Nutritional Ethics
Health & Hygiene

Chapter 2: Classical Item Statistics (Part 1.10)

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 only to the operational Regents Examination in Global History & Geography and the Transition Regents Examination in Global History and Geography – Grade 10 items.

2.1 ITEM DIFFICULTY

At the most general level, an item's difficulty is indicated by its mean score in some specified group (e.g., grade level).

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

In the mean score formula above, the individual item scores (x_i) are summed and then divided by the total number of students (n). For multiple-choice (MC) items, student scores are represented by 0s and 1s (0 = wrong, 1 = right). With 0–1 scoring, the equation above also represents the number of students correctly answering the item divided by the total number of students. Therefore, this is also the proportion correct for the item, or the p value. In theory, p values can range from 0.00 to 1.00 on the proportion-correct scale.[2] For example, if an MC item has a p value of 0.89, it means that 89 percent of the students answered the item correctly. Additionally, this value might suggest that the item was relatively easy and/or that the students who attempted the item were relatively high achievers. For constructed-response (CR) items, mean scores can range from the minimum possible score (usually zero) to the maximum possible score. To facilitate average score comparability across MC and CR items, mean item performance for CR items is divided by the maximum score possible so that the p values for all items are reported as a ratio from 0.0 to 1.0.

Although the p value statistic does not consider individual student ability in its computation

Fig 1 Sta Pb : Tha

Chapter 3: IRT Calibrations (Step 2, and 4.10)

The item response theory (IRT) model used for the Regents Examination in Global History and Geography and the Transition Regents Examination in Global History – Grade 10 is based on the work of Georg Rasch (Rasch, 1960). The Rasch model has a long-standing presence in applied testing programs. IRT has several advantages over classical test theory, and it has become the standard procedure for analyzing item response data in large-scale assessments. According to van der Linden and Hambleton (1997), "The central feature of IRT is the specification of a mathematical function relating the probability of an examinee's response on a test item to an underlying ability." Ability in this sense can be thought of as performance on the test and is defined as "the expected value of observed performance on the test of interest" (Hambleton, Swaminathan, and Roger, 1991). This performance value is often referred to as τ. Performance and τ will be used interchangeably through the remainder of this report.

A fundamental advantage of IRT is that it links examinee performance and item difficulty estimates and places them on the same scale, allowing for an evaluation of examinee performance that considers the difficulty of the test. This is particularly valuable for final test construction and test form equating, as it facilitates a fundamental attention to fairness for all examinees across items and test forms.

This chapter outlines the procedures used for calibrating the operational Regents Examination in Global History and Geography and the Transition Regents Examination in Global History and Geography – Grade 10 items. Generally, item calibration is the process of assigning a difficulty, or item "location," estimate to each item on an assessment so that all items are placed onto a common scale. This chapter briefly introduces the Rasch model, reports the results from evaluations of the adequacy of the Rasch assumptions, and summarizes the Rasch item statistics.

3.1 DESCRIPTION OF THE RASCH MODEL

The Rasch model (Rasch, 1960) was used to calibrate multiple-choice items, and the partial credit model, or PCM (Wright and Masters, 1982), was used to calibrate constructed-response items. The PCM extends the Rasch model for dichotomous (0, 1) items so that it accommodates the polytomous CR item data. Under the PCM model, for a given item i with m_i score categories, the probability of person n scoring x ($x = 0, 1, 2, \ldots m_i$) is given by

$$P_{ni}(X = x) = \frac{\exp \sum_{j=0}^{x} (\tau_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\tau_n - D_{ij})},$$

where τ_n represents examinee ability, and D_{ij} is the step difficulty of the jth

model and the single step difficulty is referred to as the item's difficulty. The Rasch model predicts the probability of person Q getting item I correct as follows:

$$P_{ni}\ X \quad 1 = \frac{\exp\ \ T_n \quad D_{ij}}{1 \quad \exp\ \ T_n \quad D_{ij}}.$$

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 computer program (Wright and Linacre, 2015), which employs unconditional (UCON), joint maximum likelihood estimation (JMLE).

3.3 CHARACTERISTICS OF THE TESTING POPULATION

The data analyses reported here are based on all students who took the Transition Regents Examination in Global History and Geography – Grade 10 in the June 2018 administration. The characteristics of this population are provided in Table 1.

3.4. ITEM DIFFICULTY-STUDENT PERFORMANCE MAPS

The distributions of the Rasch item logits (item difficulty estimates) and student performance are shown on the item difficulty-student performance map presented in Figure 2. This graphic illustrates the location of student performance and item difficulty on the same scale, along with their respective distributions and cut scores (indicated by the horizontal dotted lines). The figure shows more difficult items and higher examinee performance at the top and lower performance and easier items at the bottom.

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:

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

Marais and Andrich (2008) point out that local item dependence in the Rasch model can occur in two ways that may be difficult to distinguish. The first way occurs when the assumption 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 can 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 Global History and Geography and the Transition Regents Examination in Global History and Geography – Grade 10 items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using (T) 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 statistic is approximately $-1/(N-1)$ when no local dependence exists, where N is test length (Yen, 1993). Thus, the expected Q_3 values should be approximately −0.01 for the items on the exam. Absolute index values that are greater than 0.20 indicate a degree of local dependence that probably 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 the absolute residual correlations greater than 0.20 are also reported in this table. There was only one item pair with an absolute residual correlation greater than 0.20. The mean residual correlation was very

Table 5 Summary Statistics for Change in Grade 10 – Grade 10

Statistic Type	Value
N	946
Mean	-0.02
SD	0.03
Minimum	-0.12
P_{10}	-0.05
P_{25}	-0.03
P_{50}	-0.02
P_{75}	-0.01
P_{90}	0.02
Maximum	0.25

maintained at the target value, due to differences in the distributions of the Rasch difficulty values among the items from administration to administration.

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

All Regents Examinations are equated back to a base scale, which is held constant from year to year. Specifically, they are equated to the base scale through the use of a calibrated 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

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.

Pas4 (i(t)2 (s)14ne)10e /P <</aethe2 (i)614 (ut)a10 ( )65-10 (10 ( )r3 ( )18)-DC -29.4 -1.10 Td-2 f ssa

Reliability coefficients range from 0.0 to 1.0. The index will be 0.0 if none of the test score variances is

&RQGLWLRQDO6WDQGDUG&(RURI0HDVXUHPHQW&DUDFWHULVWLFV

The relationship between the scale score CSEM and depends both on the nature of the raw-to-scale score transformation (Kolen and Brennan, 2005; Kolen and Lee, 2011) and on whether the CSEM is derived from the raw scores or from (Lord, 1980). The pattern of CSEMs for raw scores and linear transformations of the raw score tend to have a characteristic “inverted-U” shape, with smaller CSEMs at the ends of the score continuum and larger CSEMs toward the middle of the distribution.

Achievable raw score points for these distributions are spaced equally across the score range. Kolen and Brennan (2005, p. 357) state, “When, relative to raw scores, the transformation compresses the scale in the middle and stretches it at the ends, the pattern of the conditional standard errors of measurement will be concave up (U-shaped), even though the pattern for the raw scores was concave down (inverted-U shape).”

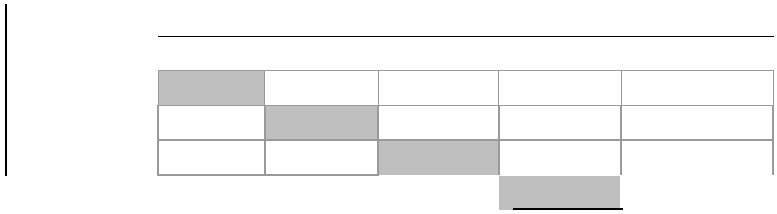
5 H V X O W V D Q G 2 E V H U Y D W L R Q V

The relationship between raw and scale scores for the Regents Examinations tends to be roughly linear from scale scores of 0 to 65 and then concave down from about 65 to 100. In other words, the scale scores track linearly with the raw scores for the first quarter of the scale score range and then are compressed relative to the raw scores for the remaining three-quarters of the range, though there are slight variations. The CSEMs for the Regents Examinations can be expected to have inverted-U shaped patterns, with some variations.

Figure 4 shows this type of CSEM variation for the Transition Regents Examination in Global History and Geography – Grade 10, in which the compression of raw score to scale scores around the cut score of 64 changes the shape of the curve slightly

		TEST ONE		
		LEVEL	LEVEL I	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

Fig 5 Pd Dd TT wh Cg



detailed attention was given to this correspondence between standards and test content during test design and construction.

The Regents Examination in Global History and Geography and the Transition Regents

4) Rubric designs that facilitate consistency of ratings (Pecheone & Chung, 2006; Wolfe & Gitomer, 2000; Cronbach, Linn, Brennan, & Haertel, 1995; Cook & Beckman, 2009; Penny, Johnson, & Gordon, 2000; Smith, 1993; Leacock, Gonzalez, and Conarroe, 2014)

The distinct steps for operational test scoring include close attention to each of these elements and begin before the operational test is even 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 are then used to build training materials for the vendor's scorers, who then score the rest of the field test responses to constructed-response items. The final model response sets for the August 2017 and January 2018 administrations of the Regents Examination in Global History and Geography, and the June 2018 administration of the Transition Regents Examination in Global History and Geography – Grade 10 are located at http://www.nysedregents.org/GlobalHistoryGeography/

includes a review of the results of classical item analyses, test reliability, and the IRT scaling and equating.

The following analyses were conducted for the Regents Examination in Global History and Geography and the Transition Regents Examination in Global History and Geography – Grade 10:

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

,WHP 'LIILFXOW\

Multiple analyses allow an evaluation of item difficulty. For this exam, Svalues and Rasch difficulty (item location) estimates were computed for MC and CR items. Items for the Regents Examination in Global History and Geography and the Transition Regents Examination in Global History and Geography – Grade 10 show a range of Svalues consistent with the targeted exam difficulty. Item Svalues in June 2018 for the Transition Regents Examination in Global History and Geography – Grade 10 range from 0.44 to 0.95, with a mean of 0.76. Furthermore, the point-biserial values (discussed in the following section) for these items indicate that they are discriminating the high performers well. Refer to Chapter 2 of this report for additional details.

,WHP 'LVFULPLQDWLRQ

How well the items on a test discriminate between high- and low-performing examinees is an important measure of the structure of a test. Items that do not discriminate well generally provide less reliable information about student performance. Tables 2 and 3 provide point-biserial values on the correct responses, and Table 2 also provides point-biserial values on the three distractors. The values for correct answers are 0.20 or higher for all items, and for all distracters, they are negative or close to zero, indicating that examinees are responding to the items as expected during item and rubric development.

'LIIHUHQWLDO ,WHP)XQFWLRQLQJ

Differential item functioning (DIF) for gender was conducted following field testing of the items in 2012 through 2017. Sample sizes for subgroups based on ethnicity and English language learner/multilingual learner status were, unfortunately, too small to reliably compute

other hand, refers to the comparison of members from the two groups who received the same score on the matching variable — the total test score in our analysis.

Eleven operational items for the June 2018 administration had DIF flags from the field test. Eight items (#s 22, SCF1, SCF3a, SCF4, 4, 1 g iteg i

History and Geography – Grade 10, both decision consistency and accuracy values for all three cut points are high, indicating very good consistency and accuracy of examinee classifications.

Re

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). 6 W D QGDUG V IU R U H G X F D W L R Q D O D Q G W H V W L Q J. Washington, DC: American Educational Research Association.

Barkaoui, Khaled. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance.

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. - R X U Q D O R I $ S S O L H G (4), 517–524. R O R J \

Messick, S. (1995). Standards of Validity and the validity of and standards in performance assessment. (G X F D W L R Q D O 0 H D V X U H P H Q W , V–8. X H V D Q G 3 U D F W L

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. - R X U Q D O R I (G X F D W 0 H D V X U H P H Q W (4), 371–389.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). & R P P R Q & R U H 6 W D 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. - R X U Q D O R I (G X F D W L R Q D O D Q G % H K D Y L R U D O 6 W D W L V W

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. 3 K L O R V R S K L F D O 7 U D Q V D F W L R Q V R I W K H 5 R \ 253–318.

Pecheone, R. L., & Chung Wei, R. R. (2007). Performance assessment for California teachers: Summary of validity and reliability studies for the 2003−04 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. 7 K H R X U Q D O R I ([S H U L P H Q W D O (G X F D W L R Q (3), 26

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. - R X U Q D O R I $ S S O L H G (4), 735–746. R O R J \

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. $ S S O L H G 3 V \ F K R O R 0 H D V X U H P H Q W , 39–49.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). + D Q G E R R N R I P R G H U Q L W H W K H R Y 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. - R X U Q D O R I (G X F D W L R Q D O 0 H D (2), 141–162. H Q W

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. - R X U Q D O R I $ S S O L H G 3 V (3), 546–561.

Weigle, S. C. (1998). Using FACETS to model rater training effects. / D Q J X D J H 7HVWLQJ, V W L Q J 15, 263–287.

43 (ab12 ( )2 (r12 (&)1 ty [<002F00440051>10.1 <004A 0 Td (43)Tj 0 Tc 00058>68 <0037>B>-6J /T

Th A. 2 T Mp fJ 2018 Ath 874 668.28 5 0 Td [A)124.48 f674.4 52.32 12.24 e52.8 12.24 f42

Table A.3 Test Map for Grade 8 Math

Position	Item Type	Max Points	Weight	Standard	Mean	Point-Biserial	Rasch Difficulty	INFIT
1	MC	1	2	4	0.40	0.31	1.2300	1.29
2	MC	1	2	3	0.77	0.42	-0.8203	1.02
3	MC	1	2	2	0.77	0.45	-0.8136	1.00
4	MC	1	2	2	0.77	0.47	-0.8337	0.98
5	MC	1	2	3	0.68	0.55	-0.2865	0.91
6	MC	1	2	2	0.66	0.51	-0.1239	0.98
7	MC	1	2	2	0.62	0.58	0.0714	0.90
8	MC	1	2	3	0.54	0.48	0.5223	1.06
9	MC	1	2	2	0.50	0.58	0.7195	0.92
10	MC	1	2	2	0.63	0.57	0.0177	0.91
11	MC	1	2	3	0.59	0.51	0.2462	1.01
12	MC	1	2	5	0.58	0.50	0.2776	1.03
13	MC	1	2	4	0.71	0.55	-0.4434	0.91
14	MC	1	2	3	0.57	0.45	0.3195	1.10
15	MC	1	2	3	0.57	0.56	0.3612	0.94
16	MC	1	2	4	0.68	0.43	-0.2691	1.07
17	MC	1	2	3	0.64	0.37	-0.0562	1.16
18	MC	1	2	2	0.55	0.26	0.4059	1.33
19	MC	1	2	2	0.68	0.48	-0.2747	1.00
20	MC	1	2	2	0.66	0.39	-0.1642	1.13
21	MC	1	2	4	0.55	0.46	0.4161	1.05
22	MC	1	2	2	0.61	0.59	0.1076	0.88
23	MC	1	2	3	0.60	0.52	0.1492	0.97
24	MC	1	2	3	0.64	0.58	-0.0669	0.89
25	MC	1	2	5	0.64	0.55149		
26	MC	1	2	5	0.53			

Th B. 3 Se 6 Tu 201 8 Ath

GENERAL RULES 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 Z K L F K R I W K should not be used to refer to the alternatives. Instead use Z K L F K 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.

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 clues to the correct answer.

5. The item should assess student understanding of the material by requiring responses that show evidence of knowledge, comprehension, application, analysis, synthesis, and/or evaluation.

6. When a stimulus is used, an introduction is required.

7. The item should clearly specify what the student is expected to do to provide an acceptable response.

8. A group of constructed response items should be arranged in logical sequence, and each item should test different knowledge, understandings, and/or skills.

9. The stimulus should provide information/data that is scientifically accurate.

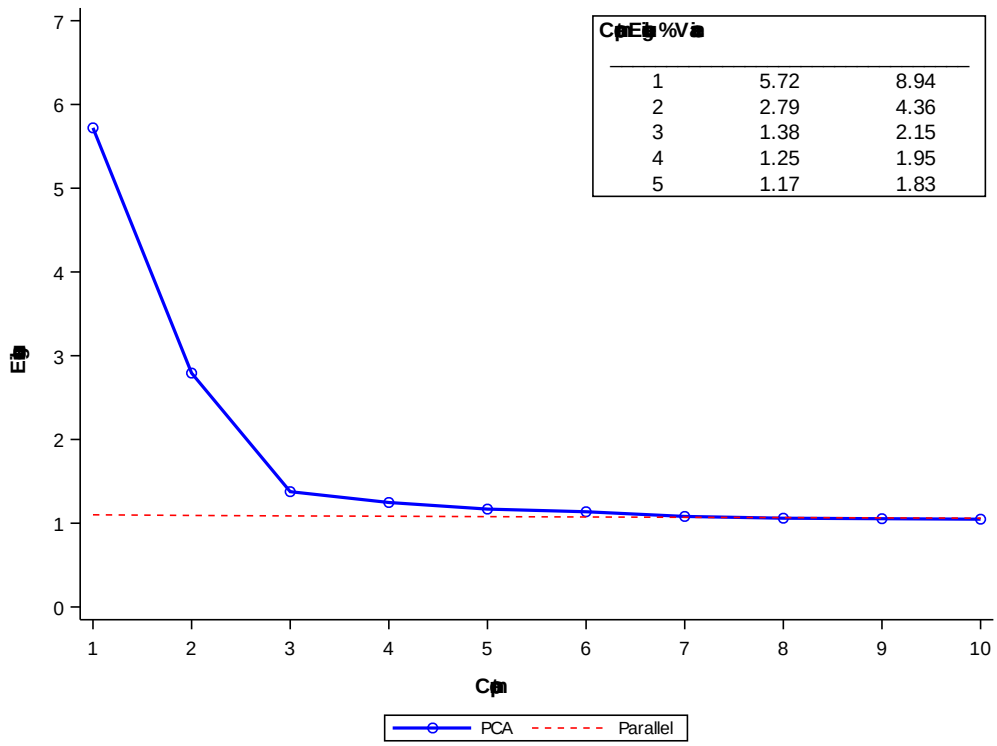
10. The source of each stimulus must be clearly identified for all material that is not original.

Table D.2 Classical Item Analysis Results for Grade 4 Reading

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

Table D.3 Distribution of Pre-Test Scores by Elementary Grade

Statistics	N	Mean	Min				



Component	Eigenvalue	% Variance
1	5.72	8.94
2	2.79	4.36
3	1.38	2.15
4	1.25	1.95
5	1.17	1.83

Fig D.3 Scree Plot Results for Grade 8 Math

Table D.4 Summary Statistics for Grade 8 Math

Statistic Type	Value
N	2,016
Mean	-0.01
SD	0.03
Minimum	-0.12
P_{10}	-0.04
P_{25}	-0.03
P_{50}	-0.02
P_{75}	0.00
P_{90}	0.02
Maximum	0.29
>|0.20|	2

The E.1 Mp -Cte Ay (2 -0.002 Tw 0.33 </MCID 2 ()1 () 5 50.9-BT 0 Tw 57-By ET 56.16 -Bt ;SN' ;M

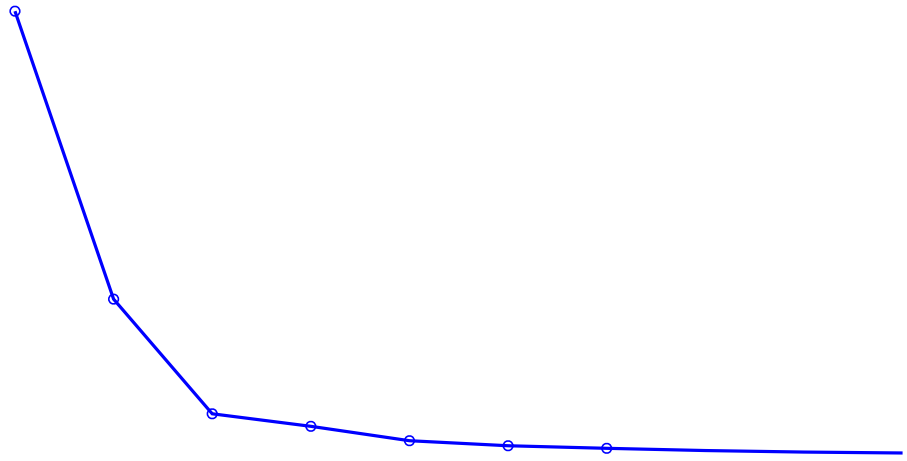


Table E.8 Grade 8 Mathematics Demographics

Demographics	Number	Mean Scale Score	SD Scale Score
All Students*	43,075	56.38	15.17
Ethnicity			
American Indian/Alaska Native	362	57.71	14.96
Asian/Native Hawaiian/Other Pacific Islander	2,312	59.19	16.34
Black/African American	14,200	55.03	14.82
Hispanic/Latino	16,053	55.54	14.69
Multiracial	552	58.81	15.67
White	9,575	58.92	15.73
English Language Learner/Multilingual Learner			
No	36,425	57.28	15.01
Yes	6,650	51.44	15.06
Economically Disadvantaged			
No	10,453	59.09	15.70
Yes	32,622	55.51	14.89
Gender			
Female	21,005	57.39	14.24
Male	22,049	55.42	15.94
Student with a Disability			
No	31,197	58.78	14.56
Yes	11,878	50.07	14.94

*Note: Twenty-one students were not reported in the Ethnicity and Gender group, but they are reflected in "All Students."