CLASSICAL AND IRT ANALYSIS OF:

FORMS OF JUSTIFICATION

– Carnegie Spring 2013 –

Tables

Figures

Sample Description

Demographic information of students' age, gender, ethnicity, and if he/she receives English Language Development, Free Lunch, or Individualized Educational Plan (Special Education) is available for the participating students.

Participated in this study 282 students, 125 of them were from grade 6, 34 from grade 7, and 121 from grade 8. The average age of these students was 12.7 years and 52.5 percent were male. White students constituted the majority of the sample (61 percent), followed by Hispanic (24 percent). Information about other ethnic groups is shown in Table 1. 43 students (15 percent) received English Language Development. Teachers did pointed 23 students as receiving Free Lunch and 28 Reduced Lunch, 92 students were classified as not receiving, and for 139 students the information is not available (N/A). Out of the 282 participating students, 21 participated in Individualized Educational Plan. Table 1 shows the number of students by ethnic origin.

Table 1. Number of students by ethnicity

Ethnicity	N. of students	% of students
White	172	61.0%
Hispanic	68	24.1%
Black/African American	12	4.3%
Multiple	19	6.7%
Asian	4	1.4%
Native Hawaiian	1	0.4%
Other	9	3.1%
Grand Total	**282**	**100%**

Table 2. Number of students by Grade and Test taken

	Reading: Forms of Justification	
	Test 1	Test 2
Grade 6	125	-
Grade 7	18	16
Grade 8	-	121
N/A	-	2
Grand Total	**143**	**139**

RESULTS

Factor Analysis

On Table 3 we report results for investigating the dimensionality of the multiple-choice items under an exploratory approach called the full information factor analyses (FIFA). The main advantage of this method is that it uses all available information in the estimation procedure (i.e., analyzes all item response patterns instead of analyzing each item separately). This approach was implemented using the Testfact software, a sophisticated program for item analysis using both classical and modern psychometric IRT methods. This software is based on the full-information item factor analysis proposed by Bock, Gibbons, and Muraki (1988).

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 (Edwards & Wirth, 2009), meaning student scores on these items are seen as indicators of overall student ability on the intended construct being measured. To determine whether an item is a component of a specific factor, a cutoff value of .3 (Lambert & Durand, 1975) is recommended as an acceptable minimum value for the coefficients.

Table 3. Factor loadings for two alternative models: 1-Factor versus 3-Factor for the MC items

Level	Item	One Factor	Three Factors*		
			Factor 1	Factor 2	Factor 3
L1	RJ1.01.L1	0.20	0.12	0.10	0.18
L2	RJ1.02.L2	0.50	0.31	0.32	0.27
L3	RJ1.03.L3	0.14	0.46	-0.01	-0.12
L1	RJ1.05.L1	0.61	0.05	0.93	0.14
L2	RJ1.06.L2	0.39	0.06	0.43	0.00
L3	RJ1.07.L3	0.26	0.34	0.02	0.17
L1	RJ1.09.L1	0.51	0.60	0.06	0.18
L2	RJ1.10.L2	0.34	0.31	0.25	-0.11
L3	RJ1.11.L3	0.38	0.08	0.43	0.11
L1	RJ1.13.L1	0.48	-0.01	0.40	0.92
L2	RJ1.14.L2	0.06	-0.02	-0.05	0.22
L3	RJ1.15.L3	0.08	0.00	0.13	-0.07
L1	RJ2.01.L1	0.73	0.80	0.07	0.19
L2	RJ2.02.L2	0.44	0.47	0.12	-0.06
L3	RJ2.03.L3	0.30	-0.06	0.72	-0.08
L1	RJ2.09.L1	0.36	0.31	0.01	0.95
L2	RJ2.10.L2	0.49	0.17	0.56	0.25
L3	RJ2.11.L3	0.50	0.15	0.61	0.08
L1	RJ2.13.L1	0.83	0.76	0.20	0.39
L2	RJ2.14.L2	0.46	0.30	0.33	0.08
L3	RJ2.15.L3	0.52	0.29	0.45	0.11
Percent of variance		20.41%	22.3%	9.5%	6.9%

* Note: Convergence not attained (300 iterations); Factor loadings may be incorrect.

For the one-factor trial run, the average factor loading was 0.41, and five items have loadings lower than 0.3. This five lowest loadings are for Justification Test 1 (RJ1). Factor loadings lower than 0.3 suggest that more than 90% of the variance in an observed variable is explained by factors other than the construct to which the variable should be theoretically related. The alternative model with 3 factors was obtained using Varimax rotation and 10 items loaded higher than 0.30 on the first factor, 10 items loaded in the second and three items on the third factor, suggesting that the data is best explained by the unidimensional model. Important to note that convergence was not obtained in this 3-factor model, hence, values of the factor loadings may be incorrect. Results from the 1-factor model are preferred instead.

Classical Test Theory

The foundation for the Classical Test Theory 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, is used as the index for the item difficulty. Hence, the item difficulty for item *j* is defined as the number of examinees with a score of 1 on item *j* divided by the total number of examinees. This statistics is, actually, an inverse indicator of item difficulty, as the higher the value of item difficulty, the easier the item. To ascertain item difficulty under the Classical Test Theory approach, the mean number of correct answers for each item was calculated. At the extremes, if all students scored correct on an item, the resulting difficulty for that item would be $x = 1$, and it would be deemed very easy. Conversely, if all students scored incorrectly on an item, it would have a difficulty of $x = 0$ and it would be deemed exceedingly difficult. A general guideline by which items are classified according to their difficulty levels is presented in Table 4.

Table 4. Classification rules for item difficulty levels

P-value (easy to hard)	Item Interpretation
100 to 96%	Inappropriately easy (unacceptable) items
90 to 95%	Very easy (possibly unacceptable) items
89 to 80%	Fairly easy (acceptable) items
79 to 40%	Hard to moderately easy (acceptable) items
39 to 30%	Difficult (acceptable) items
29 to 20%	Very difficult (possibly unacceptable) items
19 to 0%	Inappropriately difficult (unacceptable) items

* Note: Adapted from Prometric (2013)

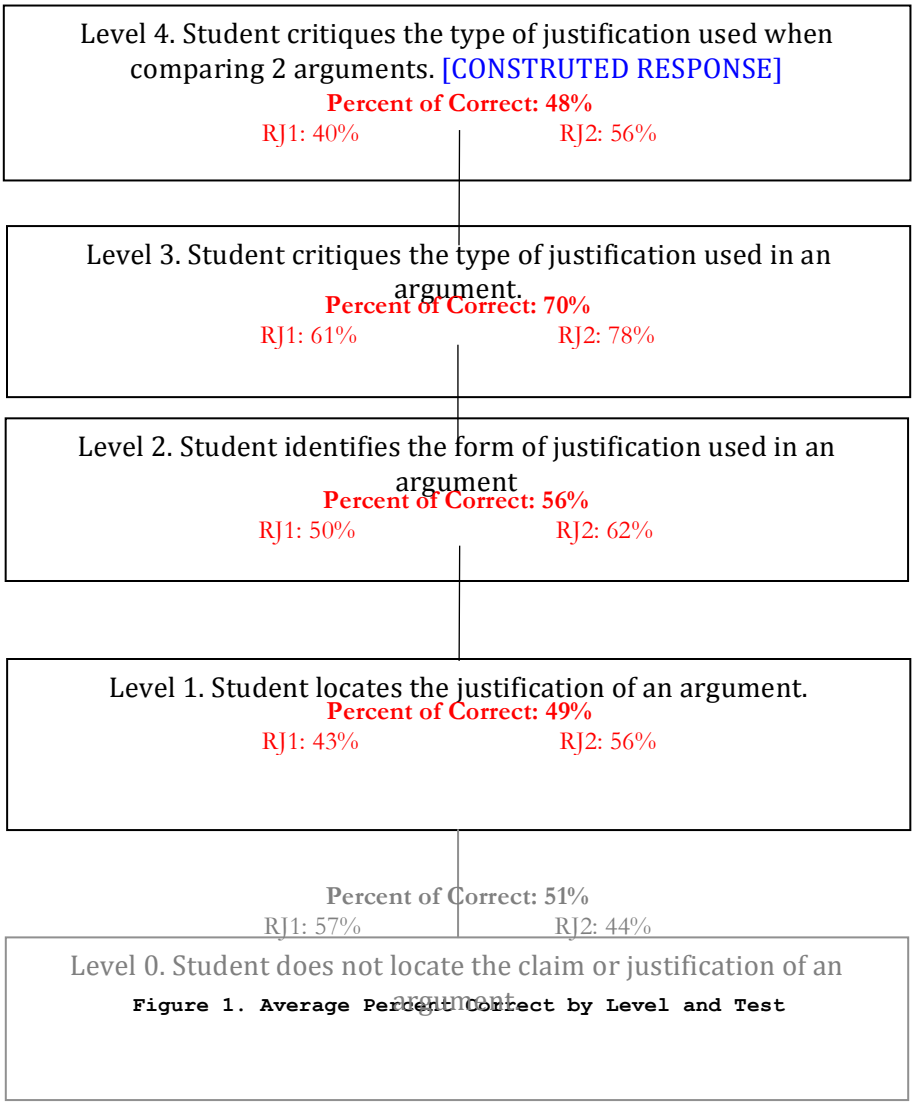


Figure 1. Average Percent Correct by Level and Test

Figure 1 shows the difficulty estimates of each level of the Reading: Forms of Justification. Results are also shown by test type (1 or 2).

Figure 2 shows the average percent of correct as function of the Level.

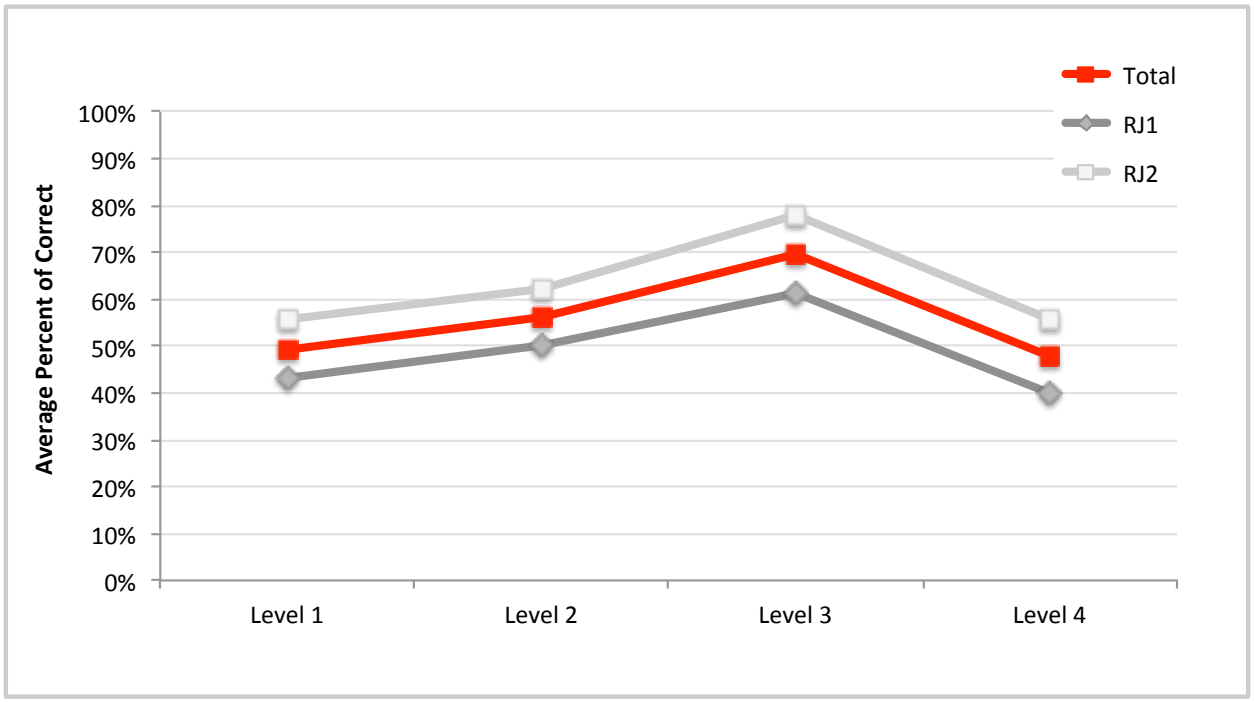


Figure 2. Average Percent Correct by Level

The total score distribution for this progress variable is presented in Figure 5.

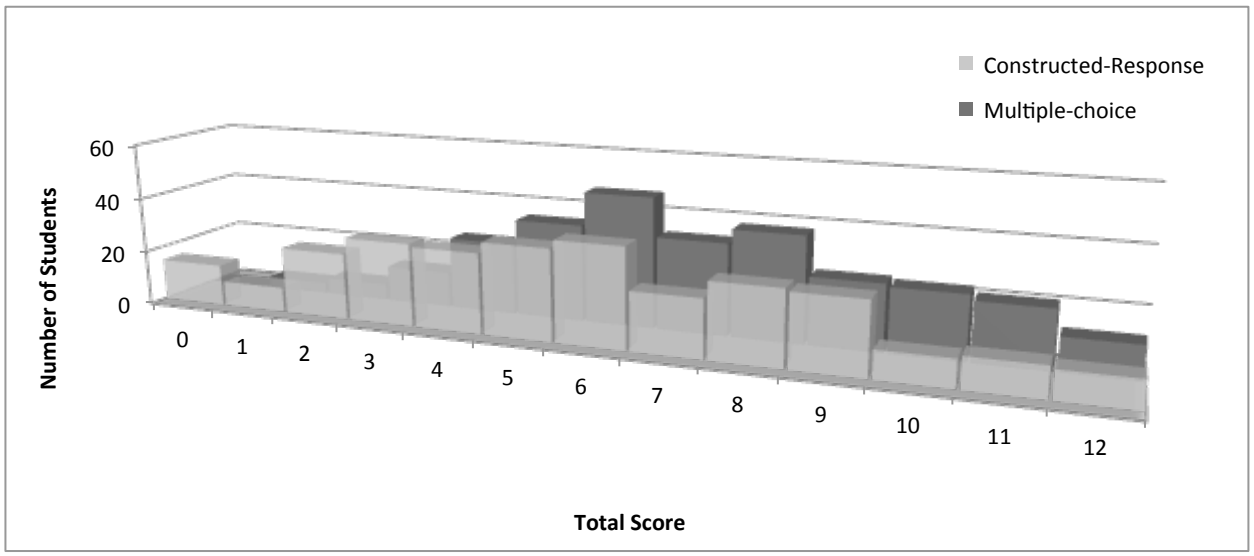


Figure 5. Total Score Distribution

Item Response Theory

In this section we present the IRT difficult estimates and the fit results on the *Reading: Forms of Justification*, Spring 2013 pilot.

The item difficulties were estimated through application of the Rasch IRT model in ConQuest. In this model, the probability of a correct response is modeled as a logistic function of the difference between the person and item parameter. When a person's ability on the latent trait is equal to the difficulty of the item, there is by definition a 0.5 probability of a correct response in the Rasch model. The point on the ability scale at which the 0.5 probability of a correct response occurs is considered the difficulty parameter. The theoretical range of the values of this parameter is $-\infty <= b <= +\infty$. However, typical values have the range $-3 <= b <= +3$, when the ability is scaled to have a mean of 0 and standard deviation of 1.

ConQuest also provides two residual statistics (MNSQ) that serves as an indication of the item fit in the following manner: weighted (infit) and unweighted (outfit). The outfit is sensitive to unexpected observations by examinees on items that are either very easy or very hard for examinees, while the infit statistic is sensitive to unexpected responses to items targeted on them – *i.e.* are close to their ability level. In this analysis, we will consider both the infit and outfit statistics. Item MNSQ value of 1 is the expected values. Values greater than 1.0 indicate unpredictability or a lack of construct homogeneity in relationship to the other items in a scale (Green, 1996). Different mean-square ranges encountered in practice have been reported. Smith, Schumacker, and Bush (1998) suggest that items be considered underfitting when the unweighted and weighted mean square values are greater than 1.3 for samples less than 500, 1.2 for samples between 500 and 1000, and 1.1 for samples larger than 1000; we will apply the cut-off of 1.3 for interpreting the results. Values less than 0.75 are considered significantly overfitting (Bond and Fox 2007) which suggests these items are redundant with other items and are not providing unique information about the construct. Typically, the *t value* indicates if the MNSQ infit and outfit statistics are significant, but Wu and Adams (2007) demonstrated that as sample size increases, the fit *t* values become progressively larger, increasing the number of items showing misfit. These authors claim that if one uses *t* values as a criterion for accepting or rejecting items on the basis of fit, one is likely to erroneously declare that most items do not fit well when the sample size is large enough. For this reason, only the unweighted and weighted fits are discussed in this report. Table 5 summarizes the item statistics that were obtained from ConQuest.

Table 5. IRT Difficulty Estimates and Fit Statistics

Level	Item	Estimate	Error	Unweighted Fit	T	Weighted Fit	T
L3	RJ2.11.L3.C	-1.67	0.10	0.74	-2.40	0.87	-0.80
L2	RJ2.10.L2.C	-1.26	0.09	0.86	-1.10	0.94	-0.40
L3	RJ1.07.L3.C	-1.25	0.09	1.00	0.10	1.00	0.00
L3	RJ2.03.L3.C	-1.06	0.09	1.01	0.10	0.99	0.00
L3	RJ1.11.L3.C	-0.86	0.08	1.00	0.10	1.01	0.20
L3	RJ2.15.L3.C	-0.40	0.09	0.95	-0.30	0.96	-0.50
L2	RJ1.06.L2.C	-0.15	0.09	1.06	0.50	1.05	1.00
L2	RJ2.02.L2.B	-0.11	0.09	0.94	-0.40	0.97	-0.40
L1	RJ2.01.L1.B	-0.07	0.09	0.84	-1.30	0.88	-1.80
L2	RJ1.14.L2.B	-0.06	0.09	1.09	0.80	1.08	1.60
L2	RJ1.10.L2.B	-0.05	0.08	1.04	0.50	1.04	0.90
L4	RJ2.12.L4	-0.01	0.07	0.95	-0.40	0.94	-0.50
L3	RJ1.15.L3.C	0.10	0.09	1.09	0.80	1.07	1.40
L1	RJ1.09.L1.B	0.18	0.08	0.94	-0.70	0.95	-1.20
L4	RJ2.16.L4	0.18	0.44	0.91	-0.70	0.93	-0.70
L4	RJ2.04.L4	0.19	0.07	1.14	1.20	1.12	1.10
L3	RJ1.03.L3.C	0.25	0.09	1.02	0.20	1.02	0.50
L1	RJ2.09.L1.B	0.26	0.09	1.06	0.50	1.05	0.70
L1	RJ1.13.L1.B	0.31	0.09	1.09	0.80	1.06	1.20
L1	RJ1.01.L1.B	0.34	0.09	1.17	1.40	1.14	2.60
L4	RJ1.04.L4	0.40	0.07	0.89	-0.90	0.90	-1.00
L1	RJ2.13.L1.B	0.43	0.09	0.83	-1.40	0.86	-2.50
L4	RJ1.08.L4	0.59	0.07	0.98	-0.10	0.99	0.00
L2	RJ2.14.L2.A	0.63	0.09	1.00	0.00	0.99	-0.20
L2	RJ1.02.L2.A	0.72	0.09	1.05	0.40	1.03	0.50
L4	RJ1.16.L4	0.72	0.07	0.99	0.00	0.99	-0.10
L1	RJ1.05.L1.B	0.79	0.09	1.07	0.60	1.06	0.80
L4	RJ1.12.L4	0.86	0.06	1.13	1.50	1.10	1.40

The difficulty of the items on this progress variable ranges from -1.67 to 0.86. From the six easiest items on this assessment, five measure Level 3 on the progress variable. Item RJ2.10.L2.C, a Level 2 item, was also easier than expected, difficulty estimate of -1.26. This result is not expected as Level 3, in theory, was supposed to be the most complex level on this progress variable (as seen on Figure 1) and the items should increase in difficult in the following order: Level 1, Level 2, and Level 3. The results of this table also display that no item shows a misfit, where no value of the weighted and unweighted fit MNSQ are greater than the critical value of 1.3. One item (RJ2.11.L3.C) had an Unweighted Fit smaller than 0.75.

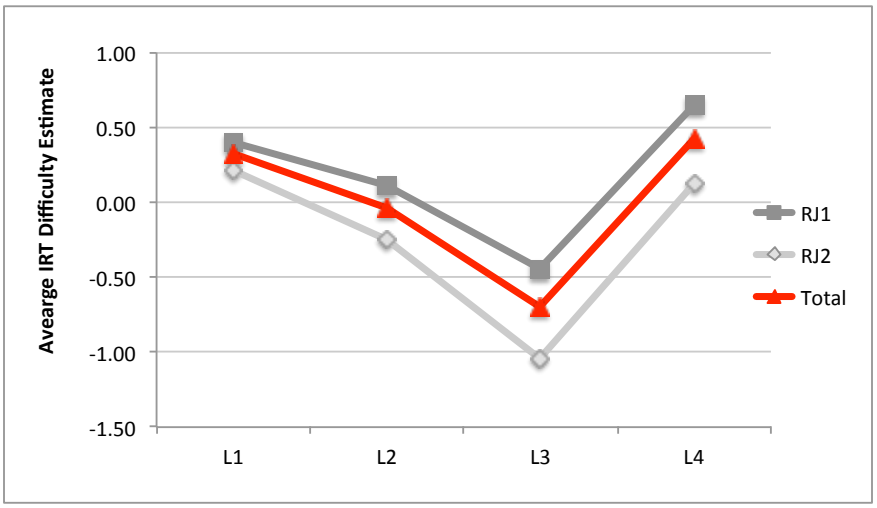


Figure 12. Average difficulty estimate by level

The EAP/PV test reliability estimate for the sample of 282 students was .75. As a general rule of thumb, a test reliability estimate of .70 is a minimum acceptable for performance of an instrument.

The next two figures provide a Rasch IRT Wright map of the students' ability and item difficulty parameters on each of the progress variables. The left-handed side of this figure represents the distribution of the measured ability of the students. The most proficient students register at the top of the graph and the least proficient at the bottom. Each "x" represents a given number of students, for example, each "x" represents 1.7 cases on Figures 13.

One strength of the Wright map is the potential to identify gaps between items along the measured construct continuum. In addition, the Wright map provides a means of evaluating the breadth of the assessment by indicating if the range of item difficulties covers the estimated students' ability. Ideally, the Wright map can be used to a) select items that measure different ranges of the continuum, b) remove items where there is redundant content coverage, as well as c) identify uncovered ranges where new items must be created (Chang & Reeve, 2005).

The first Wright Map (Figure 13) shows the latent ability distribution and item difficulty estimates for the 28 items on Forms of Justification, as a whole. This progress variable does not describe a set of skills ordered from Level 1 to Level 4. For example, Level 3 do not follow the hypothesized structure of the progress variable, being much easier than we expect it to be. The distribution of items indicates some overlapping items at points along the underlying variable (for example, there are four items at some points around 0.0) and some uncovered points such as above 0.8.

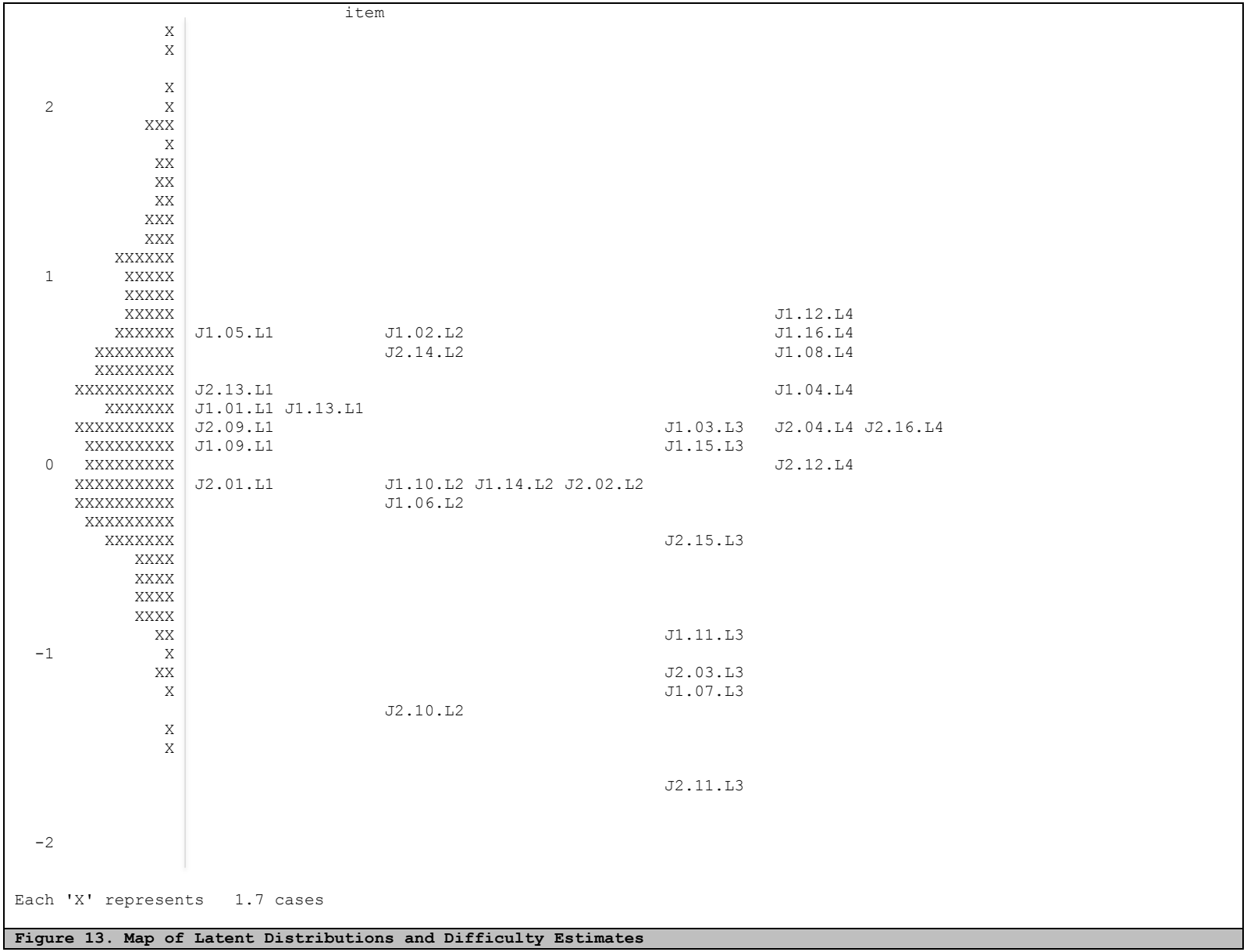


Figure 13. Map of Latent Distributions and Difficulty Estimates

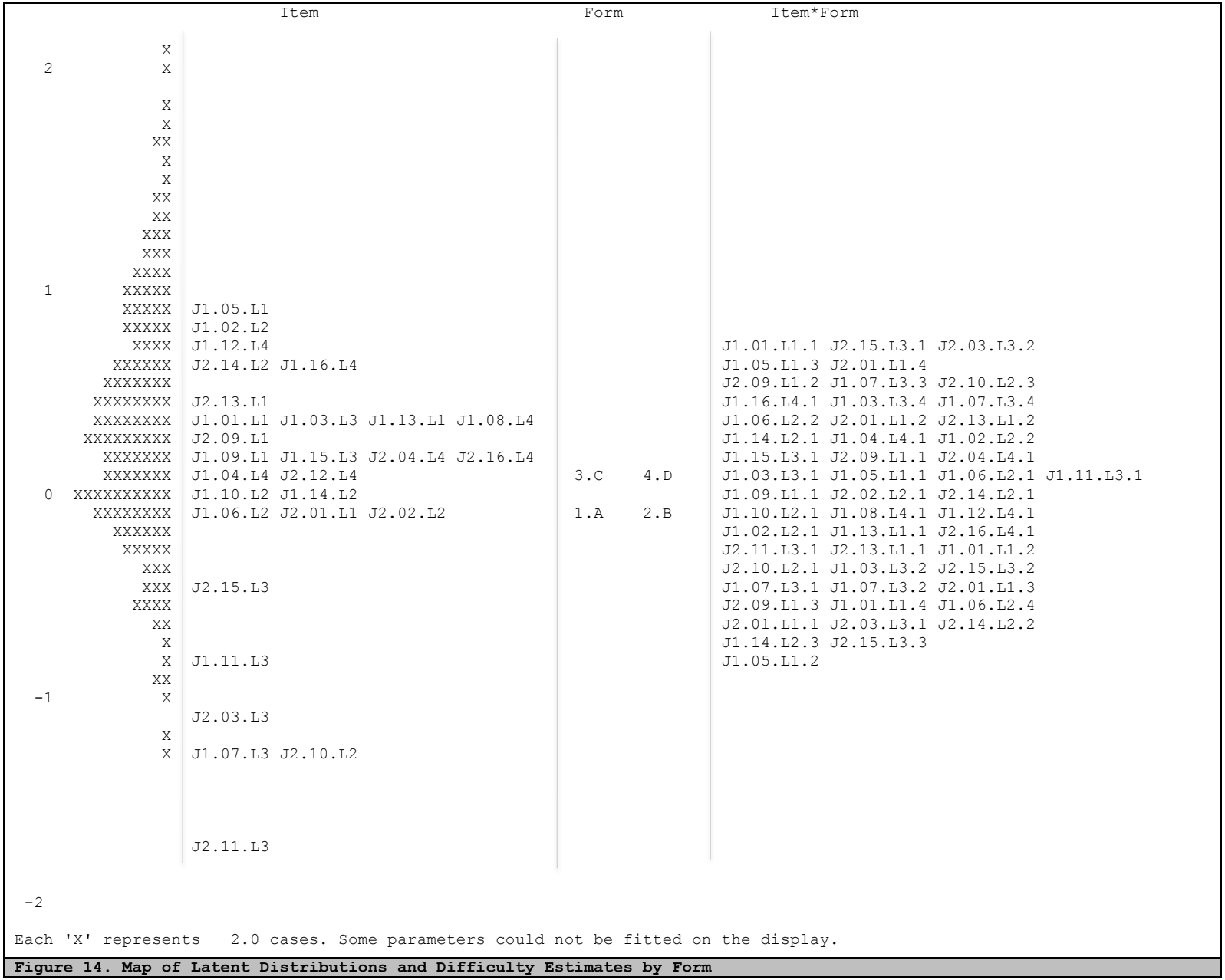


Figure 14. Map of Latent Distributions and Difficulty Estimates by Form

Jamal wonders why some earthquakes have more destructive power than others? Jamal did a little research and wrote the following argument:

Jamal's Argument:

(S1) Earthquakes that shake the ground for a longer time are more destructive. (S2) I live in California and I have felt two different earthquakes, and the one that lasted longer was a whole lot more destructive than the earthquake that only lasted a few seconds. (S3) Therefore, less destruction happens when an earthquakes only shakes the ground for a short amount of time.

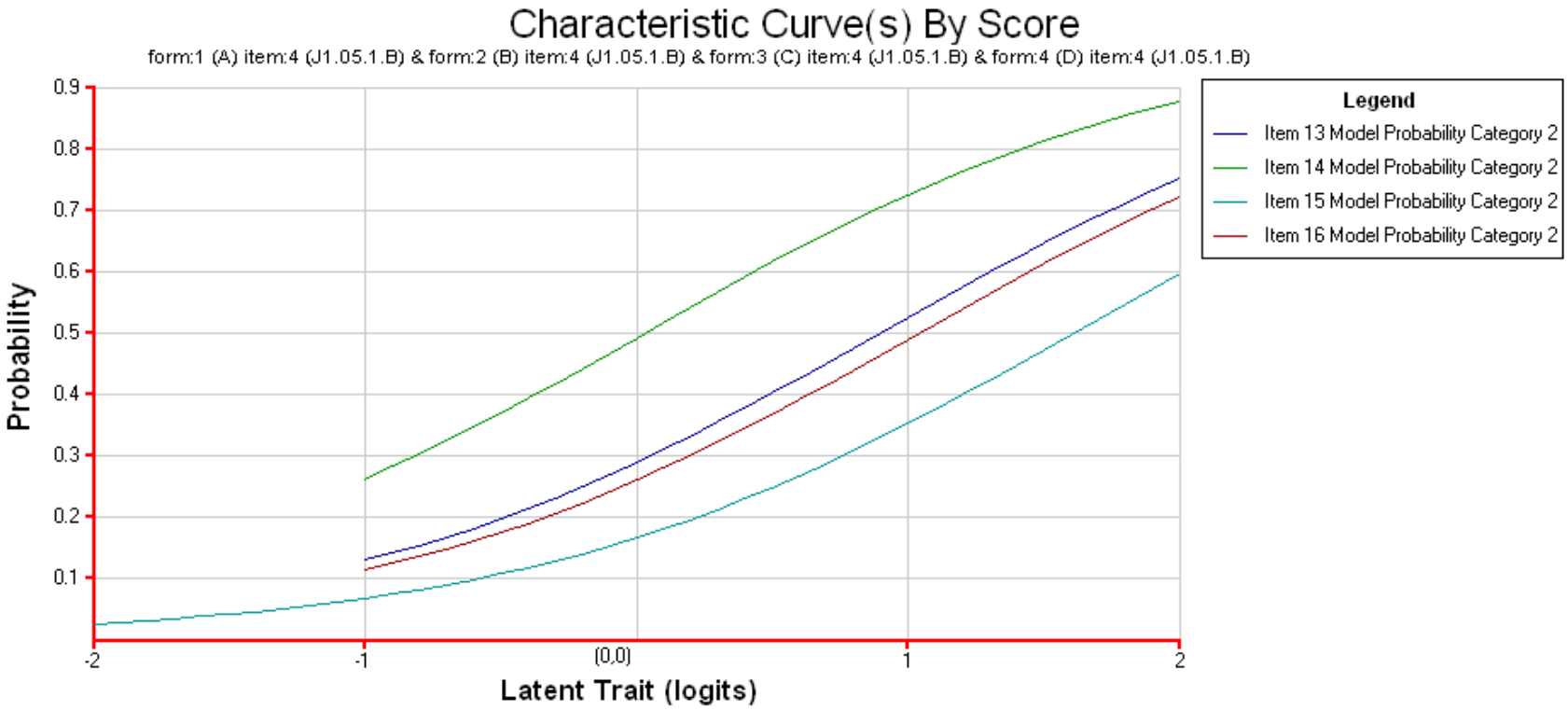
5. Read Jamal's argument closely. In which sentence does Jamal support his claim?

a. Sentence 1 (S1) only
b. Sentence 2 (S2) only
c. Sentence 1 (S1) and Sentence 2 (S2)
d. None

Question Name	Level	Key	N	Pbis
RJ1.05.L1.B	L1	B	143	0.24

% Correct	% Form A	% Form B	% Form C	% Form D
36%	36%	54%	19%	32%

IRT difficulty estimate	Unweighted Fit	T	Weighted Fit	T
0.79	1.07	0.60	1.06	0.80


Characteristic Curve(s) By Score
form:1 (A) item:4 (J1.05.1.B) & form:2 (B) item:4 (J1.05.1.B) & form:3 (C) item:4 (J1.05.1.B) & form:4 (D) item:4 (J1.05.1.B)
Probability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-2
-1
(0,0)
1
2
Latent Trait (logits)
Legend
Item 13 Model Probability Category 2
Item 14 Model Probability Category 2
Item 15 Model Probability Category 2
Item 16 Model Probability Category 2

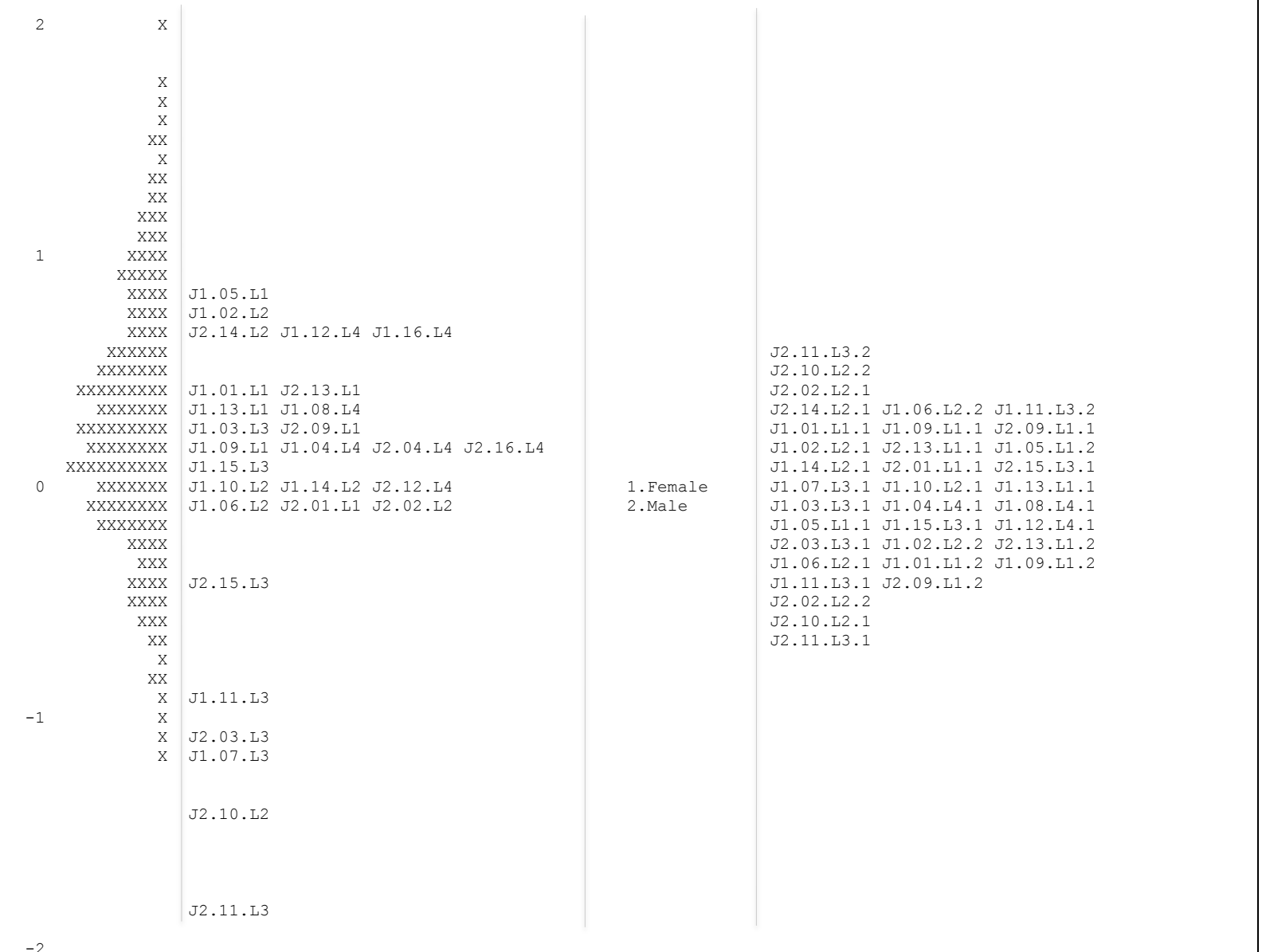


Figure 15. Map of Latent Distributions and Difficulty Estimates by Gender

Carrie wonders what makes some earthquakes release more energy than others? Carrie did a little research and wrote the following argument:

Carrie's Argument:

(S1) Earthquakes that release more energy happen less often. (S2) I know this because I have felt both an earthquake and an aftershock, and the aftershock was much less powerful than the original earthquake. (S3) This means that when an earthquake does not happen for a while it releases more energy.

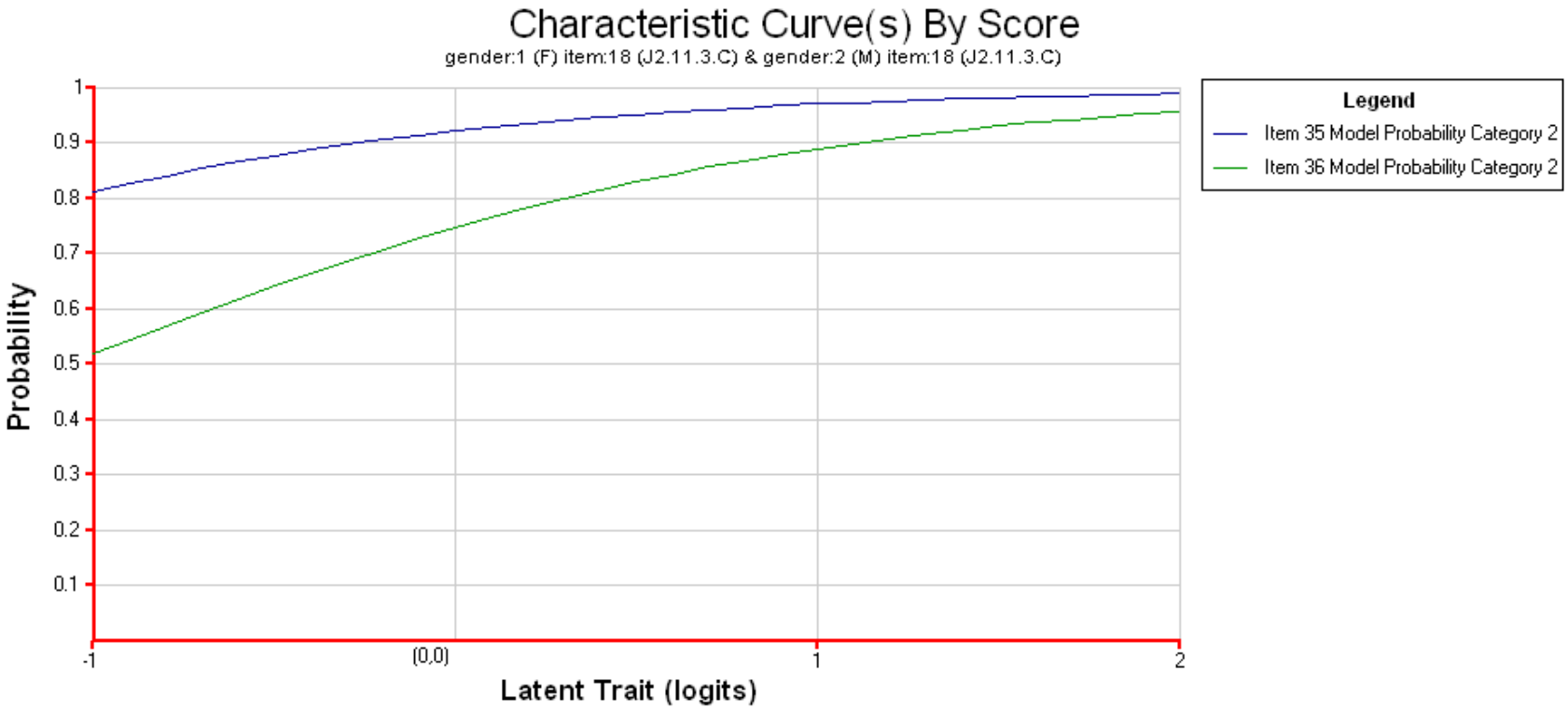
11. Carrie's argument would be stronger if she relied more on

a. a story a friend told her.
b. something a classmate said.
c. measurements from investigations.
d. a personal story about the claim.

Question Name	Level	Key	N	Pbis
RJ2.11.L3.C	L3	C	139	0.48

% Correct	% Form A	% Form B	% Form C	% Form D
86%	89%	86%	89%	82%

IRT difficulty estimate	Unweighted Fit	T	Weighted Fit	T
-1.67	0.74	-2.40	0.87	-0.80


Characteristic Curve(s) By Score
gender:1 (F) item:18 (J2.11.3.C) & gender:2 (M) item:18 (J2.11.3.C)
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-1
(0,0)
1
2
Latent Trait (logits)
Legend
Item 35 Model Probability Category 2
Item 36 Model Probability Category 2

The dimensions are:

1 "Forms of Justification"
2 "Relevant-Supporting Evidence"
3 "Sufficient Evidence"
4 "Multiple Claims"
5 "Reasoning"

Table. Difficulty Estimates for each Dimension

VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
dimension	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1 Forms of Ju..	-1.307	0.026	1.06	(0.86, 1.14)	0.9	1.00	(0.85, 1.15)	0.0
2 Relevant-Su..	0.166	0.033	0.77	(0.68, 1.32)	-1.5	0.80	(0.71, 1.29)	-1.4
3 Sufficient ..	0.055	0.033	0.79	(0.80, 1.20)	-2.2	0.82	(0.86, 1.14)	-2.8
4 Multiple Cl..	1.179	0.030	0.97	(0.86, 1.14)	-0.4	0.81	(0.70, 1.30)	-1.3
5 Reasoning	-0.093*	0.061	1.00	(0.86, 1.14)	-0.0	1.04	(0.85, 1.15)	0.5

An asterisk next to a parameter estimate indicates that it is constrained

Separation Reliability = 0.999
Chi-square test of parameter equality = 4096.44, df = 4, Sig Level = 0.000

TERM 2: dimension*step

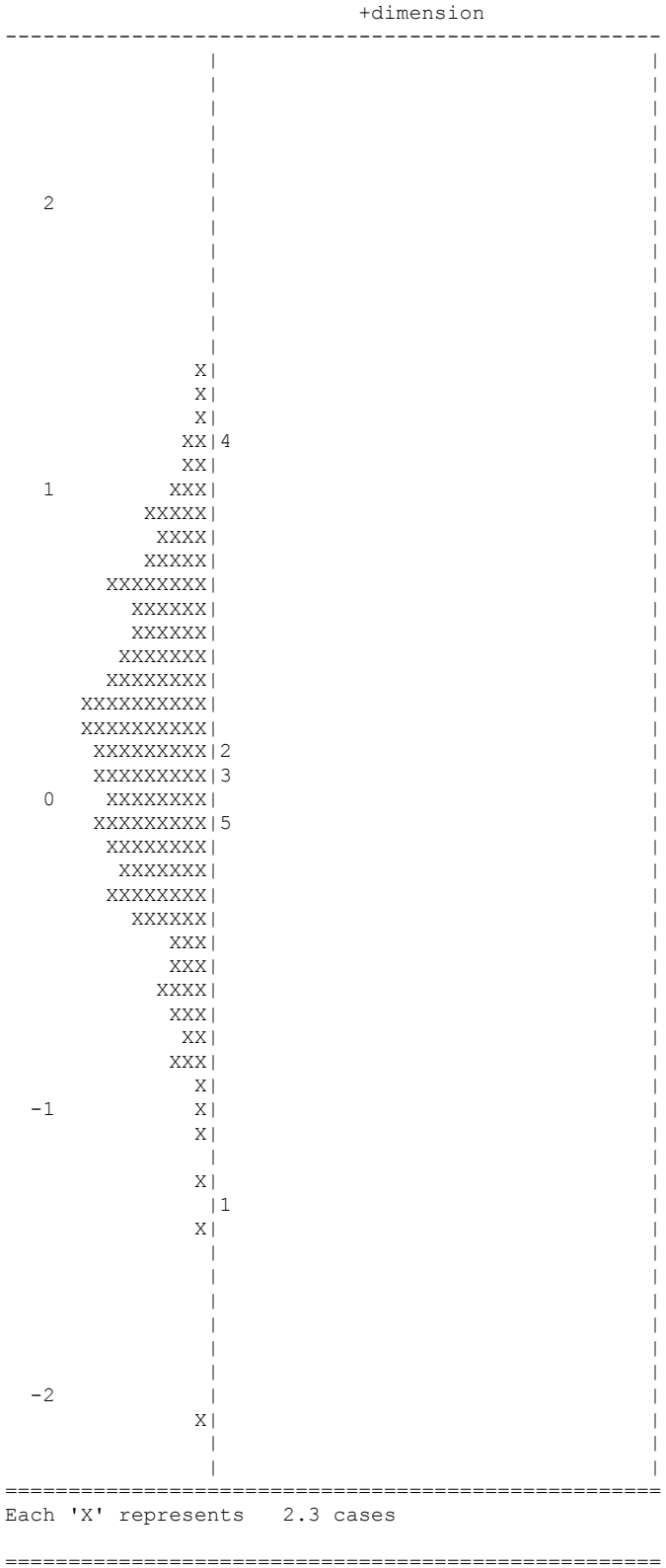
VARIABLES				UNWEIGHTED FIT			WEIGHTED FIT		
dimension	step	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1 Forms of Ju..	0			0.42	(0.86, 1.14)	-10.4	1.22	(0.54, 1.46)	0.9
1 Forms of Ju..	1	-1.715	0.075	1.16	(0.86, 1.14)	2.2	1.09	(0.85, 1.15)	1.2
1 Forms of Ju..	2	2.368	0.121	1.03	(0.86, 1.14)	0.5	1.04	(0.77, 1.23)	0.4
1 Forms of Ju..	3	-0.653*		1.03	(0.86, 1.14)	0.5	1.01	(0.86, 1.14)	0.2
2 Relevant-Su..	0			0.81	(0.68, 1.32)	-1.2	0.84	(0.72, 1.28)	-1.1
2 Relevant-Su..	1	2.132	0.388	0.97	(0.68, 1.32)	-0.1	0.95	(0.28, 1.72)	-0.0
2 Relevant-Su..	2	-2.132*		0.75	(0.68, 1.32)	-1.6	0.78	(0.71, 1.29)	-1.6

```
 3   Sufficient ..    0                              0.76 ( 0.80, 1.20) -2.5   0.82 ( 0.87, 1.13) -2.9
 3   Sufficient ..    1              0.203   0.165    0.97 ( 0.80, 1.20) -0.3   0.98 ( 0.84, 1.16) -0.2
 3   Sufficient ..    2             -0.203*           0.85 ( 0.80, 1.20) -1.5   0.89 ( 0.88, 1.12) -1.8
 4   Multiple Cl..    0                              0.72 ( 0.86, 1.14) -4.3   0.97 ( 0.69, 1.31) -0.2
 4   Multiple Cl..    1             -4.231   0.158    0.93 ( 0.86, 1.14) -1.0   0.86 ( 0.84, 1.16) -1.7
 4   Multiple Cl..    2              1.269   0.103    0.93 ( 0.86, 1.14) -0.9   0.83 ( 0.79, 1.21) -1.6
 4   Multiple Cl..    3              1.826   0.200    0.57 ( 0.86, 1.14) -7.0   0.94 ( 0.27, 1.73) -0.0
 4   Multiple Cl..    4             -1.455   0.229    1.23 ( 0.86, 1.14)  3.0   0.96 ( 0.55, 1.45) -0.1
 4   Multiple Cl..    5              2.676   0.716   13.99 ( 0.86, 1.14) 58.6   1.08 ( 0.00, 3.02)  0.4
 4   Multiple Cl..    6             -0.108   1.009    0.95 ( 0.86, 1.14) -0.7   1.19 ( 0.00, 3.12)  0.5
 4   Multiple Cl..    7              0.023*           0.18 ( 0.86, 1.14)-18.1   1.34 ( 0.00, 3.29)  0.6
 5   Reasoning        0                              1.11 ( 0.86, 1.14)  1.5   1.41 ( 0.60, 1.40)  1.8
 5   Reasoning        1             -1.930   0.122    0.73 ( 0.86, 1.14) -4.1   0.89 ( 0.80, 1.20) -1.2
 5   Reasoning        2             -0.682   0.085    1.11 ( 0.86, 1.14)  1.4   1.08 ( 0.85, 1.15)  1.1
 5   Reasoning        3             -0.724   0.069    0.83 ( 0.86, 1.14) -2.4   0.83 ( 0.88, 1.12) -3.0
 5   Reasoning        4              2.228   0.119    0.89 ( 0.86, 1.14) -1.5   0.98 ( 0.78, 1.22) -0.2
 5   Reasoning        5              1.108*           1.78 ( 0.86, 1.14)  8.8   1.09 ( 0.72, 1.28)  0.7
-------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
^ Quick standard errors have been used
===============================================================================
```

```
==================================

RELIABILITY COEFFICIENTS
------------------------

Dimension: (Dimension 1)
-----------------------
 MLE Person separation RELIABILITY:   0.649
 WLE Person separation RELIABILITY:  Unavailable
 EAP/PV RELIABILITY:                  0.623
```



1 "Forms of Justification"

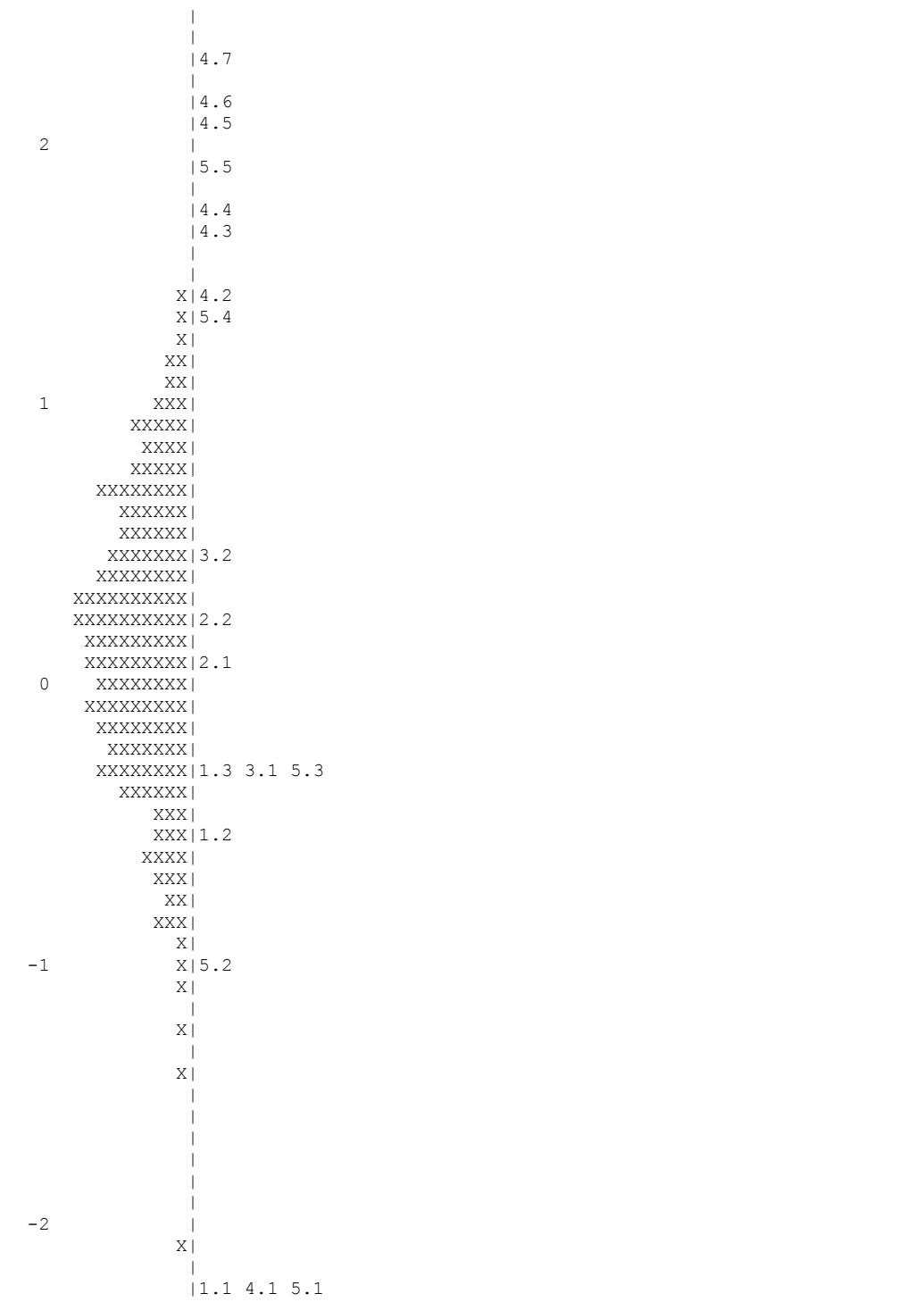
2 "Relevant-Supporting Evidence"

3 "Sufficient Evidence"

4 "Multiple Claims"

MAP OF LATENT DISTRIBUTIONS AND THRESHOLDS

Generalised-Item Thresholds



Each 'X' represents 2.3 cases
The labels for thresholds show the levels of
dimension, and step, respectively

Classical Statistics

Item 1
------
dimension:1 (Forms of Justification)
Cases for this item 952 Discrimination 0.51
Item Threshold(s): -3.05 -0.52 -0.31 Weighted MNSQ 1.00
Item Delta(s): -3.02 1.06 -1.96

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
0	0.00	19	2.00	-0.36	-11.77(.000)	-1.16	0.65
1	1.00	244	25.63	-0.35	-11.64(.000)	-0.15	0.56
2	2.00	74	7.77	-0.04	-1.18(.239)	0.07	0.57
3	3.00	615	64.60	0.45	15.46(.000)	0.28	0.51

Item 2
------
dimension:2 (Relevant-Supporting Evidence)
Cases for this item 149 Discrimination 0.64
Item Threshold(s): 0.11 0.23 Weighted MNSQ 0.80
Item Delta(s): 2.30 -1.97

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
0	0.00	85	57.05	-0.64	-10.17(.000)	-0.48	0.56
1	1.00	7	4.70	0.11	1.32(.187)	0.17	0.73
2	2.00	57	38.26	0.61	9.27(.000)	0.35	0.47

Item 3
------
dimension:3 (Sufficient Evidence)
Cases for this item 193 Discrimination 0.69
Item Threshold(s): -0.34 0.45 Weighted MNSQ 0.82
Item Delta(s): 0.26 -0.15

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
0	0.00	72	37.31	-0.68	-12.87(.000)	-0.47	0.49
1	1.00	50	25.91	0.15	2.04(.043)	0.30	0.46
2	2.00	71	36.79	0.55	9.12(.000)	0.38	0.44

Item 4
------
dimension:4 (Multiple Claims)
Cases for this item 951 Discrimination 0.36
Item Threshold(s): -3.05 1.43 1.66 1.71 2.07 2.13 2.33 Weighted MNSQ 0.81
Item Delta(s): -3.05 2.45 3.01 -0.28 3.86 1.07 1.20

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
0	0.00	42	4.42	-0.34	-11.23(.000)	-0.68	0.67
1	1.00	794	83.49	-0.02	-0.75(.454)	0.12	0.56
2	2.00	86	9.04	0.15	4.54(.000)	0.39	0.47
3	3.00	7	0.74	0.11	3.42(.001)	0.64	0.37
4	4.00	19	2.00	0.17	5.18(.000)	0.64	0.57
5	5.00	1	0.11	0.01	0.20(.840)	-0.13	0.00
6	6.00	1	0.11	0.06	1.89(.059)	1.04	0.00
7	7.00	1	0.11	0.08	2.57(.010)	1.13	0.00

Item 5
------
dimension:5 (Reasoning)
Cases for this item 952 Discrimination 0.60
Item Threshold(s): -2.27 -1.05 -0.34 1.37 1.88 Weighted MNSQ 1.04
Item Delta(s): -2.02 -0.78 -0.82 2.13 1.01

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1

0	0.00	26	2.73	-0.36	-11.85(.000)	-0.95	0.63
1	1.00	113	11.87	-0.40	-13.31(.000)	-0.41	0.52
2	2.00	198	20.80	-0.17	-5.36(.000)	-0.01	0.52
3	3.00	484	50.84	0.32	10.49(.000)	0.28	0.51
4	4.00	80	8.40	0.20	6.21(.000)	0.44	0.42
5	5.00	51	5.36	0.18	5.64(.000)	0.47	0.49

Table 87: Difficulty of each level of the rubric by Item:

TERM 4: topic*dimension*step

	VARIABLES					UNWEIGHTED FIT			WEIGHTED FIT		
	topic	dimension	step	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	Category + 1	Forms of Ju..	0			0.15	(0.86, 1.14)	-19.5	1.01	(0.27, 1.73)	0.1
1	Category + 1	Forms of Ju..	1	-0.073	0.199	0.82	(0.86, 1.14)	-2.7	0.99	(0.56, 1.44)	-0.0
1	Category + 1	Forms of Ju..	2	1.227	0.258	0.97	(0.86, 1.14)	-0.3	0.99	(0.56, 1.44)	0.0
1	Category + 1	Forms of Ju..	3	-1.155*		0.92	(0.86, 1.14)	-1.1	0.94	(0.74, 1.26)	-0.5
2	Causal + Re1	Forms of Ju..	0			0.22	(0.84, 1.16)	-14.7	0.91	(0.18, 1.82)	-0.1
2	Causal + Re1	Forms of Ju..	1	-2.404	0.128	1.21	(0.84, 1.16)	2.5	1.12	(0.88, 1.12)	1.9
2	Causal + Re1	Forms of Ju..	2	2.257	0.176	1.01	(0.84, 1.16)	0.1	1.01	(0.74, 1.26)	0.1
2	Causal + Re1	Forms of Ju..	3	0.147*		1.11	(0.84, 1.16)	1.4	1.09	(0.89, 1.11)	1.5
3	Explanatory1	Forms of Ju..	1			1.11	(0.74, 1.26)	0.9	1.10	(0.82, 1.18)	1.1
3	Explanatory1	Forms of Ju..	2	1.473	0.331	0.92	(0.74, 1.26)	-0.6	0.99	(0.46, 1.54)	0.1
3	Explanatory1	Forms of Ju..	3	-1.473*		1.21	(0.74, 1.26)	1.6	1.13	(0.82, 1.18)	1.3
4	Causal + Ex1	Forms of Ju..	0			0.98	(0.68, 1.32)	-0.1	1.10	(0.27, 1.73)	0.4
4	Causal + Ex1	Forms of Ju..	1	-1.862	0.255	1.08	(0.68, 1.32)	0.5	1.09	(0.82, 1.18)	0.9
4	Causal + Ex1	Forms of Ju..	2	2.051	0.376	1.16	(0.68, 1.32)	1.0	1.02	(0.43, 1.57)	0.2
4	Causal + Ex1	Forms of Ju..	3	-0.189*		1.09	(0.68, 1.32)	0.6	1.13	(0.76, 1.24)	1.0
5	Explanatory1	Forms of Ju..	0			0.32	(0.68, 1.32)	-5.7	0.96	(0.00, 2.01)	0.1
5	Explanatory1	Forms of Ju..	1	-1.652	0.274	0.88	(0.68, 1.32)	-0.7	0.96	(0.75, 1.25)	-0.3
5	Explanatory1	Forms of Ju..	2	2.982	0.590	0.84	(0.68, 1.32)	-1.0	0.98	(0.00, 2.05)	0.1
5	Explanatory1	Forms of Ju..	3	-1.330*		0.78	(0.68, 1.32)	-1.4	0.90	(0.73, 1.27)	-0.7
4	Causal + Ex2	Relevant-Su..	0			0.98	(0.68, 1.32)	-0.1	0.92	(0.75, 1.25)	-0.6
4	Causal + Ex2	Relevant-Su..	1	1.887	0.515	0.90	(0.68, 1.32)	-0.6	0.99	(0.11, 1.89)	0.1
4	Causal + Ex2	Relevant-Su..	2	-1.887*		1.28	(0.68, 1.32)	1.6	1.00	(0.73, 1.27)	0.1
5	Explanatory2	Relevant-Su..	0			0.88	(0.68, 1.32)	-0.7	0.86	(0.76, 1.24)	-1.2
5	Explanatory2	Relevant-Su..	1	2.259	0.590	1.29	(0.68, 1.32)	1.7	1.01	(0.00, 2.06)	0.2
5	Explanatory2	Relevant-Su..	2	-2.259*		0.78	(0.68, 1.32)	-1.4	0.87	(0.76, 1.24)	-1.1
3	Explanatory3	Sufficient ..	0			0.79	(0.74, 1.26)	-1.7	0.90	(0.77, 1.23)	-0.9
3	Explanatory3	Sufficient ..	1	-0.089	0.203	0.98	(0.74, 1.26)	-0.1	0.99	(0.84, 1.16)	-0.1
3	Explanatory3	Sufficient ..	2	0.089*		0.98	(0.74, 1.26)	-0.1	0.94	(0.85, 1.15)	-0.8
5	Explanatory3	Sufficient ..	0			0.66	(0.68, 1.32)	-2.3	0.74	(0.79, 1.21)	-2.7
5	Explanatory3	Sufficient ..	1	0.377	0.292	0.85	(0.68, 1.32)	-0.9	0.94	(0.67, 1.33)	-0.3
5	Explanatory3	Sufficient ..	2	-0.377*		0.83	(0.68, 1.32)	-1.0	0.97	(0.71, 1.29)	-0.2
1	Category + 4	Multiple Cl..	0			0.64	(0.86, 1.14)	-5.7	0.95	(0.50, 1.50)	-0.1
1	Category + 4	Multiple Cl..	1	-4.454	0.279	1.01	(0.86, 1.14)	0.2	1.01	(0.90, 1.10)	0.2
1	Category + 4	Multiple Cl..	2	0.121	0.122	1.01	(0.86, 1.14)	0.2	0.99	(0.87, 1.13)	-0.1
1	Category + 4	Multiple Cl..	3	1.923	0.216	0.53	(0.86, 1.14)	-7.8	0.96	(0.31, 1.69)	0.0
1	Category + 4	Multiple Cl..	4	-1.257	0.246	1.62	(0.86, 1.14)	7.2	1.03	(0.64, 1.36)	0.2
1	Category + 4	Multiple Cl..	5	2.954	0.729	79.62	(0.86, 1.14)	136.4	1.07	(0.00, 2.96)	0.4
1	Category + 4	Multiple Cl..	6	0.247	1.023	0.47	(0.86, 1.14)	-9.1	1.20	(0.00, 3.07)	0.5
1	Category + 4	Multiple Cl..	7	0.466*		0.16	(0.86, 1.14)	-18.9	1.48	(0.00, 3.30)	0.7
4	Causal + Ex4	Multiple Cl..	0			1.02	(0.68, 1.32)	0.2	1.10	(0.12, 1.88)	0.4
4	Causal + Ex4	Multiple Cl..	1	-3.876	0.472	0.95	(0.68, 1.32)	-0.3	1.04	(0.21, 1.79)	0.2
4	Causal + Ex4	Multiple Cl..	2	3.876*		6.05	(0.68, 1.32)	15.0	1.07	(0.00, 2.92)	0.4
1	Category + 5	Reasoning	0			0.71	(0.86, 1.14)	-4.5	1.07	(0.34, 1.66)	0.3

1	Category + 5	Reasoning	1	-1.359	0.285	0.88 (0.86, 1.14)	-1.7	0.96 (0.51, 1.49)	-0.1
1	Category + 5	Reasoning	2	-1.606	0.219	0.95 (0.86, 1.14)	-0.7	0.97 (0.78, 1.22)	-0.2
1	Category + 5	Reasoning	3	-1.582	0.139	0.95 (0.86, 1.14)	-0.6	0.96 (0.83, 1.17)	-0.5
1	Category + 5	Reasoning	4	4.547*		0.81 (0.86, 1.14)	-2.8	1.00 (0.29, 1.71)	0.1
2	Causal + Re5	Reasoning	0			0.50 (0.84, 1.16)	-7.6	0.94 (0.29, 1.71)	-0.1
2	Causal + Re5	Reasoning	1	-2.358	0.218	0.78 (0.84, 1.16)	-2.9	0.97 (0.76, 1.24)	-0.2
2	Causal + Re5	Reasoning	2	-1.076	0.153	1.01 (0.84, 1.16)	0.2	1.01 (0.89, 1.11)	0.2
2	Causal + Re5	Reasoning	3	-0.039	0.122	1.01 (0.84, 1.16)	0.2	1.01 (0.94, 1.06)	0.2
2	Causal + Re5	Reasoning	4	1.748	0.176	1.01 (0.84, 1.16)	0.2	1.02 (0.78, 1.22)	0.2
2	Causal + Re5	Reasoning	5	1.725*		2.26 (0.84, 1.16)	11.6	1.10 (0.64, 1.36)	0.6
3	Explanatory5	Reasoning	1			0.81 (0.74, 1.26)	-1.5	0.94 (0.73, 1.27)	-0.4
3	Explanatory5	Reasoning	2	-0.658	0.201	0.91 (0.74, 1.26)	-0.6	0.99 (0.75, 1.25)	-0.0
3	Explanatory5	Reasoning	3	-0.602	0.195	0.95 (0.74, 1.26)	-0.3	0.99 (0.85, 1.15)	-0.1
3	Explanatory5	Reasoning	4	0.913	0.277	1.05 (0.74, 1.26)	0.4	1.02 (0.65, 1.35)	0.1
3	Explanatory5	Reasoning	5	0.348*		0.81 (0.74, 1.26)	-1.5	1.04 (0.68, 1.32)	0.3
4	Causal + Ex5	Reasoning	0			0.80 (0.68, 1.32)	-1.2	1.05 (0.42, 1.58)	0.2
4	Causal + Ex5	Reasoning	1	-2.483	0.357	0.97 (0.68, 1.32)	-0.1	1.01 (0.85, 1.15)	0.1
4	Causal + Ex5	Reasoning	2	0.066	0.257	0.95 (0.68, 1.32)	-0.2	0.99 (0.73, 1.27)	-0.0
4	Causal + Ex5	Reasoning	3	0.720	0.318	0.70 (0.68, 1.32)	-2.0	0.96 (0.46, 1.54)	-0.1
4	Causal + Ex5	Reasoning	4	0.356	0.425	1.55 (0.68, 1.32)	2.9	1.12 (0.46, 1.54)	0.5
4	Causal + Ex5	Reasoning	5	1.342*		9.31 (0.68, 1.32)	20.2	1.25 (0.09, 1.91)	0.6
5	Explanatory5	Reasoning	0			0.52 (0.68, 1.32)	-3.6	1.05 (0.38, 1.62)	0.2
5	Explanatory5	Reasoning	1	-1.372	0.290	0.70 (0.68, 1.32)	-2.0	0.92 (0.64, 1.36)	-0.4
5	Explanatory5	Reasoning	2	-0.748	0.250	1.06 (0.68, 1.32)	0.4	1.03 (0.79, 1.21)	0.3
5	Explanatory5	Reasoning	3	0.608	0.261	0.94 (0.68, 1.32)	-0.3	0.99 (0.65, 1.35)	0.0
5	Explanatory5	Reasoning	4	0.869	0.352	0.93 (0.68, 1.32)	-0.4	1.03 (0.55, 1.45)	0.2
5	Explanatory5	Reasoning	5	0.643*		6.78 (0.68, 1.32)	16.5	1.31 (0.62, 1.38)	1.5

An asterisk next to a parameter estimate indicates that it is constrained
^ Quick standard errors have been used

Reference

Edwards, M. C. & Wirth, R. J. (2009). Measurement and the Study of Change. *Research in Human Development, 6*(2–3), 74–96. Online DOI: 10.1080/15427600902911163

Lambert, Z. V., & Durand, R. M. (1975). Some precautions in using canonical analysis. Journal of Market Research, 12, 468–475.

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