Understanding Methods of Measuring Vocabulary Ability Through Classical Test Theory, Factor Analysis and Item Response Theory

Introduction:

From weekly vocabulary tests in elementary school to the SAT Critical Reading section in late high school, testing of vocabulary pervades a good chunk of testing throughout primary education. According to William Nagy, a comprehensive long-term approach to instruction of vocabulary can help students immensely in their reading comprehension skills. This is because an effective vocabulary program teaches students how to use definitional and contextual clues to piece together meanings of unknown words, which they can apply to comprehension problems and elsewhere in order to break down larger concepts (Nagy 2005). Thus, a thorough instruction of vocabulary is important for the progression of a student's education.

There are many ways to approach measuring strength of vocabulary. One way, a rudimentary method, would be to give words and simply ask for their definitions. This would be simple because it would not reinforce the skills that are applicable in other areas such as reading comprehension; rather, it would be a test of simple memorization. A better way to test vocabulary would be to construct a test that emphasizes the use of definition and context clues to figure out the meaning of a word (Nagy 2005). A good example of a vocabulary test structured like this is the Critical Reading section of the SAT, where students are given a sentence with one word taken out, and five choices of words to fill in. This allows students to use the contextual clues in the sentence as well as the definition clues from the other word choices to come up with their answer.

The goal of this research project was to create a scale that measured the trait of vocabulary strength, where increasing knowledge of words and their meanings indicates a stronger vocabulary. This was to be achieved by finding data on an existing vocabulary test given to a number of subjects and paring it down via Classical Test Theory, Factor Analysis and Item Response Theory methods to produce a shorter, more unidimensional scale.

There have been other studies that tackle this psychometric view of vocabulary. Bogue et al. examined ten standardized vocabulary tests designed for children at varying levels of primary education via classical test theory methods such as validity and reliability. They found that none

of the ten tests chosen met their established criteria for either measure, proving the necessity of using Classical Test Theory concepts to create a scale that actually measures the trait it is intended to measure (Bogue et al., 2014).

Another study performed by Simos et al. on vocabulary psychometrics strays closer to the objective of this project. Simos et al. looked at the Peabody Picture Vocabulary Test–Revised (PPVT-R) Greek Vocabulary exam and attempted to create a short form of the test. First, Factor Analysis was performed on the items of the PPVT-R to confirm that the data measured an underlying factor of vocabulary strength. Then, an IRT model was applied on the data in the form of a Rasch model to pare down the data and create a shorter, unidimensional scale from the larger dataset. Finally, the shortened test was administered to a new sample of students, where scores stayed consistent, which was exhibited by CTT concepts such as test–retest reliability and convergent, discriminant and predictive validity (Simos et al., 2011).

Clearly, others have had success in doing what this research project set out to do, as researchers are no strangers to applying Classical Test Theory, Factor Analysis, and Item Response Theory ideas to vocabulary tests to either gauge their psychometric properties or create new scales. Hopefully, this project will bring a new lens to the idea by using each of the three concepts in slightly different ways to ultimately create a shorter, unidimensional scale that measures strength of vocabulary.

Methods:

Data Description

The data chosen was from the Vocabulary IQ Test from the Open Psychometrics raw data bank. In this dataset, there are 45 vocabulary questions along with 30 personality/demographic questions, answered by 12,173 participants. The personality questions were answered on a 5-level scale from disagree to agree, and ranged from questions relating to habits (e.g., does the subject packs heavy when traveling) to questions directly relating to personality (e.g., does the subject tend to follow rules). The demographic questions had anywhere from 3-4 choices (e.g.., whether the subject grew up in a rural, suburban, or urban area), along with a question relating to age.

Each of the 45 vocabulary questions consisted of 5 words. From these 5 choices, the subject selected the two words that mean the same thing. For example, if the words were “tiny”, “faded”, “new”, “large”, and “big”, then the answers here would be “large” and “big”. The questions varied in difficulty, with some questions requiring an inherently more expansive vocabulary to answer. For example, deciding which two words share a meaning in the group “stanchion”, “strumpet”, “pole”, “pale”, and “forestall” required a breadthful knowledge of vocabulary, since these are not very commonly used words. Thus, this data should ideally measure the trait of strength of vocabulary (defined earlier as level of expansiveness of vocabulary) well.

Data Manipulation

First, the data was subset to include only the first 45 columns. These are only the columns which contain direct vocabulary questions, since these are the direct indicators of vocabulary strength. Thankfully, there were no missing values in the data to be dealt with.

The raw data was not quite ripe for analysis (Appendix A). For each question in the data, the response was denoted with a number that corresponds to a binary representation of the choices. For example, if someone picked the first and second words out of the five choices, their answer corresponded to a 11000 coding, which then translated to a 3 in the raw data. For example, in question 1, the correct answer was the fourth and fifth choice of words, which would correspond to 00011 and thus a 24 by the codebook. Since many people got question 1 right, its mean was close to 24. To prevent difficulties in analysis, the data was completely recoded in a binary fashion, where a 1 denoted a correct response and a 0 denoted an incorrect response (Appendix B).

It was verified that no variables had correlations outside of the -1 to 1 range, no variables correlated 1.0 with any other variables, and no variables correlated zero with all other variables (Appendix E).

In looking at the means (Appendix C), there were quite a few questions whose means were above 0.9, meaning that for those questions, over 90% of people got them right. This likely meant that these questions would not be great indicators of strength of vocabulary, since an

overwhelming majority of people got those questions right. In line with this thinking, the data was pared down so that any question with a mean response above 0.9 (or in other words, any question which over 90% of responders got right) was taken out. This reduced the data from 45 columns to 33 columns, meaning were are now only 33 vocabulary questions in the dataset (Appendices F & G).

Factor Analysis

The first step in analysis was a series of minimum residual factor analyses performed on the data. A polychoric correlation matrix was calculated in R using the "psych" library. Then, the eigenvalues of this polychoric correlation matrix were found. Four factor analyses were done: one with one factor and no rotation, one with two factors and a geomin rotation, one with two factors and a varimax rotation and one with three factors and a bifactor rotation. For the two-factor solutions, initially, an oblimin rotation was done, but yielded a factor loading larger than 1; thus, the geomin rotation was done instead. All factor analyses were done using the "psych" package, where the default minimum residual factoring method was used. The data was analyzed based on their factor loadings, and ultimately, the scale was reduced based on the one-factor rotation.

Item Response Theory

To continue analysis and shorten the scale further, Information Response Theory models were implemented. These models were implemented in R using the "ltm" package. Two IRT models were implemented: one 2PL model and one 3PL model. An ANOVA test in the "ltm" package was performed to compare the two models and decide which was more ripe for analysis; based on this, the 2PL model was chosen to continue with analysis. Using the difficulty and discrimination coefficients of the model, plots of the Test Information Function, Test Response Function and Item Characteristic Curves, the scale was further shortened. All plots were created using the "ltm" package except for the Test Response Function curve, which was created with the "irtoys" package in R.

Conclusions

Finally, a number of summary statistics were calculated on both the final and original scales to show that the final scale had more internal consistency than the original one (the final scale being the scale with the pared down data after factor analysis and IRT model implementation, and original scale being the original 45-item dataset). The mean correlation between items was the average of the off-diagonal elements of the correlation matrix. The inverse alpha was the quotient of the mean covariance between items (off-diagonal elements of the variance-covariance matrix) and mean item variance. The mean factor loading was the average of the factor loadings on a one-factor model without rotation. To calculate these values, minimum residual factor analyses were done on both the original and final scale. The mean discrimination parameter was the mean of discrimination parameters for all items under a 2PL model. To calculate this value for the original scale, a 2PL model was created with the initial dataset.

The ratio of the largest eigenvalue to the sum of eigenvalues was calculated for the final and original scales to confirm that the final was more unidimensional than the original. This value was the quotient of largest eigenvalue and the sum of eigenvalues from the eigendecomposition of each scale's respective polychoric matrix.

Results:

Factor Analysis

Figure 1: Eigenvalues of Polychoric Correlation Matrix

```
 [1] 17.394  1.592  1.155  0.997  0.888  0.800  0.690  0.674  0.660  0.630  0.576  0.541  0.480  0.467  0.430  0.408  0.388  0.381  0.369  0.357  0.336  0.309  0.305  0.281  0.271
[26]  0.256  0.250  0.246  0.235  0.207  0.197  0.174  0.057
```

Figure 2: Factor Loading Matrix, One Factor

	MR1
Q2	0.73
Q4	0.53
Q6	0.80
Q7	0.16
Q8	0.82
Q9	0.80
Q12	0.80
Q13	0.80
Q15	0.63
Q18	0.72
Q19	0.77
Q20	0.83
Q21	0.80
Q24	0.89
Q25	0.57
Q26	0.33
Q27	0.68
Q28	0.76
Q29	0.69
Q30	0.56
Q33	0.79
Q34	0.86
Q35	0.59
Q36	0.67
Q37	0.85
Q38	0.72
Q39	0.83
Q40	0.72
Q41	0.82
Q42	0.83
Q43	0.73
Q44	0.53
Q45	0.49

Figure 3: Factor Loading Matrix, Two Factors, Geomin Rotation

	MR1	MR2
Q2	0.82	-0.25
Q4	0.57	-0.12
Q6	0.78	0.08
Q7	0.21	-0.17
Q8	0.85	-0.09
Q9	0.82	-0.05
Q12	0.86	-0.16
Q13	0.83	-0.09
Q15	0.63	0.01
Q18	0.70	0.09
Q19	0.76	0.05
Q20	0.80	0.09
Q21	0.73	0.21
Q24	0.96	-0.21
Q25	0.55	0.07
Q26	0.43	-0.31
Q27	0.52	0.51
Q28	0.68	0.26
Q29	0.70	-0.04
Q30	0.58	-0.05
Q33	0.72	0.21
Q34	0.86	0.01
Q35	0.52	0.23
Q36	0.54	0.43
Q37	0.89	-0.10
Q38	0.64	0.26
Q39	0.84	-0.02
Q40	0.63	0.27
Q41	0.81	0.05
Q42	0.79	0.13
Q43	0.62	0.36
Q44	0.50	0.09
Q45	0.50	-0.02

Figure 4: Factor Loading Matrix, Two Factors, Varimax Rotation

	MR1	MR2
Q2	0.37	0.70
Q4	0.31	0.46
Q6	0.63	0.50
Q7	0.01	0.24
Q8	0.54	0.64
Q9	0.55	0.60
Q12	0.48	0.68
Q13	0.52	0.62
Q15	0.46	0.44
Q18	0.58	0.44
Q19	0.59	0.50
Q20	0.65	0.51
Q21	0.71	0.40
Q24	0.51	0.78
Q25	0.45	0.34
Q26	0.04	0.46
Q27	0.81	0.10
Q28	0.71	0.34
Q29	0.47	0.51
Q30	0.37	0.43
Q33	0.69	0.39
Q34	0.62	0.59
Q35	0.57	0.24
Q36	0.75	0.15
Q37	0.55	0.67
Q38	0.68	0.31
Q39	0.58	0.59
Q40	0.69	0.30
Q41	0.62	0.54
Q42	0.68	0.48
Q43	0.75	0.24
Q44	0.44	0.30
Q45	0.34	0.36

Figure 5: Factor Loading Matrix, Three Factors, Bifactor Rotation

	MR1	MR2	MR3
Q2	0.74	-0.32	0.10
Q4	0.53	-0.08	0.16
Q6	0.79	0.09	0.11
Q7	0.15	-0.08	0.19
Q8	0.82	-0.10	0.13
Q9	0.80	-0.08	0.09
Q12	0.79	-0.10	0.25
Q13	0.80	-0.15	0.06
Q15	0.62	0.05	0.16
Q18	0.73	-0.01	-0.05
Q19	0.79	-0.12	-0.12
Q20	0.83	0.00	-0.01
Q21	0.81	0.04	-0.17
Q24	0.90	-0.30	0.08
Q25	0.56	0.15	0.17
Q26	0.31	-0.09	0.54
Q27	0.68	0.48	-0.08
Q28	0.75	0.22	0.00
Q29	0.67	0.06	0.27
Q30	0.55	0.05	0.26
Q33	0.78	0.23	0.10
Q34	0.87	-0.13	-0.06
Q35	0.59	0.21	0.00
Q36	0.67	0.36	-0.10
Q37	0.84	-0.07	0.22
Q38	0.73	0.09	-0.19
Q39	0.84	-0.17	-0.07
Q40	0.71	0.29	0.06
Q41	0.82	0.00	0.05
Q42	0.84	-0.04	-0.14
Q43	0.74	0.18	-0.24
Q44	0.53	0.07	0.02
Q45	0.49	-0.02	0.07

Figure 6: Final Loading Matrix after Factor Analysis

	MR1
Q2	0.74
Q6	0.79
Q8	0.82
Q9	0.81
Q12	0.80
Q13	0.82
Q18	0.73
Q19	0.79
Q20	0.83
Q21	0.81
Q24	0.91
Q28	0.74
Q33	0.76
Q34	0.87
Q37	0.84
Q38	0.72
Q39	0.84
Q41	0.83
Q42	0.84
Q43	0.73

Item Response Theory

Figure 7: Coefficients for 2PL Model

	Dffclt	Dscrmn
Q2	-1.596	1.997
Q6	-1.069	2.178
Q8	-0.987	2.380
Q9	-0.867	2.201
Q12	-1.412	2.333
Q13	-1.207	2.423
Q18	-0.002	1.692
Q19	-0.141	2.076
Q20	-0.763	2.468
Q21	-0.035	2.305
Q24	-0.978	3.627
Q28	-0.490	1.802
Q33	-1.446	1.967
Q34	-0.681	2.947
Q37	-1.265	2.713
Q38	0.486	1.796
Q39	-0.563	2.544
Q41	-0.475	2.345
Q42	-0.041	2.547
Q43	0.827	1.962

Figure 8: Coefficients for 3PL Model

	Gussng	Dffclt	Dscrmn
Q2	0.04	-1.56	2.01
Q6	0.06	-0.98	2.29
Q8	0.00	-0.98	2.35
Q9	0.00	-0.86	2.19
Q12	0.00	-1.42	2.29
Q13	0.00	-1.21	2.37
Q18	0.00	0.01	1.70
Q19	0.00	-0.13	2.07
Q20	0.02	-0.72	2.53
Q21	0.02	0.01	2.48
Q24	0.00	-0.97	3.54
Q28	0.14	-0.22	2.46
Q33	0.25	-1.08	2.43
Q34	0.02	-0.64	3.05
Q37	0.00	-1.27	2.67
Q38	0.01	0.50	1.92
Q39	0.00	-0.56	2.51
Q40	0.00	0.09	1.55
Q41	0.00	-0.46	2.36
Q42	0.00	-0.03	2.60
Q43	0.01	0.82	2.09

Figure 9: Test Information Function curve for 2PL Model

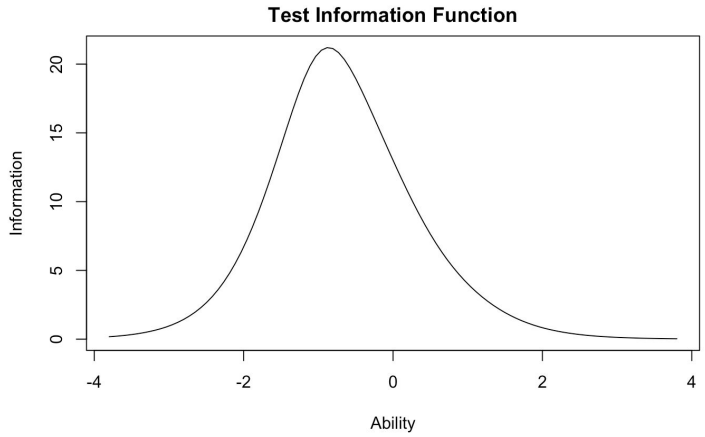


Figure 10: Test Response Function curve for 2PL Model

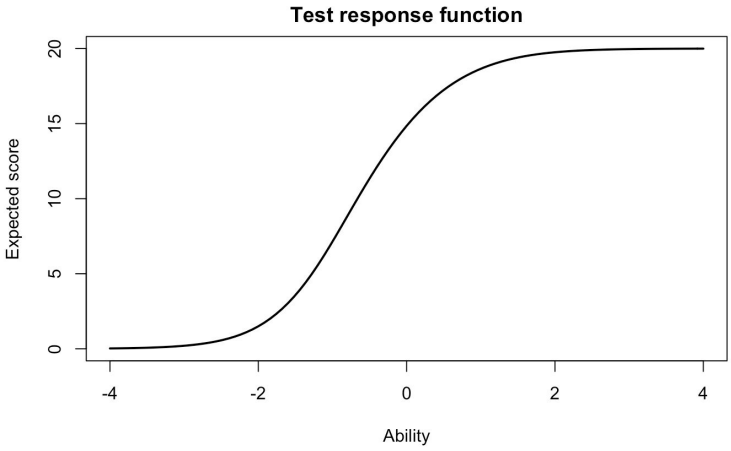


Figure 11: Item Characteristic Curve for Question 43

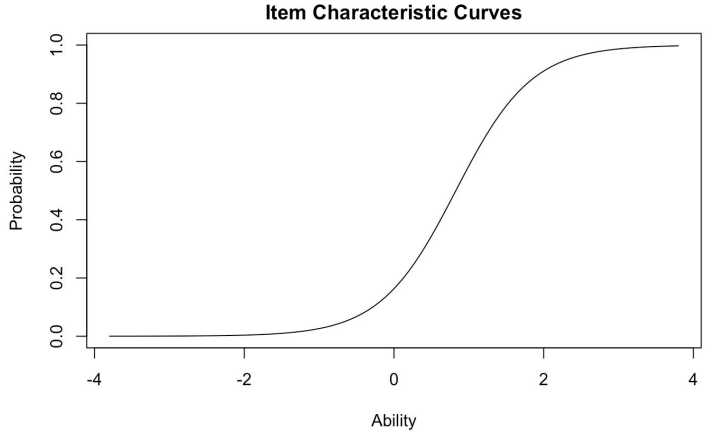


Figure 12: Coefficients for 2PL Model, Reduced Data

	Dffclt	Dscrmn
Q6	-1.06	2.19
Q8	-0.96	2.47
Q9	-0.85	2.26
Q12	-1.40	2.36
Q13	-1.19	2.48
Q19	-0.14	2.07
Q20	-0.75	2.48
Q21	-0.04	2.31
Q24	-0.95	3.85
Q34	-0.67	3.09
Q37	-1.25	2.71
Q39	-0.55	2.62
Q41	-0.47	2.38
Q42	-0.04	2.63

Figure 13: Test Information Function curve for 2PL Model, Reduced Data

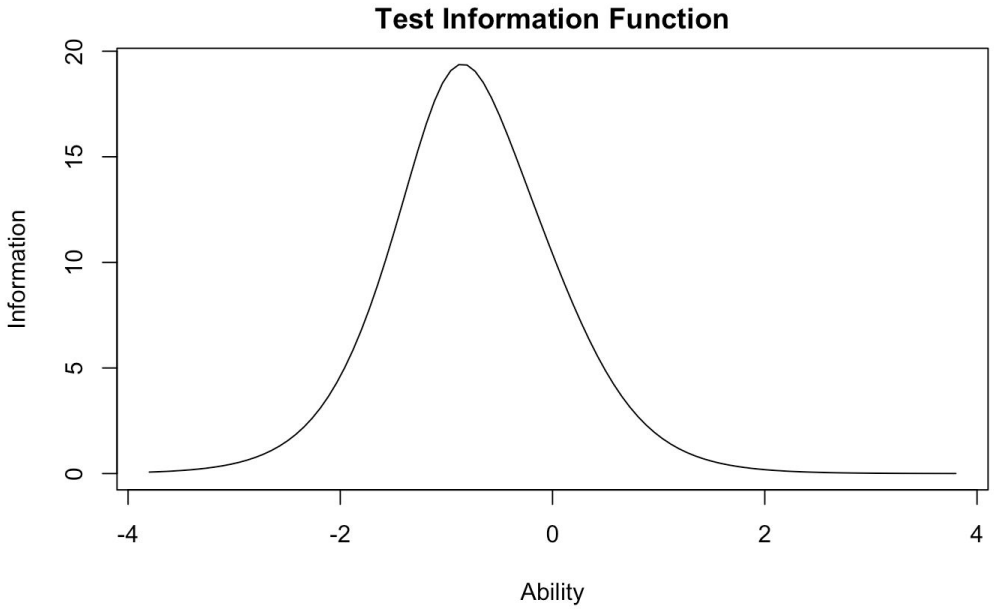
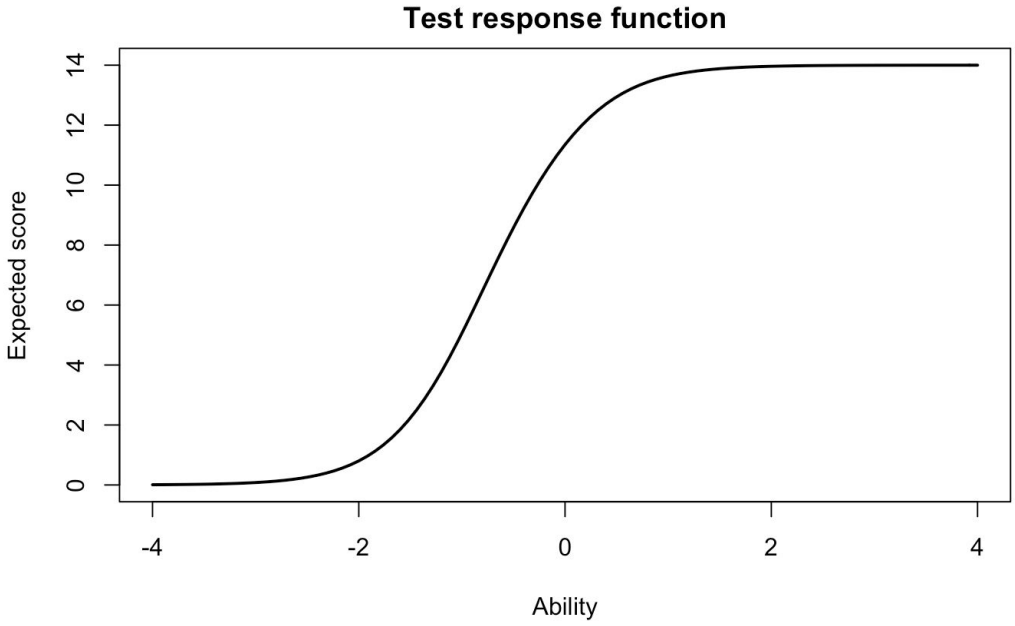


Figure 14: Test Response Function curve for 2PL Model, Reduced Data



Conclusion

Figure 15: Table of Summary Statistics for Original and Final Scales

Statistic	Original Scale	Final Scale
Mean Correlation Between Items	0.24	0.41
Inverse Alpha	0.23	0.41
Mean Factor Loadings	0.68	0.82
Mean Discrimination Parameter	1.78	2.54
Ratio of Largest Eigenvalue to Sum of All Eigenvalues	0.49	0.70

Discussion

Data

The participants could use definitional and contextual clues on items in the test, which are crucial in the testing of vocabulary according to Nagy. For example, deciding which two words share a meaning in the group "fulminant", "doohickey", "ligature", "epistle", and "letter" from Question 38 required a knowledge of the definitions of at least some of those words, since these are not all very commonly used words. Beyond definitional clues, this test also offers participants contextual clues, as they can use the meanings of words they do know out of the five choices to make deductions about words they might not. For example, Question 10 featured the words "entrapment", "partner", "fool", "companion" and "mirror". If one did not know the answer here, they could maybe use contextual clues to piece together the answer. Most people know that a mirror is an object, or a noun. Most people would also deduce that a word with the suffix "-ment" is likely a noun; thus, entrapment is a noun. It is very obvious that "mirror" does not mean the same thing as "entrapment"; thus, the participant is left with three choices ("partner", "fool", and "companion"), where the two words that share a definition are fairly clear through rudimentary definitional clues.

Factor Analysis

Four factor analyses were performed since there was one eigenvalue far greater than the rest, but still three eigenvalues greater than 1 (Figure 1). It was clear through the loading matrices in Figures 2-5 that the one factor solution was ideal, as no other solution yielded such high loadings on any factor as the one factor solution did. The two factors in a geomin rotation correlated 0.263, which was insignificant enough to deem an orthogonal rotation better fit for the data than an oblique one. Though the varimax rotation showed better loadings on the second factor, there was still no comparison to the single-factor solution.

In examining the single-factor solution, there were many strong loadings, meaning that the factor correlated well with many of the items of the vocabulary test. It was decided that the dataset would be cut down to only those questions whose factor loadings were above 0.7, as this is a generally accepted statistical standard for a strong positive correlation. The dataset was then

reduced to 20 questions (or columns) from the initial 33, whose factor loadings are shown in Figure 6.

Item Response Theory

As seen in Figure 8, the guessing parameter in the 3PL model did not seem to impact the discrimination and difficulty all that much, especially considering that over half of the guessing parameters were about zero, which is what they are considered to be in the 2PL model, whose coefficients are shown in Figure 7. An ANOVA test (Appendix J) confirmed that these models were nearly identical, so the simpler 2PL model was chosen for analysis. However, the 3PL model did give some interesting insight. Figure 8 shows that question 28 had a guessing parameter of 0.25, indicating that this item was vastly different from the others in the dataset. Considering the surplus of items still in the dataset, it was eliminated from the scale.

In looking at Figures 9 and 10, and through the difficulty and discrimination parameters, this scale was very good at discriminating between people with a below-average strength of vocabulary. The Test Information Function curve in Figure 9 shows that this scale gave us the most information for those with about a -1 ability level. The data for the Test Response Function, plotted in Figure 10, showed that a person with average ability (b-value of 0) would have an expected score of 14.85 (15, since one cannot get 14.85 questions correct) out of 20 questions, or 75% (Appendix L). The test was fairly simple, as indicated by the fact that there were only 3 questions that had above-average difficulty (3 questions with difficulty parameters greater than 0) and one question that had average difficulty (difficulty parameter equals 0) (Figure 7). The rest of the questions had less than average difficulty (difficulty parameters below 0). However, the scale had strong discriminant power, as no discrimination coefficient was less than 1 (Figure 7). To improve the scale by paring it, it was decided that all items with average and above average difficulty would be taken out. Since the scale was already good at discriminating between participants with lower vocabulary ability, difficult items such as Question 43, whose Item Characteristic Curve is shown in Figure 11, would not help improve the scale as they were better at discriminating people with higher vocabulary ability.

After removing the questions with average and above-average difficulty, items were further removed if their discrimination parameter values were less than 2, so as to make the scale stronger in discriminating between those with below average vocabulary skill. The dataset was reduced to 14 columns (or questions) from 21. The coefficients for the 2PL model with the new data are shown in Figure 12 and plots of the Test Information Function curve and Test Response Function curve for a 2PL model with the new data are shown in Figures 13 and 14.

Now, the scale showed stronger discriminatory power amongst those with below average vocabulary skill. Compared to the 2PL model of the previous dataset, the whole Test Information Function curve was shifted to the left; 79.69% of the information was contained in the interval $-6 \leq \text{Ability} \leq 0$ in the Test Information Curve of the new data in Figure 13 (Appendix P), compared to 77.35% of the information in the same interval in the previous Test Information Curve in Figure 9 . Also, the Test Response Function curve showed that one with an average vocabulary ability would do better on this test than the last one. The data for the Test Response Function curve plotted in Figure 14 showed that a person with average ability (b-value of 0) would have an expected score of 11.35 (11, since one cannot get 11.35 questions correct) out of 14 questions, or 78.57% (Appendix O), which was higher than the expected score of a participant with average ability from the first 2PL model (75%).

Conclusions

The summary statistics in Figure 15 show that the final scale with 14 items was more internally consistent in many different ways. The data itself showed this, as the mean correlation between all items rose from 0.24 to 0.41, meaning that each of the items were more interdependent on each other in the final scale. Classical Test Theory methods exhibit improved internal consistency in the new scale, as the inverse alpha score rose from 0.23 to 0.41, implying that the final scale was more reliable. Factor Analysis methods showed that the mean factor loadings jumped from 0.68 in the initial to 0.82 in the final. This meant that the ostensible factor of vocabulary strength correlated better with the items in the final scale than those in the initial scale, thus making it more internally consistent. Finally, Item Response Theory techniques also

showed that the mean discrimination parameter grew from 1.78 to 2.54; thus, the final scale discriminated better amongst participants than the original scale.

Additionally, Figure 15 shows that the final scale is more unidimensional than the original scale. The ratio of the first eigenvalue to the sum of all eigenvalues was closer to 1 (0.70) for the final scale than it was for the initial scale (0.49). This means that one factor in the final scale explained 70% of the variance in the items, while one factor explained only 49% of the variance in the items in the original scale, thus proving that the final scale was more unidimensional.

Overall, the scale to measure strength of vocabulary was reduced from 45 questions to 14 questions, while increasing internal consistency and unidimensionality. The scale also retained its offering of definitional and contextual clues through the nature of the questions. Unfortunately, the scale only showed power in discriminating amongst those with below-average ability, as previously stated. This makes the scale less favorable than it would be if it had a greater range of questions in terms of difficulty.

Perhaps researchers could give harder questions next time so that there are more difficult items that show up to the right of the 0 ability mark on the Item Characteristic Curves Appendices K&N). Then, these models and criteria for item selection could be applied on the set of items with a wider range of difficulty. This could create an internally consistent, unidimensional scale that tests vocabulary strength well for those at varying vocabulary strength levels. As it stands right now, the scale is internally consistent and unidimensional, and follows the criteria for a meaningful vocabulary test in terms of test-taking strategy, but is not as universal in terms of application as one might hope a vocabulary strength scale would be.

References

Bogue, E. L., DeThorne, L. S., & Schaefer, B. A. (2014). A Psychometric Analysis of Childhood Vocabulary Tests. *Contemporary Issues in Communication Science and Disorders,41*, 55-69.

Nagy, W. (2005). Why Vocabulary Instruction Needs to Be Long-Term and Comprehensive. In *Teaching and Learning Vocabulary: Bringing Research to Practice*. Routledge.

Simos, P. G., Sideridis, G. D., Protopapas, A., & Mouzaki, A. (2011). Psychometric Evaluation of a Receptive Vocabulary Test for Greek Elementary Students. *Assessment for Effective Intervention,37*, 34-49.

Appendices

Appendix A: Preview of Raw Data

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	24	3	10	24	9	-1	5	10	18	10
2	24	3	10	3	9	12	17	10	20	10
3	24	3	10	5	9	9	10	10	17	10
4	24	3	10	5	9	9	17	10	0	10
5	24	3	10	5	9	9	17	10	17	10
6	24	3	10	18	9	9	17	10	17	10

Appendix B: Preview of Rekeyed Data

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	1	1	1	0	1	0	0	1	0	1
2	1	1	1	0	1	0	1	1	0	1
3	1	1	1	1	1	1	0	1	1	1
4	1	1	1	1	1	1	1	1	0	1
5	1	1	1	1	1	1	1	1	1	1
6	1	1	1	0	1	1	1	1	1	1

Appendix C: Means of Each Column, Full Rekeyed Data

```
  Q1   Q2   Q3   Q4   Q5   Q6   Q7   Q8   Q9  Q10  Q11  Q12  Q13  Q14  Q15  Q16  Q17  Q18  Q19  Q20  Q21  Q22  Q23
0.98 0.88 0.97 0.43 0.97 0.79 0.46 0.78 0.74 0.96 0.97 0.86 0.83 0.96 0.79 0.95 0.92 0.50 0.54 0.73 0.51 0.97 0.95
 Q24  Q25  Q26  Q27  Q28  Q29  Q30  Q31  Q32  Q33  Q34  Q35  Q36  Q37  Q38  Q39  Q40  Q41  Q42  Q43  Q44  Q45
0.80 0.55 0.87 0.29 0.63 0.80 0.86 0.97 0.90 0.85 0.71 0.74 0.54 0.85 0.36 0.67 0.48 0.64 0.51 0.27 0.45 0.75
```

Appendix D: Standard Deviations of Each Column, Full Rekeyed Data

```
  Q1   Q2   Q3   Q4   Q5   Q6   Q7   Q8   Q9  Q10  Q11  Q12  Q13  Q14  Q15  Q16  Q17  Q18  Q19  Q20  Q21  Q22  Q23
0.13 0.33 0.16 0.49 0.17 0.41 0.50 0.42 0.44 0.21 0.18 0.34 0.38 0.20 0.41 0.21 0.27 0.50 0.50 0.45 0.50 0.18 0.23
 Q24  Q25  Q26  Q27  Q28  Q29  Q30  Q31  Q32  Q33  Q34  Q35  Q36  Q37  Q38  Q39  Q40  Q41  Q42  Q43  Q44  Q45
0.40 0.50 0.33 0.45 0.48 0.40 0.34 0.16 0.29 0.35 0.45 0.44 0.50 0.36 0.48 0.47 0.50 0.48 0.50 0.44 0.50 0.43
```

Appendix E: Preview of Correlation Matrix, Full Rekeyed Data

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25
Q1	1.00	0.24	0.50	0.07	0.42	0.16	0.07	0.16	0.15	0.37	0.42	0.20	0.20	0.38	0.16	0.35	0.28	0.09	0.09	0.13	0.08	0.40	0.31	0.17	0.09
Q2	0.24	1.00	0.27	0.19	0.20	0.30	0.09	0.37	0.36	0.17	0.27	0.37	0.40	0.22	0.24	0.17	0.40	0.27	0.31	0.35	0.28	0.21	0.20	0.44	0.17
Q3	0.50	0.27	1.00	0.11	0.38	0.20	0.10	0.20	0.18	0.32	0.43	0.24	0.24	0.33	0.19	0.31	0.35	0.12	0.14	0.18	0.12	0.38	0.30	0.23	0.11
Q4	0.07	0.19	0.11	1.00	0.08	0.24	0.05	0.27	0.28	0.05	0.12	0.24	0.21	0.10	0.21	0.08	0.20	0.24	0.27	0.24	0.23	0.10	0.10	0.29	0.21
Q5	0.42	0.20	0.38	0.08	1.00	0.16	0.08	0.17	0.17	0.32	0.36	0.23	0.20	0.31	0.16	0.28	0.26	0.10	0.12	0.16	0.11	0.32	0.28	0.19	0.09
Q6	0.16	0.30	0.20	0.24	0.16	1.00	0.08	0.42	0.42	0.13	0.20	0.41	0.39	0.20	0.29	0.12	0.29	0.32	0.35	0.45	0.38	0.21	0.17	0.45	0.30
Q7	0.07	0.09	0.10	0.05	0.08	0.08	1.00	0.09	0.09	0.07	0.08	0.09	0.08	0.09	0.07	0.09	0.10	0.05	0.06	0.08	0.06	0.06	0.09	0.10	0.05
Q8	0.16	0.37	0.20	0.27	0.17	0.42	0.09	1.00	0.46	0.14	0.22	0.45	0.43	0.20	0.30	0.15	0.38	0.33	0.38	0.46	0.38	0.19	0.18	0.52	0.28
Q9	0.15	0.36	0.18	0.28	0.17	0.42	0.09	0.46	1.00	0.12	0.21	0.41	0.42	0.18	0.31	0.14	0.35	0.36	0.39	0.44	0.38	0.17	0.16	0.49	0.28
Q10	0.37	0.17	0.32	0.05	0.32	0.13	0.07	0.14	0.12	1.00	0.33	0.18	0.16	0.28	0.14	0.24	0.20	0.08	0.08	0.11	0.07	0.31	0.24	0.12	0.07
Q11	0.42	0.27	0.43	0.12	0.36	0.20	0.08	0.22	0.21	0.33	1.00	0.27	0.26	0.30	0.20	0.30	0.37	0.13	0.15	0.19	0.13	0.37	0.31	0.26	0.12
Q12	0.20	0.37	0.24	0.24	0.23	0.41	0.09	0.45	0.41	0.18	0.27	1.00	0.40	0.25	0.30	0.19	0.36	0.28	0.31	0.37	0.30	0.23	0.22	0.49	0.24
Q13	0.20	0.40	0.24	0.21	0.20	0.39	0.08	0.43	0.42	0.16	0.26	0.40	1.00	0.21	0.28	0.17	0.39	0.32	0.35	0.48	0.37	0.20	0.19	0.47	0.23
Q14	0.38	0.22	0.33	0.10	0.31	0.20	0.09	0.20	0.18	0.28	0.30	0.25	0.21	1.00	0.18	0.26	0.26	0.12	0.14	0.17	0.13	0.30	0.26	0.22	0.13
Q15	0.16	0.24	0.19	0.21	0.16	0.29	0.07	0.30	0.31	0.14	0.20	0.30	0.28	0.18	1.00	0.14	0.28	0.28	0.29	0.31	0.27	0.18	0.16	0.33	0.22
Q16	0.35	0.17	0.31	0.08	0.28	0.12	0.09	0.15	0.14	0.24	0.30	0.19	0.17	0.26	0.14	1.00	0.26	0.07	0.10	0.13	0.08	0.28	0.24	0.17	0.07
Q17	0.28	0.40	0.35	0.20	0.26	0.29	0.10	0.38	0.35	0.20	0.37	0.36	0.39	0.26	0.28	0.26	1.00	0.24	0.27	0.34	0.23	0.27	0.23	0.50	0.19
Q18	0.09	0.27	0.12	0.24	0.10	0.32	0.05	0.33	0.36	0.08	0.13	0.28	0.32	0.12	0.28	0.07	0.24	1.00	0.38	0.37	0.39	0.12	0.11	0.37	0.26
Q19	0.09	0.31	0.14	0.27	0.12	0.35	0.06	0.38	0.39	0.08	0.15	0.31	0.35	0.14	0.29	0.10	0.27	0.38	1.00	0.39	0.45	0.13	0.13	0.45	0.26
Q20	0.13	0.35	0.18	0.24	0.16	0.45	0.08	0.46	0.44	0.11	0.19	0.37	0.48	0.17	0.31	0.13	0.34	0.37	0.39	1.00	0.46	0.17	0.15	0.48	0.28
Q21	0.08	0.28	0.12	0.23	0.11	0.38	0.06	0.38	0.38	0.07	0.13	0.30	0.37	0.13	0.27	0.08	0.23	0.39	0.45	0.46	1.00	0.11	0.13	0.40	0.25
Q22	0.40	0.21	0.38	0.10	0.32	0.21	0.06	0.19	0.17	0.31	0.37	0.23	0.20	0.30	0.18	0.28	0.27	0.12	0.13	0.17	0.11	1.00	0.31	0.19	0.11
Q23	0.31	0.20	0.30	0.10	0.28	0.17	0.09	0.18	0.16	0.24	0.31	0.22	0.19	0.26	0.16	0.24	0.23	0.11	0.13	0.15	0.13	0.31	1.00	0.20	0.11
Q24	0.17	0.44	0.23	0.29	0.19	0.45	0.10	0.52	0.49	0.12	0.26	0.49	0.47	0.22	0.33	0.17	0.50	0.37	0.45	0.48	0.40	0.19	0.20	1.00	0.29
Q25	0.09	0.17	0.11	0.21	0.09	0.30	0.05	0.28	0.28	0.07	0.12	0.24	0.23	0.13	0.22	0.07	0.19	0.26	0.26	0.28	0.25	0.11	0.11	0.29	1.00

Appendix F: Means of Each Column, Data Without Means>0.9

```
  Q2   Q4   Q6   Q7   Q8   Q9  Q12  Q13  Q15  Q18  Q19  Q20  Q21  Q24  Q25  Q26  Q27  Q28  Q29  Q30  Q33  Q34  Q35
0.88 0.43 0.79 0.46 0.78 0.74 0.86 0.83 0.79 0.50 0.54 0.73 0.51 0.80 0.55 0.87 0.29 0.63 0.80 0.86 0.85 0.71 0.74
 Q36  Q37  Q38  Q39  Q40  Q41  Q42  Q43  Q44  Q45
0.54 0.85 0.36 0.67 0.48 0.64 0.51 0.27 0.45 0.75
```

Appendix G: Standard Deviations of Each Column, Data Without Means>0.9

```
  Q2   Q4   Q6   Q7   Q8   Q9  Q12  Q13  Q15  Q18  Q19  Q20  Q21  Q24  Q25  Q26  Q27  Q28  Q29  Q30  Q33  Q34  Q35
0.33 0.49 0.41 0.50 0.42 0.44 0.34 0.38 0.41 0.50 0.50 0.45 0.50 0.40 0.50 0.33 0.45 0.48 0.40 0.34 0.35 0.45 0.44
 Q36  Q37  Q38  Q39  Q40  Q41  Q42  Q43  Q44  Q45
0.50 0.36 0.48 0.47 0.50 0.48 0.50 0.44 0.50 0.43
```

Appendix H: Preview of Polychoric Correlation Matrix, Data Without Means>0.9

	Q2	Q4	Q6	Q7	Q8	Q9	Q12	Q13	Q15	Q18	Q19	Q20	Q21	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q33	Q34	Q35	Q36
Q2	1.00	0.40	0.54	0.19	0.63	0.62	0.64	0.67	0.46	0.57	0.63	0.62	0.59	0.72	0.33	0.34	0.33	0.47	0.53	0.42	0.49	0.64	0.38	0.38
Q4	0.40	1.00	0.44	0.08	0.48	0.48	0.50	0.41	0.38	0.37	0.41	0.40	0.36	0.55	0.32	0.17	0.32	0.40	0.43	0.30	0.42	0.47	0.29	0.29
Q6	0.54	0.44	1.00	0.15	0.65	0.65	0.67	0.63	0.49	0.56	0.59	0.68	0.66	0.69	0.51	0.31	0.58	0.60	0.56	0.43	0.66	0.69	0.46	0.57
Q7	0.19	0.08	0.15	1.00	0.16	0.15	0.18	0.15	0.13	0.08	0.09	0.13	0.09	0.18	0.08	0.16	0.10	0.10	0.14	0.19	0.07	0.14	0.03	0.04
Q8	0.63	0.48	0.65	0.16	1.00	0.69	0.72	0.68	0.51	0.57	0.63	0.69	0.65	0.76	0.47	0.32	0.49	0.59	0.58	0.47	0.60	0.73	0.47	0.53
Q9	0.62	0.48	0.65	0.15	0.69	1.00	0.68	0.67	0.51	0.59	0.62	0.67	0.61	0.74	0.46	0.26	0.49	0.60	0.54	0.45	0.61	0.70	0.47	0.47
Q12	0.64	0.50	0.67	0.18	0.72	0.68	1.00	0.66	0.52	0.56	0.60	0.63	0.59	0.76	0.47	0.37	0.44	0.58	0.56	0.51	0.63	0.70	0.46	0.49
Q13	0.67	0.41	0.63	0.15	0.68	0.67	0.66	1.00	0.48	0.59	0.63	0.74	0.69	0.72	0.41	0.36	0.50	0.55	0.55	0.46	0.58	0.71	0.37	0.48
Q15	0.46	0.38	0.49	0.13	0.51	0.51	0.52	0.48	1.00	0.48	0.50	0.51	0.46	0.54	0.38	0.30	0.45	0.49	0.47	0.39	0.48	0.48	0.39	0.38
Q18	0.57	0.37	0.56	0.08	0.57	0.59	0.56	0.59	0.48	1.00	0.57	0.59	0.57	0.66	0.40	0.14	0.49	0.57	0.52	0.37	0.57	0.61	0.42	0.47
Q19	0.63	0.41	0.59	0.09	0.63	0.62	0.60	0.63	0.50	0.57	1.00	0.62	0.65	0.76	0.39	0.18	0.48	0.59	0.48	0.38	0.59	0.70	0.48	0.50
Q20	0.62	0.40	0.68	0.13	0.69	0.67	0.63	0.74	0.51	0.59	0.62	1.00	0.72	0.72	0.45	0.27	0.57	0.63	0.56	0.46	0.63	0.72	0.45	0.58
Q21	0.59	0.36	0.66	0.09	0.65	0.61	0.59	0.69	0.46	0.57	0.65	0.72	1.00	0.70	0.38	0.15	0.58	0.62	0.50	0.44	0.64	0.71	0.48	0.62
Q24	0.72	0.55	0.69	0.18	0.76	0.74	0.76	0.72	0.54	0.66	0.76	0.72	0.70	1.00	0.51	0.39	0.46	0.59	0.58	0.49	0.62	0.83	0.45	0.46
Q25	0.33	0.32	0.51	0.08	0.47	0.46	0.47	0.41	0.38	0.40	0.39	0.45	0.38	0.51	1.00	0.23	0.44	0.46	0.42	0.33	0.48	0.49	0.39	0.38
Q26	0.34	0.17	0.31	0.16	0.32	0.26	0.37	0.36	0.30	0.14	0.18	0.27	0.15	0.39	0.23	1.00	0.15	0.21	0.36	0.33	0.29	0.26	0.14	0.17
Q27	0.33	0.32	0.58	0.10	0.49	0.49	0.44	0.50	0.45	0.49	0.48	0.57	0.58	0.46	0.44	0.15	1.00	0.60	0.47	0.34	0.59	0.53	0.44	0.64
Q28	0.47	0.40	0.60	0.10	0.59	0.60	0.58	0.55	0.49	0.57	0.59	0.63	0.62	0.59	0.46	0.21	0.60	1.00	0.53	0.44	0.62	0.62	0.52	0.58
Q29	0.53	0.43	0.56	0.14	0.58	0.54	0.56	0.55	0.47	0.52	0.48	0.56	0.50	0.58	0.42	0.36	0.47	0.53	1.00	0.42	0.56	0.52	0.38	0.46
Q30	0.42	0.30	0.43	0.19	0.47	0.45	0.51	0.46	0.39	0.37	0.38	0.46	0.44	0.49	0.33	0.33	0.34	0.44	0.42	1.00	0.49	0.45	0.36	0.39
Q33	0.49	0.42	0.66	0.07	0.60	0.61	0.63	0.58	0.48	0.57	0.59	0.63	0.64	0.62	0.48	0.29	0.59	0.62	0.56	0.49	1.00	0.63	0.57	0.58
Q34	0.64	0.47	0.69	0.14	0.73	0.70	0.70	0.71	0.48	0.61	0.70	0.72	0.71	0.83	0.49	0.26	0.53	0.62	0.52	0.45	0.63	1.00	0.48	0.53
Q35	0.38	0.29	0.46	0.03	0.47	0.47	0.46	0.37	0.39	0.42	0.48	0.45	0.48	0.45	0.39	0.14	0.44	0.52	0.38	0.36	0.57	0.48	1.00	0.49
Q36	0.38	0.29	0.57	0.04	0.53	0.47	0.49	0.48	0.38	0.47	0.50	0.58	0.62	0.46	0.38	0.17	0.64	0.58	0.46	0.39	0.58	0.53	0.49	1.00

Appendix I: Correlation Matrix of Factors, Bifactor Analysis with Geomin Rotation, Data Without Means>0.9

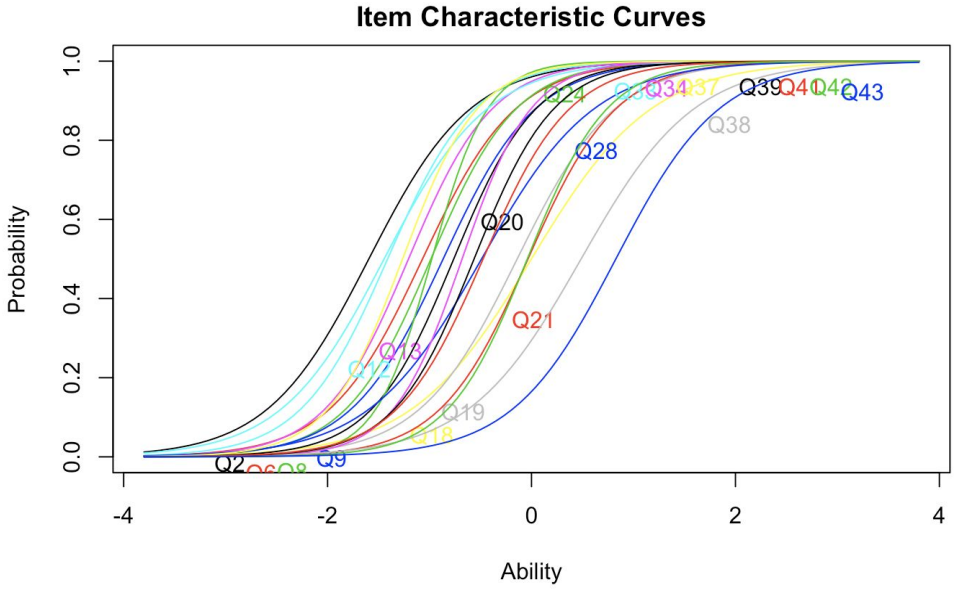
```
     [,1]  [,2]
[1,] 1.000 0.263
[2,] 0.263 1.000
```

Appendix J: ANOVA Test between 2PL and 3PL Models on Data After Factor Analysis

	AIC <dbl>	BIC <dbl>	log.Lik <dbl>	LRT <fctr>	df <fctr>	p.value <fctr>
IRTmodel	196404.6	196700.8	-98162.28			
IRTmodel2	196286.5	196730.9	-98083.25	158.07	20	<0.001

Note that “IRTmodel” is the 2PL model, and “IRTmodel2” is the 3PL model.

Appendix K: Item Characteristic Curves, All Items, 2PL Model on Data After FA



Appendix L: Preview of Data for Test Response Function, 2PL Model on Data After FA

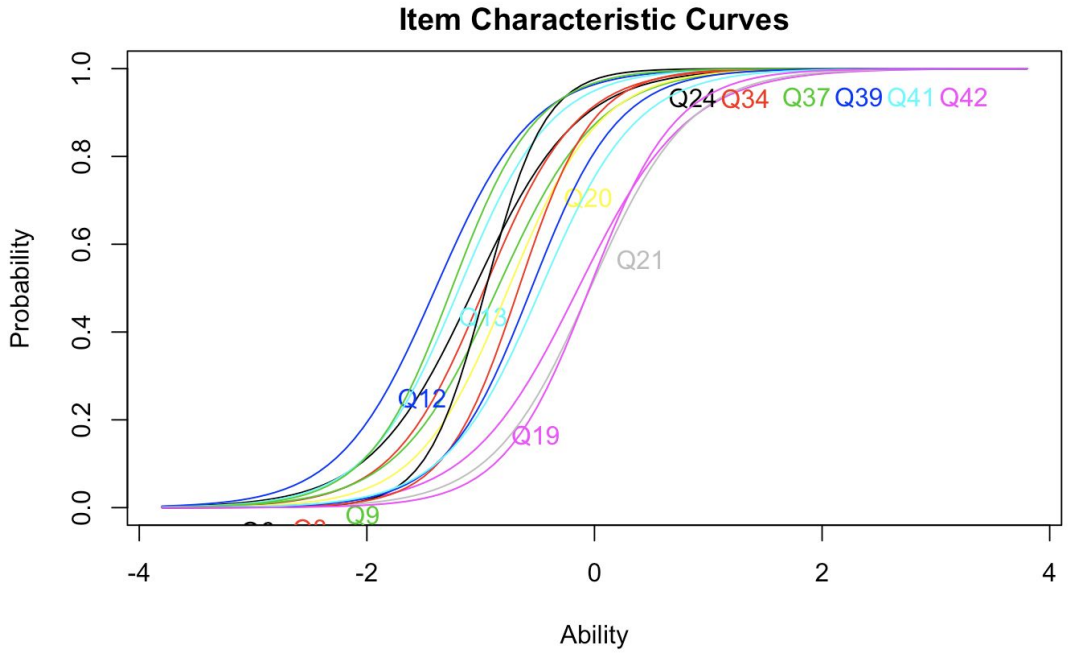
35	-1.28	4.97524463
36	-1.20	5.55582289
37	-1.12	6.16938633
38	-1.04	6.81072124
39	-0.96	7.47346143
40	-0.88	8.15044287
41	-0.80	8.83418955
42	-0.72	9.51742601
43	-0.64	10.19350397
44	-0.56	10.85667484
45	-0.48	11.50220068
46	-0.40	12.12633881
47	-0.32	12.72624649
48	-0.24	13.29984536
49	-0.16	13.84567592
50	-0.08	14.36276496
51	0.00	14.85052252
52	0.08	15.30867589
53	0.16	15.73723746
54	0.24	16.13649403
55	0.32	16.50700156
56	0.40	16.84957127
57	0.48	17.16523972
58	0.56	17.45522292
59	0.64	17.72086067
60	0.72	17.96355963
61	0.80	18.18474304

Appendix M: Information in Ability Range -6 ≤ b ≤0, 2PL Model on Data After FA

```
Call:
ltm(formula = new3 ~ z1, IRT.param = TRUE)

Total Information = 46.3
Information in (-6, 0) = 35.81 (77.35%)
Based on all the items
```

Appendix N: Appendix J: Item Characteristic Curves, All Items, 2PL Model on Final Data



Appendix O: Preview of Data for Test Response Function, 2PL Model on Final Data

38	−1.04	4.786774788
39	−0.96	5.336956281
40	−0.88	5.903809964
41	−0.80	6.479138302
42	−0.72	7.054728932
43	−0.64	7.622977740
44	−0.56	8.177303259
45	−0.48	8.712325480
46	−0.40	9.223861834
47	−0.32	9.708814842
48	−0.24	10.165010539
49	−0.16	10.591026385
50	−0.08	10.986035310
51	0.00	11.349684747
52	0.08	11.682019648
53	0.16	11.983445747
54	0.24	12.254718135
55	0.32	12.496935233
56	0.40	12.711520810
57	0.48	12.900184656
58	0.56	13.064861756
59	0.64	13.207636780
60	0.72	13.330663839
61	0.80	13.436091010

Appendix P: Information in Ability Range -6 ≤ b ≤0, 2PL Model on Final Data

```
Call:
ltm(formula = new4 ~ z1, IRT.param = TRUE)

Total Information = 35.57
Information in (-6, 0) = 28.34 (79.69%)
Based on all the items
```

Appendix Q: Means of Each Column, Final Data

Q6	Q8	Q9	Q12	Q13	Q19	Q20	Q21	Q24	Q34	Q37	Q39	Q42	Q43
0.79	0.78	0.74	0.86	0.83	0.54	0.73	0.51	0.80	0.71	0.85	0.67	0.51	0.27

Appendix R: Standard Deviations of Each Column, Final Data

Q6	Q8	Q9	Q12	Q13	Q19	Q20	Q21	Q24	Q34	Q37	Q39	Q42	Q43
0.41	0.42	0.44	0.34	0.38	0.50	0.45	0.50	0.40	0.45	0.36	0.47	0.50	0.44