CLASSICAL AND RASCH ANALYSIS OF:

FORMS OF JUSTIFICATION

This work was funded by the Carnegie Corporation of New York through grant #B 8780 to the Jacqueline Barber of the Learning Design Group; The Lawrence Hall of Science, and David Pearson, Graduate School of Education, both at the University of California, Berkeley and Katherine McNeill at the Lynch School of Education, Boston College.

Contents

Executive Summary 4
Introduction 5
Method 6
Study Design 6
Step 1: Construct Map 6
Steps 2 and 3: Item Design and Outcome Space 7
Step 4: Measurement Model 8
Data Collection 13
Results 15
Classical Test Theory 15
Rasch Modelling 17
References 21
Appendix A 23
Appendix B 26
Appendix C 29

Tables

Table 1. *Forms of Justification Construct Map for the Reading Modality* 7
Table 2. *Classification Rules for Item Difficulty Levels* 8
Table 3. *Sample Size per Testlet Combination* 14
Table 4. *Number of Students by Grade* 14
Table 5. *Number of Students by Ethnicity* 14
Table 6. *Proportion-Based Item Difficulties* 15
Table 7. *EFA Factor Loadings* 16
Table 8. *Raw and Polychoric Alpha* 17
Table 9. *Model Comparison Tests* 17
Table 10. *Estimated Item Locations* 18
Table 11. *Model Fit for Testlets* 23
Table 12. *Model Fit for Items* 23
Table 13. *Model Fit for Items*Testlet* 24
Table 14. *Model Fit for Items*Testlet*Step* 25

Figures

Figure 1. *Argument Assessments* 5
Figure 2. *Scree Plot for EFA* 16
Figure 3. *Wright Map for Testlet* Item*Step (FOJ)* 19
Figure 4. *Item Characteristic Curves for Level-1 Item* 26
Figure 5. *Item Characteristic Curves for Level-2 Item* 27
Figure 6. *Item Characteristic Curves for Free-Response Item* 28

Executive Summary

A series of procedures, using classical test theory (CTT) and the Rasch model, were completed to understand the psychometric characteristics of an instrument developed to measure students' ability to critique the Forms of Justification used in written arguments. The CTT analyses utilized factor analysis and measures of internal reliability. The Rasch model utilized a multi-faceted partial credit model. The CTT had mixed interpreations, while internal reliability was adequate, the factor analysis suggesed a two-dimensional solution. Despite the potential two-factor solution, we assumed a single factor solution and conducted a Rasch model. The results of the Rasch model indicated good overall fit to the model and significant interaction effects with testlets, which indicated that the items are behaving differently between testlets. The biggest differences in item difficulties between testlets appear to be between content areas (Volcanoes vs. Earthquakes).

Introduction

This Technical Report details the psychometric properties for an assessment of middle school students' ability to construct and critique scientific arguments. By argument, we mean a justified claim. A claim is a statement that answers a question; justifications are used to support the claim. An argument marshals the justification to support the claim.

There are numerous aspects of students' understanding of argumentation that could be assessed; it is difficult to know which are the most important. We assessed specific characteristics (or constructs) of the structure of an argument (i.e. forms of justification and relevant-supporting evidence) in reading (i.e. forms of justification, relevant-supporting evidence, sufficiency of evidence, multiple views, and reasoning) in our written assessments as well the process of constructing interpersonal and intrapersonal arguments within our verbal assessments.

Figure 1 summarizes the ten assessments we have developed to date. This report focuses on the items in our assessment of students' ability to identify, interpret, compare, and critique the forms of justification in written arguments.

Figure 1. *Argument Assessments*

Construct (characteristics of arguments)	Reading	Writing	Talking
Forms of Justification	✔	✔	
Relevant-Supporting Evidence	✔	✔	
Sufficiency of Evidence		✔	
Multiple Claims		✔	
Reasoning		✔	
Intrapersonal			✔
Interpersonal			✔

Forms of Justification

This report focuses on the psychometric characteristics of the instrument designed to measure the construct *Forms of Justification* within the reading modality (the highlighted cell in Figure 1).

This construct recognizes that students use many different forms of justifications, including empirical evidence (McNeill & Krajcik, 2012), science ideas (Osborne, Erduran, & Simon, 2004), appeals to authority, plausible mechanisms, and prior experiences (Sandoval & Cam, 2011). While students may use many different forms of justification, some forms are more accepted within the scientific community. One example is the preference for

empirical data as a form of justification (Sandoval & Cam, 2011). While another form of justification, such as science ideas or an appeal to authority, used in combination with empirical evidence can serve to further strengthen the argument, the value of these other forms of justification diminish when they stand-alone. In this project we present a way to measure students' abilities to critique the forms of justification in reading, writing, and talking tasks. It is this ability that is of interest in this assessment.

Method

Study Design

Assessment design process followed the guidelines of the Berkeley Evaluation and Assessment Research (BEAR) Assessment System (Wilson, 2005; 2009). This process is comprised of iterative steps with four building blocks: 1) construct maps, 2) items design, 3) outcome space, and 4) measurement model. First, construct maps, which are theoretical models of cognition that extend from low to high ability, are derived based on the construct that is to be measured. The second stage of development, item design, produces the means in which the theoretical construct is measured. The third stage, the outcome space, addresses the possible ways students answer the items. For multiple-choice questions, the outcome space is often focused on developing strong answer choices based on theory. In comparison, the outcome space for constructed response items addresses the qualitative classification of student responses using rubrics or scoring guides. The final stage of development is the measurement model, which translates item scores to the original construct map.

Each of the four building blocks will be detailed in the context of forms of justification.

Step 1: Construct Map

The research on students' abilities to read and critique forms of justification is sparse. Students whose ability is at the *identifies* level can identify the specific form of justification used when given the following options: empirical evidence, science ideas, plausible mechanisms, appeals to authority, and personal stories. While there is no empirical research to support this level, we rely on the expertise of the team members whose experiences suggest that this is challenging for many students. The results of one research group, however, suggest that students tend to struggle with locating justifications (Norris & Phillips, 1994). Our empirical results suggest that the *identifies* level is of lower difficulty than the *locates* level. At the highest level, the students critique the quality of the form of justification. *Critique* is a difficult skill and, in line with other researchers, we reserve it for the most difficult level (e.g., Osborne, Erduran, & Simon, 2004). The students whose ability is at the *compare & critique* level do so in regards to the quality of two arguments.

Table 1. *Forms of Justification Construct Map for the Reading Modality*

	Level	Description
High ↑	3 - Compares & Critiques	Student critiques the form of justification used when comparing two arguments.
	2 - Interpret	Student distinguishes between what is the claim versus the justification
	1 - Identifies	Student identifies the form of justification used in an argument.
Low	0 - Does not Identify	Student does not identify the form of justification used in an argument.

Steps 2 and 3: Item Design and Outcome Space

Three item templates (two multiple-choice and one constructed-response) were produced based on the construct map. Based on the item templates, a set of items (which will be referred to as *testlets*) was derived for each of four topics (2 for *Earthquakes*, and 2 for *Volcanoes*).

The two multiple-choice items in each testlet shared a common prompt which presented a hypothetical argument. These hypothetical arguments consisted of three sentences: 1) claim, 2) a justification, and 3) reasoning. To identify the sentences, we used the notation "(S1)", "(S2)", and "(S3)" to indicate sentence 1, 2, and 3 respectively. To make this notation as clear as possible, we included a legend along with the hypothetical argument. This notation is aligned with how other assessment systems are developing items (e.g. PARCC). All four testlets are available in Appendix C.

Step 4: Measurement Model

The measurement model pertains to the inferences that can be drawn from the data that has been collected. Inherent to this step is proper fit of the data to the statistical model. Two competing statistical models, classical test theory and Rasch modelling, were implemented for comparison.

Classical Test Theory

Item Difficulty. The foundation for the Classical Test Theory (CTT) rests on aspects of a total test score made up of multiple items. The success rate of a particular pool of examinees on an item, well-known as the *p*-value of the item (not to be confused with the *p*-value associated with hypothesis testing), is used as the index for the item difficulty. The difficulty for a particular item is defined as the proportion of examinees that correctly answered the item. This statistic is an inverse indicator of item difficulty, as the higher the *p*-value, the easier the item. A general guideline by which items are classified according to their difficulty levels is presented in Table 2.

For dichotomous items, the simple proportion of correct responses achieved by the sample on the item will be presented as the *p*-value of the item. For partial credit items, the raw average of the scores divided by the maximum score will be presented as the *p*-value of the item. This effectively allows the partial credit item to be placed on the same scale as the dichotomous items for comparison. A set of *p*-values will be provided for each of the four testlets.

Table 2. *Classification Rules for Item Difficulty Levels*

***p*-value (easy to hard)**	**Item Interpretation**
.96 to 1.00	Inappropriately easy (unacceptable) items
.90 to .95	Very easy (possibly unacceptable) items
.80 to .89	Fairly easy (acceptable) items
.40 to .79	Hard to moderately easy (acceptable) items
.30 to .39	Difficult (acceptable) items
.20 to .29	Very difficult (possibly unacceptable) items
.00 to .19	Inappropriately difficult (unacceptable) items

*Adapted from Prometric (2013)

Unidimensionality. The goal of exploratory factor analysis (EFA) is to determine the smallest number of latent factors that can reasonably explain the correlations among the observed variables (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Given this goal, EFA places no a priori expectations on the number of latent factors present or the expected relationships between the observed measures and these latent factors (Brown, 2006).

We begin the analysis with an EFA because students' skills necessary to construct and critique arguments were originally theorized to be multidimensional. However, the specific skills that combine together to create the ability were viewed as unidimensional.

Therefore, EFA was the logical first step in determining a final set of items that would conform to the simple structure required to estimate interpretable scores.

EFA was conducted on each of the FOJ related items across all four testlets (i.e., 3 items for 4 testlets, for a total of 12 generalized items) through Mplus (v. 7.11). A 1-factor and 2-factor EFA model was produced utilizing the mean- and variance-adjusted weighted least squares estimator (WLSMV), then compared using the resulting scree plot and rotated factor loadings (via the GEOMIN oblique rotation). WLSMV estimation was preferred in this particular application due the use of the ordered probit link function was deemed appropriate for the ordinal nature of the data (see Muthèn & Muthèn, 1998-2001; Appendix 8, p. 370).

Satisfactory model fit was determined by the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) both being greater than .90, and a Root Mean Square Error of Approximation (RMSEA) of .06 or less (Hu & Bentler, 1999; Byrne, 2010). The commonly used chi-square goodness-of-fit test was not utilized because of its known sensitivity to sample size (Wu, Li, & Zumbo, 2007).

Factor loadings are used to provide information about how strongly items are related to the latent construct. Items with stronger relationship to the construct are deemed more reliable indicators of that construct (i.e., student scores on these items are seen as indicators of overall student ability on the intended construct being measured) (Edwards & Wirth, 2009). To determine whether an item is a component of a specific factor, a cut-off value of .30 is recommended as an acceptable minimum value for the coefficients

Reliability. Internal reliability of the items was determined using Cronbach's α based on tetrachoric, and polychoric correlations between the dichotomous and ordinal scaled items. This method was preferred over the traditional Cronbach's α, which assumes continuous scaling of the variables (Gadermann, Guhn, & Zumbo, 2012). A single indicator of reliability will be presented across all four testlets.

Rasch Modelling

Model Comparison. ConQuest (v. 2.0) implements the multidimensional random coefficients multinomial logit model (MRCMLM), which is a generalization of many Rasch-family models, such as the simple logistic model (Rasch, 1960), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), and the mult-faceted model (Linacre, 1989b) (Wu, Adams, Wilson, & Haldane, 2007). The MRCMLM was used to specify a multi-faceted partial credit model, with facets being each of the four testlets. As each participant took up to two testlets during administration, the measurement model effectively treated the four testlets as four different raters (with varying degrees of rating "harshness"), with up to two raters scoring each individual. Given that the items and their respective scoring rubrics were designed to have very similar characteristics across testlets, the multi-faceted model was be used to determine general differences between testlets and interactions between items and testlets. Further, the use of the partial credit model allowed both the multiple choice and constructed response scores to be examined in a single unidimensional model.

All Rasch-family models are rooted in Rasch's (1960) simple dichotomous model, which can be written in the following format:

$$\ln\left(\frac{\pi_{n,i,1}}{\pi_{n,i,0}}\right) = \theta_n - \delta_i \quad (1)$$

where

$\pi_{n,i,1}$ = probability of participant n scoring 1 on dichotomous item i,
$\pi_{n,i,0}$ = probability of participant n scoring 0 on dichotomous item i,
θ_n = "ability" of participant n, and
δ_i = "difficulty" of item i.

The left side of the equation is the natural-logarithm of the odds ratio of participant n correctly answering item i (i.e., getting a score of 1) and the same participant incorrectly answering the same item. A log-odds unit (or logit) is continuous with a range of (-∞, ∞), hence, the coefficients on the right side of the equation (θ_n and δ_i) are also values along the same logit continuum. Further, the probability of participant n answering item i correctly is 50% when θ_n and δ_i are equal, since:

$$\ln\left(\frac{\pi_{n,i,1}}{\pi_{n,i,0}}\right) = \ln\left(\frac{\pi_{n,i,1}}{1-\pi_{n,i,1}}\right) = \ln\left(\frac{0.5}{1-0.5}\right) = 0 \quad (2)$$

The dichotomous Rasch model can also be used to model K-category score (i.e., polytomous) items using the partial credit model (Masters, 1982), which can be written as:

$$\ln\left(\frac{\pi_{n,i,k}}{\pi_{n,i,k-1}}\right) = \theta_n - \left(\beta_i + \tau_{i,k}\right) \quad (3)$$

where

$\pi_{n,i,k}$ = probability of participant n scoring k on item i,
$\pi_{n,i,0}$ = probability of participant n scoring $k-1$ on item i,
θ_n = "ability" of participant n,
β_i = "overall difficulty" of item i, and
$\tau_{i,k}$ = "step difficulty" of scoring k relative to scoring $k-1$ for item i $(k = 1, 2, \ldots, K_i)$.

Since there are $K-1$ possible transitions (e.g., 0 to 1, 1 to 2, 2 to 3) in a K category item, a step difficulty τ is modeled for each of $K-1$ steps for each item. Further, the additive value of β_i and $\tau_{i,k}$ equals the item difficulty δ_i from the dichotomous model, with $\tau_{i,k}$ being the deviance from the overall difficulty β_i. The $\tau_{i,k}$ coefficient is the location along the logit continuum that the adjacent categories k and $k-1$ are equally probable (i.e., 50%) (Bond & Fox, 2007).

In addition to polytomous scoring, the current measurement model also takes into account many-faceted data using Linacre's multi-faceted partial credit model (1989b), which can be written as:

$$\ln\left(\frac{\pi_{n,i,j,k}}{\pi_{n,i,j,k-1}}\right) = \theta_n - \left(\beta_i + \alpha_j + \gamma_{i,j} + \tau_{i,j,k}\right) \quad (4)$$

where

$\pi_{n,i,j,k}$ = probability of participant n scoring k on item i,
$\pi_{n,i,j,k-1}$ = probability of participant n scoring $k-1$ on item i,
θ_n = "ability" of participant n,
α_j = "overall severity" of facet j,
$\gamma_{i,j}$ = item-by-facet interaction for item i and for facet j, and
$\tau_{i,j,k}$ = step-by-facet interaction for scoring k relative to scoring $k-1$ for each item i and for each facet j $(k = 1, 2, \ldots, K_i)$.

As noted earlier, the facets are testlet versions in the current model. The coefficient α_j will be used to determine if the "severity" (i.e., testlet difficulty) is different between the testlets. The coefficient γ_{ij} accounts for interaction effects between a facet and an item's overall difficulty. Lastly, the coefficient $\tau_{i,j,k}$ accounts for inconsistencies in item step difficulties between facets that are not accounted for in the previous coefficients.

Equation (4) will serve as the unconstrained model (Model 1), which will be compared to a nested model replacing $\tau_{i,j,k}$ with $\tau_{i,k}$ (i.e., removing the modeling of possibly varying step difficulties between facets). The nested model (Model 2) can be written as:

$$\ln\left(\frac{\pi_{n,i,j,k}}{\pi_{n,i,j,k-1}}\right) = \theta_n - \left(\beta_i + \alpha_j + \gamma_{i,j} + \tau_{i,k}\right) \quad (5)$$

where $\tau_{i,k}$ is the "step difficulty" of scoring k relative to scoring $k-1$ for item i, where $k = 1, 2, \ldots, K_i$. Since models 1 and 2 are nested, the inclusion of the addition of the step-by-facet interaction in Model 1 significantly improves model fit. The relative change in the model fit deviance statistic (also known as the G^2 statistic) has been shown to be distributed approximately chi-square with degrees of freedom equaling the difference in the number of parameters estimated in the two models (Briggs & Wilson, 2003). Thus, a significant value in this test would indicate significantly improved fit between models (4) and (5), and indicating that the additional step-by-facet interaction in model (5) is necessary. If the test does not indicate significance, the more parsimonious constrained model will be used instead of the unconstrained model.

A third model which removes any testlet interactions will be used to determine if an item-by-testlet interaction effect significantly improves model fit. This model can be written as:

$$\ln\left(\frac{\pi_{n,i,j,k}}{\pi_{n,i,j,k-1}}\right) = \theta_n - \left(\beta_i + \alpha_j + \tau_{i,k}\right) \quad (6)$$

Equation (6) (Model 3) removes the $\gamma_{i,j}$ term from Model 2, and, thus, is the most constrained model of the three. Model 3 will be compared to Model 2 using the deviance statistic to determine if the absence of the $\gamma_{i,j}$ term significantly worsens model fit.

Lastly, the main effect of "overall severity" (i.e., difference between testlets) will be examined for the most parsimonious model. This can be tested using separation reliability and the chi-square test of parameter equality[1] (i.e., p<.05 indicates significantly different testlet severities).

Model Fit. ConQuest provides two residual statistics that serve as an indication of the model fit: the weighted mean-square (infit) and the unweighted mean-square (outfit). For items, the outfit statistic is sensitive to unexpected observations by examinees on items that are either very easy or very hard for examinees (such as lucky guesses), while the infit statistic is sensitive to unexpected responses to items whose difficulties are close to student abilities (such as some students with specialized knowledge leading to a different rank ordering of difficulty in the items) (Bond & Fox, 2007). For facets (i.e., testlets), the outfit is sensitive to unexpected ratings from a generally consistent rater, while the infit is less sensitive to the same conditions (Eckes, 2011). As such, while both infit and outfit will be considered, the indicated misfit shown by the infit statistic will be considered more problematic to overall model fit than the outfit statistic.

Smith, Schumacker, and Bush (1998) suggest that items be considered under-fitting when the outfit and infit values are greater than 1.30 for samples less than 500, 1.20 for samples between 500 and 1000, and 1.10 for samples larger than 1000. Bond and Fox (2007) suggested that values less than .75 are over-fitting which can lead to falsely inflated reliability statistics. Typically, a *t*-statistic indicates whether or not an item's infit or outfit significantly deviates from one. However, Smith et al. (2008) have demonstrated that as sample size increases, the fit *t*-statistics will indicate misfit even for well-fitting items. For this reason, only the infit and outfit statistics will be discussed in this report.

Lastly, the discrimination index provided by ConQuest will be examined as an indicator of model fit. Item discrimination can be described as the magnitude of the contribution of a particular item to the entirety of the scale. This is analogous to the factor loadings resulting from factor analysis. As such, item discrimination that is negligibly small or even negative is problematic to the model as a whole. The discrimination index is the correlation between the total score and the score of a particular item. Items exhibiting particularly low or negative values of the discrimination index (i.e., <.20) will be considered as indicators of misfit to the Rasch model.

Reliability. ConQuest provides the person-separation reliability (PSR) statistic. The PSR is analogous to Cronbach's α as a measure of internal reliability (Wright & Masters, 1982). Larger PSR indicates a greater level of confidence that the person ability estimates are distinct and that the difference in ability estimates between individuals is not merely due to error. As with Cronbach's α, a PSR index that is greater than .70 was considered

[1] ConQuest was used to produce full standard errors (as opposed to quick standard errors, which typically under-estimates the true standard error) for a more accurate test of parameter equality. Item difficulties were constrained to equal zero for model identification purposes.

satisfactory. Expected-a-posteriori plausible values (EAP/PV) estimates of person ability estimates (which utilize the normal distribution as a Bayesian prior) were used in computing the PSR.

As the PSR tends to increase with an increase in the number of items and an increase in variance in item difficulties, the indicated reliability with current three-item set (which were theorized to share similar difficulty levels across testlets) was not expected to be particularly high, especially when the differences between testlets were not significant. However, the PSR was useful in recommendations for future iterations of the instrument.

Item Characteristic Curves. Item characteristic curves (ICCs) provide a graphical representation of the probability of a particular score for a particular item (on the Y-axis) for corresponding values of participant ability (on the X-axis). As each of the 12 generalized items (i.e., 3 items for 4 testlets, for a total of 12) have varying degrees of difficulty, an ICC will be presented for each of the generalized items.

Wright Map. The Wright map provides a visual assessment of the location of item difficulties, testlet severities, and the distribution of participants mapped onto a common logit continuum. The Wright map will be used to provide comparisons between items (and item steps) and testlets. In the case of dichotomous items, the item difficulty (i.e., the location in which a participant which equal ability has a 50% probability of correctly answer the item) will be displayed on the Wright map. In the case with polytomous items, the Thurstone thresholds will be displayed (for each of $K - 1$ steps). Thurstone thresholds represent the location in which the probability of scoring at or above a particular rating is 50% for a participant of equal ability.

Data Collection

Sampling. Each participant in this pilot was assigned up to two testlets, and each testlet was completed on a separate day. A particular form was assigned to each teacher's class, with the goal of having a minimum of 50 students complete each form. The assessment was presented and answers collected via an online survey program (i.e., Qualtrics), and the students used either tablets or computers to answer the items.

After accounting for survey incompleteness (i.e., participants with <50% completion on the entirety of the survey, including both the Forms of Justification and Relevant Supporting Evidence items, were removed) and parental opt-outs, the analyzed sample was 525. Table 3 summarizes the number of students who took each testlet. Omitted questions were considered incorrect (i.e., a score of zero).

Demographics. In terms of grade level, there was a slight trend that favored upper grade level in the sample (see Table 4). Gender of participants was roughly equal, with males and females accounting for 48.2% (253) and 51.0% (268) of the sample, respectively (gender information was not available for 4 participants). The average age of the participants was 12.65.

For ethnicity, the majority of the participants indicated that they were ethnically white (63.6%), followed by Hispanic or Latino (19.4%). The remaining ethnic categories accounted for less than 5% of the sample (see Table 5). A relatively small portion of the sample (6.1%, or 32) was identified as receiving some form of English Language

Development[2]. Further, 11.8% (62) of individuals were identified as being participants of the free or reduced lunch program[3]. Lastly, 12.6% (66) participants were identified as being provided either Individual Education or 504 plans[4].

Table 3. *Sample Size per Testlet Combination*

	Testlet				
Strand	**Earthquakes 1**	**Earthquakes 2**	**Volcanoes 1**	**Volcanoes 2**	**# Unique**
1	22	.	.	.	22
2	79	79	.	.	79
3	80	.	80	.	80
4	77	.	.	77	77
5	.	23	.	.	23
6	.	74	74	.	74
7	.	72	.	72	72
8	.	.	15	.	15
9	.	.	69	69	69
10	.	.	.	14	14
Total	*258*	*248*	*238*	*232*	*525*

Table 4. *Number of Students by Grade*

Grade	**N. of students**	**% of students**
5th	2	0.4
6th	140	26.7
7th	182	34.7
8th	197	37.5
Unknown	4	0.8
Total	*525*	*100.0*

Table 5. *Number of Students by Ethnicity*

Ethnicity	**N. of students**	**% of students**
Asian	21	4.0
American Indian or Native American	2	0.4
Black/African American	18	3.4
Hispanic or Latino (a)	102	19.4
Native Hawaiian or Pacific Islander	22	4.2
White	334	63.6
Other	14	2.7
Multiple	8	1.5
Unknown	4	0.8
Total	*525*	*100.0*

[2] ELD/ESL information was not available for 7.8% of the sample (or 41 individuals).
[3] FL/RL information was not available for a relatively large portion of the sample (19.8%, or 104) due to issues regarding the disclosure of economic status information implemented at certain school districts.
[4] IEP/504 information was not available for 4.9% (or 26 individuals) of the sample.

Results

Classical Test Theory

Item Difficulty. The *p*-values of each testlet and overall total are shown in Table 6. In general, the level-1 item (FOJ_L1) appears to be an easier than the level-2 item (FOJ_L2), which is appropriate given the construct map. There is some indication of a testlet-by-item interaction effect, as the level-2 item (FOJ_L2) appears to be harder than the free-response item (FOJ_FR) for the two Earthquakes testlets, while the reverse is true for the two Volcanoes items.

Table 6. *Proportion-Based Item Difficulties*

		Score (Count)				
Testlet	Item	0	1	2	3	*p*-value
Earthquakes 1 (*n*=258)	FOJ_L1	45	213	.	.	0.826
	FOJ_L2	159	99	.	.	0.384
	FOJ_FR	42	90	77	49	0.505
Earthquakes 2 (*n*=248)	FOJ_L1	52	196	.	.	0.790
	FOJ_L2	172	76	.	.	0.306
	FOJ_FR	37	96	78	37	0.488
Volcanoes 1 (*n*=238)	FOJ_L1	51	187	.	.	0.786
	FOJ_L2	115	123	.	.	0.517
	FOJ_FR	92	58	59	29	0.368
Volcanoes 2 (*n*=232)	FOJ_L1	60	172	.	.	0.741
	FOJ_L2	120	112	.	.	0.483
	FOJ_FR	63	70	61	38	0.440
Total* (*n*=976)	FOJ_L1	208	768	.	.	0.787
	FOJ_L2	566	410	.	.	0.420
	FOJ_FR	234	314	275	153	0.452

*The total is based on all surveys taken, ignoring the repeated measures of the 525 total unique participants.

Unidimensionality. The scree plot for the EFA (shown on Figure 2) indicated a two-factor solution. Accordingly, the model fit statistics were not satisfactory for the one-factor solution (RMSEA=.052, CFI=0.788, TLI=0.741), while the two-factor solution showed satisfactory model fit (RMSEA=.023, CFI=0.967, TLI=0.950). The rotated factor loadings (shown on Table 7) indicated low factor loadings for the level-2 items (FOJ_L2) from each of the testlets. The factor loadings from the two-factor solution indicated a second factor based on the level-2 items. There was no evidence of correlation between the two factors in the two-factor solution (r=0.041, p>.05).

The current results of the EFA do not support the assumption of unidimensionality. The reason for the lack of scaling with the level-2 items should be investigated for future iterations. However, we choose to move forward with a 1-factor solution based on our underlying theory of the construct (McNeil & Krajcik, 2012).

Figure 2. *Scree Plot for EFA*

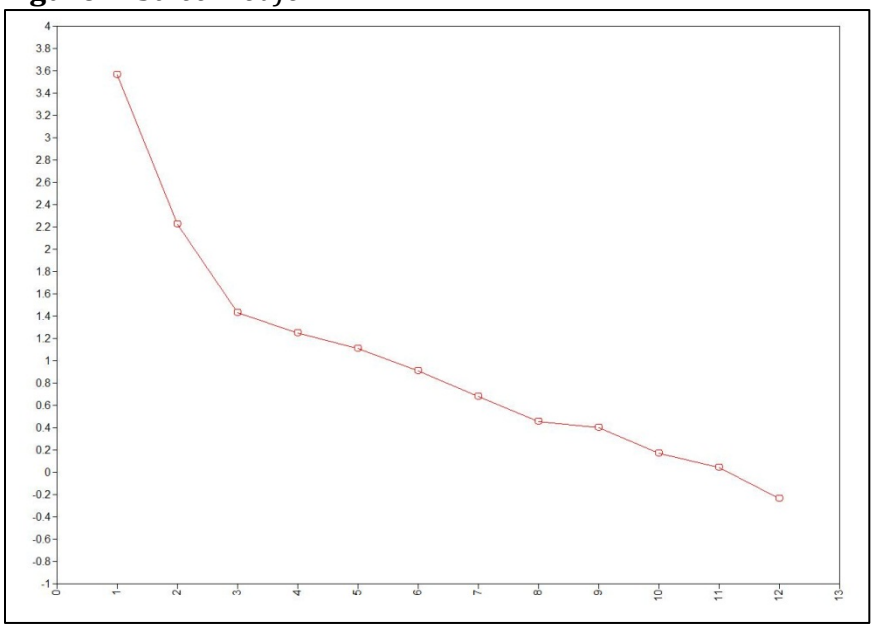


Table 7. *EFA Factor Loadings*

		1-Factor Solution	2-Factor Solution	
Item	Testlet	Factor 1	Factor 1	Factor 2
FOJ_L1	E1	0.451 *	0.332 *	0.351 *
FOJ_L2	E1	0.304 *	0.918 *	0.001
FOJ_FR	E1	0.671 *	0.114	0.670 *
FOJ_L1	E2	0.502 *	0.126	0.481 *
FOJ_L2	E2	0.231 *	0.783 *	-0.025
FOJ_FR	E2	0.687 *	0.053	0.703 *
FOJ_L1	V1	0.519 *	-0.105	0.539 *
FOJ_L2	V1	0.328 *	0.473 *	0.258 *
FOJ_FR	V1	0.744 *	-0.048	0.757 *
FOJ_L1	V2	0.296 *	0.117	0.269 *
FOJ_L2	V2	0.379 *	0.454 *	0.313 *
FOJ_FR	V2	0.748 *	-0.128	0.800 *

*Statistically significant <.05.

Reliability. The raw Cronbach's α coefficients and the polychoric Cronbach's α coefficients are shown in Table 8. Both indicators of reliability were satisfactory, though the polychoric variant is notably higher. There were no obvious indicators of scaling issues based on the "if removed" reliabilities. This implies a reliable instrument with all items contributing to the measurement of the construct.

Table 8. *Raw and Polychoric Alpha*

Item	Testlet	Raw α If Removed	Polychoric α If Removed
FOJ_L1	E1	0.70	0.76
FOJ_L2	E1	0.70	0.76
FOJ_FR	E1	0.66	0.75
FOJ_L1	E2	0.71	0.77
FOJ_L2	E2	0.70	0.76
FOJ_FR	E2	0.66	0.75
FOJ_L1	V1	0.70	0.76
FOJ_L2	V1	0.69	0.76
FOJ_FR	V1	0.65	0.75
FOJ_L1	V2	0.69	0.75
FOJ_L2	V2	0.70	0.76
FOJ_FR	V2	0.65	0.75
Scale α		0.70	0.77

Rasch Modelling

Model Comparison. A unidimensional model was fit for all three multi-faceted Rasch models, and their respective deviance statistics are shown in Table 9. A test of model fit difference between Model 1 (unconstrained) and Model 2 (item-by-facet) was significant ($\chi^2(6)=25.62$, $p<.01$). Further, a test of model fit difference between Model 2 and Model 3 (step-by-facet) was also significant ($\chi^2(6)=60.60$, $p<.01$). These results indicate that model fit significantly decreases with the exclusion of both the item-by-facet and step-by-facet interaction effects. Thus, both constrained models were rejected in favor of the unconstrained model. This implies that when the order and range between items, as well as, the order and range between steps within items, were allowed to vary across testlets was best and was therefore the model that we used.

Table 9. *Model Comparison Tests*

	Model 1	Model 2	Model 3
Deviance	4755.117	4780.736	4841.330
Parameters	21	15	9

Model Fit and Reliability. Based on Model 1, the infit and outfit statistics for all main effect and interaction terms were generally satisfactory (see Appendix A for fit statistics). Overall, the infit and outfit statistics were higher for the level-2 item (particularly for the main effect of item), which corresponds to the two-factor solution found during the EFA. There was no evidence of model misfit due item discrimination (see Table 11). Meaning that we have no evidence that the assumption of uniform item discrimination is violated. The separation reliability of the testlets was 0.566, which was found to be non-significant ($\chi^2(3)= 4.61$, $p=0.203$). This result indicates an absence of a main effect of testlet (i.e., on average, the testlet "severities" are equal). The EAP/PV person-separation reliability was 0.496. This implies that the reliability in the rank ordering of students is relatively low, so a

subsequent assessment using the same items might result in a drastically different ranking of the same students ability.

As shown in Appendix B, the patterns of the ICCs for the free-response item (FOJ_FR) are especially different between the two Earthquakes and the two Volcanoes testlets. Specifically, the score of "1" for both Volcanoes testlets is less probable than the two Earthquakes testlets. For Volcanoes 1, the probability of a score of "1" is always lower than the either of its adjacent scores. This is evidence that a change in the scoring rubric may be necessary.

Wright Maps. The Wright map for each item step by testlet combination for Model 1 is shown on Figure 3. The Wright map on Figure 3 maps the Thurstone threshold locations shown on Table 11 (the right-hand side) next to a histogram of the participant ability estimates (the left-hand side). In general, the item difficulties for the level-1 item are very similar across the four testlets, ranging from -.89 to -1.41. The level-2 item difficulties are less similar across the four testlets, ranging from .38 to 1.23. The level-2 items are consistently more difficult than the level-1 items, which is consistent with the design of the instrument based on the construct map.

In general, the pattern of Thurstone thresholds suggest that the testlets within each of the Volcanoes and Earthquakes content areas are more similar when compared between content areas (i.e., Volcanoes vs. Earthquakes).

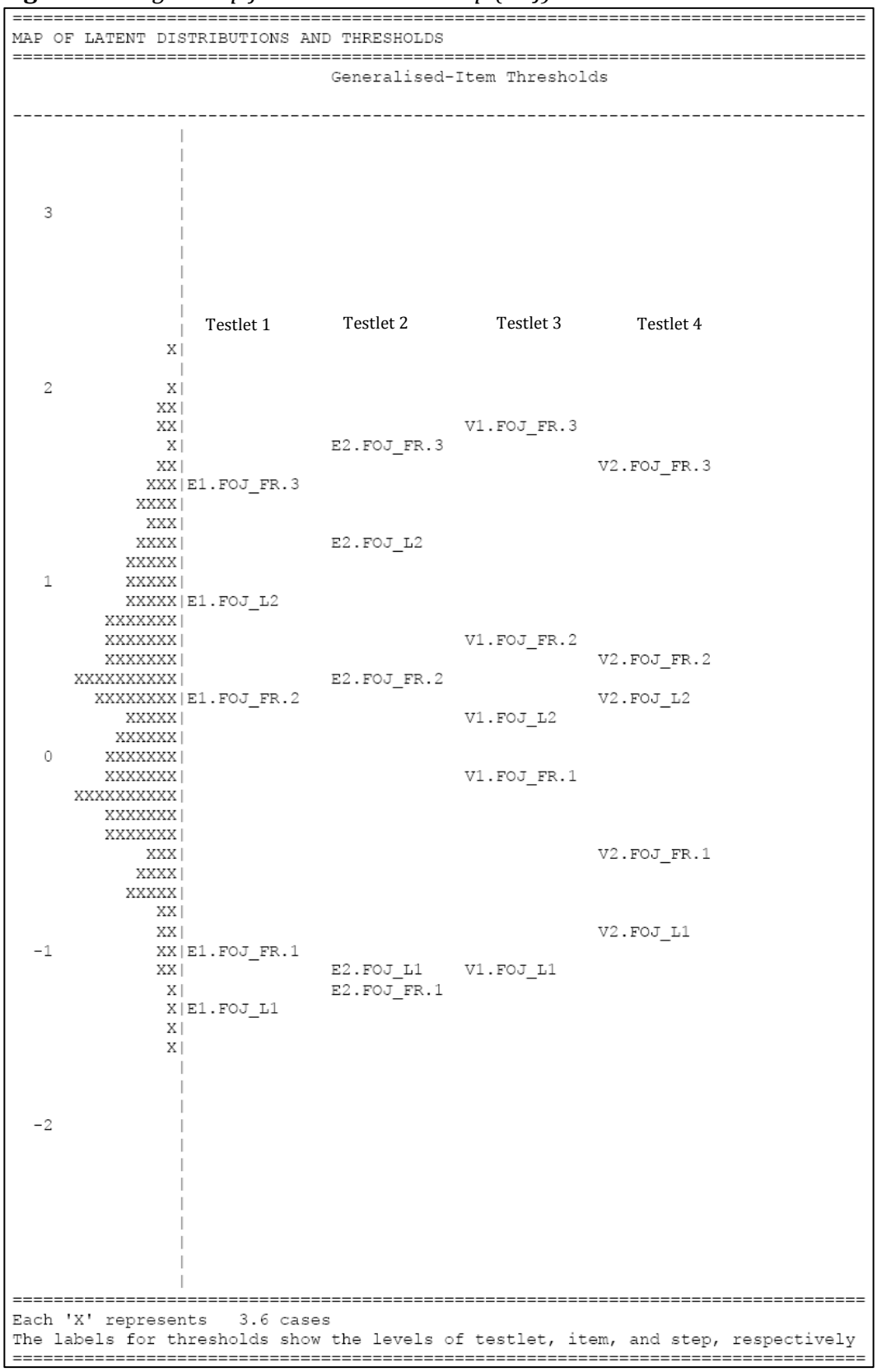
Table 10. *Estimated Item Locations*

			Threshold(s):*			Delta(s):**		
Testlet	Item	Discrim.	1	2	3	1	2	3
E1	FOJ_L1	0.40	-1.41	.	.	-1.41	.	.
E2	FOJ_L1	0.46	-1.19	.	.	-1.19	.	.
V1	FOJ_L1	0.45	-1.19	.	.	-1.19	.	.
V2	FOJ_L1	0.38	-0.89	.	.	-0.89	.	.
E1	FOJ_L2	0.39	0.87	.	.	0.87	.	.
E2	FOJ_L2	0.33	1.23	.	.	1.23	.	.
V1	FOJ_L2	0.44	0.21	.	.	0.21	.	.
V2	FOJ_L2	0.44	0.38	.	.	0.38	.	.
E1	FOJ_FR	0.76	-1.08	0.37	1.53	-0.87	0.48	1.21
E2	FOJ_FR	0.75	-1.25	0.43	1.78	-1.08	0.53	1.52
V1	FOJ_FR	0.78	-0.05	0.70	1.87	0.49	0.45	1.60
V2	FOJ_FR	0.76	-0.50	0.54	1.62	-0.15	0.52	1.29

* Thurstone thresholds represent the location in which the probability of scoring at or above a particular rating is 50% for a participant of equal ability.

**Deltas correspond to the sum of the subtractive coefficients in the multi-faceted Rasch model (i.e., the coefficients within the parentheses), or the location of the intersection points between ICCs of adjacent steps for polytomous items.

Figure 3. *Wright Map for Testlet* Item*Step (FOJ)*



Conclusion. Within each testlet, each level-1 item is easier than the level-2 item, and the highest level of the rubric for the level-3 item is always the most difficult item. The variation across testlets largely comes in the first two levels of the scoring rubric for item 3. Which lead to the results of the multi-faceted Rasch model which, generally indicate inconsistency of scoring characteristics between testlets. As shown by the Wright map and the ICCs, the differences appear primarily between the Volcanoes and Earthquakes content areas (i.e., the testlets within each content area are more similar). The level-1 item appears to be the most consistent among the items in the scale. These results correspond to the initial model comparison results, which indicated significant worsening in model fit when interaction terms with testlet were removed.

One possible reason for the observed variation is the lack of evidence of unidimensionality, as shown through exploratory factor analysis. In particular, the level-2 item appears to be loading onto its own factor, as opposed to factoring with the level-1 and the free-response items. However, based on the satisfactory fit of the Rasch model, and the polychoric Cronbach's α coefficient, there may be a case to suggest that the scale is sufficiently unidimensional.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Briggs, D. C. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, 4(1), 87-100.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M. (2010). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Taylor and Francis Group.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Lang.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, 6(2–3), 74–96.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.

Gadermann, A. M. , Guhn, M. , & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, & Evaluation*, 17(3).

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.

Linacre, J. M. (1989b). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

McNeill, K. L., & Krajcik, J. (2012). *Supporting grade 5–8 students in constructing explanations in science: The claim, evidence and reasoning framework for talk and writing*. Upper Saddle River, NJ: Pearson.

Muthèn, L., & Muthèn, B. (1998-2001). *Mplus user's guide*. Los Angeles, CA: Muthèn & Muthèn.

Norris, S. P., & Phillips, L. M. (1994). Interpreting pragmatic meaning when reading popular reports of science. *Journal of Research in Science Teaching*, 31(9), 947-967.

Osborne, J. F., Erduran, S., & Simon, S. (2004a). Enhancing the quality of argument in school science. *Journal of Research in Science Teaching*, 41(10), 994-1020.

Prometric (2013). *Internal psychometric guidelines for classical test theory*. Retrieved online on May 15, 2013 from https://www.prometric.com/en-us/news-and-resources/reference-materials/Pages/Internal-Psychometric-Guidelines-for-Classical-Test-Theory.aspx.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Sandoval, W. A., & Cam, A. (2011). Elementary children's judgments of the epistemic status of sources of justification. *Science Education*, 95(3), 383-408.

Smith A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Reseach Methodology*, 8(33).

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2, 66-78.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah. NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716–730.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wu, A .D., Li, Z. , & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis. *Practical Assessment Research & Evaluation*, 12(3).

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ConQuest* (Version 2.0) [Computer Software]. Camberwell, Australia: ACER.

Appendix A

Table 11. *Model Fit for Testlets*

			Outfit			Infit		
Testlet	Estim.	Error	MNSQ	CI	*t*	MNSQ	CI	*t*
E1	-0.091	0.072	1.00	(0.83, 1.17)	0.1	1.01	(0.84, 1.16)	0.1
E2	0.119	0.073	0.96	(0.82, 1.18)	-0.4	0.97	(0.83, 1.17)	-0.4
V1	-0.044	0.073	0.90	(0.82, 1.18)	-1.1	0.91	(0.83, 1.17)	-1.1
V2	0.016*	.	0.88	(0.82, 1.18)	-1.4	0.88	(0.83, 1.17)	-1.5

*Constrained parameter estimate.

Table 12. *Model Fit for Items*

			Outfit			Infit		
Item	Estim.	Error	MNSQ	CI	*t*	MNSQ	CI	*t*
FOJ_L2	0.671	0.056	1.25	(0.88, 1.12)	2.6	1.27	(0.90, 1.10)	5.0
FOJ_L1	-1.168	.	1.04	(0.88, 1.12)	0.5	1.04	(0.87, 1.13)	0.6
FOJ_FR	0.497*	.	1.02	(0.88, 1.12)	0.3	1.02	(0.88, 1.12)	0.3

*Constrained parameter estimate.

Table 13. *Model Fit for Items*Testlet*

				Outfit			Infit		
Testlet	Item	Estim.	Error	MNSQ	CI	*t*	MNSQ	CI	*t*
E1	FOJ_L2	0.286	0.095	1.11	(0.83, 1.17)	1.2	1.04	(0.91, 1.09)	0.9
E1	FOJ_L1	-0.151	0.108	0.97	(0.83, 1.17)	-0.4	0.99	(0.79, 1.21)	0.0
E1	FOJ_FR	-0.135*	.	0.98	(0.83, 1.17)	-0.2	0.98	(0.85, 1.15)	-0.3
E2	FOJ_L2	0.438	0.099	1.06	(0.82, 1.18)	0.7	1.06	(0.88, 1.12)	1.0
E2	FOJ_L1	-0.141	0.107	0.90	(0.82, 1.18)	-1.1	0.97	(0.82, 1.18)	-0.3
E2	FOJ_FR	-0.297*	.	0.94	(0.82, 1.18)	-0.6	0.95	(0.84, 1.16)	-0.6
V1	FOJ_L2	-0.419	0.096	1.02	(0.82, 1.18)	0.2	1.01	(0.91, 1.09)	0.3
V1	FOJ_L1	0.027	0.107	0.97	(0.82, 1.18)	-0.2	1.04	(0.82, 1.18)	0.4
V1	FOJ_FR	0.392*	.	0.81	(0.82, 1.18)	-2.2	0.84	(0.84, 1.16)	-2.0
V2	FOJ_L2	-0.304*	.	1.07	(0.82, 1.18)	0.8	1.07	(0.91, 1.09)	1.5
V2	FOJ_L1	0.265*	.	1.03	(0.82, 1.18)	0.4	1.00	(0.85, 1.15)	0.0
V2	FOJ_FR	0.039*	.	0.95	(0.82, 1.18)	-0.5	0.91	(0.84, 1.16)	-1.1

*Constrained parameter estimate.

Table 14. *Model Fit for Items*Testlet*Step*

					Outfit			Infit		
Testlet	Item	Step	Estim.	Error	MNSQ	CI	*t*	MNSQ	CI	*t*
E1	FOJ_FR	0	.	.	0.96	(0.83, 1.17)	-0.4	1.02	(0.78, 1.22)	0.2
E1	FOJ_FR	1	-1.143	0.133	1.02	(0.83, 1.17)	0.2	1.01	(0.91, 1.09)	0.3
E1	FOJ_FR	2	0.205	0.168	1.00	(0.83, 1.17)	0.1	1.01	(0.90, 1.10)	0.3
E1	FOJ_FR	3	0.938*	.	0.76	(0.83, 1.17)	-2.9	0.94	(0.82, 1.18)	-0.6
E2	FOJ_FR	0	.	.	0.97	(0.82, 1.18)	-0.3	0.98	(0.77, 1.23)	-0.2
E2	FOJ_FR	1	-1.404	0.136	0.92	(0.82, 1.18)	-0.8	0.94	(0.93, 1.07)	-1.6
E2	FOJ_FR	2	0.207	0.167	0.91	(0.82, 1.18)	-1.0	0.95	(0.90, 1.10)	-0.9
E2	FOJ_FR	3	1.196*	.	1.01	(0.82, 1.18)	0.1	1.04	(0.77, 1.23)	0.4
V1	FOJ_FR	0	.	.	0.87	(0.82, 1.18)	-1.5	0.91	(0.87, 1.13)	-1.4
V1	FOJ_FR	1	-0.359	0.157	1.03	(0.82, 1.18)	0.4	1.02	(0.85, 1.15)	0.3
V1	FOJ_FR	2	-0.398	0.196	0.82	(0.82, 1.18)	-2.1	0.96	(0.86, 1.14)	-0.5
V1	FOJ_FR	3	0.757*	.	0.51	(0.82, 1.18)	-6.6	0.88	(0.74, 1.26)	-0.9
V2	FOJ_FR	0	.	.	0.91	(0.82, 1.18)	-0.9	0.98	(0.85, 1.15)	-0.3
V2	FOJ_FR	1	-0.707	0.146	1.02	(0.82, 1.18)	0.3	1.02	(0.89, 1.11)	0.3
V2	FOJ_FR	2	-0.034	0.186	1.06	(0.82, 1.18)	0.6	0.97	(0.87, 1.13)	-0.4
V2	FOJ_FR	3	0.741*	.	0.74	(0.82, 1.18)	-3.0	0.97	(0.78, 1.22)	-0.3

*Constrained parameter estimate.

Appendix B

Figure 4. *Item Characteristic Curves for Level-1 Item*

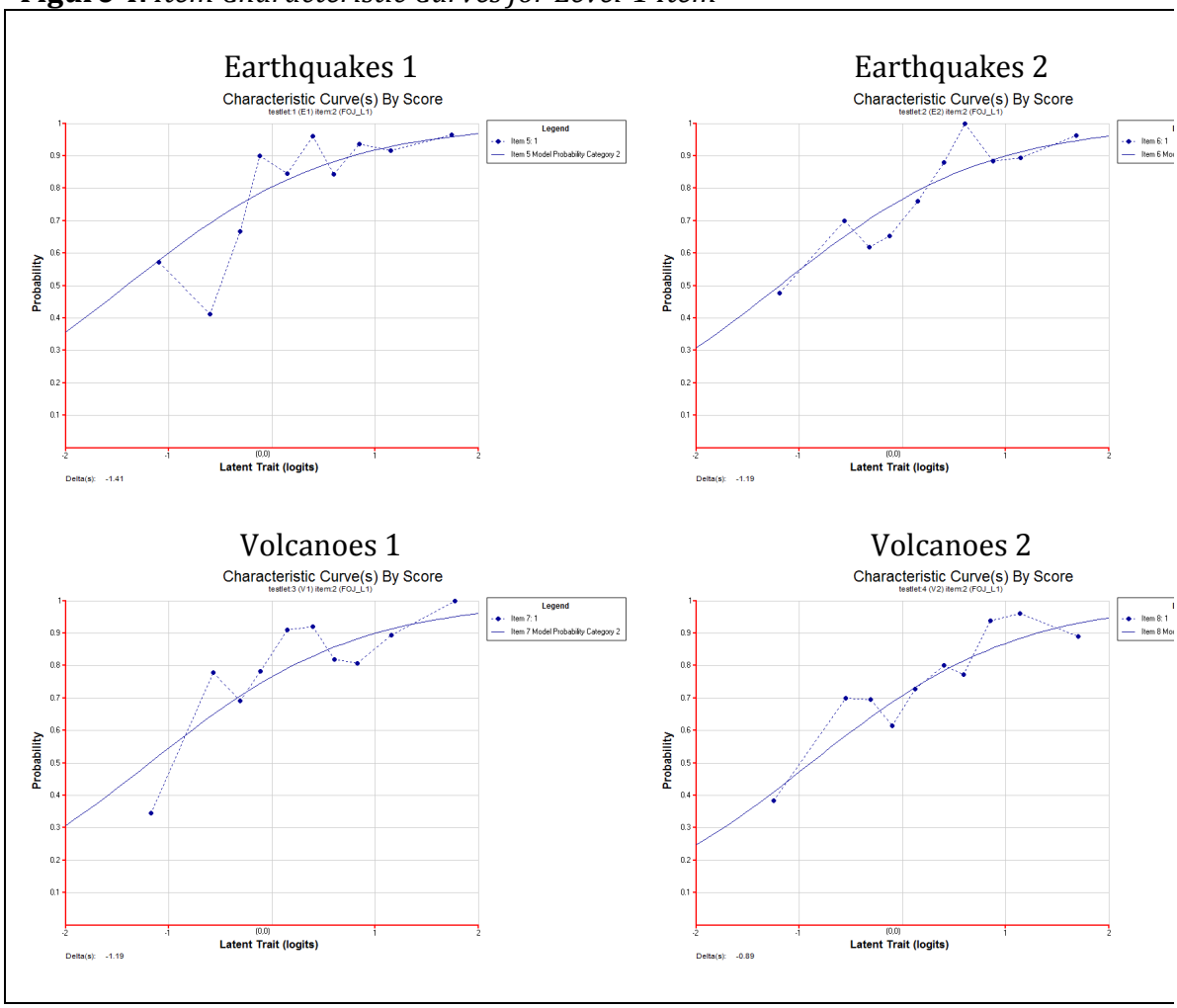


Figure 5. *Item Characteristic Curves for Level-2 Item*

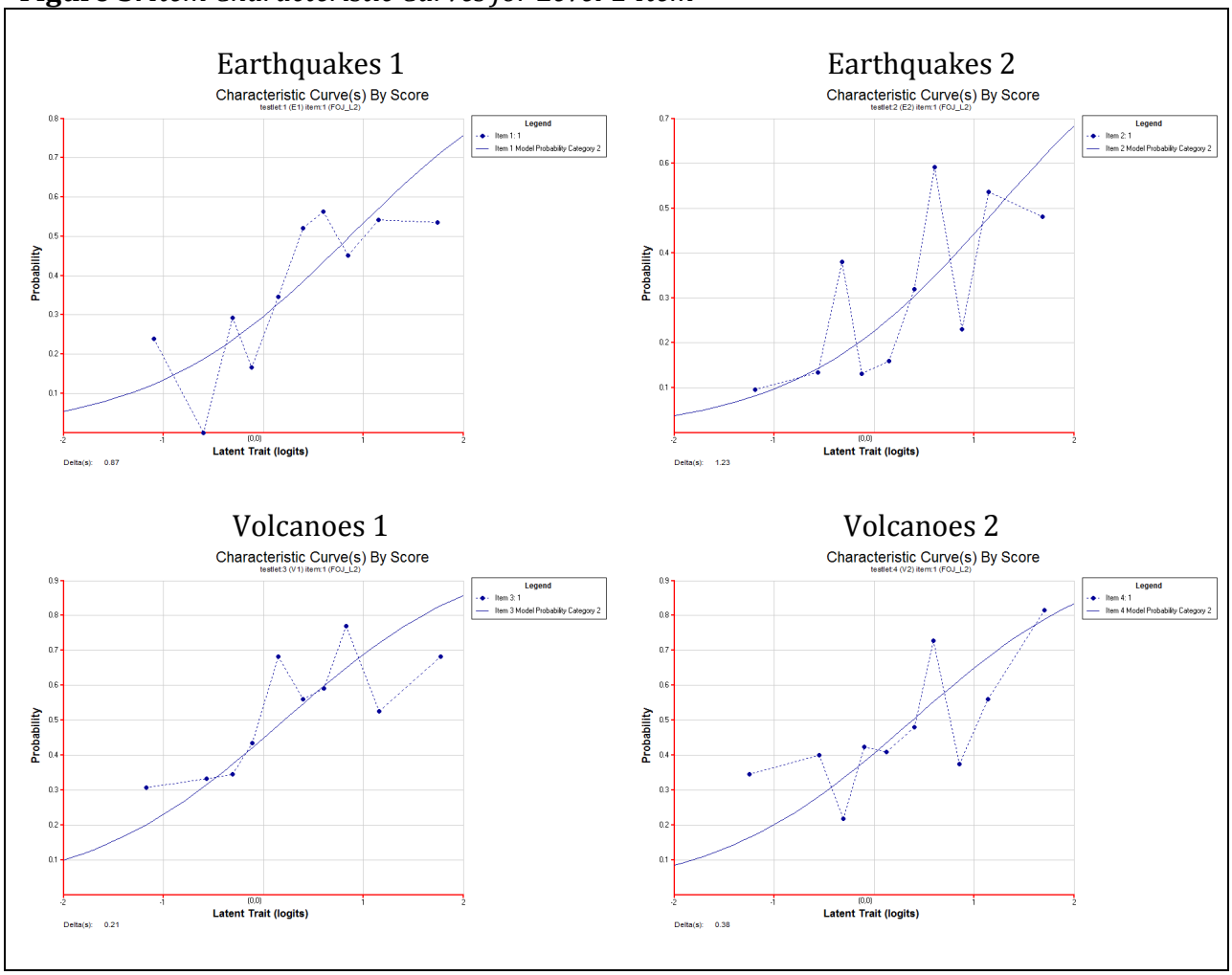
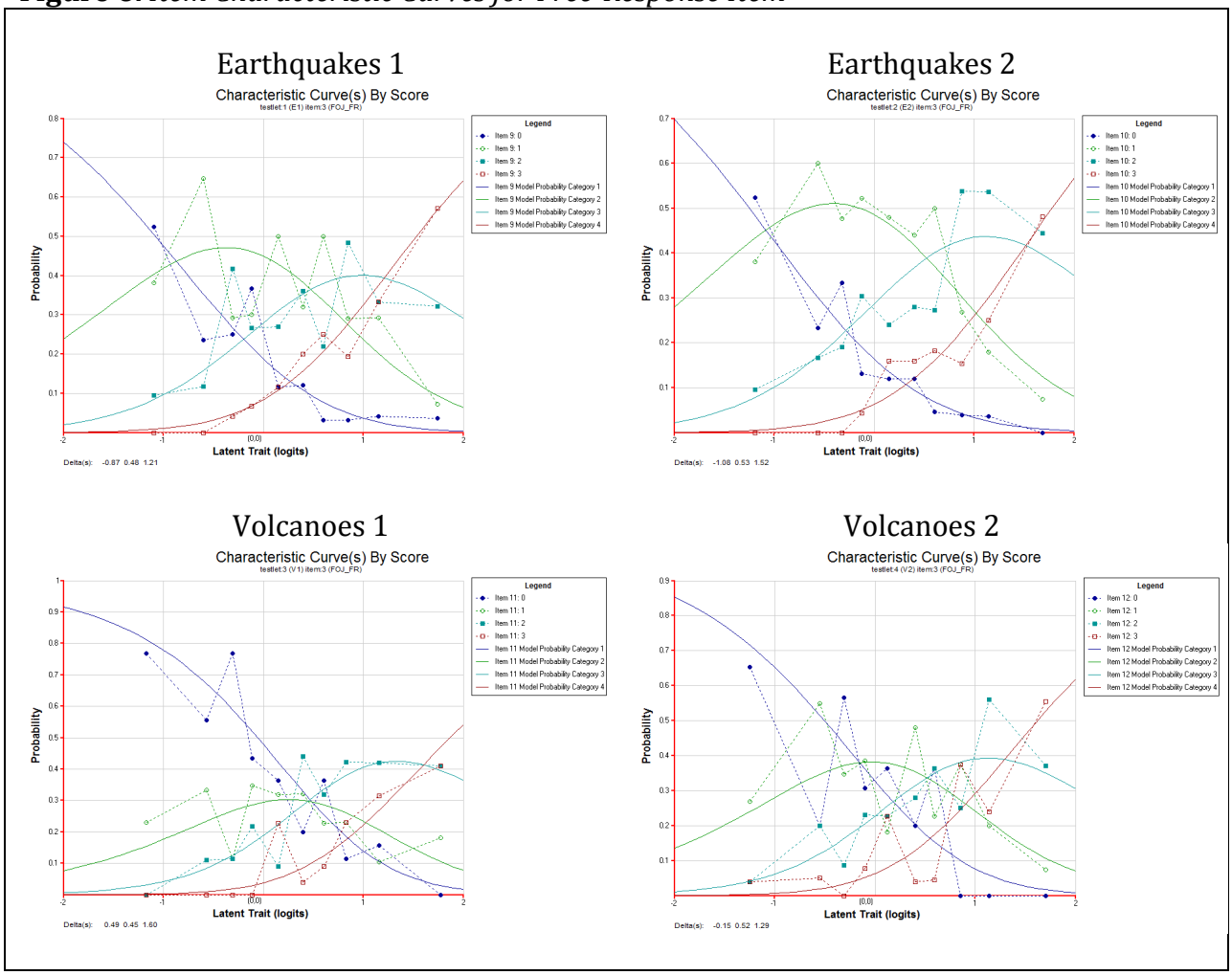


Figure 6. *Item Characteristic Curves for Free-Response Item*



Appendix C

Reading: Forms of Justification

Earthquake 1

Jamal wonders: What is related to the strength of an earthquake? Jamal did a little research and wrote the following argument. Read Jamal's argument closely.

Jamal's Argument:

(S1) = Sentence 1, (S2) = Sentence 2, (S3) = Sentence 3

(S1) Earthquakes that shake the ground for a longer time can be stronger. (S2) I felt an earthquake last year that was so strong I couldn't stand up, and it lasted so long that I thought it was never going to stop shaking! (S3) Therefore, I think that weaker earthquakes can happen when an earthquake only shakes the ground for a short amount of time.

Q1. In which sentence does Jamal support his claim?

- ❍ Sentence 1 (S1) only
- ❍ Sentence 2 (S2) only
- ❍ Sentence 1 (S1) and Sentence 2 (S2)
- ❍ There is no justification in the argument

Q2. Jamal supports his claim with...

- ❍ something an expert said.
- ❍ data from science investigations.
- ❍ comparisons to a personal experience.
- ❍ science ideas he already learned.

Tamara is also in Jamal's class. To help the class write better arguments, Mr. Sanchez put Jamal and Tamara's arguments on the board. He asked the class to compare them:

Jamal's Argument:

Earthquakes that shake the ground for a longer time can be stronger. I felt an earthquake last year that was so strong I couldn't stand up, and it lasted so long that I thought it was never going to stop shaking! Therefore, I think that weaker earthquakes can happen when an earthquake only shakes the ground for a short amount of time.

Tamara's Argument:

Earthquakes that shake the ground for a longer time can be stronger. The Northridge Earthquake shook the ground for 18 seconds and had a strength of 9 at the Earth's surface, and the Coalinga Earthquake shook the ground for 9 seconds and had a strength of 6. The Coalinga Earthquake shook the ground for less time and was weaker.

Q3. Which student, Jamal or Tamara, better supports his or her argument? Why? Please write your answer in the box below.

Reading: Forms of Justification

Earthquake 2

Carrie wonders: What is related to how much energy an earthquake releases? Carrie did a little research and wrote the following argument. Read Carrie's argument closely.

Carrie's Argument:

(S1) = Sentence 1, (S2) = Sentence 2, (S3) = Sentence 3

(S1) Earthquakes that release more energy tend to happen less often. (S2) I have lived in Northern California my whole life and have only felt one earthquake, but the earthquake I felt was really strong and was very scary! (S3) When an earthquake releases a lot of energy, there is usually a long time before the next large earthquake happens at the same location.

Q4. In which sentence does Carrie support her claim?

- ❍ Sentence 1 (S1) only
- ❍ Sentence 2 (S2) only
- ❍ Sentence 1 (S1) and Sentence 2 (S2)
- ❍ There is no justification in the argument

Q5. Carrie supports her claim with...

- ❍ something an expert said.
- ❍ data from science investigations.
- ❍ comparisons to a personal experience.
- ❍ something she learned in science class.

Eric is also in Carrie's class. To help the class write better arguments, Mrs. Johnson put Carrie and Eric's arguments on the board. She asked the class to compare them:

Carrie's Argument:

Earthquakes that release more energy tend to happen less often. I have lived in Northern California my whole life and have only felt one earthquake, but the earthquake I felt was really strong and was very scary! When an earthquake releases a lot of energy, there is usually a long time before the next large earthquake happens at the same location.

Eric's Argument:

Earthquakes that release more energy tend to happen less often. There are about 12,000 earthquakes that release a small amount of energy each year, but only one earthquake each year releases a large amount of energy. When an earthquake releases a lot of energy, there is usually a long time before the next large earthquake happens at the same location.

Q6. Which student, Carrie or Eric, better supports his or her argument? Why? Please write your answer in the box below.

Reading: Forms of Justification

Volcano 1

Cara wonders: What is related to the height of ash clouds that are released from volcanic eruptions? Cara did a little research and wrote the following argument. Read Cara's argument closely.

Cara's Argument:

(S1) = Sentence 1, (S2) = Sentence 2, (S3) = Sentence 3

(S1) Volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. (S2) For instance, the Mount St. Helens Volcano had an explosive power of 5 and the ash cloud was about 22,000 meters high, and the Kilauea Volcano had an explosive power of 0 and its ash cloud is only about 100 meters high. (S3) The Mount St. Helens eruption was more powerful, which made the volcanic ash go much higher into the sky.

Q7. In which sentence does Cara support her claim?

- ❍ Sentence 1 (S1) only
- ❍ Sentence 2 (S2) only
- ❍ Sentence 1 (S1) and Sentence 2 (S2)
- ❍ There is no justification in the argument

Q8. Cara supports her claim with...

- ❍ an expert's opinion.
- ❍ data from science investigations.
- ❍ comparisons to a personal experience.
- ❍ science ideas she already learned.

Jose is also in Carrie's class. To help the class write better arguments, Mrs. Warren put Cara and Jose's arguments on the board. She asked the class to compare them:

Cara's Argument:

Volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. For instance, the Mount St. Helens Volcano had an explosive power of 5 and the ash cloud was about 22,000 meters high, and the Kilauea Volcano had an explosive power of 0 and its ash cloud is only about 100 meters high. The Mount St. Helens eruption was more powerful, which made the volcanic ash go much higher into the sky.

Jose's Argument:

Volcanoes that have more explosive power usually produce ash clouds that reach higher into the sky. On our field trip to Mount St. Helens, a scientist said that when the Mount St. Helens Volcano erupted in 1980, the eruption was so large that the ash cloud blocked the sun for several days! This only happened because the eruption was so powerful that it was able to push the volcanic gases and dust really high into the sky.

Q9. Which student, Cara or Jose, better supports her or his argument? Why?
Please write your answer in the box below.

Reading: Forms of Justification

Volcano 2

Allen wonders: What is related to the thickness of magma? Allen did a little research and wrote the following argument. Read Allen's argument closely

Allen's Argument:

(S1) = Sentence 1, (S2) = Sentence 2, (S3) = Sentence 3

(S1) Volcanoes that release very hot magma often have the thinnest magma. (S2) The Fernandina Volcano in the Galapagos Islands has magma that reaches temperatures of more that 1,000 °C, and it also has some of the thinnest magma ever discovered. (S3) The Fernandina Volcano is a good example because it has high temperature magma that is also very runny.

Q10. In which sentence does Allen support his claim?

- ❍ Sentence 1 (S1) only
- ❍ Sentence 2 (S2) only
- ❍ Sentence 1 (S1) and Sentence 2 (S2)
- ❍ There is no justification in the argument

Q11. Allen supports his claim with...

- ❍ something an expert said.
- ❍ data from science investigations.
- ❍ comparisons to a personal experience.
- ❍ Something he learned in science class.

Deshawn is also in Allen's class. To help the class write better arguments, Mr. Tao put Allen and Deshawn's arguments on the board. He asked the class to compare them:

Allen's Argument:

Volcanoes that release very hot magma often have the thinnest magma. The Fernandina Volcano in the Galapagos Islands has magma that reaches temperatures of more that 1,000 °C, and it also has some of the thinnest magma ever discovered. The Fernandina Volcano is a good example because it has high temperature magma that is also very runny.

Deshawn's Argument:

Volcanoes that release very hot magma often have runny magma. When I visited a gift store in Ecuador they sold lava rocks that had once been magma that flowed all the way to the ocean and cooled. To flow all the way to the ocean, the magma must have been very thin and runny, and I know that Ecuador's volcanoes have some of the hottest magma.

Q12. Which student, Allen or Deshawn, better supports his or her argument? Why? Please write your answer in the box below.