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

& K D S W H U , Q W U R G X F W L R Q

1.1 INTRODUCTION

This technical report for the Regents Examination in Algebra I will provide New York State with documentation of the purpose of the Regents Examination, scoring information, evidence

1.3 TARGET POPULATION (STANDARD 7.2)

The examinee population for the Regents Examination in Algebra I is composed of students who have completed a course in Algebra I.

Table 1 provides a demographic breakdown of all students who took the August 2017, January 2018, and June 2018 Regents Examination in Algebra I. All analyses in this report are based on the population described in Table 1. -1 (E)-3 (o)-33 (k 20.(S (E)-e)6 ((d)-3 (b)-3/ 612 79 q 0 041 (t)8

&KDSWHU &ODVVLFDO ,WHP 6WDWLVWL

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 Algebra I items.

2.1 ITEM DIFFICULTY

$W WKH PRVW JHQHUDO OHYHO DQ LWHP¶V GLIILFXOW\ LV group (e.g., grade level).

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

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 proportiove8IQ q 0 T Q q 0n3F( )8 (co)- n BT /TT1 -42.45 (e)

perform well on the Regents Examination in Algebra I would be more likely to answer any given item correctly, while low-performing students (i.e., those who perform poorly on the exam

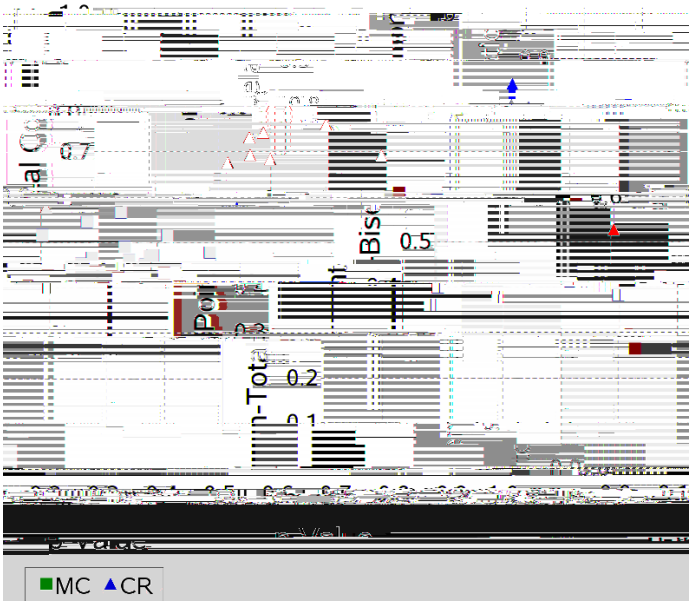

Figure 1 Scatter Plot: Regents Examination in Algebra I

Table 4 Descriptive Statistics in p-value and Point-Biserial Correlation: Regents Examination in Algebra I

Statistics	N	Mean	Min	Q1	Median	Q3	Max

2.4 OBSERVATIONS AND INTERPRETATIONS

The p-values for the MC items ranged from about 0.27 to 0.80, while the mean proportion-correct values for the CR items (Table 3) ranged from about 0.23 to 0.51. From the difficulty distributions illustrated in the plot, a wide range of item difficulties appeared on each exam, which was one test development goal.

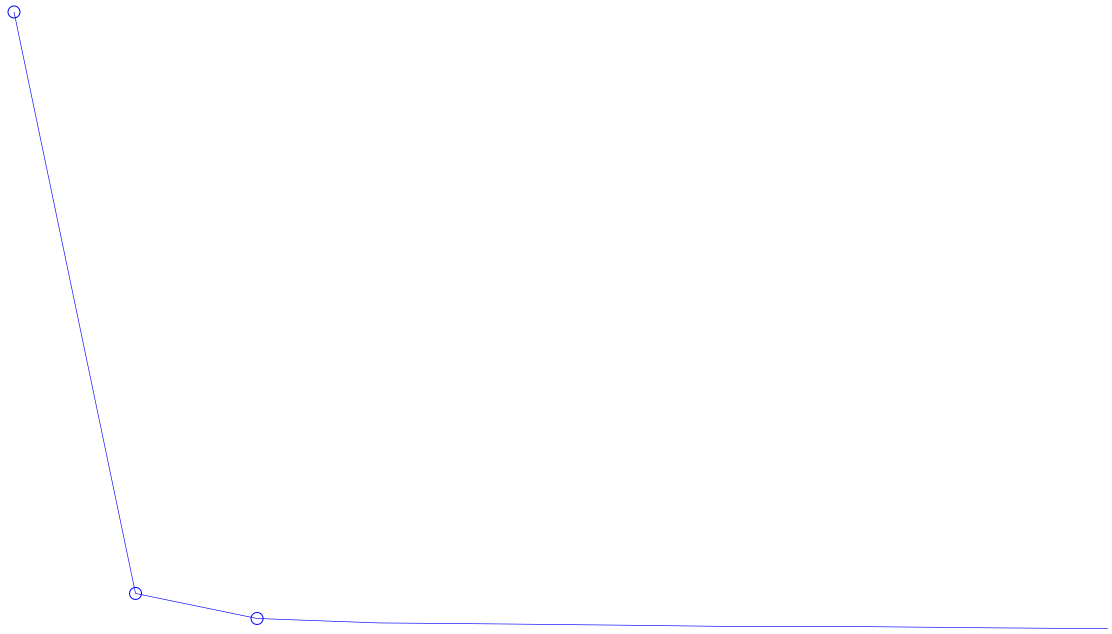

Figure 3 Scree Plot: Regents Examination in Algebra I

Local Independence

Local independence (LI) is a fundamental assumption of IRT. This means that, for statistical purposes, an examinee ¶V response to any one item should not depend on W K H H [D P L Q H H ¶V response to any other item on the test. In formal statistical terms, a test X that is comprised of items X_1, X_2 «X_n is locally independent with respect to the latent variable if, for all x = (x_1, x_2 «x_n) and ,

$$P \quad | T \quad \prod_{i=1} P(X_i \quad x_i | T \quad .$$

This formula essentially states that the probability of any pattern of responses across all items (**x**), after conditioning on the H [D P L Q H H ¶V W(T) as measured by the test, should be equal to the product of the conditional probabilities across each item (i.e., the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of local independence. A weak form of local independence (WLI) is proposed by McDonald (1979). The

distinction is important because many indicators of local dependency are actually framed by WLI. For WLI, the conditional covariances of all pairs of item responses, conditioned on the abilities, are assumed to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on 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:

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

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

Table 5

Table 6 Summary of INFIT

The relationship between raw and scale scores is explicated in the scoring tables for each administration. These tables for the August 2017, January 2018, and June 2018 administrations can be found in Appendix B. These tables are the end product of the following scaling procedure.

All Regents Examinations are equated back to a base scale, which is held constant from year to year. Specifically, they are equated to the base scale through the use of a calibrated LWHP SRRO 7KH 5DVFK GLIILFXOWLHV IURP WKH LWHPV¶ LQL test are used to equate the scale for the current administration to the base administration. For this examination, the base administration was the June 2014 administration. Scale scores from the August 2017, January 2018, and June 2018 administrations are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2016 administration.

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

$$SS = m_0 + m_1(RS) + m_2(RS)^2 + m_3(RS)^3 + m_4(RS)^4,$$

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

&KDSWHU 5HOLDELOLW\ 6WDQGDUG

Test reliability is a measure of the internal consistency of a test (Cronbach, 1951). It is a

would be pure random noise (i.e., all measurement error). If the index achieved a value of 1.0, scores would be perfectly consistent (i.e., contain no measurement error). Although values of 1.0 are never achieved in practice, it is clear that larger coefficients are more desirable because they indicate that the test scores are less influenced by random error.

Coefficient Alpha

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

A general computational formula for Coefficient Alpha is as follows:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_{Yi}^2}{\sigma_X^2}\right)$$

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

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

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

Traditional Standard Error of Measurement Confidence Intervals

The SEM is an index of the random variability in test scores reported in actual score units, which is why it has such great utility for test score users. SEMs allow statements regarding the SUHFLVLRQ RI LQGLYLGXDO WHVW VFRUHV 6(0V KHOS SODFH FRQILGHQFH EDQGV 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 Algebra I are provided in Table 7.

Table 7 Reliabilities and Standard Errors of Measurement: Regents Examination in Algebra I

Subject	Coefficient Alpha	SEM

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

$$\sigma_T = \sqrt{\sigma_x^2 - \sigma_x^2(1 - \rho_{xx})}.$$

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 VWXGHQW WKH PHDQ RI WKH GLVWULEXWLRQ RI WKH VWXGHQW¶V UHVXOWV ZRXOG EH WKH VWXGHQW¶V true score (, 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 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 V¶V estimated . 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 and the scale score is not expressible in a simple mathematical form because it is a blend of the fourth-degree polynomial relationship between the raw and

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

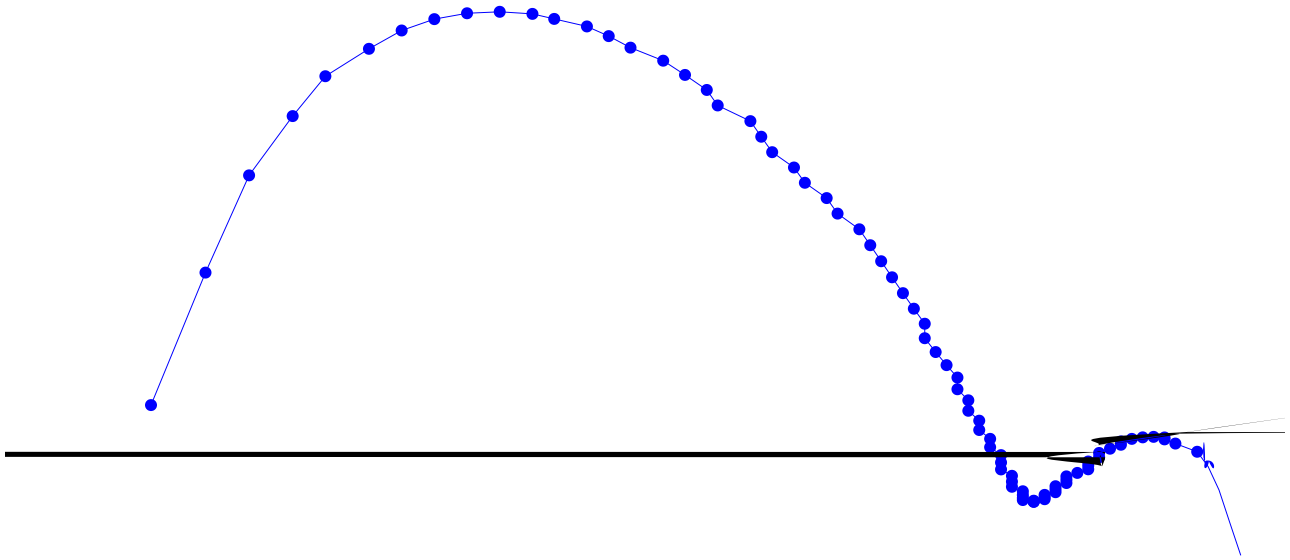

Figure 4 Conditional Standard Error Plot: Regents Examination in Algebra I

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 (o)-3 ((e1o)-)8 (a)-33 (ce)-r.92 re r2 (rd)Tj ETr (S)3.99e-3

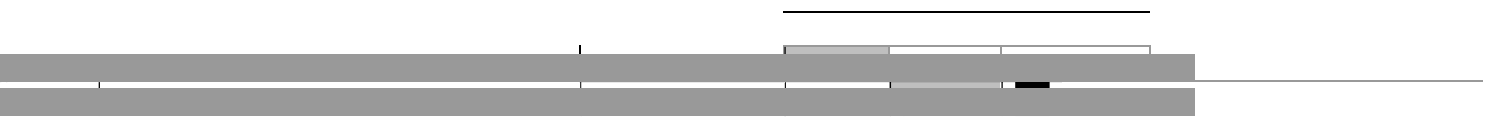

Table 9 Group Means: Regents Examination in Algebra I

Demographics	Number	Mean Scale Score	SD Scale Score
Ethnicity			
English Language Learner/Multilingual Learner			
Economically Disadvantaged			
Gender			
Student with a Disability			

*Note: Nineteen students were not reported in the Ethnicity and Gender group, EXW WKH\ DUH UHIOHFW Students. ´

4.5 STATE PERCENTILE RANKINGS

State percentile rankings based on raw score distributions are noted in Table 10. The percentiles are based on the distribution of all students taking the Regents Examination in Algebra I for the June 2018 administration. Note that the scale scores for the Regents Examination range from 0 to 100, but some scale scores may not be obtainable depending on the raw score-to-scale score relationship for a specific administration. The percentile rank(sca)-raw n BT

Table 10 State Percentile Ranking for Raw Score ±Regents Examination in Algebra I

	Percentile Rank	Scale Score	Percentile Rank	

Restating the purpose and uses of the Regents Examination in Algebra I, this exam measures examinee achievement against the New York State learning standards. The exam is prepared by teacher examination committees and New York State Education Department subject matter and testing specialists, 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, in order 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 Algebra I is intended for use in satisfying state testing requirements for students who have finished a course in Algebra I. A passing score on the exam counts toward requirements for a high school diploma, as described in the New York State diploma requirements: http://www.nysed.gov/common/nysed/files/programs/curriculum-instruction/currentdiplomarequirements2.pdf. Results of the Regents Examination in Algebra I may also be used to satisfy various locally established requirements throughout the state.

The validity of score interpretations for the Regents Examination in Algebra I 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 in order 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 disc(d6 (1)-fic-5 (t)61 (d)-3 (n)6 (d)-3 (a)--3 (isc)12 ((g1 12 Tf 287.45)Tj ET Q q 0 .01 T

Career (PARCC, 2014). The model content frameworks are located at http://www.parcconline.org/resources/educator-resources/model-content-frameworks/mathematics-model-content-framework. The standards for mathematics are located at http://www.engageny.org/resource/new-york-state-p-12-common-core-learning-standards-for-mathematics. Clarifications for Algebra I standards are located at http://www.engageny.org/resource/regents-exams-mathematics-algebra-i-standards-clarifications.

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 Algebra I is essentially the design document for constructing the exam. It provides an explicit definition of the construct 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 11 displays domain titles along with their cluster, standard, and targeted proportions of conceptual categories on the exam.

Table 11 Test Blueprint, Regents Examination in Algebra I

Conceptual Category

Only New York State-certified educators may participate in this process. The New York State Education Department asks for nominations from districts, and all recruiting is done with diversity of participants in mind, including diversity in gender, ethnicity, geographic region, and teaching experience. Educators with item-writing skills from around the state are retained to write all items for the Regents Examination in Algebra I, under strict guidelines that leverage best practices (see Appendix C). State educators also conduct all item quality and bias reviews, in order to ensure that item content is appropriate to the construct being measured and fair for all students. Finally, educators use the defi* n BT BT /TT1iFirds, (n)-stoctpn51 (th)-6caas, (n BT B)4 (ty

Math Items provide detailed information about how items are developed for the Regents Examinations. The guidelines are included in Appendix C.

Item Review Process

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

All reviewers participate in rigorous training designed to assist in a consistent interpretation of the standards throughout the item review process. This is a critical step in item development because consistency between the standards and what the items are asking examinees is a

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.

Refer to the following link for more detail on the item review criteria: https://www.engageny.org/resource/regents-exams-

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

1. Quality training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, Wong, and Kwong, 2010; Gorman & Rentsch, 2009; Schleicher, Day, Bronston, Mayes, and Riggo, 2002; Woehr & Huffcutt, 1994; J3l(m)-6 .9 792 re W*Ri-2 (won)-5-2 (.57(2)-3 (0)s re W* n BT /TT1

Because the rater is often unaware of such bias, this type of variation may be the most challenging source of variation in scoring to control and measure. Rater biases can appear as severity or leniency in applying the scoring rubric. Bias also includes phenomena such as the halo effect, which occurs when good or poor performance on one element of the rubric encourages inaccurate scoring of other elements. These types of rater bias can be effectively controlled by training practices with a strict focus on rubric requirements.

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

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

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

- The rubric should clearly specify the criteria for awarding each credit.
- The rubric should be aligned to what is asked for in the item and correspond to the knowledge or skill being assessed.
- Whenever possible, the rubric should be written to allow for alternate 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 Algebra I. Operational raters are selected based on expertise in the exam subject and are assigned a specific set of items to score. No more than one-third of the items on the test are assigned to any one rater. This 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 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 that the item developer intended them to. As documented in Table 2, the point-biseri

responding to the items as expected during item development. Refer to Chapter 2 of this report for additional details.

Differential Item Functioning

Differential item functioning (DIF) for gender was conducted following field testing of the items in 2012 ±2017. Sample sizes for subgroups based on ethnicity and English language learner/multilingual learner status were, unfortunately, too small to reliably compute DIF statistics, so only gender DIF analyses were conducted. The Mantel-Haenszel ï 6 and standardized mean difference were used to detect items that may function differently for any of these subgroups. The Mantel-Haenszel ï 6 is a conditional mean comparison of the ordered response categories for reference and focal groups combined over values of the matching Y D U L D E O H V F R U H ³ 2 U G H U H G ´ P H D Q V W K D W D U H V S R Q V H H D than a response earning a s F R U H , Pa I ³ ´ L V E H W W H U W K D Q ³ ´ D Q G V R R other hand, refers to the comparison of members from the two groups who received the same score on the matching variable ² the total test score in our analysis.

Two operational items for the June 2018 administration had DIF flags from the field test. Two items (#s 13, 20) had a moderate DIF favoring male students. One item (# 2) had a strong DIF favoring female students. The items were subsequently reviewed by content specialists. They were unable to identify content-based reasons why the items might be functioning differently between male students and female students, and they did not see any issue with using them for the operational exam.

Full differential item functioning results are reported in Appendix E of the field test reports for 2013 through 2015 and Appendix F of the 2016 and 2017 technical reports.

IRT Model Fit

Model fit for the Rasch method used to estimate location (difficulty) parameters for the items on the Regents Examination in Algebra I 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

dl i3 Td (d)-5 d 2 O q 0 0 612 792 re W* n BT /TT1 12 Tf 57.6 301.73366.91[23 Td (d)-T O q 0 0 612 792 re W* n BT

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 fn n BT /TT1 9.012 Tf 1 re W* n BT /T-

Paomem5

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

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

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

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

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 í 4 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 k8 ( )-t-3 (ry)k q 0 0e61 (co)6 (b1 1gET Q q 0 0 612 792 re)Tj R9.45 Tdw-3 (li)4 (e)-3 (d)

rv

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.

Table A.1 Test Map for August 2017 Administration

Position	Item Type	Max Points	Weight	Cluster	Mean	Point-Biserial	Rasch Difficulty	INFIT

Table A.2

Table A.3 Test Map for June 2018 Administration

Position	Item Type	Max Points	Weight	Cluster	Mean	Point-Biserial	Rasch Difficulty	INFIT

Table B.2

Table B.3 Score Table for June 2018 Administration

Raw Score	Ability	Scale Score

Raw Score	Ability	Scale Score

Raw Score	Ability	Scale Score

GUIDELINES FOR WRITING MULTIPLE-CHOICE MATH ITEMS

1. **The item measures the knowledge, skills, and proficiencies characterized by the standards within the identified cluster.**

2. **The focus of the problem or topic should be stated clearly and concisely.**
   The stem should be meaningful and convey the central problem. A multiple-choice item functions most effectively when a student is required to compare specific alternatives related to the stem. It should not be necessary for the student to read all of the alternatives to understand an item. (Hint: Cover the alternatives and read the stem on its own. Then ask yourself if the question includes the essential elements or if the essential elements are lost somewhere in the alternatives.)

3. **Include problems that come from a real-world context or problems that make use of multiple representations.**
   When using real-world problems, use formulas and equations that are real-world (e.g., the kinetic energy of an object with mass, m, and velocity, V, is k = ½ mv2). Use real-world statistics whenever possible.

4. **The item should be written in clear and simple language, with vocabulary and sentence structure kept as simple as possible.**
   Each multiple-choice item should be specific and clear. The important elements should generally appear early in the stem of an item, with qualifications and explanations following. Difficult and technical vocabulary should be avoided, unless essential for the purpose of the question.

5. **The stem should be written as a direct question or an incomplete statement**
   Direct questions are often more straightforward. However, an incomplete statement may be used to achieve simplicity, clarity, and effectiveness. Use whichever format seems more appropriate to present the item effectively.

6. **The stem should not contain irrelevant or unnecessary detail.**
   Be sure that sufficient information is provided to answer the question, but avoid excessive G H W D L O R U ³Z L Q G R Z G U H V V L Q J ´

7. **The phrase** which of the following **should not be used to refer to the alternatives; instead, use** which **followed by a noun.**
   In the stem, which of the following requires the student to read all of the alternatives before knowing what is being asked and assessed. Expressions such as which statement, which expression, which equation, and/or which graph are acceptable.

8. **The stem should include any words that must otherwise be repeated in each alternative.**
In general, the stem should contain everything the alternatives have in common or as much as possible of their common content. This practice makes an item concise. Exceptions include alternatives containing units and alternatives stated as complete sentences.

9. **The item should have one and only one correct answer.**
Items should not have two or more correct alternatives. All of the above and none of the above are not acceptable alternatives.

10. **The distractors should be plausible and attractive to students who lack the knowledge, understanding, or ability assessed by the item.**
Distractors should be designed to reflect common errors or misconceptions of students.

11. **The alternatives should be grammatically consistent with the stem.**
Use similar terminology, phrasing or sentence structure in the alternatives. Alternatives must use consistent language, including verb tense, nouns, singular/plurals, and declarative statements. Place a period at the end of an alternative only if the alternative by itself is a complete sentence.

12. **The alternatives should be parallel with one another in form.**
The length, complexity and specificity of the alternatives should be similar. For example, if the stem refers to a process, then all the alternatives must be processes. Avoid the use of absolutes such as always and never in phrasing alternatives.

13. **The alternatives should be arranged in logical order, when possible.**
When the alternatives consist of numbers and letters, they should ordinarily be arranged in ascending or descending order. An exception would be when the number of an alternative and the value of that alternative are the same. For example: (1) 1 (2) 2 (3) 0 (4) 4.

14. **The alternatives should be independent and mutually exclusive.**
Alternatives that are synonymous or overlap in meaning often assist the student in eliminating distractors. á B: T; L =4 E Í ¶ á@5 @á … á ë Å E >á • < á ë Å A

15. **The item should not contain extraneous clues to the correct answer.**
Any aspect of the item that provides an unintended clue that can be used to select or eliminate an alternative should be avoided. For example, any term that appears in the stem should not appear in only one of the alternatives.

16. **Notation and symbols as presented on Common Core examinations should be used consistently.**
For example, AB

Sensitivity/Bias	Yes	No	n/a	Explain or Describe

Math Art	Yes	No	n/a	Explain or Describe
1. The artwork clearly relates to the item and is important as an aspect of the problem-solving experience.				
2. The details in the artwork accurately and appropriately portray numbers/concepts contained in text or in lieu of text. Items should be drawn to scale as much as possible. By default, we do not include the text ³1RGUDZQ WR VF every item; however, if a figure is drawn and there is a distortion in the figure, it should be indicated under the art that the ILJXUH LV ³QRW VFDOH´ 7KH GH distortion should not be actively misleading.				
3. Graphics are clear (symbols are highly distinguished, free from clutter, at a reasonable scale, etc.).				

Math Art

Item Alignment	Yes	No	n/a

Item Alignment	Yes	No	n/a	Explain or Describe
4. The item requires students to show understanding of key aspects of the standard. ,I ³1R ´ ZKLFK D\ not attended to? For constructed response items, it is important that the item be solved through an understanding of the key point of the standard. For example, if the language of the standard FDOOV IRU ³VKRZ items should actually involve proof to be aligned, not simply the ability to solve a related problem or perform a related manipulation.				
5. Does the question lend itself to being answered using a below-grade-level standard rather than the skills/concepts references in the on-grade-level standard?				
6. The item requires the student to use skills referenced in the primary standard and any additional standards listed.				

Item Alignment	Yes	No	n/a	Explain or Describe
7. The item includes grade/course-appropriate standard numbers/variables (e.g., students are asked to solve questions using numbers/variables that are grade-appropriate). Note: This includes the parameters outlined in the PARCC Pathways document for guidance on how some standards are split across A1 and A2.				

8. The item is aligned to the correct primary Multiple Representations(s).
 , I ³ 1 R ´ L Q G L F D V
MR code(s).

Application/Modeling Items	Yes	No	n/a

Mathematic Correctness	Yes	No	n/a	Explain or Describe
1. The stem addresses a central math concept, either implicitly or explicitly.				
2. The math presented in stem is clear, accurate, and conceptually plausible.				
3. At least one strategy exists that is on grade level to solve the problem.				
4. If there is more than one strategy, regardless of the strategy employed, the same correct answer will be achieved.				
5. There is a rationale for the correct response that is aligned to the language of the Standards and that demonstrates knowledge and/or aemonstrates kn				

Mathematic Correctness	Yes	No	n/a	Explain or Describe
7. For MCQs: Is answer Choice 2 plausible or the correct answer? If not, why?				
8. For MCQs: Is answer Choice 3 plausible or the correct answer? If not, why?				
9. For MCQs: Is answer Choice 4 plausible or the correct answer? If not, why?				

Constructed Response and All Regents	Yes	No	n/a	Explain or Describe

Overarching Comments	Yes	No	n/a	Explain or Describe
1. The item is aligned to standard.				

2. The item is rigorous.

   The math should be sound, tight, challenging, and at

Guidelines for Writing Constructed-Response Math Items

1. **The item measures the knowledge, skills, and proficiencies characterized by the standards within the identified cluster.**

2. **The focus of the problem or topic should be stated clearly and concisely**.
   The item should be meaningful, address important knowledge and skills, and focus on key concepts.

3. **Include problems that come from a real-world context or problems that make use of multiple representations.**
   When using real-world problems, use formulas and equations that are real-world (e.g., the kinetic energy of an object with mass, m, and velocity, V is k = ½ mv2). Use real-world statistics whenever possible.

4. **The item should be written with terminology, vocabulary and sentence structure kept as simple as possible. The item should be free of irrelevant or unnecessary detail.**
   The important elemen

9. The item should be written so that the student does not have to identify units of measurement in the answer, unless the question is testing dimensional analysis.
)RU H[DPSOH FRQVLGHU WKH TXHVWLRQ ³$ FLUFOH KDV the number of centimeters in the length of the arc intercepted by a central angle measuring UDGLDQV ´ 6WXGHQWV ZRXOG UHFHLYH FUHGLW I EH SHQDOL]HG IRU ZULWLQJ ³ FP ´

10. The item should be written to require a specific form of answer.
3KUDVHV OLNH è³ Ĺ tQ the nearest tenth ´ DQG ³LQ VLPSOHVW UDGL simplify the writing of the rubric for these types of items.

11. Items that require students to explain in words are encouraged.
One of the emphases of the Common Core standards is to foster student ability to communicate mathematical thinking. An example is to have students construct viable arguments such as to make conjectures, analyze situations or justify conclusions. These items would require students to demonstrate precision of knowledge in their responses.

12. Items may be broken into multiple parts that may be labeled a**,** b**,** c**, etc.**
Clear division of the parts of the problems may simplify the writing of the rubric for these types of items.

13. Notation and symbols as presented on Common Core examinations should be used consistently.
For example, AB means the length of line segment AB, ‘$means line segment AB, m áA means the number of degrees in the measure of angle A, etc.

Table D.1

Table D.2 Constructed-Response Item Analysis Summary: Regents Examination in Algebra I

Item	Min. score	Max. score	Number of Students	Mean	SD	p-Value	Point-Biserial

Table D.3 Descriptive Statistics in p-value and Point-Biserial Correlation: Regents Examination in Algebra I

p							

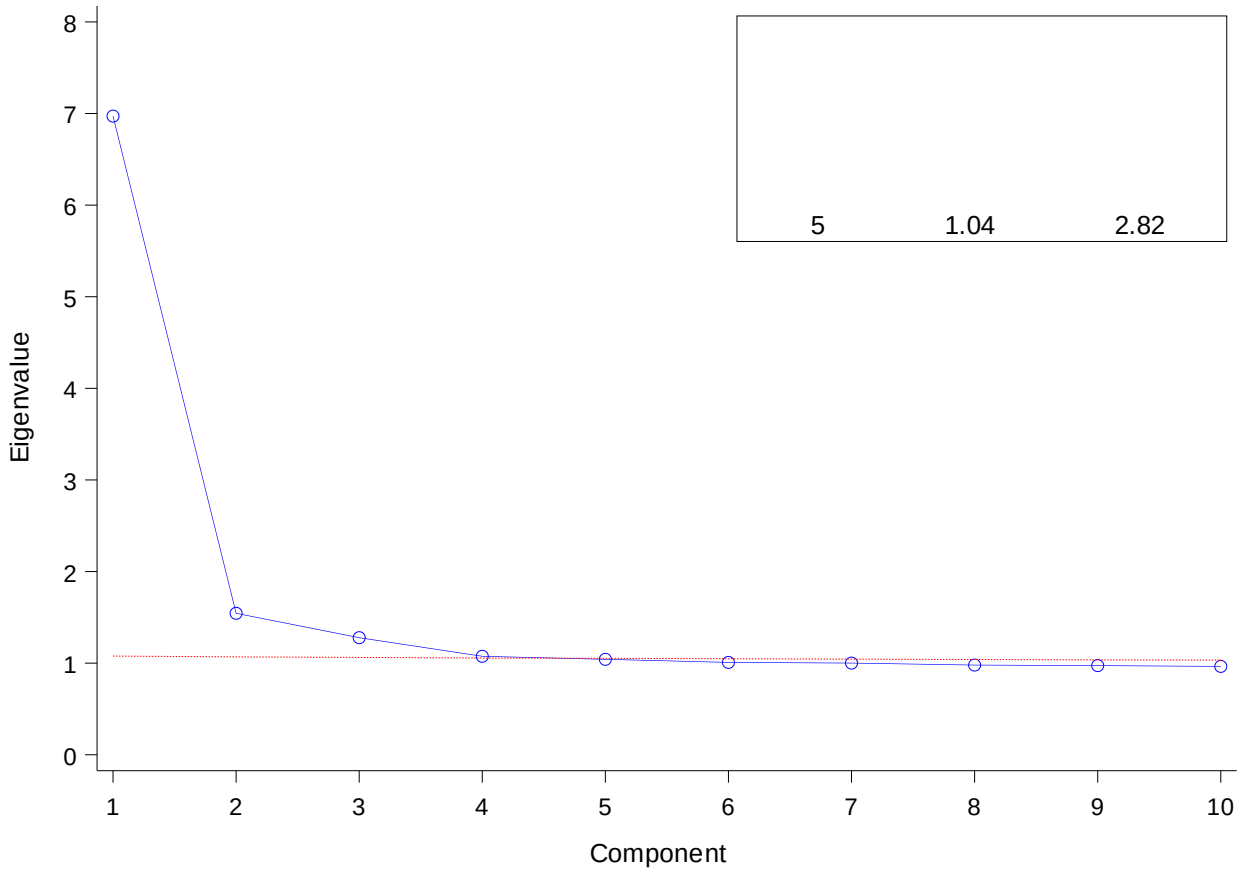

Figure D.3 Scree Plot: Regents Examination in Algebra I

Table D.4 Summary of Item Residual Correlations: Regents Examination in Algebra I

Table D.5 Summary of INFIT Mean Square Statistics: Regents Examination in Algebra I

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]

Table D.6 Reliabilities and Standard Errors of Measurement: Regents Examination in Algebra I

Subject	Coefficient Alpha	SEM

Table D.7 Decision Consistency and Accuracy Results: Regents Examination in Algebra I

Statistic	1/2	2/3	3/4	4/5

Figure D.4 Conditional Standard Error Plot: Regents Examination in Algebra I

Table E.1 Multiple-Choice Item Analysis Summary: Regents Examination in Algebra I

Item	Number	p-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3

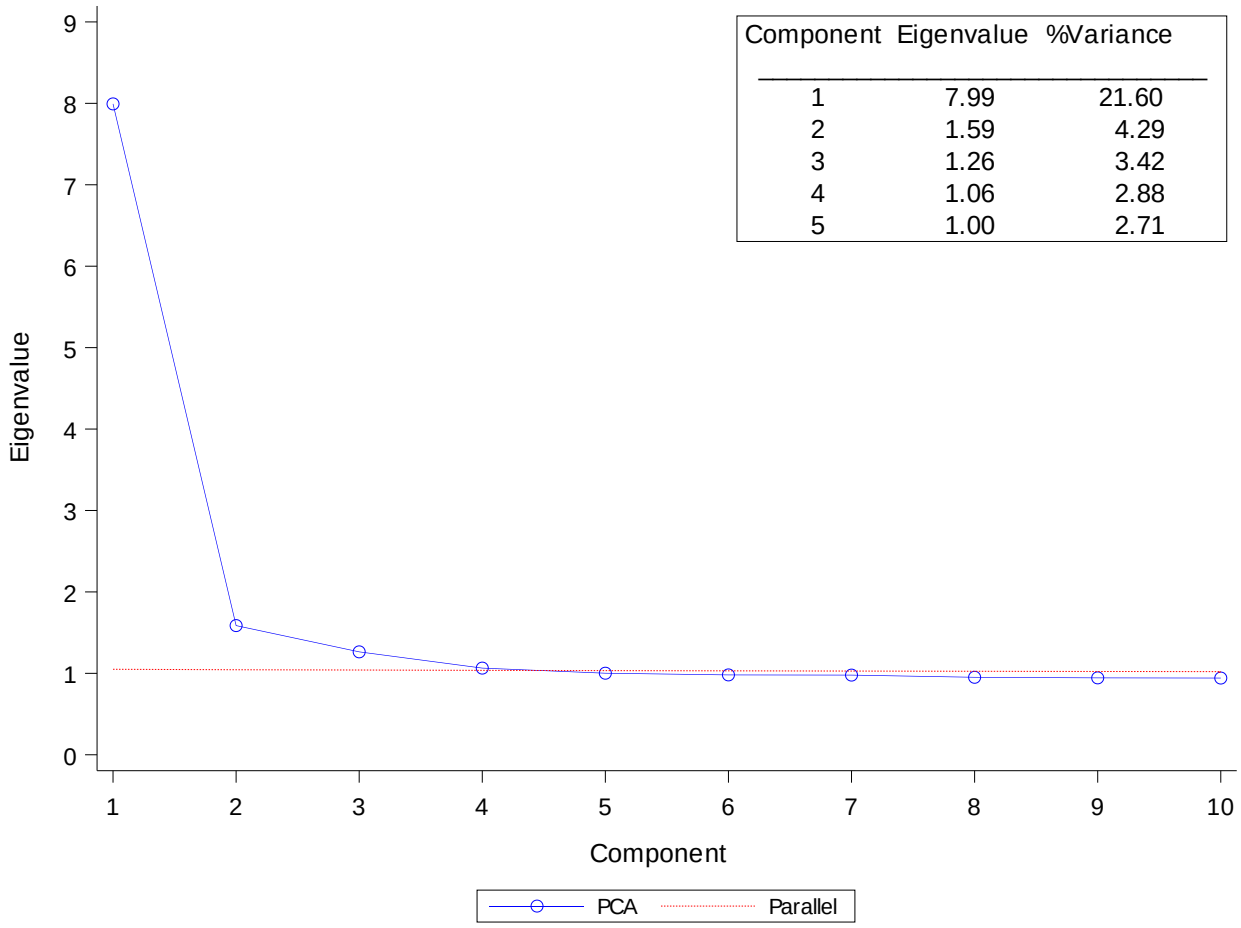

Figure E.3 Scree Plot: Regents Examination in Algebra I

Table E.4 Summary of Item Residual Correlations: Regents Examination in Algebra I

Statistic Type	Value

Table E.5 Summary of INFIT Mean Square Statistics: Regents Examination in Algebra I

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]

Figure E.4 Conditional Standard Error Plot: Regents Examination in Algebra I