New York State Regents Examination in Physical Setting/Earth Science

2018 Technical Report

Prepared for the New York State Education Department
by Pearson

March 2019

Contents

List of Tables

1.3 TARGET POPULATION (STANDARD 7.2)

The examinee population for the Regents Examination in Physical Setting/Earth Science is composed of students who have completed a course in Earth Science.

Table 1 provides a demographic breakdown of all students who took the August 2017, January 2018, and June 2018 Regents Examination in Physical Setting/Earth Science. All analyses in this report are based on the population described in Table 1. Annual Regents [redacted]e Student Information Repository System (SIRS) as of the reporting deadline. The results include those exams administered August 2017, January 2018, and June 2018 (see http://data.nysed.gov/).
If a student takes the same exam 4 Tc f4 (,)8 ( )]TJ -0.006 Tc 0.006 Tw 1.12 >>BDC hE BT s48 581.5.a

Table 3 Constructed- u0 T3w y:S 6en (6t)on 6at 6 i- 683 0 Td [(1 (N)-8 (95 2.9 5)-9.7 P5.1 h (6ysi-Td [(

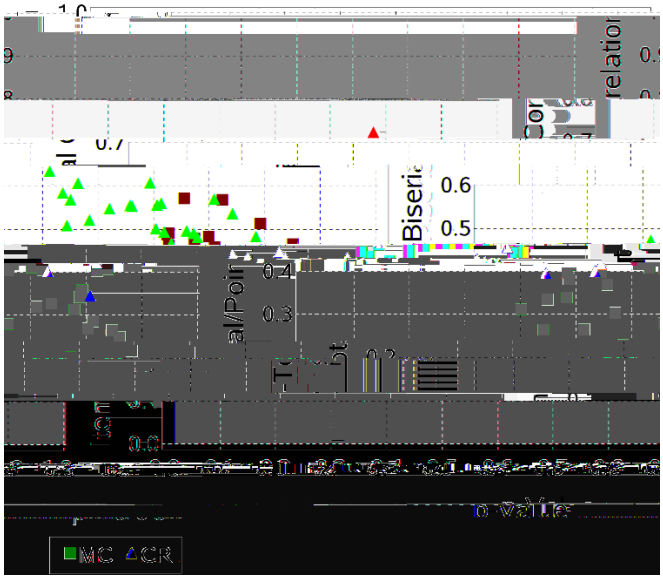


Figure 1 Scatter Plot: Regents Examination in Physical Setting/Earth Science

Table 4 Descriptive Statistics in p -value and Point -Biserial Correlation: Regents Examination in Physical Setting/Earth Science

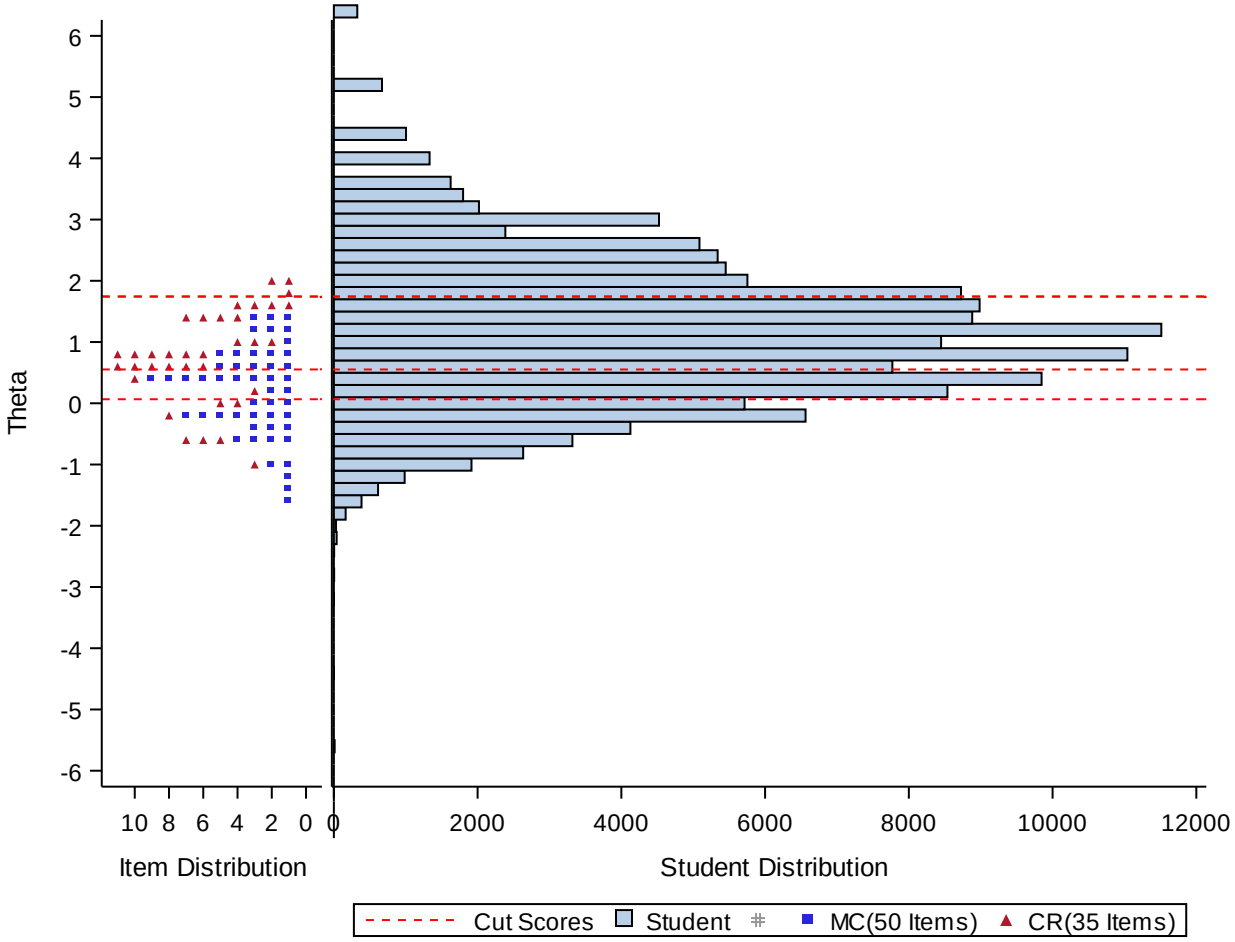
Statistics	N	Mean	Min	Q1	Median	Q3	Max
p-value	86	0.64	0.31	0.s6vm.9			

$$P_{ni}/X \mid 1\emptyset = \frac{\exp/\chi_n 4 D_{ij}\emptyset}{12 \exp/\chi_n 4 D_{ij}\emptyset}.$$

The Rasch model places both performance and item difficulty (estimated in terms of log-odds or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of examinee performance and item difficulty that are theoretically invariant across random samples of the same examinee population.

3.2 SOFTWARE AND ESTIMATION ALGORITHM

Item calibration was implemented via the WINSTEPS 3.60


Theta
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
10 8 6 4 2 0
Item Distribution
0 2000 4000 6000 8000 10000 12000
Student Distribution
Cut Scores
Student
MC(50 Items)
CR(35 Items)

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 2004 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 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 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 65 and 85 were set to correspond to those raw score cuts. A third-degree polynomial is required to fit a line exactly to four arbitrary points (e.g., the raw scores corresponding to the four critical scale scores of 0, 65, 85, and 100). The general form of this best-fitting line is:

$$SS = m3 \times RS^3 + m2 \times RS^2 + m1 \times RS + m0,$$

where SS is the scaled score, RS is the raw score, and m0 through m3 are the transformation constants that convert the raw score into the scale score (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 m3 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 difficulty parameters for the items on the base form were then used to derive a raw score-to-Rasch student ability (theta score) relationship. This allowed the relationship between the Rasch theta score and the scale score to be known, mediated through their common relationship with the raw scores.

In succeeding years, each test form was selected from the pool of items that had been tested in previous years' field tests, each of which had known Rasch item difficulty parameter(s). These known parameters were then used to construct the relationship between the raw and Rasch theta scores for that particular form. Because the Rasch difficulty parameters are all on a common scale, the Rasch theta scores were also on a common scale with previously administered forms. The remaining step in the scaling process was to find the scale score equivalent for the Rasch theta score corresponding 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, and 85.

Reliability coefficients range from 0.0 to 1.0. The index will be 0.0 if none of the test score variances is true. If all test score variances were true, the index would equal 1.0. Such scores would be pure random noise (i.e., all measurement error). If the index achieved a value of 1.0, scores would be perfectly consistent (i.e., contain no measurement error). Although values of 1.0 are never achieved in 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 by 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 forat rel itheimix (ar)b-4 (A)-5tenc8pspp (t-)10 (t)2 (enc)4 (8p1 (m)-

error (Harvill, 1991). Additionally, the value of the SEM takes the group variation (i.e., score standard deviation) into account. Consider that a SEM of 3 on a 10-point test would be very different Tc 00.022 Tc 00

The relationship between and the scale score is not expressible in a simple mathematical form because it is a blend of the third-degree polynomial relationship between the raw and scale scores, along with the nonlinear relationship between the expected raw and scores. In addition, as the exam is equated from year to year, the relationship between the raw and scale scores moves away from the original third-degree polynomial relationship to one that is also no longer expressible in

		TEST ONE		
		LEVEL	LEVEL I	MARGINAL
TEST TWO	LEVEL I	π11	π12	π1●
	LEVEL II	π21	π22	π2●
	MARGINAL	π●1	π●2	1

Figure 5 Pseudo- Decision Table for Two Hypothetical Categories

		TEST ONE				
		LEVEL	LEVEL I	LEVEL II	LEVEL I\	MARGINAL
TEST TWO	LEVEL I	π11	π12	π13	π14	π1●
	LEVEL II	π21	π22	π23	π24	π2●
	LEVEL III	π31	π32	π33	π34	π3●
	LEVEL IV	π41	π42	π43	π44	π4●
	MARGINAL	π●1	π●2	π●3	π●4	1

Figure 6 Pseudo- Decision Table for Four Hypothetical Categories

If a student is classified as being in one category based on Test One's score, how probable would it be that the student would be reclassified as being in the same category if he or she

Table 9 Group Means: Regents Examination in Physical Setting/Earth Science

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	147,580	75.93	16.58
Ethnicity			
American Indian/Alaska Native	866	69.95	16.95
Asian/Native Hawaiian/Other Pacific Islander	12,142	81.09	15.32
Black/African American	20,101	64.95	17.05
Hispanic/Latino	31,951	68.90	16.77
Multiracial	2,673	76.76	16.02
White	79,841	80.77	14.04
English Language Learner/Multilingual Learner			
No	140,661	76.85	16.00
	6,919		17.42

Chapter 5: Validity (Standard 1)

The Regents Examination in Physical Setting/Earth Science measures student achievement on the NYS P–12 Curriculum Standards for Physical Setting/Earth Science. The Physical Setting/Earth Science standards can be found at

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 Physical Setting/Earth Science, 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.

Initial item development is conducted under the criteria and guidance provided by the Department. Both multiple-choice and constructed-response items are included in the Regents Examination in Physical Setting/Earth Science, in order to ensure appropriate coverage of the construct domain.

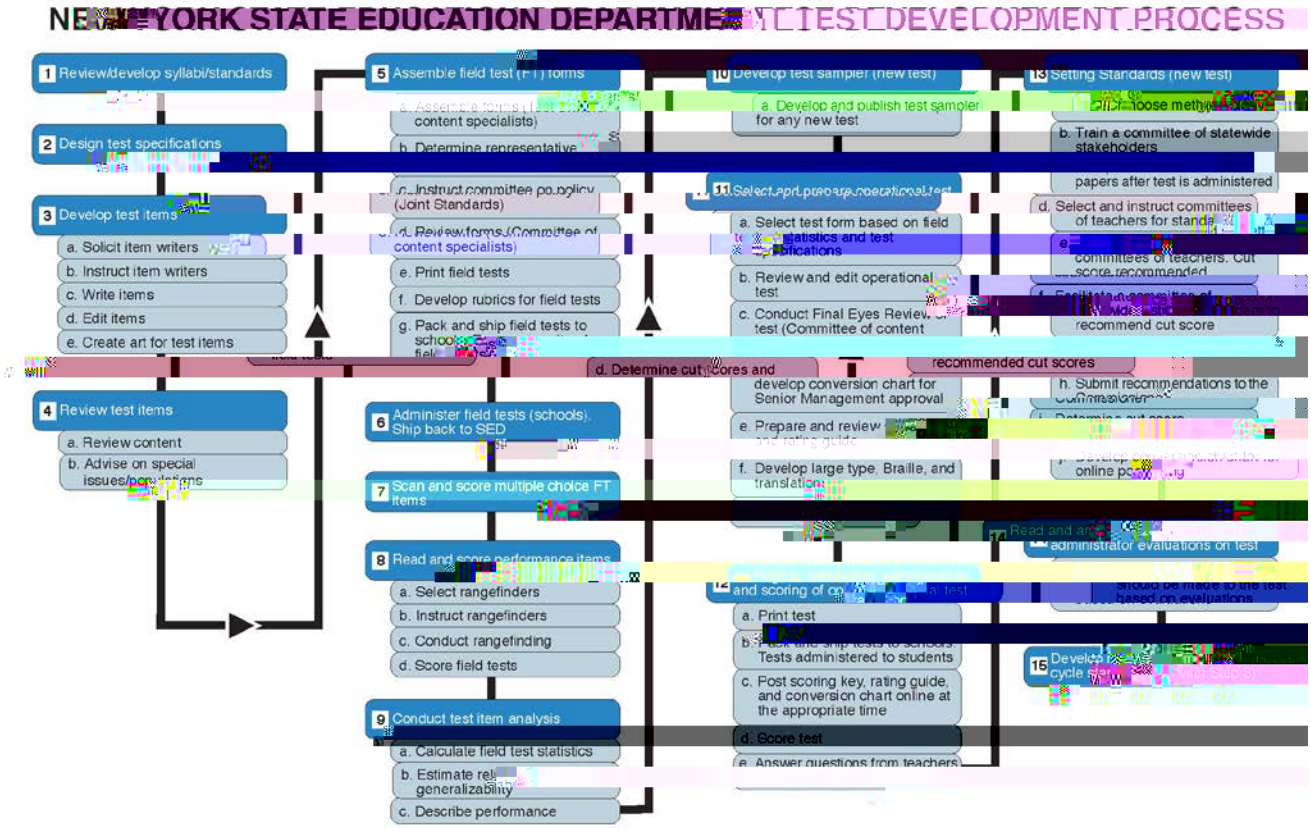


Figure 7 New York State Education Department Test Development Process

Item Review Process

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

All

solicits thoughtful, specific responses from stakeholders about individual standards within the NYS P–12 Standards.

5.2 EVIDENCE BASED ON RESPONSE PROCESSES

The second source of validity evidence is based on examinee response processes. This standard requires evidence that examinees are responding in the manner intended by the test items and rubrics and that raters are scoring those responses in a manner that is consistent with the rubrics. Accordingly, it is important to control and monitor whether 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 and attention given to reducing the potential for construct irrelevance in response processes by attending to the quality and alignment of test content to the test blueprint and to the item development guidelines (Appendix C). Further evidence is documented in the test administration and scoring procedures, as well as in the results of statistical analyses, which are covered in the following two sections.

Administration and Scoring

Adherence to standardized administration procedures is fundamental to the validity of test scores and their interpretation, as such procedures allow for adequate and consistently applied conditions for scoring the work of every student who takes the examination. For this reason, guidelines, which are contained in the *School Administrator's Manual, Secondary Level Examinations* (http://www.p12.nysed.gov/assessment/manuals/), have been developed and implemented for the New York State Regents testing program. All secondary-level Regents

3.

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 of New York State universities and colleges finds that, beyond the explicitly defined use as a testing requirement toward graduation for students who have completed a course in Physical Setting/Earth Science, the exam is most commonly used to inform admissions and course placement decisions. Such uses

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

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

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

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

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

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

Kolen, M. J., & Brennan, R. L. (2005). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J., & Lee, W. (2011). Psychometric Properties of Raw and Scale Scores on Mixed-Format Tests. *Educational Measurement: Issues and Practice 30*(2), 15–24.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129–140.

Leacock, Claudia, Gonzalez, Erin, Conarroe, Mike. (2014). *Developing effective scoring rubrics for AI short answer scoring*. CTB McGraw-Hill Education Innovative Research and Development Grant.

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,* 54–72.

McDonald, R.

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 York, NY: Springer-Verlag.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement*, *37*(2), 141–162.

Wang, X. M., Wong, K. F. E., & Kwong, J. Y. Y. (2010). The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology, 95*(3), 546–561.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*

Position	Item Type	Max

Position	Item Type	Max Points	Weight	Standard	PI	Mean	Point-Biserial	RID	INFIT
77	CR	1	1	4	1.2j	0.45	0.55	0.8535	0.92
78	CR	1	1	4	2.1v	0.34	0.58	1.4439	0.86
79	CR	1	1	4	2.1u	0.35	0.50	1.3641	0.97
80	CR	1	1	4	2.2c	0.41	0.62	1.0272	0.81
81	CR	1	1	4	2.1a	0.44	0.50	0.9082	0.96
82	CR	1	1	4	3.1c	0.64	0.56	-0.1101	0.91
83	CR	1	1	4	3.1c	0.70	0.57	-0.4338	0.87
84	CR	1	1	4	3.1c	0.59	0.62	0.1769	0.83
85	CR	1	1	4	3.1c	0.31	0.52	1.6575	0.95

Table A.2 Test Map for January 2018 Administration

Position	Item Type	Max Points	Weight	Standard	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	4	1.1	0.63	0.45	-0.0156	0.98
2	MC	1	1	4	1.2	0.40	0.38	1.1330	1.11
3	MC	1	1	4	1.2	0.50	0.42	0.6229	

Position	Item Type	Max Points	Weight	Standard	PI	Mean	Point-Biserial	RID	INFIT
39	MC	1	1	4	1.1	0.53	0.47	0.4792	0.98
40	MC	1	1	4	1.1	0.72	0.48	-0.4759	0.94
41	MC	1	1	4	1.1	0.55	0.31	0.3501	1.15

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

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

Table B.1 Score Table for August 201 7 Administration

Raw Score	Ability	Scale Score
0	-5.5710	0.000

Raw Score	Ability	Scale Score
41	0.3864	61.923

Raw Score	Ability	Scale Score
82	4.0781	98.232

Table B.2 Score Table for January 201 8 Administration

Raw Score	Ability	Scale Score
0	-5.4873	0.000
1	-4.2704	1.827
2	-3.5586	3.640
3	-3.1343	5.445
4	-2.8278	7.243
5		

Raw Score	Ability	Scale Score
41	0.3541	61.213
42	0.4078	62.390
43	0.4615	63.504
44	0.5152	64.656
45	0.5690	65.761

Raw Score	Ability	Scale Score
82	4.0165	98.125
83	4.4415	98.743
84	5.1540	99.365
85	6.3714	100.000

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. Avoid using negatively stated stems.

6. The phrase *which of the following* should not be used to refer to the alternatives. Instead use *which* followed by a noun.

7. The stem should include any words that must otherwise be repeated in each alternative.

8. The item should have one and only one correct answer (key).

9. The distractors should be plausible and attractive to students who lack the knowledge, understanding, or ability assessed by the item.

10. The alternatives should be grammatically consistent with the stem.

11. The alternatives should be parallel with one another in form.

12. The alternatives should be arranged in logical order, when possible.

13. The alternatives should be independent and mutually exclusive.

14. The item should not contain extraneous clues to the correct answer.

15. Items should be written in the third person. Use generic terms instead of proper nouns, such as first names and brand names.

CHECKLIST OF TEST CONSTRUCTION PRINCIPLES
(Multiple-Choice Items)

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

GUIDELINES FOR WRITING CONSTRUCTED-RESPONSE ITEMS

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

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

3. The item should be written in the third person. Use generic terms instead of proper nouns such as first names and brns

Appendix D: Tables and Figures for August 2017 Administration

Table D.1 Multiple -

Item	Number of Students	p-Value	SD	Point-

Table D.2 Constructed- Response Item Analysis Summary: Regents Examination in Physical Setting/Earth Science

Item	Min. score	Max. score	Number of Students	Mean	SD	p-Value	Point-Biserial
51	0	1	11,185	0.13	0.33	0.13	0.23
52	0	1	11,185	0.28	0.45	0.28	0.33
53	0	1	11,185	0.49	0.50	0.49	0.27
54	0	1	11,185	0.51	0.50	0.51	0.29
55	0	1	11,185	0.74	0.44	0.74	0.37
56	0	1	11,185	0.31	0.46	0.31	0.43
57	0	1	11,185	0.62	0.49	0.62	0.41

Figure D.1 Scatter Plot：Regents Examination in Physical Setting/Earth Science

Table D.3 Descriptive Statistics in p -value and Point-Biserial Correlation: Regents Examination in Physical Setting/Earth Science

Statistics	N	Mean	Min	Q1	Median	Q3	Max

Table D.8 Group Means: Regents Examination in Physical Setting/Earth Science

Demographics	Number	Mean Scale Score	SD Scale Score
All Students	11,185	57.28	14.25

Aener

Appendix E: Tables and Figures for January 201 8 Administration

Table E.1 Multiple -Choice Item Analysis Summary: Regents Examination in Physical Setting/Earth Science

Item	Number of Students	p-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	9,449	0.46	0.50	0.30	-0.07	-0.09	-0.22
2	9,449	0.29	0.46	0.27	-0.12	-0.12	

Item	Number of Students	p-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
34							

Table E.2 Constructed- Response Item Analysis Summary: Regents Examination in Physical Setting/Earth Science

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

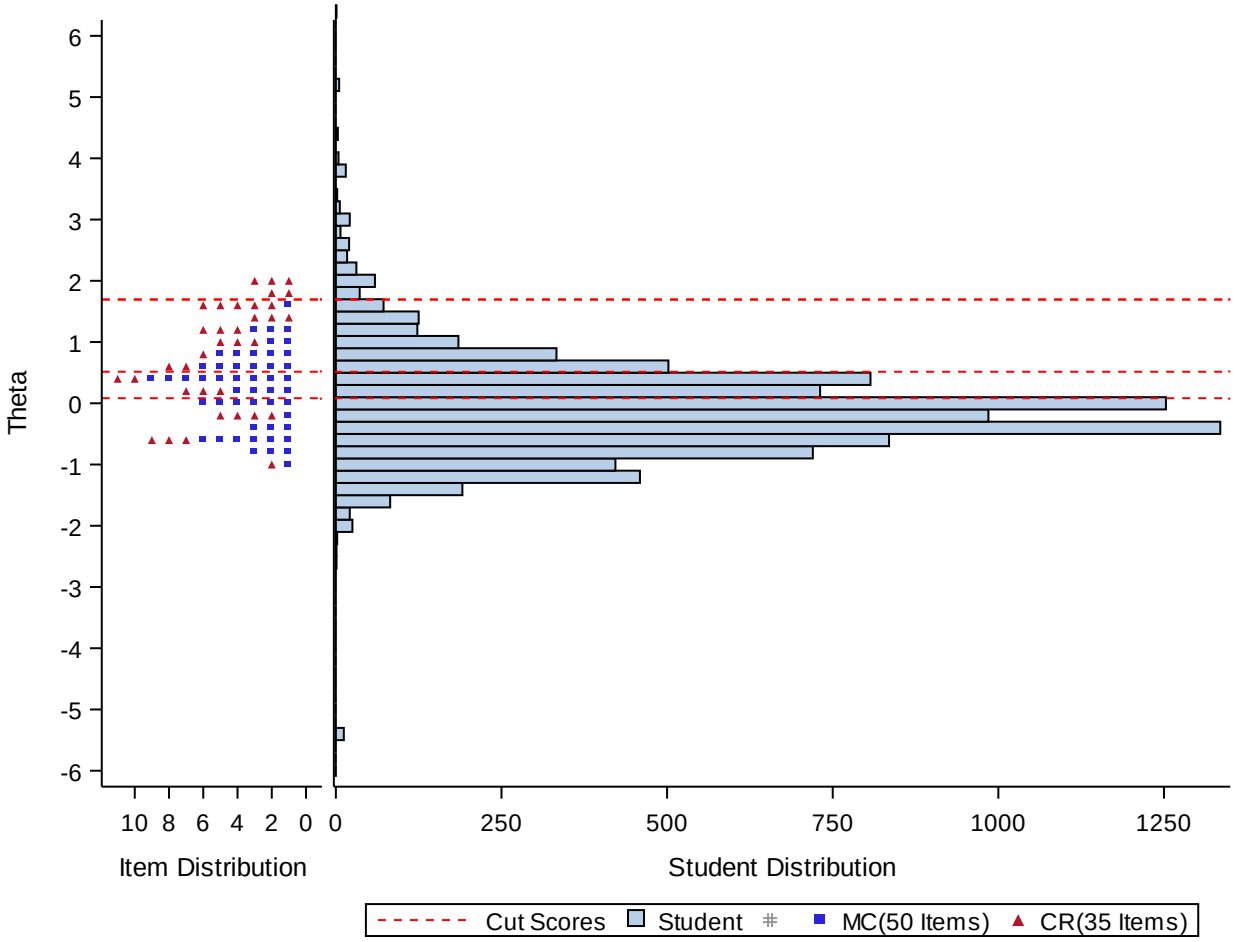


Figure E.2 Student Performance Map: Regents Examination in Physical Setting/Earth Science

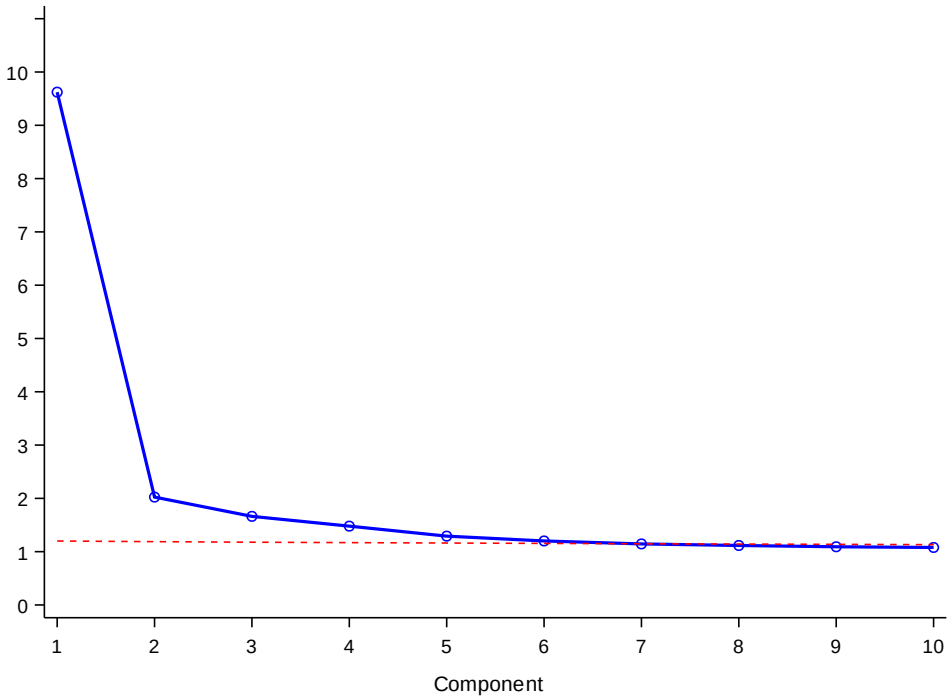

10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Component

Table E.5 Summary of INFIT Mean Square Statistics: Regents Examination in Physical Setting/Earth Science

		INFITMean Square	
		Mean	SD