New York State Regents Examination in Physical Setting/Earth Science

2021 Technical Report

Prepared for the New York State Education Department
by Pearson

March 2022

Contents

COPYRIGHT II

CHAPTER 1: INTRODUCTION 1

CHAPTER 2: CLASSICAL ITEM STATISTICS (STANDARD 4.10) 4

CHAPTER 3: IRT CALIBRATIONS, EQUATING, AND SCALING (STANDARDS 2 AND 4.10) 9

List of Tables

P

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 Physical Setting/Earth Science is intended for use in satisfying state testing requirements for students who have finished a course in Physical Setting/Earth Science. 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 Physical Setting/Earth Science 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 can be found here: http://www.nysed.gov/common/nysed/files/programs/state-assessment/memo-june-august-2021-assessm0 Tc 0 Tt2aeln0 (n).u54na527.16 4u Tt2ETEMC /Span 0 Tnan /MCID 2 B78Nh84ipR(sse)

Table 1 Total Examinee Population: Regents Examination in Physical Setting/Earth Science

	June Admin.*
[redacted]	

Chapter 2: Classical Item Statistics (Standard 4.10)

This chapter provides an overview of the two most familiar item-level statistics obtained from classical item analysis: item difficulty and item discrimination. The following results pertain to the operational (m)-3 (di)P3gessri Ex( ) (i)-3 (a)9.9 ionnte(ov)4 (P)1 (e o)10y4 (i)6 (c)4 (al)6 (i)6 (tS)1

poorly on the exam overall) would be more likely to answer the same item incorrectly. Pearson's product-moment correlation coefficient (also commonly referred to as a point-biserial correlation) between item scores and test scores is used to indicate discrimination (Pearson, 1896). The correlation coefficient can range from ī1.0 to +1.0. If high-scoring students tend to get the item correct while low-scoring students do not, the correlation between the item score and the total test score will be both positive and noticeably large in its magnitude (i.e., above zero), meaning that the item is likely discriminating well between high- and low-performing students. Point-biserial values are computed for each answer option, including correct and incorrect options (commonly referred to as "distractors"). Point-biserial values for each distractor are an important part of the analysis. The point-biserial values on the distractors are typically negative. Positive values can indicate that higher-performing students are selecting an incorrect answer or that the item key for the correct answer should be checked.

Table 2 and Table 3 provide the point-biserial values on the correct response and three distractors (Table 2, only) for th-0.002c8J10 (r(ng)8J10 t)2 (0.004 Tc 6(in)-622c8J100)PoTc 6(in)-622

Item	Number of Students	*p*-Value	SD	Point-Biserial

( ) $\frac{(\quad)}{(\quad)}$.

A parallel analysis (Horn, 1965) can be further helpful in distinguishing 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 Physical Setting/Earth Science, 100 random data sets of sizes equal to the original data were created. For each random data set, a PCA was performed, and the resulting eigenvalues stored. Then for each component, the upper 95th percentile value of the distribution of the 100 eigenvalues from the random data sets was plotted. Given the size of the data generated for the parallel analysis, the reference line is essentially equivalent to plotting a reference line for an eigenvalue of 1.

Figure 3 shows the PCA and parallel analysis results for the Regents Examination in Physical Setting/Earth Science. The results include the eigenvalues and the percentage of variance explained for the first five components, as well as the scree plots. The scree plots show the eigenvalues plotted by component number and the results of a parallel analysis. Although the total number of components in the PCA is the same as the total number of items in a test, Figure 3 shows only the first 10 components. This view is sufficient for interpretation because components are listed in descending eigenvalue order. The fact that the eigenvalues for components 2 through 10 are much lower than the first component demonstrates that there is only one dominant component, showing evidence of unidimensionality.

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 than those from the parallel analysis.

As shown in Figure 3, the primary dimension explained 22.97 percent of the total variance for the Regents Examination in Physical Setting/Earth Sentanc.he in1valuo f the se3ohe

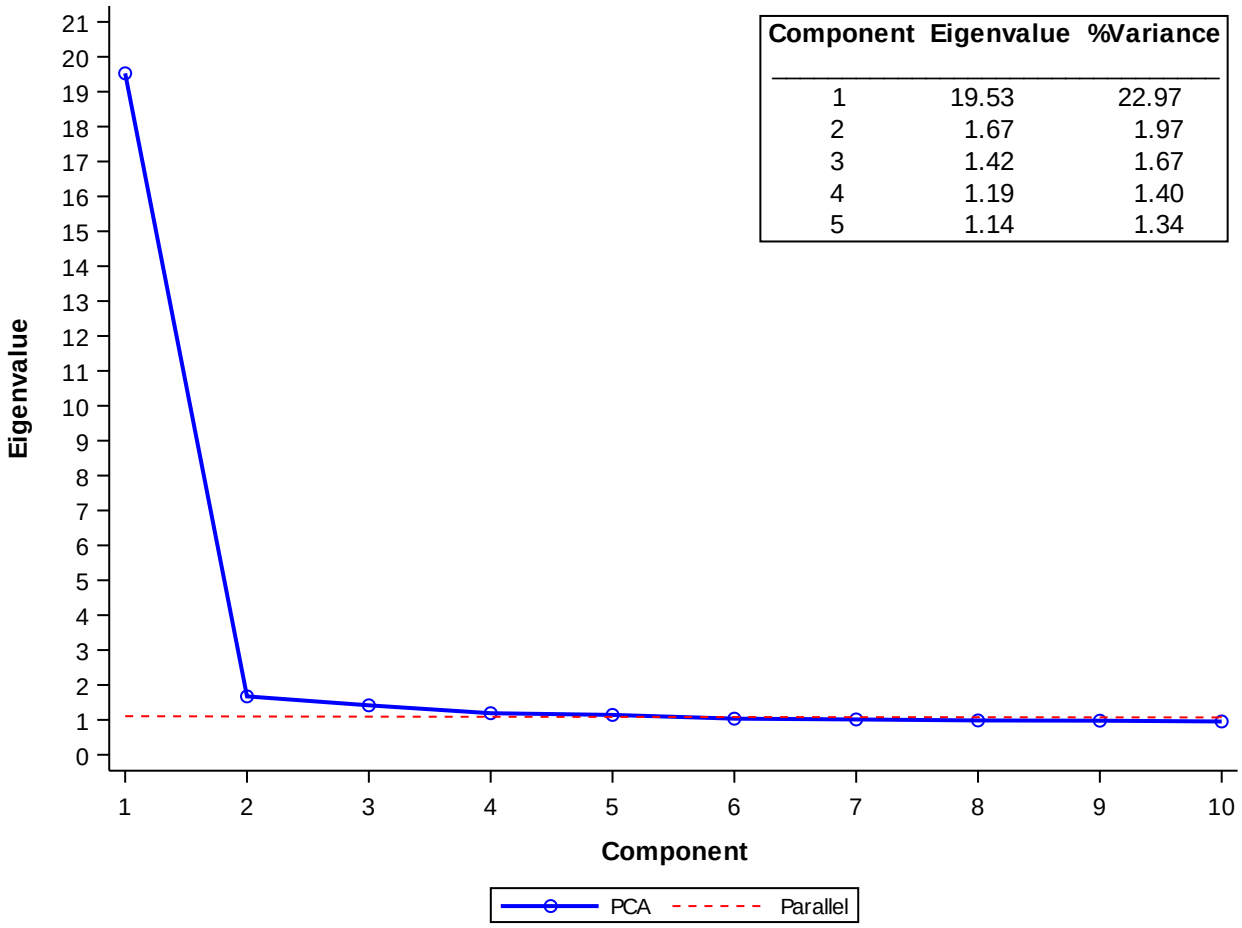

Figure 3 Scree Plot: Regents Examination in Physical Setting/Earth Science

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 θ if, for all $x = (x_1, x_2, \ldots x_n)$ and θ,

$$P(\quad = \quad | \theta) = \prod_{i=1}^{I} P(X_i = x_i | \theta).$$

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

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

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

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

Residual item correlations, provided in WINSTEPS for each item pair, were used to assess the local dependence between the Regents Examination in Physical Setting/Earth Science 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 are determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. It is noted that the raw score residual correlation essentially corresponds to Yen's Q_3 index, a popular statistic used to assess local independence. The expected value for the Q_3 statistic is approximately $-1/(k-1)$ when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected Q_3 values should be approximately -0.01 for the items on the exam. Index values that are greater than 0.20 indicate a degree of local dependence that should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS was used for these analyses. Table 5 shows the summary statistics — mean, standard deviation, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, P_{90}) — for all the residual correlations for each test. The total number of item pairs (N) and the number of pairs with residual correlations greater than 0.20 are also reported in this table. There is no item pair with residual correlations greater than 0.20. The mean residual correlations are slightly negative, and the values were close to -0.01. The vast majority of the correlations are very small, suggesting that local item independence generally holds for the Regents Examination in Physical Setting/Earth Science.

this examination, the base administration was the June 2004 administration. Scale scores from the June 2021 administrations are on the same scale and can be directly compared to scale scores on all previous administrations back to the June 2004 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

with the minimum score; if any raw scores other than zero have scale scores that round to zero, their scale scores are instead set equal to one.

With regard to the cuts, if two or more scale scores round to 55, 65, or 85, the lowest raw score's scale score is set equal to 55, 65, or 85 and the scale scores corresponding to the higher raw scores are set to 56, 66, or 86, as appropriate. If no scale score rounds to these critical cuts, then the raw score with the largest scale score that is less than the cut is set equal to the cut. The overarching principle, when two raw scores both round to either scale score cut, is that the lower of the raw scores is always assigned to be equal to the cut so that students are never penalized for this ambiguity.

The New York State Regents Examination in Physical Setting/Earth Science contains usually both a written and a lab (or performance) component. The lab component remains constant from year to year, while the written component is refreshed with each successive administration. However, the June 2021 examination does not contain a lab component due to COVID-19. Only the written component was addressed in the present report. During a typical school year, students taking the assessment receive two scale scores, one from each of the two components. The scale score of the performance component is similar to that of the written component in that it ranges from 0 to 100, and it was subjected to a standard setting that determined the raw score points that were fixed to the scale scores of 65 and 85. A student's final scale score on the assessment is a weighted sum of written and performance scale scores and is determined via the following formula:

$$SS_{Total} = 0.15*SS_{Performance} + 0.85*SS_{Written,}$$

where SS_{Total} is the total scale score and $SS_{Performance}$ and $SS_{Written}$ are the scale scores from the performance and written components of the test, respectively.

Although a number of procedures are available, a well-known method developed by Livingston and Lewis (1995) that implements a specific true score model is used.

Several factors might affect decision consistency and accuracy. One important factor is the reliability of the scores. All other things being equal, more reliable test scores tend to result in more similar reclassifications and less measurement error. Another factor is the location of the cut scoreblelocatheard .9 (d (.)2 0 (ur)7)Tj095 0 .15 Td2925 105o(2 (i)6 (oe2 (m)710 (r)4 (ai)6e1.15 Tdc

When compared with the number of expected test takers, based on recent examination administrations, approximately 20% of students took the Regents Examination in Physical Setting/Earth Science 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 a Physical Setting/Earth Science course during the 2020-2021 school year.

The overall mean scale score was computed based on all students who took the Regents Examination in Physical Setting/Earth Science. The result is reported in Table 9.

Table 9 Mean: Regents Examination in Physical Setting/Earth Science

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. It provides teachers and students with important information about student learning and performance against the established curriculum standards. Results of this exam may be used to identify student strengths and needs to guide classroom teaching and learning. The exam also provides students, parento(eng)10curc

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 Physical Setting/Earth Science, under strict guidelines that leverage best practices (see Appendix C). State educators also conduct all item quality and bias reviews to ensure that item content is appropriate to the construct being measured and fair for all students. Finally, educators use the defined standards, test blueprint targets, and statistical information generated during field testing to select the highest quality items for use in the operational test.

Figure 7 summarizes the full test development process, with steps 3 and 4 addressing initial item development and review. This figure also demonstrates the ongoing nature of ensuring the content validity of items through field test trials, and final item selection for operational testing.

Initial item development is conducted under the criteria and guidance provided by the Department. Both MC and CR items are included in the Regents Examination in Physical Setting/Earth Science to ensure appropriate coverage of the construct domain.

5.2 EVIDENCE BASED ON RESPONSE PROCESSES

The second source of validity evidence is based on examinee response processes. This standard requires evidence that examinees are responding in the manner intended by the test items and rubrics, and that raters are scoring those responses in a manner that is consistent with the rubrics. It is important to control and monitor whether or not 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 Physical Setting/Earth Science 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 of and attention given to reducing the potential for construct irrelevance in response processes by attending to the quality and alignment of test content to the test blueprint and to the item development guidelines (Appendix C). Further evidence is documented in the test administration and scoring procedures, as well as in the results of statistical analyses, which are covered in the following two sections.

Administration and Scoring

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

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

may be assigned to review the consistency of scoring for the lead's assigned staff against model responses and to be available for consultation throughout the scoring process.

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

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

In support of the goal of valid score interpretations for each examinee, such scoring training procedures are implemented for the Regents Examination in Physical Setting/Earth Science. Operational raters are selected based on expertise in the exam subject and are assigned a specific set of items to score. No more than approximately one-half of the items on the test are assigned to any one rater. This increases the consistency of scoring across examinee responses by allowing each rater to focus on a subset of items. It also ensures that no one rater is allowed to score the entire test for any one student. This practice reduces the effect of any potential bias of a single rater on individual examinees. Additionally, raters anc b (ees)4y[(em)sTJ0.0

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

Item Difficulty

Multiple analyses allow for an evaluation of item difficulty. For this exam, *p*-values and Rasch difficulty (item location) estimates were computed for MC and CR items. Items for the June 2021 Regents Examination in Physical Setting/Earth Science show a range of *p*-values consistent with the targeted exam difficulty. Item *p*-values ranged from 0.35 to 0.87, with a mean of 0.59. The diffic

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 6. The mean INFIT value is 1.00, with all but one item falling in a targeted range of [0.7, 1.3]. As the range of [0.7, 1.3] is used as a guide for ideal fit, fit values outside of the range are considered individually. These results indicate that, for most items, the Rasch model fits the Regents Examination in Physical Setting/Earth Science item data well.

Test Reliability]ftucTj25.43i1n.004Tg(h)10 (4(ndn(gh t)1 (.)( )Td(s.038 T0 i(g)6gTj))0 (al)215u0.03eden9 (

placement decisions. Such uses can be considered reasonable, assuming that the competencies demonstrated in the Regents Examination in Physical Setting/Earth Science 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 Physical Setting/Earth Science 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 Physical Setting/Earth Science 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).

650 T05 Tw 1.11 0 Td[()923pC (a)7–5 (2)-7 (6)3 (5)]TJ0

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.

o

4Tj EM-41.68wspons79pC (a)7 (i)6n (t)2 (e)9.932J0 Tc2

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-

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– g0.0185 TdCe38 0 -0.004 Tw 0.56 0 Td[(g0.018.0.56 0

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.

sk(eas)4 (e)11 (t7 (c))4 (o)10 (f)a g2(pp)en2(pp)er (oma3

Position	Item Type	Max. Points	Weight	Standard	PI	Mean	Point-Biserial	RID	INFIT

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

Table B.1 Score Table for June 2021

Appendix C: Item Writing Guidelines

GUIDELINES FOR WRITING MULTIPLE-CHOICE ITEMS

1. The item should focus on a single issue, problem, or topic stated clearly and concisely in the stem.

2. The item should be written in clear and simple language, with vocabulary and sentence structure kept as simple as possible.

3. The stem should be written as a direct question or an incomplete statement.

4. The stem should not contain irrelevant or unnecessary detail.

5. The stem should be stated positively u

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		