Technical Report:

Psychometric Analysis of LOCUS as

Implemented through the Supporting Teachers' Enactment

of the Probability and Statistics Standards (STEPSS) Project

Submitted on September 27, 2019

A. Corinne Huggins-Manley
University of Florida

Tim Jacobbe
University of Florida

Corresponding author: A. Corinne Huggins-Manley; 1215 Norman Hall, Gainesville, FL 32611; (352) 273- 4342; amanley@coe.ufl.edu

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant "Supporting Teacher Enactment of the Probability and Statistics Standards (STEPSS)" to Florida State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Table of Contents

List of Abbreviations

3PL----------	3-parameter Logistic Model
CFA----------	Confirmatory Factor Analysis
CFI----------	Comparative Fit Index
CTT----------	Classical Test Theory
IRT----------	Item Response Theory
LOCUS-------	Levels of Conceptual Understanding of Statistics
RMSEA-------	Root Mean Square Error of Approximation
TLI----------	Tucker-Lewis Index

Executive Summary

This report contains descriptive statistics, dimensionality analysis, and item response theory (IRT) analysis of a 23 item Levels of Conceptual Understanding of Statistics (LOCUS) assessment given to 2,536 examinees. All items were scored correct and incorrect, and there was no missing data because skipped items were scored as 0 prior to receipt of the data. Overall, the results demonstrate that 20 of the items measure a common trait, and that the 20-item assessment is best suited for estimating the ability of higher performing examinees. This conclusion stems from the relatively high difficulty of the items and the fact that, while overall reliability is a bit low for making important decisions about examinees based on their scores, ability scores for higher performing examinees have more acceptable levels of reliability. These conclusions are drawn with the important caveat that "higher performing examinees" is a relative term. If, for example, the examinees are from a sample of gifted students in a state, then the analysis indicates that the assessment is best suited for students of very high ability (i.e., the top performing gifted students). Yet if the sample of examinees is from a school of students who are academically at-risk, then the analysis indicates that the assessment is best suited for students who are on the higher end of at-risk ability levels. In contrast, if the sample is randomly representative of a population of examinees in the US, then the assessment is best suited for examinees who have higher than average ability in the US. In sum, the assessment is best suited for students above the average ability of the sample, and that must be interpreted in light of the overall ability of the sample itself. It is also important to interpret this information in light of examinee motivation to do well on an assessment. It should also be noted that approximately 14,000 students were scheduled to take the assessments while only 2,536 were completed. Data was also not available regarding how long each test taker spent on the assessments when administered. Interpretations

must factor in the manner in which the data were collected as part of this project. Sometimes an assessment can appear hard simply because examinees had low motivation to perform well. In addition, scoring skipped items as incorrect in some ways changes the nature of the ability that is being estimated, as it is unknown how the students would perform on the items that they skipped. Information on the sample of examinees and their motivation to do well on the assessment is critical for interpreting the results in this report.

Descriptive Statistics

Table 1 shows item descriptive statistics. First, the frequencies and percentages of examinees within each item category are reported. Based on incorrect responses, this information shows that the items were relatively difficult for the sample of examinees. Second, classical test theory (CTT) item difficulty statistics are reported. These can be interpreted as the proportion of individuals who answered the item correctly. These difficulty values have a possible range of 0 to 1, with lower values indicating more difficult items. Again, one can see that the items were relatively difficult for the sample of examinees. Third, CTT item discrimination statistics are provided. For each item, these can be interpreted as the correlation between the item's responses and the total test score, after the particular item has been removed from the calculation of the total test score. For example, for item Q1, the responses of examinees to that item are correlated at $r = .27$ to the test scores coming from all the other items. Items with 0 discrimination would indicate that the data on that item is not related to data on other items. Table 1 shows that the items on the assessment have low to moderate discrimination (which is expected for multiple choice items (Crocker & Algina, 1986)), with three items showing essentially no discrimination power (i.e., items Q13, Q18, and Q20).

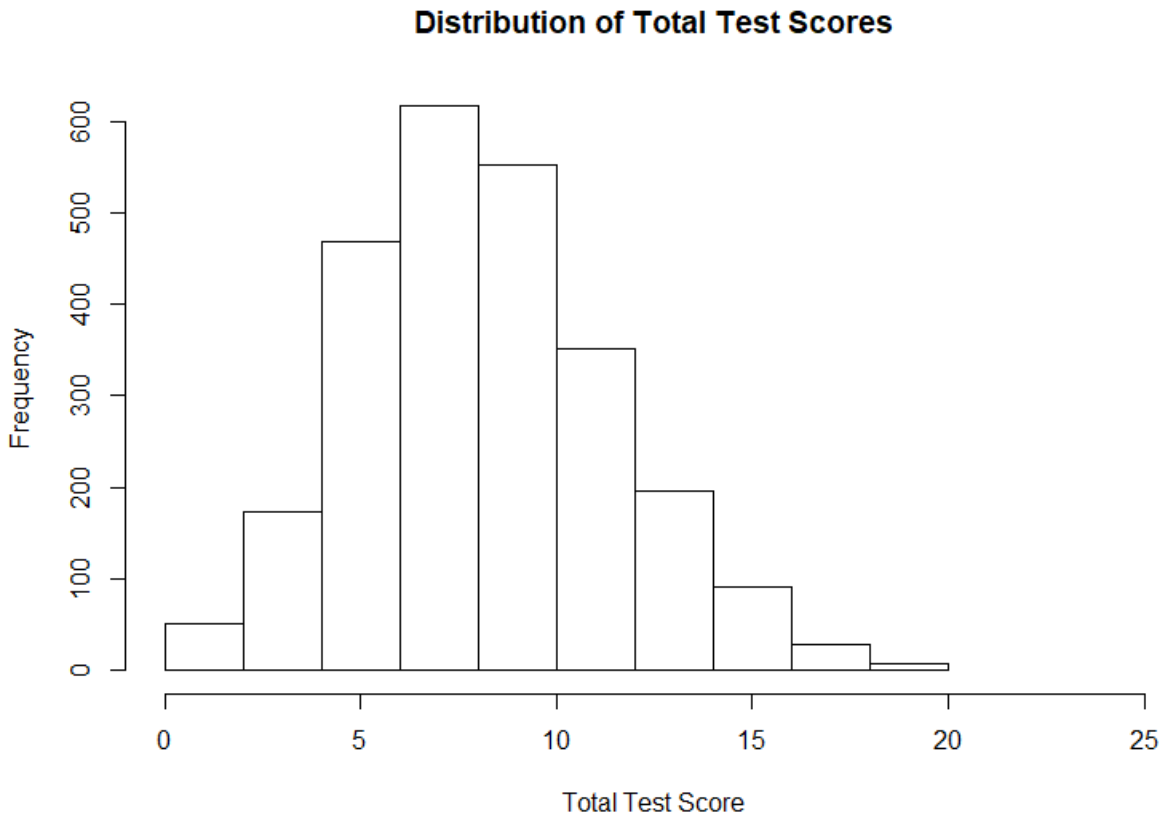
Cronbach's alpha for the full set of 23 items was estimated at $\alpha = .58$. This serves as an estimate of internal consistency, or the degree to which item responses are consistent across items. Generally, $\alpha = .58$ is considered low for making decisions from the total test scores.

Figure 1 shows the distribution of examinee total test scores. The possible range for total test scores is 0 to 23. The figure shows that total test scores were positively skewed with the average score closer to the lower end of the scale and, in general, more examinees toward the lower end of the scale as opposed to the higher end of the total test score scale.

Table 1: Item Descriptive Statistics

Item	*f* correct % correct	*f* incorrect % incorrect	CTT Item Difficulty	CTT Item Discrimination
Q1	1486 58.60%	1050 41.40%	.59	.27
Q2	1741 68.65%	795 31.35%	.69	.30
Q3	1622 63.96%	914 36.04%	.64	.32
Q4	1100 43.38%	1436 56.62%	.43	.31
Q5	1847 72.83%	689 27.17%	.73	.31
Q6	888 35.02%	1648 65.00%	.35	.19
Q7	774 30.52%	1762 69.48%	.31	.09
Q8	764 30.13%	1772 69.87%	.30	.23
Q9	1015 40.00%	1521 60.00%	.40	.23
Q10	771 30.40%	1765 69.60%	.30	.11
Q11	884 34.86%	1652 65.14%	.35	.15
Q12	697 27.48%	1839 72.52%	.27	.10
Q13	602 23.74%	1934 76.26%	.24	-.01
Q14	1310 51.66%	1226 48.34%	.52	.27
Q15	753 29.70%	1783 70.31%	.30	.21
Q16	860 33.91%	1676 66.09%	.34	.12
Q17	719 28.35%	1817 71.65%	.28	.21
Q18	527 20.78%	2009 79.22%	.21	-.06
Q19	488 19.24%	2048 80.76%	.19	.12
Q20	631 24.88%	1905 75.12%	.25	.03
Q21	680 26.81%	1856 73.19%	.27	.10
Q22	1148 45.27%	1388 54.73%	.45	.26
Q23	553 21.81%	1983 78.19%	.22	.16

Figure 1: Distribution of Total Test Scores



Dimensionality Analysis

For the current LOCUS data, theoretically, one could expect a single dimension underlying the assessment data (i.e., all items measure the same latent trait, that of Conceptual Understanding of Statistics). A unidimensional categorical confirmatory factor analysis (CFA) was fit to the full item data in *M*plus (Muthén & Muthén, 2012) using weighted least squares estimation with adjusted means and variances. Model goodness of fit indices were inadequate with respect to the comparative fit index (CFI = .88) and Tucker-Lewis index (TLI = .87) (Hu & Bentler, 1999). However, the root mean square of approximation (RMSEA) was satisfactory at RMSEA = 0.027. Based on the above item descriptive analysis, and on standardized factor loadings from the unidimensional CFA, items Q13 (standardized loading = -0.04), Q18 (standardized loading = -0.14), and Q20 (standardized loading = 0.00) were removed from the item set. The unidimensional CFA was fit to the revised set of items and all three fit indices were acceptable (RMSEA = .026, CFI = .91, TLI = .90; Hu & Bentler, 1999). The omega coefficient (McDonald, 1999) was calculated to evaluate reliability of the single dimension, with a result of ω = .74. For interpretation, 74% of all the variance in total test scores (across the 20 remaining items) is associated with the latent trait as opposed to random measurement error. This may be interpreted as adequate for some low-stakes decisions based on the test scores, but inadequate for high-stakes decisions.

In sum, there was one dimension underlying the LOCUS data once three problematic items were removed. Those items were flagged in various ways as having item response sets that were not related to the other items. The remaining twenty items measured a single underlying trait with an estimated reliability of .76.

Item Response Theory Analysis

The 3-parameter logistic (3PL; Birnbaum, 1968) model was chosen amongst the family of IRT models because it is appropriate for binary data and it models probabilities of item response in a manner that takes guessing into account. It is possible to guess the correct answer on the multiple choice LOCUS items, and hence this model is theoretically appropriate. The 3PL model assumes unidimensionality, which was evaluated in the above section and holds for twenty of the LOCUS items. The 3PL model also assumes local independence. The model was fit to the 20-item data in the mirt package (Chalmers, 2012), and standardized residual correlations were evaluated for significance. No residuals correlations were significant at the $\alpha = .001$ level, indicating that the assumption of local independence was met. Item fit and person fit were also examined to ensure that the model adequately captured the data. According to Orlando and Thissen's (2000) chi-square test, no item models were significantly misfit to the data at the $\alpha = .01$ level. For person fit, Drasgow, Levine, and Williams' (1985) Zh statistic flagged only 2% of the sample for significant person misfit at the $\alpha = .05$ level, indicating acceptable person fit across the sample. Based on all information above, the 3PL model fit well to the data and assumptions were upheld. Hence, the remainder of this section details the results of the IRT analysis. For interpretation purposes, know that the latent ability trait (θ) is scaled as a z-score, with 0 representing the average ability of the sample of examinees, and each integer unit away from 0 indicating a standard deviation unit distance from 0.

Table 2 shows the item parameter estimates for the 20 items retained in the analysis. The difficulty parameters (*b*) can be interpreted as the value along the latent trait for which there is a strong match between the difficulty of the item and the ability of the examinee. For example, item Q14's difficulty is $b = 0.08$, and hence this item's difficulty is targeted at examinees near the average ability of the sample. Hence, item Q14 would provide the most reliable information

for examinees of average ability in the sample. The item discrimination parameters (*a*) can be interpreted as the ability of the item to differentiate between persons who differ in ability. Higher numbers indicate higher discriminatory power. The item guessing parameters (*c*) can be interpreted as the probability of a low ability examinee guessing the correct answer to the item. Values close to 0 indicate low guessing probabilities on the item.

Figure 2 shows plots of the item parameters through item characteristic curves. Each window shows the plot for a single item, with the x-axis representing the latent ability trait (θ) and the y-axis representing the probability of a correct response on an item. Lines shifted farther to the right represent more difficult items, steeper lines represent more discriminating items, and lines with a higher lower asymptotes represent items with more guessing. Overall, the guessing parameter estimates align with expectations of guessing probabilities for multiple choice items, and the discrimination parameters are all positive with moderate to large discriminatory power. With respect to item difficulty, the items appear to be relatively hard for the sample of examinees.

Figure 3 shows an item person map, which is a visual of the alignment (or lack of alignment) between the distribution of examinee ability and the distribution of item difficulty on the assessment. The histogram in Figure 3 represents the latent ability score distribution, and the bolded item numbers show where the item difficulties lie in relation to the examinee ability distribution. This figure shows that the majority of items are hard for examinees, in that they are targeted to ability levels on the higher end of the sample distribution. Only three items (Q2, Q3, and Q5) have difficulty values that are meaningfully below 0. Restated, only three items have a difficulty that can provide reliable information about examinees with ability that is below the sample average.

Because the items are, overall, targeted toward the higher end of the sample ability distribution, reliability of ability scores is much higher for examinees scoring above the sample average than for examinees scoring below the sample average. Figure 4 shows the conditional reliability of ability estimates. For examinees who are two standard deviation units above the sample mean, conditional reliability is approximately $r = .80$. However, for examinees with ability up to three standard deviations below the mean, conditional reliability ranges from approximately $r = .10$ to $r = .70$. Marginal reliability is $r = .68$, which is a weighted average of the conditional reliability values plotted in Figure 4. This level of marginal reliability would be considered low but possibly acceptable for some low-stakes decisions, and too low for high-stakes decisions. However, when considering conditional reliability, decisions about examinees who are scoring above the mean are more appropriate as their estimated scores contain more reliable variance (i.e., less error in the score estimates).

IRT true scores can provide information about the expected total test score for examinees, with the expectation coming from the IRT item parameters. Each item characteristic curve in Figure 2 provides conditional probabilities of correctly responding to a test item. If one takes a single latent ability trait (θ) value, obtains the probability of correct response on each of the test items for that trait value, and then sums all the probabilities across the set of test items, one would obtain the IRT true score for a person with that trait value. Figure 5 shows the distribution of IRT true scores for each examinee in the sample data. Similar to the distribution of total test scores, the IRT true score is positively skewed with a center near a score of 7 on the true score scale range, indicating that more examinees are expected to have lower total test scores on the test as compared to higher total test scores on the test.

For reporting purposes, a scale was developed such that the latent ability trait (θ) was converted to a normal distribution with a mean of 50 and a standard deviation of 10. This was achieved by multiplying the trait scores (θ's) by 10, then adding 50, then rounding to the closest integer. For scale scores, an examinee having a score of 50 would indicate that he or she scored at the average score of the sample of examinees. A score of 60 would indicate that he or she scored one standard deviation unit above the average score of the sample of examinees, whereas a score of 30 would indicate that he or she scored two standard deviation units below the average score of the sample of examinees. Figure 6 shows the sample distribution of scale scores.

Overall, the IRT results demonstrate that the assessment is best suited for estimating the ability of higher performing examinees, with the important caveat that "higher performing examinees" is a relative term. If, for example, the examinees are from a sample of gifted students in a state, then the analysis indicates that the assessment is best suited for students of very high ability (i.e., the top performing gifted students). Yet if the sample of examinees is from a school of students who are academically at-risk, then the IRT analysis indicates that the assessment is best suited for students who are on the higher end of at-risk ability levels. In contrast, if the sample is randomly representative of a population of examinees in the US, then the assessment is best suited for examinees who have above average ability in the US.

In sum, the assessment is best suited for students above the average ability of the sample, and that must be interpreted in light of the ability of the sample itself. It should also be noted that approximately 14,000 students were scheduled to take the assessments while only 2,536 were completed. Data was also not available regarding how long each test taker spent on the assessments when administered. Interpretations must factor in the manner in which the data were collected as part of this project. Sometimes an assessment can appear hard simply because

examinees had low motivation to perform well. In addition, scoring skipped items as incorrect in some ways changes the nature of the ability that is being estimated, as it is unknown how the students would perform on the items that they skipped. Information on the sample of examinees and their motivation to do well on the assessment is critical for interpreting the results in this report.

Table 2: 3PL Item Parameter Estimates

Item	**b (difficulty)**	**a (discrimination)**	**c (guessing)**
Q1	0.41	1.25	0.31
Q2	-0.77	1.30	0.02
Q3	-0.45	1.49	0.06
Q4	0.87	2.04	0.24
Q5	-0.34	2.12	0.31
Q6	1.51	0.90	0.15
Q7	2.56	1.24	0.25
Q8	1.48	1.91	0.19
Q9	1.14	0.99	0.17
Q10	2.57	0.92	0.21
Q11	1.88	0.96	0.21
Q12	2.76	0.94	0.20
Q14	0.08	0.93	0.06
Q15	1.76	0.93	0.12
Q16	2.18	0.34	0.01
Q17	1.89	0.91	0.12
Q19	2.78	1.32	0.15
Q21	2.23	2.35	0.24
Q22	0.51	1.09	0.10
Q23	1.94	2.43	0.17

Figure 2: Item Characteristic Curves

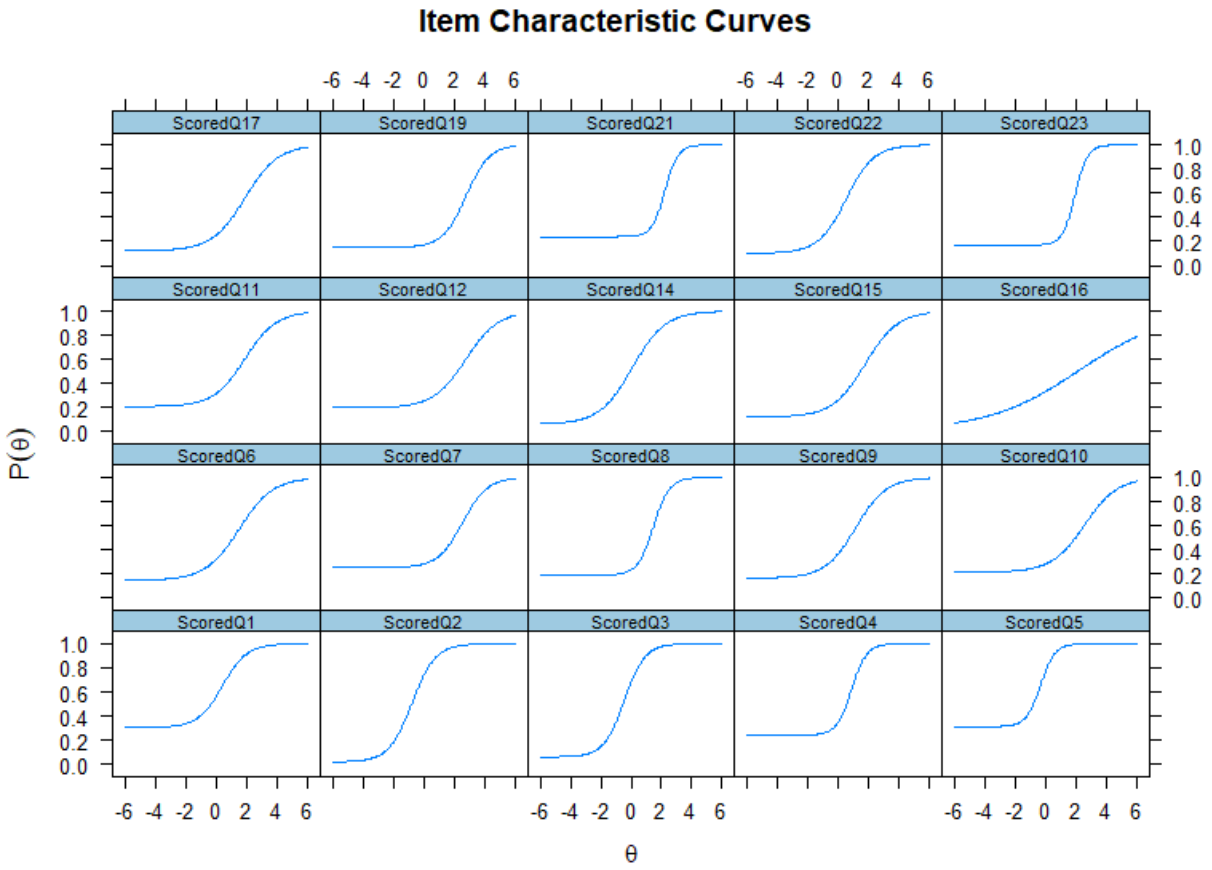


Figure 3: Item Person Map

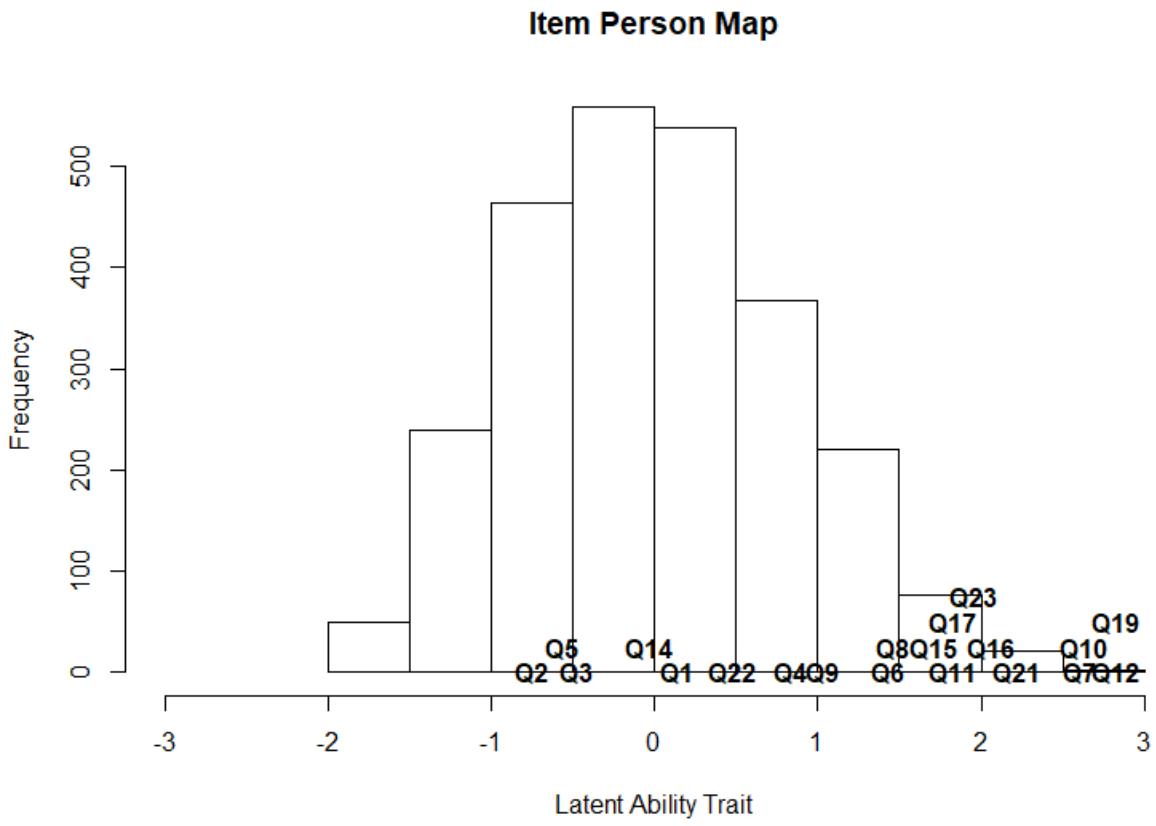


Figure 4: Conditional Reliability

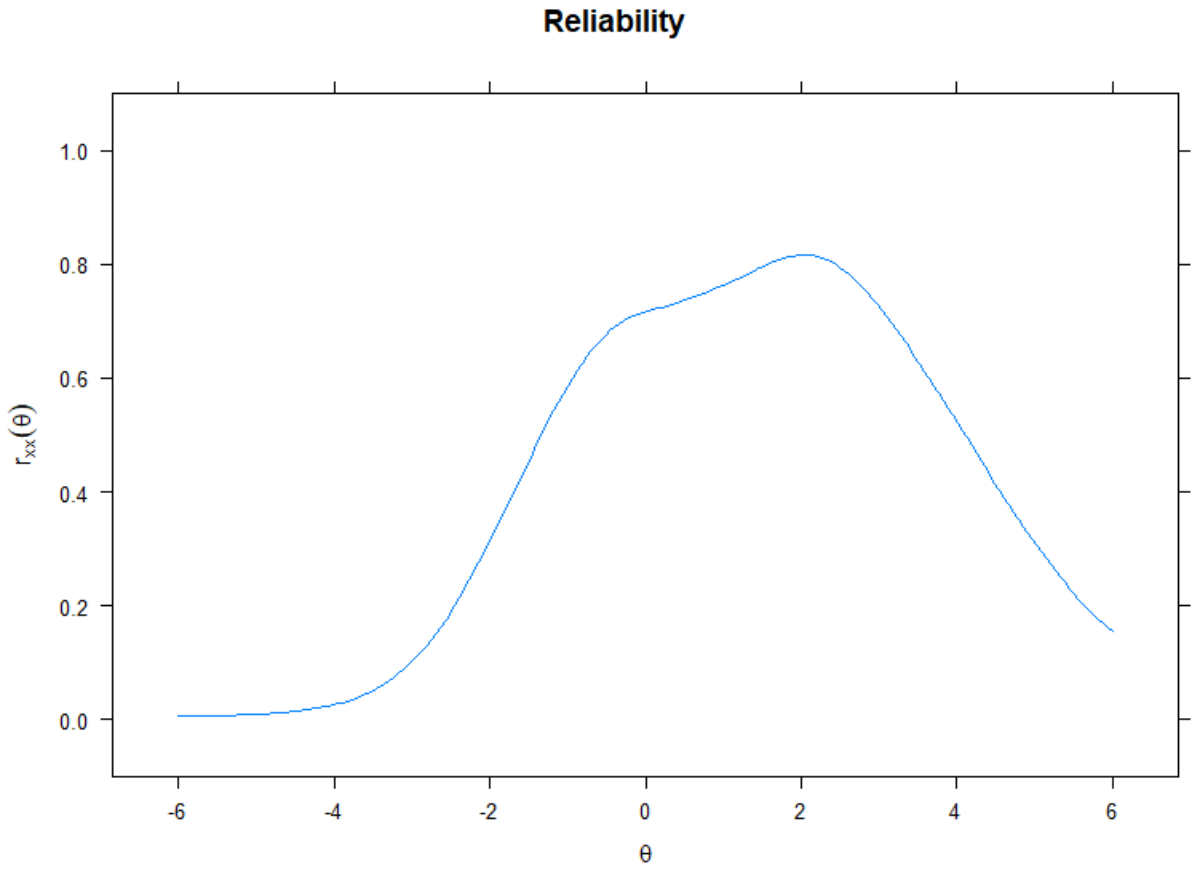


Figure 5: IRT True Score Distribution

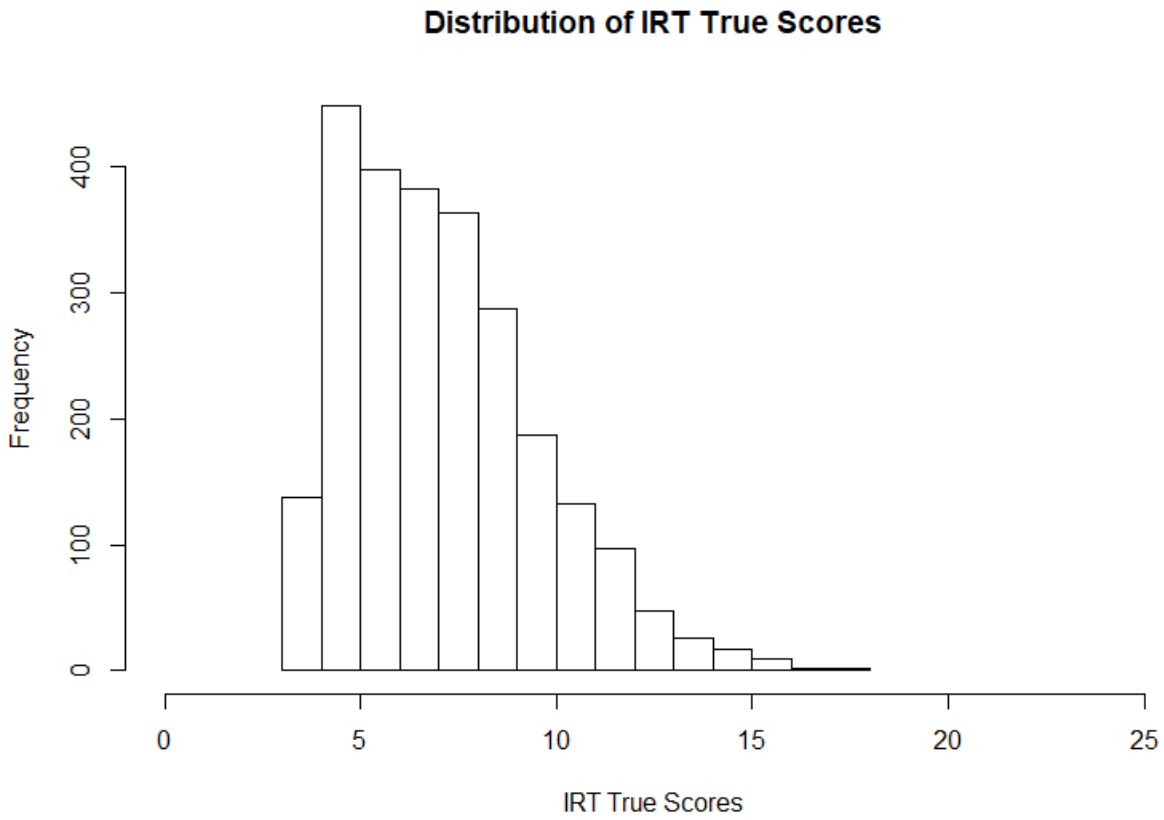
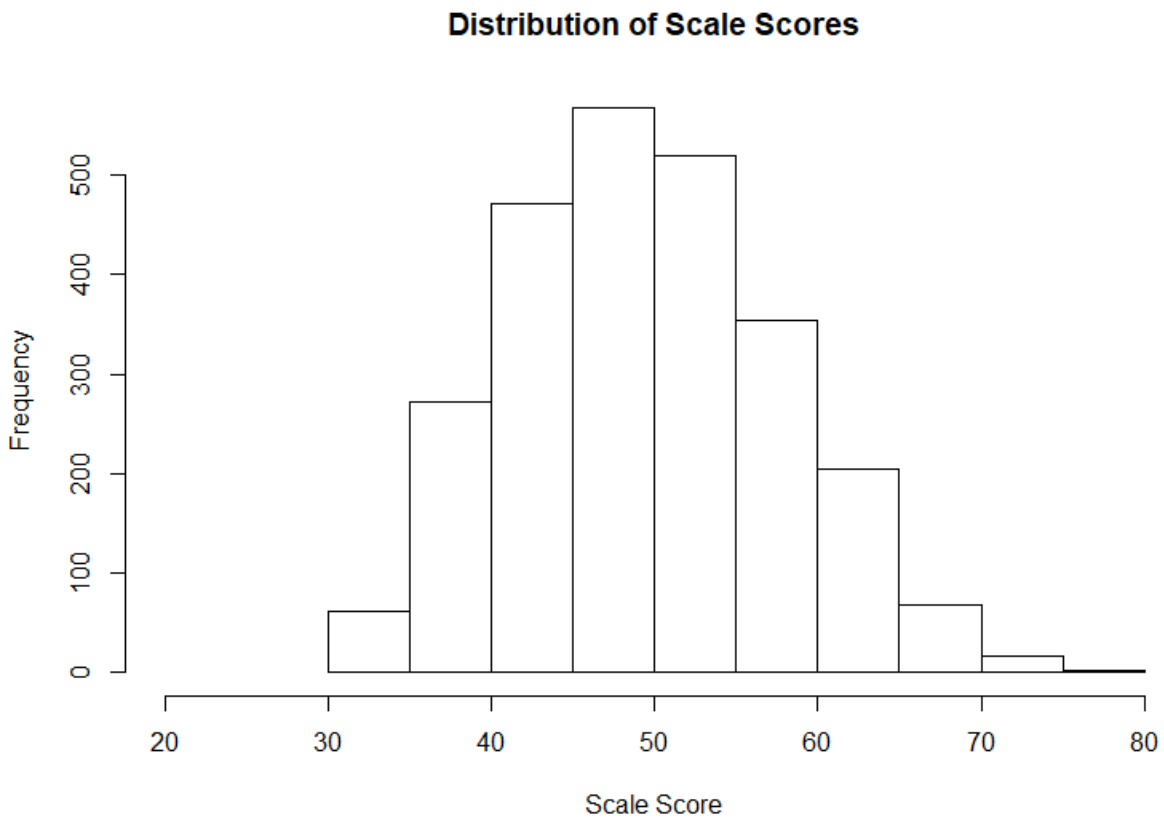


Figure 6: Scale Score Distribution



References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick's (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1-29.

Crocker, L. C., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

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

McDonald, R. P. (1999). *Test theory: A unified treatment.* Hillsdale, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Orlando, M. & Thissen, D. (2000). Liklihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64.