-.22	-.17	-.15
40	25268	.33	.47	.27	-.25	-.17	.06
41	25268	.52	.50	.24	.00	-.15	-.19
42	25268	.53	.50	.31	-.16	-.18	-.12
43	25268	.43	.50	.24	-.08	-.19	-.02
44	25268	.58	.49	.41	-.17	-.22	-.20
45	25268	.41	.49	.37	-.21	-.16	-.07
46	25268	.48	.50	.23	-.10	-.12	-.05
47	25268	.63	.48	.33	-.09	-.21	-.16
48	25268	.33	.47	.28	-.05	-.11	-.17
49	25268	.68	.47	.35	-.20	-.19	-.13
50	25268	.48	.50	.21	-.21	.03	-.14

Table C 2 Constructed-Response Item Analysis Summary: Regents Examination in United States History

Item	Min. score	Max. score	Number of Students	Mean	SD	*p* Value	Point Biserial
51	0	5	25268	1.64	1.16	.33	.47
52	0	2	25268	1.58	.64	.79	.28
53	0	1	25268	.78	.41	.78	.28
54	0	1	25268	.82	.39	.82	.30
55	0	1	25268	.72	.45	.72	.29
56	0	2	25268	1.60	.65	.80	.38
57	0	1	25268	.85	.36	.85	.30
58	0	1	25268	.84	.36	.84	.32
59	0	1	25268	.90	.31	.90	.28
60	0	1	25268	.84	.37	.84	.33
61	0	1	25268	.87	.34	.87	.32
62	0	2	25268	1.49	.74	.75	.41
63	0	5	25268	1.95	1.08	.39	.44

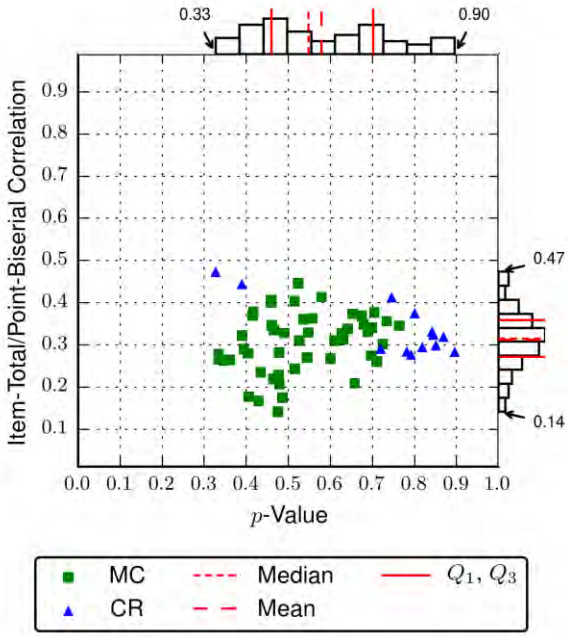

Figure C 1 Scatterplot: Regents Examination in United States History

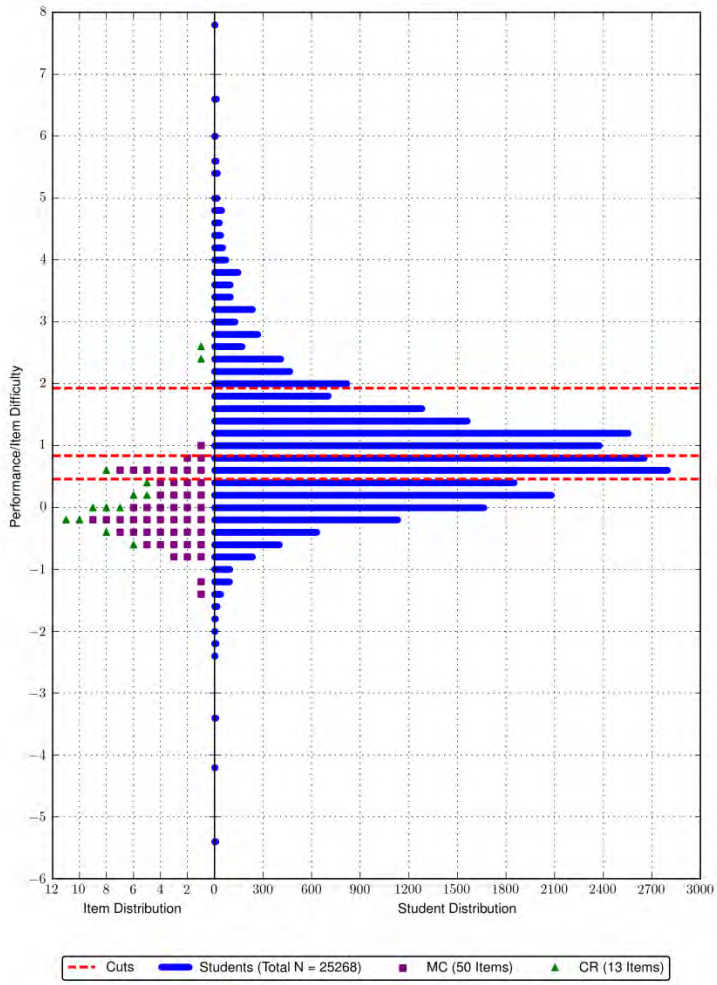

Figure C 2 Student Performance Map: Regents Examination in United States History

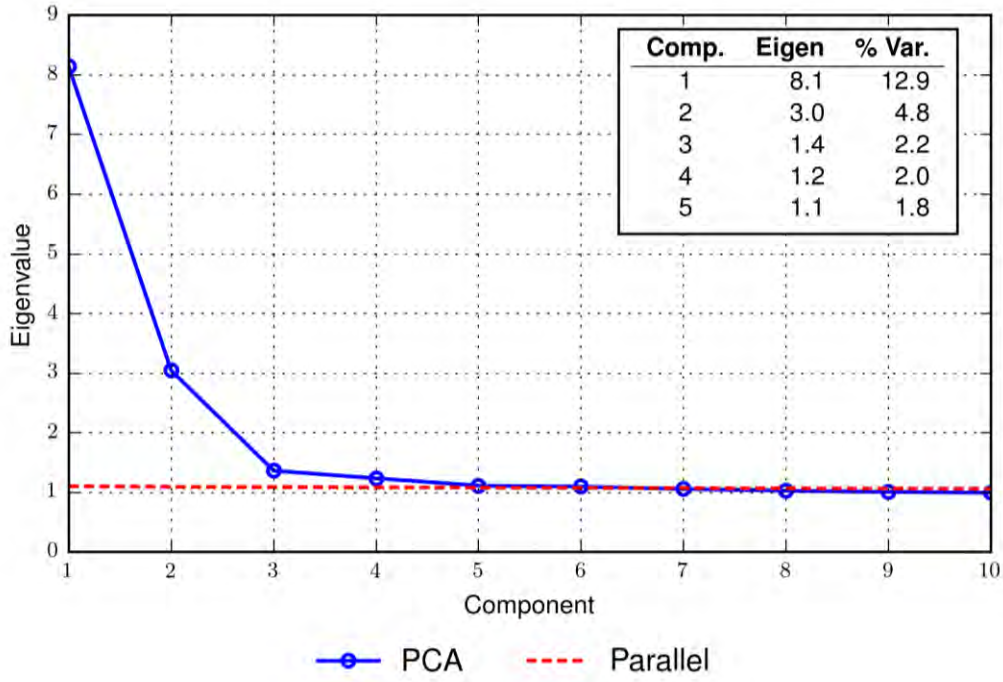

Figure C 3 Scree Plots: Regents Examination in United States History

Table C 3 Summary of Item Residual Correlations: Regents Examination in United States History

Statistic Type	Value
N	1953
Mean	-0.01
SD	0.05
Minimum	-0.13
P_{10}	-0.06
P_{25}	-0.04
P_{50}	-0.02
P_{75}	0.00
P_{90}	0.02
Maximum	0.35
>|0.20|	14

Table C 4 Summary of Infit Mean Square Statistics: Regents Examination in United States History

	Infit Mean Square				
	Mean	**SD**	**Min**	**Max**	**[0.7, 1.3]**
United States History	1.12	0.25	0.63	2.20	53/63

Table C 5 Reliabilities and Standard Errors of Measurement: Regents Examination in United States History

Subject	**Coefficient Alpha**	***SEM***
United States History	0.88	4.82

Table C 6 Decision Consistency and Accuracy Results: Regents Examination in United States History

Statistic	**1/2**	**2/3**	**3/4**
Consistency	0.85	0.85	0.93
Accuracy	0.90	0.89	0.95

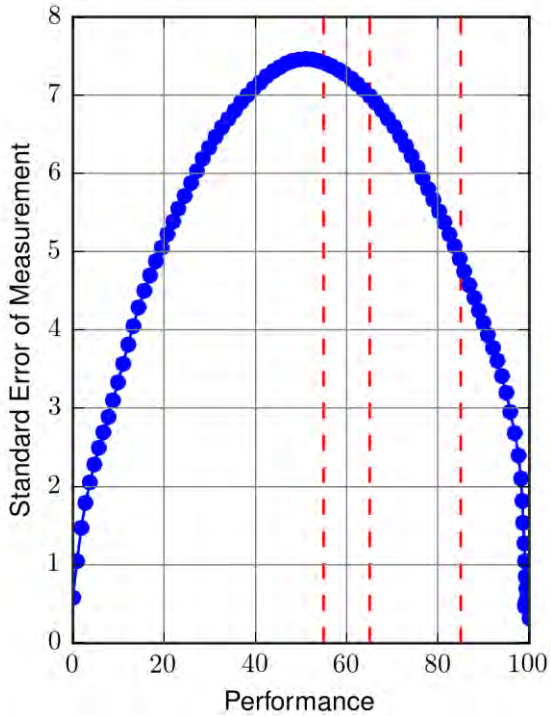

Figure C 4 Conditional Standard Error Plots: Regents Examination in United States History

Table C 7 Group Means: Regents Examination in United States History

Demographics	Number	Mean Scale score	Standard error of group
All Students*	25268	61.79	18.27
Ethnicity			
American Indian/Alaska Native	173	60.16	18.14
Asian/Native Hawaiian/Other Pacific Islander	1851	63.37	18.87
Black/African American	8362	58.59	17.48
Hispanic/Latino	8913	59.60	17.31
Multiracial	177	68.91	18.11
White	5788	69.15	18.49
English Language Learner			
No	21995	63.03	18.13
Yes	3273	53.46	16.96
Economically Disadvantaged			
No	7960	65.49	19.42
Yes	17308	60.10	17.46
Gender			
Female	12430	62.68	17.44
Male	12834	60.94	19.00
Student with Disabilities			
No	19160	64.44	17.74
Yes	6108	53.49	17.40

*Note: 4 students were not reported in the Ethnicity and Gender group but they are reflected in “All Students”.