New York State Regents Examination in

Algebra I epa(r.7 t) 20(m)-39 entn r.7 sn

Copyright

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

Contents

S

List of Tables

matter and testing specialists. Further, it provides teachers and students with important information about student learning and performance against the established curriculum standards. Results of this exam may be used to identify student strengths and needs in order to guide classroom teaching and learning. The exam also provides students, parents, counselors, administrators, and college admissions officers with objective and easily understood achievement information that may be used to inform empirically-based educational and vocational decisions about students.

As a state-provided objective benchmark, the Regents Examination in 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.

For the 2020-2021 school year, due to the COVID-19 pandemic, eligible students could be exempted from the associated diploma requirement for this Regents examination. Details of those exemptions may be found here: http://www.nysed.gov/common/nysed/files/programs/state-assessment/memo-june-august-2021-assessments.pdf and http://www.nysed.gov/common/nysed/files/programs/state-assessment/faq-june-august-2021-assessments.pdf.

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. For the 2020-2021 school year, due to the COVID-19 pandemic, those students who were receiving entirely remote instruction were not required to come to school for the sole purpose of taking a Regents examination. Moreover, this examination was only to be administered where schools and districts could ensure the health and safety of students and teachers. Eligible students were exempted from the associated diploma requirement for this Regents examination. Details of those exemptions can be found here: http://www.nysed.gov/common/nysed/files/programs/state-assessment/memo-june-august-2021-assessments.pdf and http://www.nysed.gov/common/nysed/files/programs/state-assessment/faq-june-august-2021-assessments.pdf.

Table 1 provides a demographic breakdown of all students who took the June 2021 Regents Examination in Algebra I. All analyses in this report are based on the population described in Table 1. Annual Regents Examination results reported in the New York State Report Cards are those reported in the Student Information Repository System (SIRS) as of the reporting deadline. As noted previously, for a typical year, the results would include those exams administered in August, January, and June; though for 2021, this assessment was administered once, in June (see http://data.nysed.gov/). Typically, if a student takes the same exam multiple times in the year, only the highest score is included

in these results. Item-level data used for the analyses in this report are reported by districts on a similar timeline, yet through a different collection system.

When compared with the number of expected test takers, based on recent examination administrations, approximately 20% of students took the Regents Examination in Algebra I due to circumstances related to the COVID-19 pandemic.

Table 1 Total Examinee Population: Regents Examination in Algebra I

	June Admin*	
Demographics	Number	Percent
All Studetuds Nuuuw 16 >>6.7 Td>>52DC (u)00A,DC Tm [1(u)00m [1(uw 1.265.7 Td>>842DC (u)00A)5.7.93.7 Td>>E		

Item	Number	p-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
15	52,842	0.62	0.49	0.55	-0.25	-0.29	-0.28
16	52,842	0.44	0.50	0.29	-0.09	-0.14	-0.19
17	52,842	0.60	0.49	0.53	-0.23	-0.18	-0.37
18	52,842	0.48	0.50	0.52	-0.19	-0.22	-0.28
19	52,842	0.55	0.50	0.48	-0.28	-0.29	-0.10
20	52,842	0.39	0.49	0.38	-0.07	-0.17	-0.22
21	52,842	0.61	0.49	0.49	-0.11	-0.31	-0.27
22	52,842	0.46	0.50	0.44	-0.31	-0.11	-0.16
23	52,842	0.44	0.50	0.37	-0.04	-0.18	-0.24
24	52,842	0.44	0.50	0.43	-0.20	-0.20	-0.15

Table 3 Constructed- Response Item Analysis Summary: Regents Examination

2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOT

Figure 1 shows a scatter plot of item discrimination values (ym004 Tc 0.018 Tww 0.3 0 Td18.

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

Figure 3 shows

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's response to any one item should not depend on the examinee's response to any other item on the test. In formal statistical terms, test X, which comprises items X_1, X_2, …X_n is locally independent with respect to the latent variable if, for all $x = (x_1, x_2, \ldots x_n)$ and ,

.

This formula essentially states that the probability of any pattern of responses across all items (x), after conditioning on the examinee's true score ( ) as measured by the test, should be equal to the product of the conditional probabilities across each item (i.e., the

Table 5 Summary of Item Residual Correlations: Regents Examination in Algebra I

Statistic Type	Value
N	666
Mean	-0.02
SD	0.03
Minimum	-0.12
P_{10}	-0.06
P_{25}	-0.04
P_{50}	-0.02
P_{75}	0.00
P_{90}	0.02
Maximum	0.11
>|0.20|	0

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 *T*relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure (i.e., informative, on-target responses).

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

Table 6 presents the summary statistics of INF()7 (es)4 (e)10 (nt)2 (s)4 (ITd ( )Tj 0.004 Tc988 T

Table 6 Summary of INFIT Mean Square Statistics: Regents Examination in Algebra I

The relationship between raw and scale scores is explicated in the scoring table for this administration. The table for the June 2021 administration may be found in Appendix B. This table is the end product of the following scaling procedure.

All Regents Examinations are equated back to a base scale, which is held constant from year to year. Specifically, they are equated to the base scale through the use of a calibrated item pool. The Rasch difficulties from the items' initial administration in a previous year's field test are used to equate the scale for the current administration to the base administration. For this examination, the base administration was the June 2014 administration. Scale scores from the June 2021 administration are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2014 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:

$$55 = I4\ Û45^{8} + I3\ Û45^{7} + I2\ Û45^{6} + I1\ Û45 + I0,$$

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.

The Rasch difes 8eles

to each raw score point on the new form, using the theta-to-scale score relationship established in the base year. This was done via linear interpolation.

This process results in a relationship between the raw scores on the form and the overall scale scores. The scale scores corresponding to each raw score are then rounded to the nearest integer for reporting on the conversion chart (posted at the close of each administration). The only exceptions are for the minimum and maximum raw scores and the raw scores that correspond to the scaled cut scores of 55, 65, 80, and 85.

The minimum (zero) and maximum possible ram and2 (het)2 (a)-04 (ut)2 (s(m)7 (ax)4 (i)1(s)4 (c(

Chapter 4: Reliability (Standard 2)

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

Although values of 1.0 are never achieved in practice, larger coefficients are more desirable because they indicate that the test scores are less influenced by random error.

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:

where N is the number of parts (items), is the variance of the observed total test scores, and is the variance of part i.

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

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

Traditional Standard Error of Measurement

The standard error

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.[5] 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
Algebra I	0.93	5.65

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:

.

(he)68ct <<60.-59052 34 -33.2sg0(u)1dnd T* (l0 (bep.1)14(e.)2 (ear-34 -33 w)6 (i)6 (( )]Tden.23s)’)1

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

/EMC /P <</MCID 3o>>BDC

When compared with the number of expected test takers, based on recent examination administrations, approximately 20% of students took the Regents Examination in Algebra I due to circumstances related to the COVID-19 pandemic. Therefore, mean scale scores based on demographic variables (e.g., race/ethnicity, gender, etc.) were not calculated as the sample of students who took the June 2021 administration of the examination were not representative of all students enrolled in an Algebra I course during the 2020-2021 school year.

The overall mean scale score was computed based on all students who took the Regents Examination in Algebra I. The result is reported in Table 9.

Table 9 Mean: Regents Examination in Algebra I

Demographics	Number	Mean Scale Score	SD Scale Score

Chapter 5: Validity (Standard 1)

This exam measures examinee achievement against the New York State Learning Standards and was prepared by teacher examination committees and New York State Education Department subject matter and testing specialists. Further, it provides teachers and students with important information about student learning and performance against the established curriculum standards. Results of this exam may be used to identify student strengths and needs, in order to guide classroom teaching and learning. The

The Regents Examination in Algebra I measures student achievement on the NYS P–12 Learning Standards for Mathematics, consistent with the Model Content Frameworks for Mathematics provided by the Partnership for the Assessment of Readiness for College and Career (PARCC, 2014). The model content frameworks are located at https://www.engageny.org/resource/grades-9-12-mathematics-curriculum-map-and-course-overviews

Only New York State-certified educators may participate in this process. The New York State Education Department asks for nominations from districts, and all recruiting is done with diversity of participants in mind, including diversity in gender, ethnicity, geographic region, and teaching experience. Educators with item-writing skills from throughout the state are retained to write all items for the Regents Examination in 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 defined standards, test blueprint targets, and statistical information generated during field testing, in order 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.

Figure 7 New York State Education Department Test Development Process

Initial item development is conducted under the criteria and guidance provided by multiple documents, including the blueprint, item writing criteria, and a content verification checklist. Both MC and CR items are included in the Regents Examination in Algebra I, a

The fourth category of the item review, criteria framework, 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, MC 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-mathematics-item-criteria-checklist.

Ongoing attention is also given to the relevance of the standards used to guide curriculum and assessment. Consistent with a desire to assess this relevance, the New York State Education Department (NYSED)

Plans). Full test administration procedures are available at http://www.nysed.gov/state-assessment/high-school-regents-examinations.

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

1. Quality training (Hoyt & Kerns, 1999; Lumley & McNamara, 1995; Wang, Wong, & Kwong, 2010; Gorman & Rentsch, 2009; Schleicher, Day, Bronston, Mayes, & Riggo, 2002; Woehr & Huffcutt, 1994; Johnson, Penny, & Gordon, 2008; Weigle, 1998)
2. Detection and correction of rating bias (McQueen & Congdon, 1997; Congdon & McQueen, 2000; Myford & Wolfe, 2009; Barkaoui, 2011; Patz, Junker, Johnson, & Mariano, 2002)
3. Consistency or reliability of ratings (Congdon & McQueen, 2000; Harik, Clauser, Grabovsky, Nungester, Swanson, & Nandakumar, 2009; McQueen & Congdon, 1997; Myford & Wolfe, 2009; Mero & Motowidlo, 1995; Weinrott & Jones, 1984)
4. Rubric designs that facilitate consistency of ratings (Pecheone & Chung, 2007; Wolfe & Gitomer, 2000; Cronbach, Linn, Brennan, & Haertel, 1995; Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000; Smith, 1993; Leacock, Gonzalez, & Conarroe, 2014).

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

To review and select these anchor papers, New York State 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 CR items. The final model response sets for the June 2021 administration of the Regents Examination in Algebra I are located at http://www.nysedregents.org/algebraone/.

During the range-finding and field test-scoring processes, it is important to be aware of and to control for sources of variation in scoring. One possible source of variation in CR scores is unintended rater bias associated with items and examinee responses. The rater is often unaware of such bias, so this type of variation may be the most challenging source of variation in scoring to control and measure. Rater biases can appear as severity or leniency in applying the scoring rubric. Bias also includes phenomena such as the halo effect, 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 New York State educators begins with a review and discussion of actual student work on CR 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 CR items. When monitoring and correction do not occur during scoring, construct-irrelevant variation may be introduced. Accordingly, a scoring lead may be assigned to review the consistency of scoring for the lead's the .i 0.0040 (d)- (e l)6. Tc 08a(c)4c6iewore6 3f (t)st1(ant)12 ( )-13(e)10 (n

effect of any potential bias of a single rater on individual examinees. Additionally, raters are not allowed to score the responses of their own students.

Statistical Analysis

One statistic that is useful for evaluating the response processes for MC items is an item's point-biserial correlation on the distractors. A high point-biserial on a distractor may indicate that students are not able to identify the correct response for a reason other than the difficulty of the item. A finding of poor model fit for an item may also support a finding that examinees are not responding in the way in which the item developer intended. As

Table 2 and Table 3 provide point-biserial values on the correct responses; Table 2 also provides point-biserial values on the three distractors. The values for correct answers are 0.29 or higher for all items, indicating that most items are discriminating well between

Classification Consistency and Accuracy

A decision consistency analysis measures the agreement between the classifications based on two non-overlapping, equally difficult forms of the test. If two parallel forms of the test were given to the same students, the consistency of the measure would be reflected by the extent to which the classification decisions

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 Algebra I.

In terms of predictive validity, time is a fundamental constraint on gathering evidence. The gold standard for supporting the validity of predictive statements about test scores requires empirical evidence of the relationship between test scores and future performance on a defined characteristic. To the extent that the objective of the standards is to prepare students for meeting graduation requirements, it will be important to gather evidence of this empirical relationship over time.

5.5 EVIDENCE BASED ON TESTING CONSEQUENCES

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

Cronbach (1988), on the other hand, argues that negative consequences could invalidate test use. From this perspective, the test user is obligated to make the case for test use and to ensure appropriate and supported uses. Regardless of perspective on the nature of consequential validity, it is important to caution against uses that are not supported by the validity claims documented for this test. For example, use of this test to predict examinee scores on other tests is not directly supported by either the stated purposes or 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 Algebra I, the exam is most commonly used to inform admissions and course placement decisions. Such uses can be considered reasonable, assuming that the competencies demonstrated in the Regents Examination in Algebra I 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 Algebra I and to assess their appropriateness for the inferences being made about course placement.

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

References

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

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

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

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

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

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

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), Test Validity'i o24 0 8d [(0)7 (2p)10 p (.)]T (297)J 0 Tc 0 Tw 3.446

Tc 0 Tw 1.67 0 Td ( )T121.-)TP1 0 Tc -Tc -0.002 Tw 1.5 0 Td [(3)10Hitpeces,
(Eortivern I. (&)1ag(ent) (i11i)6.1 ionbac alh/TT2 1 Tf 0 Tc 0 Tw 5.456 0.004 Tc

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

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

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

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

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

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

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

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

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). Assessing performance. New York, NY: The Guilford Press.

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

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

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

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

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

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

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. Journal of Applied Measurement, 9(3), 200-15.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. Multivariate Behavioral Resev 12, 1te,

McDonald1 R. (M4J(M)17 (ul,f -0.0-2 (.)C0 (el)2 (.)gas)4 (on)5 (or)7 (mP)(he s)4 (J(M)17 2 (i)6 (s)4 ((t)12 7)5

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

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

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

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

REVIEW CRITERIA CHECKLIST FOR POTENTIAL MATH ITEMS

The following list of criteria will be used to train item writers and then to review items for possible inclusion on test forms.

Language Use23412s th23412m e23412st te23412rms2.16 25 1oe512 1o2101 0T* [(pos)412				

Sensitivity/Bias	Yes	No	N/A	Explain or Describe
1. The item is free of content that might be deemed offensive to groups of students, based upon culture, religion, race, ethnicity, gender, geographic location, ability, socioeconomic status, etc.				

Math Art	Yes	No	N/A	Explain or Describe
4. Visual load requirements are reasonable (interpreting graphic does not confuse underlying construct) and as simple as possible to present the prompt. “Visual load” refers to the amount of visual/graphic				

Item Alignment	Yes	No	N/A	Explain or Describe
4. The item requires students to show understanding of key aspects of the standard. If “No,” which aspects are not attended to? For constructed-response				

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 Representation(s). If "No," indicate the correct MR code(s).				
9. The item expects students to use a formula that is: - from a standard for an earlier grade level (i.e., prior knowledge); - part of the current mathematics curriculum; - not from another content area (e.g., physics). If "No," the formula should be in the item stem. For example, the formula for kinetic energy from physics should be included in the item stem.				

P

Application/Modeling Items	Yes	No	N/A	Explain or Describe
5. Figures/numbers/concepts used in modeling/application as well as in the response are realistic (e.g., downloads cost 99 cents, the side of a house isn't 2x-32 long).				
6. Modeling scenario is presented in the most realistic and simple manner possible.				
7. Modeling/application scenario does not assume outside knowledge (e.g., approximate weight of paper, definition of a micron).				
8. Modeling/application scenario provides all necessary information for student to apply math concepts.				
9. Item does not clue students to which math strategy is needed to solve, but rather allows the student to choose a strategy to solve the item correctly. For example, we should not tell students to use Pythagorean theorem, but rather allow them to decide which approach to solving is appropriate.				

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?				
