New York State Regents Examination in Living Environment

Copyright

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

Contents

Chapter 1: Introduction

1.1 INTRODUCTION

This technical

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 Regents Examination in Living Environment items.

2.1 ITEM DIFFICULTY

Table 3 Constructed-Response Item Analysis Summary: Regents Examination in Living Environment

Item	Min. Score	Max. Score	Number of Students	Mean	SD	p-Value	Point-Biserial
44	0	1	229,272	0.68	0.47	0.68	0.50
45	0	1	229,272	0.83	0.37	0.83	0.42
46	0	1	229,272	0.79	0.41	0.79	0.40
48	0	1	229,272	0.71	0.45	0.71	0.46
51	0	1	229,272	0.68	0.47	0.68	0.49
52	0	1	229,272	0.69	0.46	0.69	0.62
53	0	1	229,272	0.44	0.50	0.44	0.58
54	0	1	229,272	0.52	0.50	0.52	0.52
55	0	1	229,272	0.66	0.47	0.66	0.50
56	0	1	229,272	0.31	0.46	0.31	0.55
57	0	1	229,272	0.65	0.48	0.65	0.47
58	0	1	229,272	0.83	0.37	0.83	0.37
59	0	1	229,272	0.61	0.49	0.61	0.55
60	0	1	229,272	0.60	0.49	0.60	0.51
61	0	1	229,272	0.50			

2.3 DISCRIMINATION ON DIFFICULTY SCATTER PLOT

Figure 1 shows a scatter plot of item discrimination values (y-axis) and item difficulty values (x-axis). The distributions of p-value and point-biserial values, including mean, minimum, Q1, median, Q3, and maximum, are presented in Table 4.

Figure 1 Scatter Plot: Regents Examination in Living Environment

Table 4 Descriptive Statistics in p-value and Point-Biserial Correlation: Regents Examination in Living Environment

Statistics	Number of Items	Mean	Min	Q1	Median	Q3	Max
p-value	85	0.65	0.30	0.53	0.66	0.77	0.92

Table 5 Summary of Item Residual Correlations: Regents Examination in Living Environment

Statistic Type	Value
N	3,570
Mean	-0.01
SD	0.02
Minimum	-0.09

Table 6 Summary of INFIT Mean Square Statistics: Regents Examination in Living Environment

	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Living Environment	85	1.00				

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 2018, January 2019, and June 2019 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.

The minimum (zero) and maximum possible raw scores are assigned scale scores of 0 and 100, respectively. In the event that there are raw scores less than the maximum with scale scores that round to 100, their scale scores are set equal to 99. A similar process is followed 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,

Traditional Standard Error of Measurement Confidence Intervals

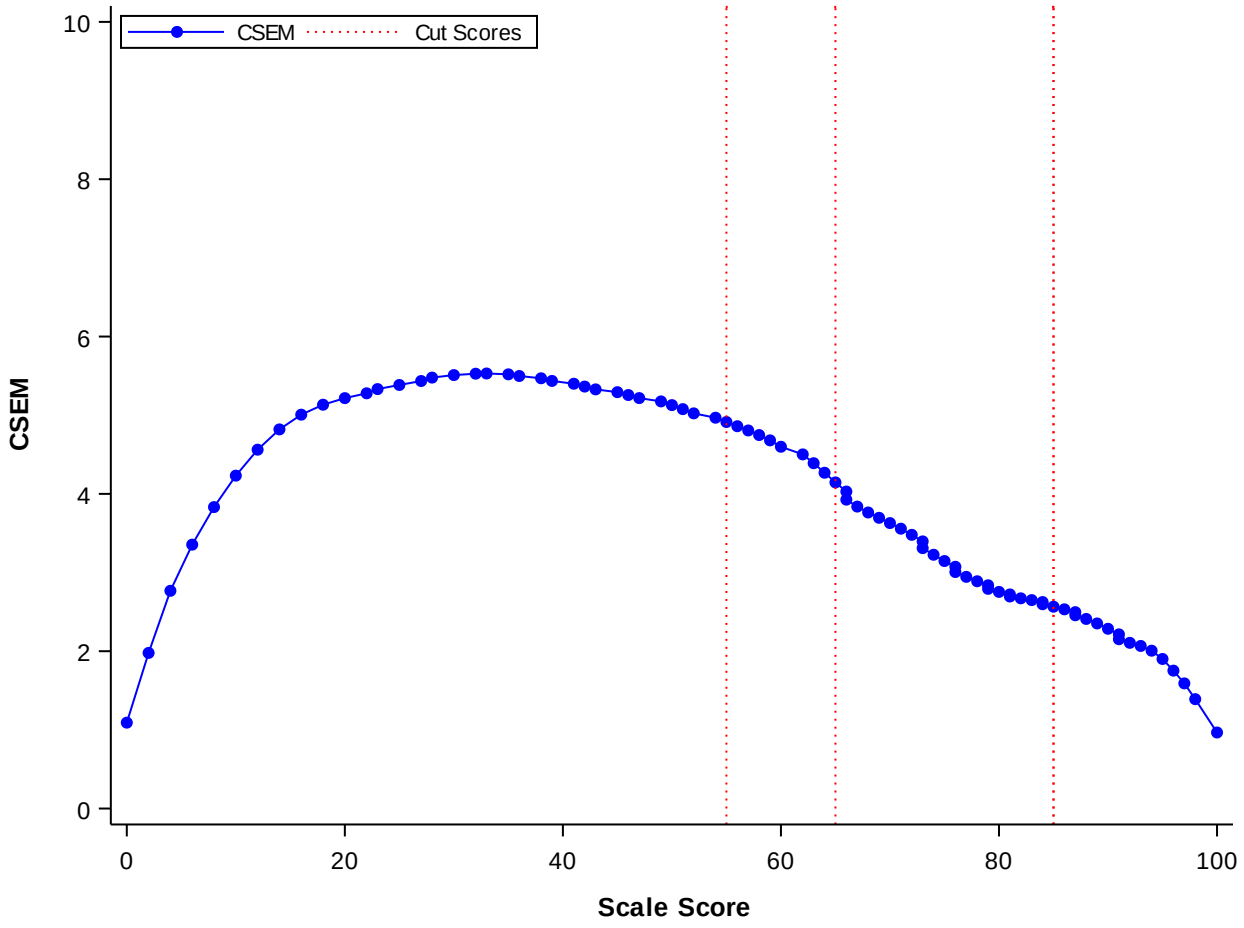


Figure 4 Conditional Standard Error Plot: Regents Examination in Living Environment

4.3 DECISION CONSISTENCY AND ACCURACY (STANDARD 2.16)

In a standards-based testing program, there is interest in knowing how accurately students are classified into performance categories. In contrast to the Coefficient Alpha, which is concerned with the relative rank-ordering of students, it is the absolute values of student scores that are important in decision consistency and accuracy.

Classification consistency refers to the degree to which the achievement level for each student can be replicated upon retesting by using an equivalent form (Huynh, 1976). Decision consistency answers the following question: What is the agreement in classifications between the two non-

		TEST ONE		
		LEVEL I	LEVEL II	MARGINAL
TEST TWO	LEVEL I	M_{11}	M_{12}	$M_{1\cdot}$
	LEVEL II	M_{21}	M_{22}	$M_{2\cdot}$
	MARGINAL	$M_{\cdot 1}$	$M_{\cdot 2}$	1

Figure 5 Pseudo-Decision Table for Two Hypothetical Categories

Several f

Table 9 Group Means: Regents Examination in Living Environment

Demographics	Number	

Table 10 State Percentile Ranking for Scale Score: Regents Examination in Living Environment

Scale Score	Percentile Rank	Scale Score	Percentile Rank	Scale Score	Percentile Rank	Scale Score	Percentile Rank
0	1	26	1	52	12	78	47
1	1	27	1	53	13	79	49
2	1	28	1	54	13	80	52
3	1	29	1	55	14	81	55
4	1	30	1	56	15	82	58
5	1	31	1	57	16	83	60
6	1	32	1	58	17	84	63
7	1	33	2	59	18	85	67
8	1	34	2	60	19	86	70
9	1	35	2	61	19	87	73
10	1	36	3	62	20	88	76
11	1	37	3	63	22	89	79
12	1	38	3	64	23	90	82
13	1	39	4	65	24	91	86
14	1	40	4	66	25	92	89
15	1	41	5	67	27	93	91
16	1	42	5	68	28	94	93
17							

The Regents Examin(x)146f1277Tn

and statistical information generated during field testing, to select the highest quality items for use in the operational test.

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

Initial item development is conducted under the criteria and guidance provided by the Department. Both multiple-choice and constructed-response items are included in the Regents Examination in Living Environment, 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.

fundamental form of evidence of the validity of the intended score interpretations. Another integral component of this item review process is to review the scoring rules, or “rubrics,” for their clarity and consistency in what the examinee is being asked to demonstrate by responding to each item. Each of these elements

The controls and monitoring in place for the Regents Examination in Living Environment 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 Examinations are administered under these standard conditions to support valid inferences for all students. These standard procedures also cover testing580.009.31Re t

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 constructed-response 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, NYS educators serve as table leaders during the range-finding session. In the range-finding process, committees of educators receive a set of student papers for each field-tested question. Committee members familiarize themselves with each item type and score a number of responses that are representative of each of the different score points. After the independent scoring is completed, the committee reviews and discusses their results and determines consensus scores for the student responses. During this process, atypical responses are important to identify and annotate for use in training and live scoring. The range-finding results areg (er)7 (ts]TJ0 Tce i)6 (04 Tc -i)6 48 Tw 12 (er)77ull( )]TJrainin

The following analyses were conducted for the Regents Examination in Living Environment:

- item difficulty
- 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 Regents Examination in Living Environment show a range of p-values consistent with the targeted exam difficulty. The item p-values range from 0.30 to 0.92, with a mean of 0.65. The difficulty distribution illustrated in Figure 1 shows a wide range of item difficulties on the exam.

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. F7 (u)10Tuse. Fnval

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, 35o6 (l)6 (M)3 1Bn,n-0.006 Tc2Tj /TT1 1 T.75 -23-

Appendix A: Operational Test Maps

Table A.1 Test Map for August 2018 Administration

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
1	MC	1	1	4	1	1.1b	0.78	0.48	-1.3422	0.96
2	MC	1	1	4	5	5.2c	0.77	0.54	-1.1652	0.87
3	MC	1	1	4	2	2.2b	0.78	0.43	-1.2232	0.97
4	MC	1	1	4	2	2.2c	0.78	0.45	-1.3233	1.01
5	MC	1	1	4	1	1.1f	0.40	0.42	0.805	1.04
6	MC	1	1	4	5	5.1a	0.77	0.28	-1.1869	1.16
7	MC	1	1	4	2	2.1a	0.73	0.47	-0.9755	0.97
8	MC	1	1	4	2	2.0	0.52	0.45	0.1918	1.01
9	MC	1	1	4	2	2.1a	0.87	0.43	-1.9884	0.90
10	MC	1	1	4	3	3.1f	0.71	0.48	-0.7905	0.97
11	MC	1	1	4	4	4.0	0.47	0.46	0.4611	1.06
12	MC	1	1	4	4	4.1f	0.66	0.52	-0.5568	

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
37	MC	1	1	1	1	1.2a	0.42	0.41	0.6813	1.09
38	MC	1	1	1	1	1.2a	0.63	0.44	-0.4254	1.04
39	MC	1	1	4	5	5.1f	0.58	0.37	-0.1487	1.15
40	MC	1	1	4	5	5.0	0.75	0.45	-1.105	0.97
41	MC	1	1	1	3	3.1	0.63	0.42	-0.3743	1.05
42	MC	1	1	4	7	7.3	0.55	0.49	0.0219	0.96

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
77	CR	1	1	Lab		1	0.73	0.55	-0.9125----	

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
79										

Table A.3 Test Map for June 2019 Administration

Position	Item Type	Max

Position	Item Type	Max Points	Weight	Standard	Key Idea	PI	Mean	Point-Biserial	RID	INFIT
39	MC	1	1	Appendix A			0.75	0.48	-0.9813	0.89
40	MC	1	1	4	2	2.2c	0.84	0.46	-1.7061	0.91
41	MC	1	1	4	1					

Position	Item Type	Max Points	Weight	Standard	Key

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

Table B.1 Score Table for August 2018 Administration Administration

Appendix C: Item Writing Guidelines

GENERAL RULES FOR WRITING MULTIPLE-CHOICE ITEMS

1.

CHECKLIST OF TEST CONSTRUCTION PRINCIPLES
(Multiple-Choice Items)

	YES	NO
1.	Is the item significant?	

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 brand names.

4. The item should not contain extraneous cl2LBody /AM./11.42 0 Td[(0 Td( )Tj- Tc 0 Tw 17.2T (ac)10

Appendix D: Tables and Figures for August 2018 Administration

Table D.1 Multiple-Choice Item Analysis Summary: Regents Examination in Living Environment

Item	Number of Students	p-Value	SD	Point-Biserial	Point-Biserial Distractor 1	Point-Biserial Distractor 2	Point-Biserial Distractor 3
1	24,233	0.57	0.50	0.32	-0.16	-0.18	-0.17
2	24,233	0.42	0.49	0.40	-0.17	-0.25	-0.08
3	24,233	0.56	0.50	0.32	-0.13	-0.17	-0.15
4	24,233	0.66	0.47	0.25	-0.09	-0.17	-0.10
5	24,233	0.23	0.42	0.16	0.06	-0.07	-0.18
[redacted]							-0.14
							-0.13
							-0.16
							-0.11
							-0.10
							-0.03
							-0.15
							-0.16
							-0.14

Item	Number of Students

Table D.2 Constructed-Response Item Analysis Summary: Regents Examination in Living Environment

Item	Min. score	Max. score	Number of Students	Mean	SD	p-Value	Point-Biserial
44	0	1	24,233	0.49	0.50	0.49	0.30
45	0	1	24,233	0.75	0.43	0.75	0.29
46	0	1	24,233	0.56	0.50	0.56	0.46
48	0	1	24,233	0.28	0.45	0.28	0.33
51	0	1	24,233	0.51	0.50	0.51	0.40
52	0	1	24,233	0.21	0.41	0.21	0.42

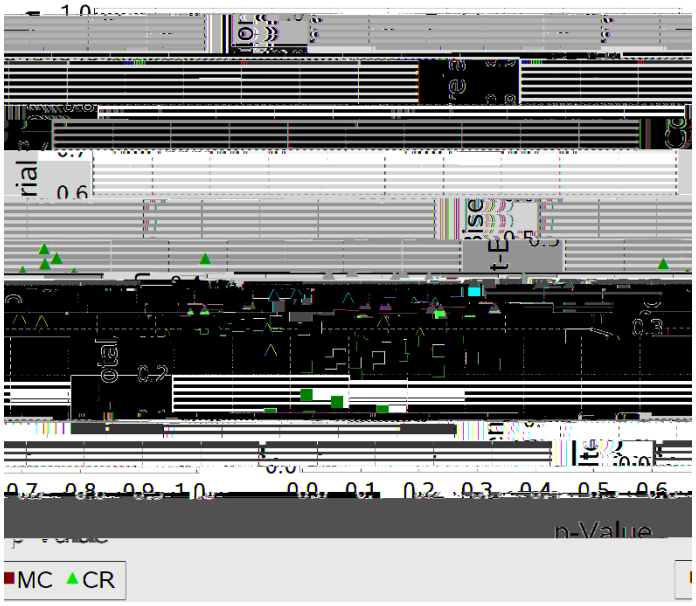
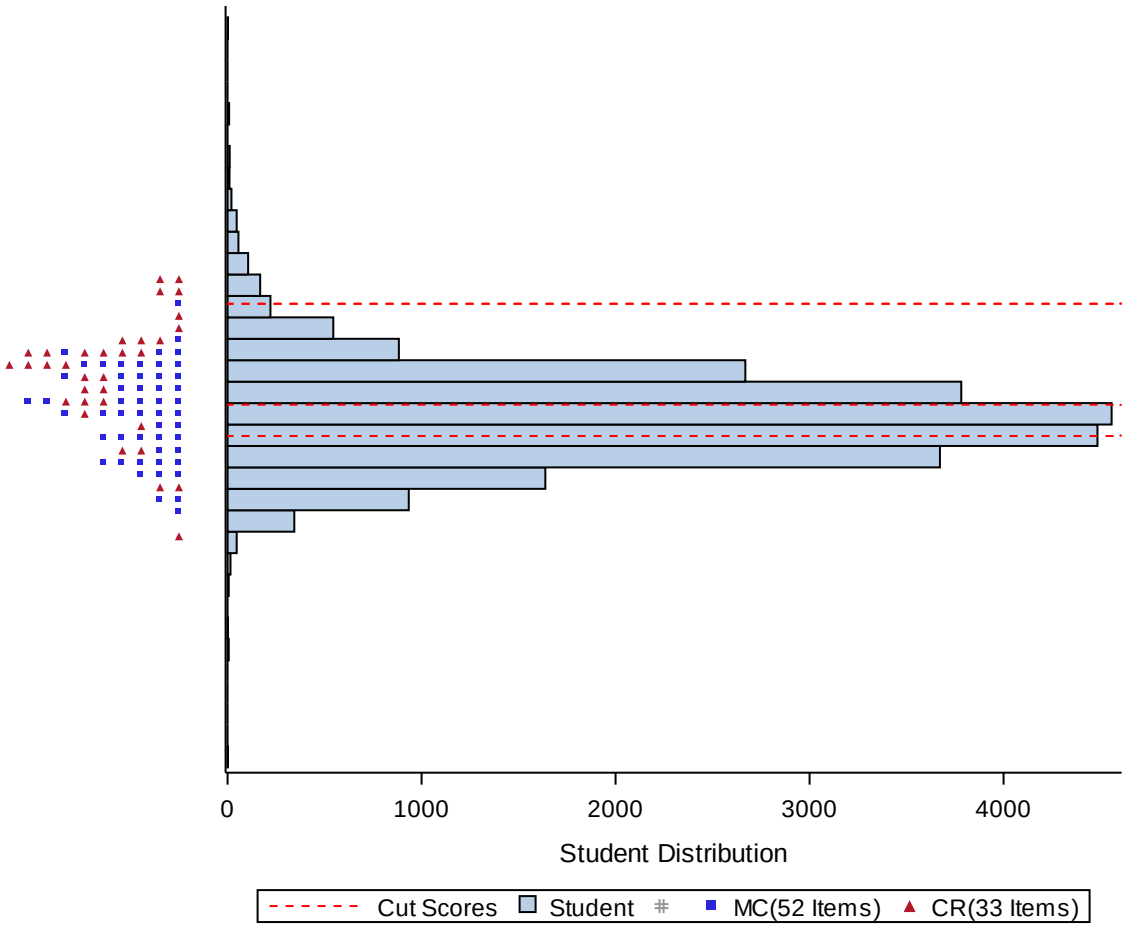


Figure D.1 Scatter Plot: Regents Examination in Living Environment

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

Statistics	N	Mean	Min	Q1	Median	Q3	Max
p-value	85	0.43	0.07	0.32	0.41	0.55	0.80
Point-Biserial	85						


0
1000
2000
3000
4000
Student Distribution
Cut Scores
Student
MC(52 Items)
CR(33 Items)

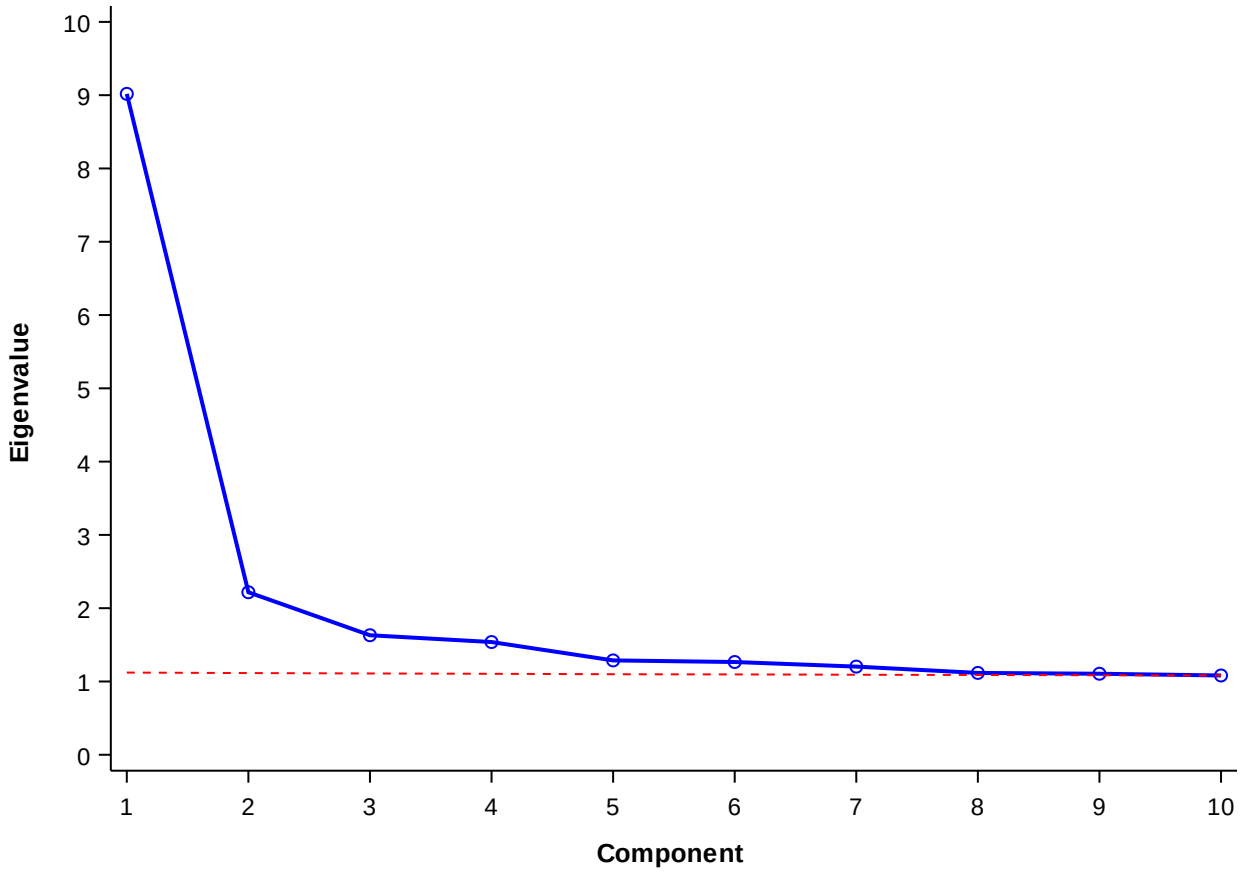

Eigenvalue
10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Component

Table D.5 Summary of INFIT Mean Square Statistics: Regents Examination in Living Environment

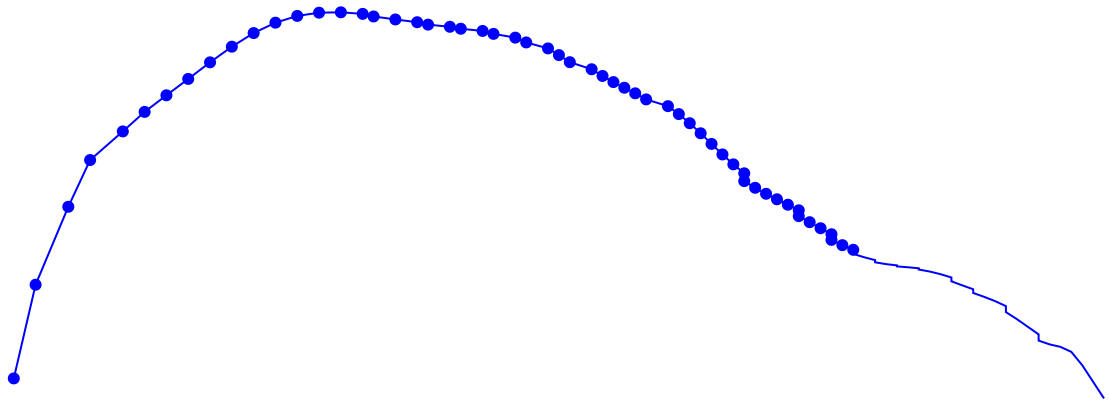
	INFIT Mean Square					
	N	Mean	SD	Min	Max	[0.7, 1.3]
Living Environment	85	1.00	0.06	0.88	1.16	[85/85]

Table D.6 Reliabilities and Standard Errors of Measurement: Regents Examination in Living Environment

Subject	Coefficient Alpha	SEM
Living Environment	0.89	4.14

Table D.7 Decision Consistency and Accuracy Results: Regents Examination in Living Environment

Statistic	1/2	2/3	3/4
Consistency	0.85	0.87	0.99
Accuracy	0.89	0.91	0.99



Item	Number of

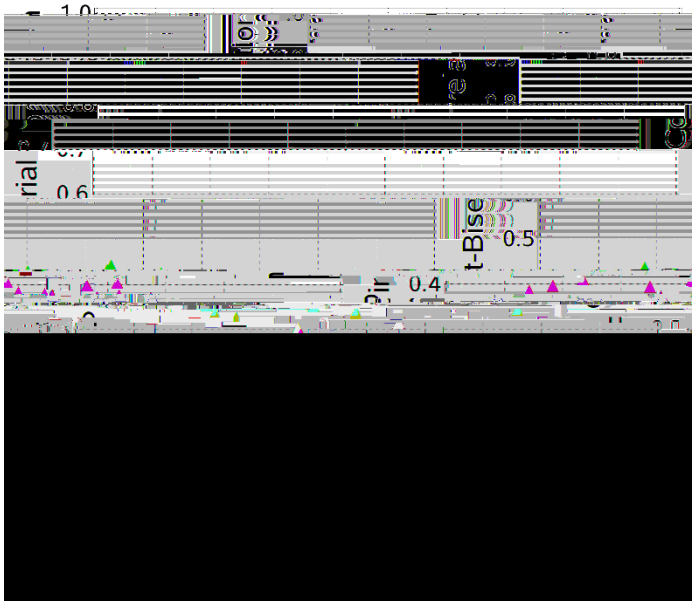
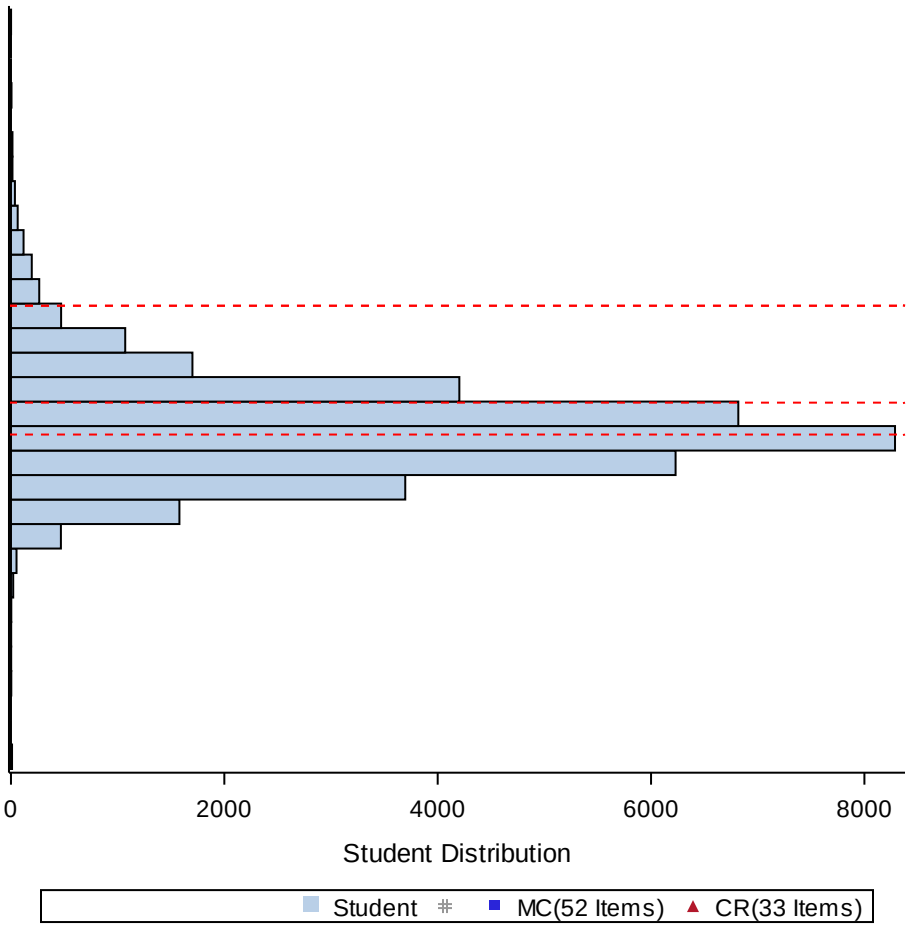


Figure E.1 Scatter Plot: Regents Examination in Living Environment

Table E.3 Descriptive Statistics in p-value and Point-


0
2000
4000
6000
8000
Student Distribution
Student
MC(52 Items)
CR(33 Items)

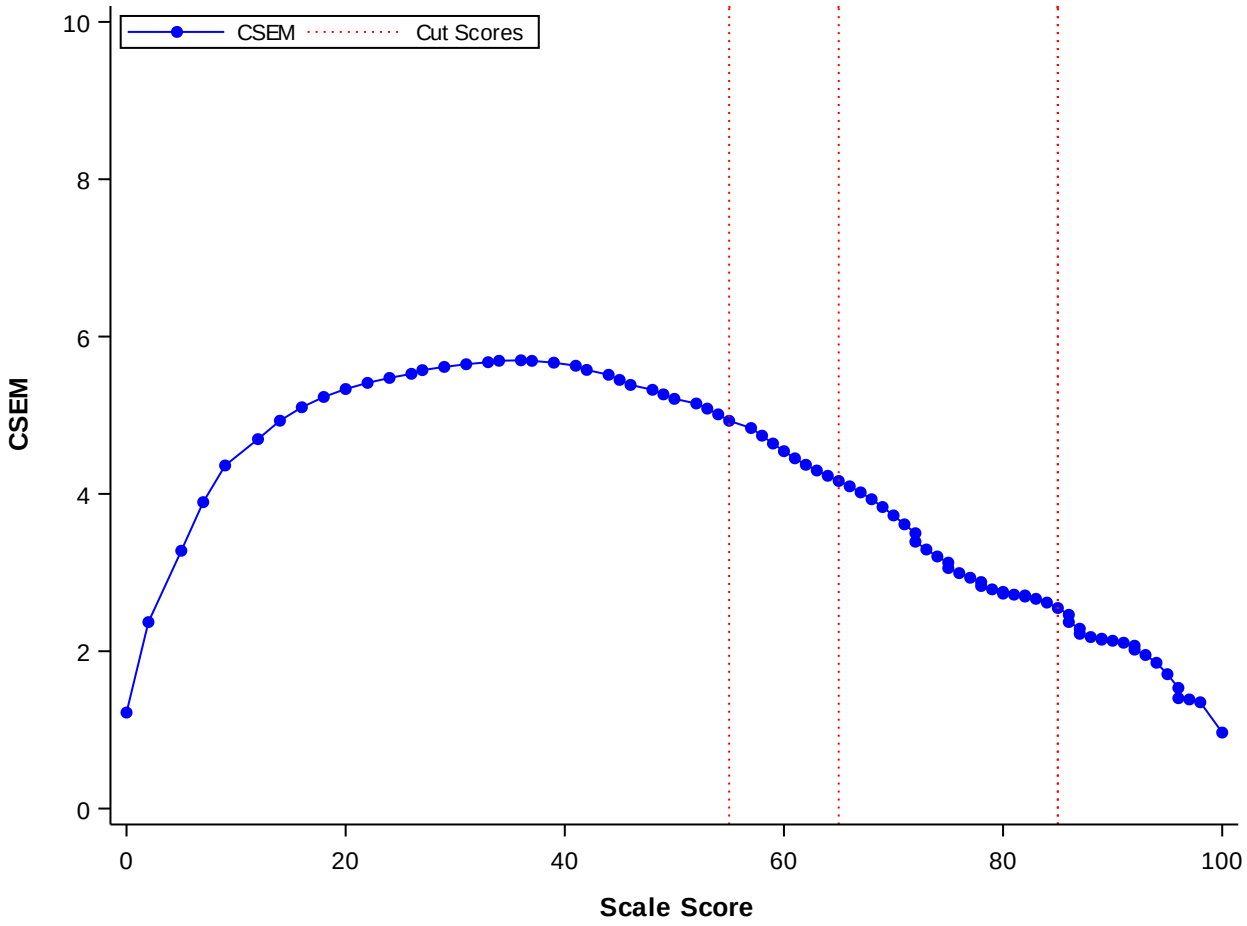


Figure E.4 Conditional Standard Error Plot: Regents Examination in Living Environment

Table E.8 Group Means: Regents Examination in Living Environment

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	35,328	55.40	13.97 c 0 Tw 25.[illegible]4344 0 Tw 1.108 [illegible]