MC Test Analysis Report: Example Report

Jane Q. Doe

March 15, 2018

Contents

Introduction

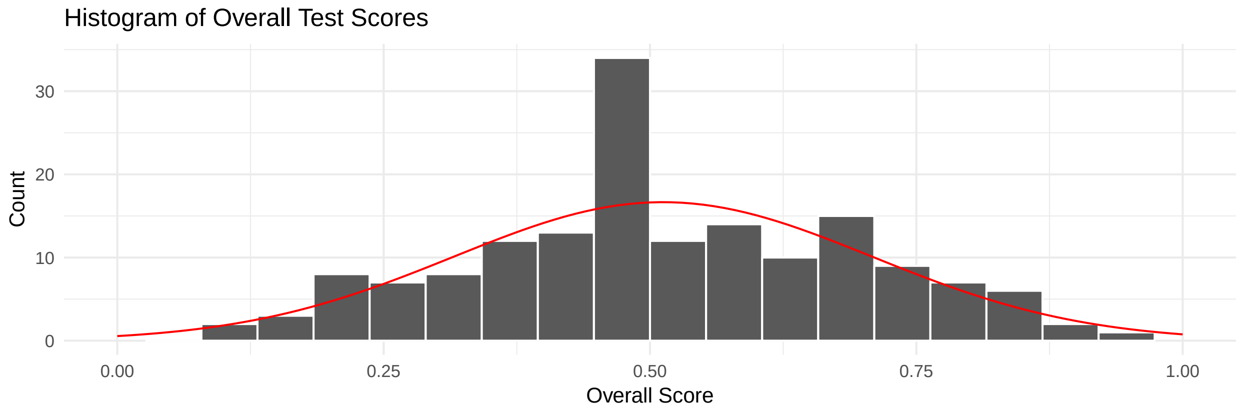
The purpose of this generated report is to provide the analytical framework proposed in the paper "An Analytic Framework for Evaluating the Validity of Concept Inventory Claims" (Jorion et al., 2015) from the University of Chicago, while providing extra statistical routines based on Classical Test Theory. Within the contents of this report, you will find graphical representations such as plots, graphs, and tables all intended to support an analysis of a multiple-choice test based on the framework proposed in Jorion's paper.

Test Overview and Descriptions

Answer Key

Question	Answer	Title	Concept
Q1	4	Question 1	A
Q2	4	Question 2	A
Q3	4	Question 3	A
Q4	1	Question 4	A
Q5	2	Question 5	A
Q6	2	Question 6	B
Q7	3	Question 7	B
Q8	3	Question 8	B
Q9	1	Question 9	B
Q10	4	Question 10	B
Q11	4	Question 11	C
Q12	3	Question 12	C
Q13	1	Question 13	C
Q14	3	Question 14	C
Q15	1	Question 15	C
Q16	3	Question 16	D
Q17	3	Question 17	D
Q18	1	Question 18	D
Q19	3	Question 19	D
Q20	3	Question 20	D

Overall Score Histogram



Option Selection by Item

The following table presents the percentage of students selecting each option by item.

Table 2: Option selection by item

Question	Title	Answer	Concept	1	2	3	4	Missing
Q1	Question 1	4	A	16	15	17.5	50	1.5
Q2	Question 2	4	A	18.5	10.5	13	57.5	0.5
Q3	Question 3	4	A	11.5	18.5	18	51	1
Q4	Question 4	1	A	54	17.5	12.5	14.5	1.5
Q5	Question 5	2	A	11	67	9	11.5	1.5
Q6	Question 6	2	B	22.5	43.5	14	18	2
Q7	Question 7	3	B	8	6.5	79	6.5	0
Q8	Question 8	3	B	14.5	15	45.5	25	0
Q9	Question 9	1	B	65.5	8	11.5	14.5	0.5
Q10	Question 10	4	B	8.5	7.5	5.5	78	0.5
Q11	Question 11	4	C	24	16.5	21	36.5	2
Q12	Question 12	3	C	19.5	17	46	17	0.5
Q13	Question 13	1	C	51	16	14.5	16.5	2
Q14	Question 14	3	C	13	16	56	14.5	0.5
Q15	Question 15	1	C	47.5	17	16.5	18.5	0.5
Q16	Question 16	3	D	28	30	16	24	2
Q17	Question 17	3	D	29.5	26	19.5	22.5	2.5
Q18	Question 18	1	D	62	14	8	16	0
Q19	Question 19	3	D	21.5	16.5	42	20	0
Q20	Question 20	3	D	21	26	26	26.5	0.5

Classic Test Theory

Summary

The following tables provide common statistical parameters used in Classic Test Theory (CTT).

Cronbach Alpha The coefficient of internal reliability, indicating how closely related the set of items are as a group.

Cronbach Alpha without item (WOI) The Cronbach Alpha calculated for the test without including the item of interest.

Subscale Alpha The Cronbach Alpha for the subscale or concept group. The value of alpha is influenced by test length, so it is expected that a low number of items per subscale will result in a lower subscale alpha value.

Difficulty Index Measures the proportion of students who answered the test item accurately. Higher values close to 1 are indicative of less difficult items (more students answered the item correctly), while lower values close to 0 are associated with more difficult items.

Discrimination Index Measures the ability of the item to discriminate between high and low scoring students. Positive values indicate that the students who scored well on the overall test tended to answer this question correctly, while students who scored poorly on the overall test were likely to answer this question incorrectly. Negative values indicate the opposite – low-scoring students were more likely to answer the question correctly, while high-scoring students tended to choose the wrong answer – and suggest that the item should be reviewed. Values near zero suggest the item does not differentiate between high- and low-performing students.

Item Variance Measures the spread among item responses.

Point-Biserial Correlation Coefficient (PBCC) Measures the Pearson correlation between a dichotomous variable, in this case the dichotomously scored item (correct/incorrect), and a continuous variable, in this case the overall test score.

Modified Point-Biserial Correlation Coefficient (Modified PBCC) Measures PBCC where item scores are correlated to overall test scores without considering the given item in the overall test score.

Test Summary

Table 3: Classic Test Theory Summary

	Value
Avg. Overall Score	0.5122
Cronbach Alpha	0.7468
Avg. Difficulty Index	0.5122
Avg. Discrimination Index	0.4514
Avg. PBCC	0.4148
Avg. Modified PBCC	0.309
Avg. Item Variance	0.2241

Test Summary by Concept Group

Table 4: Classic Test Theory Summary by Concept Group

	Subscale		Avg	Avg	Avg	Avg Item
Concept	Alpha	Avg Difficulty	Discrimination	PBCC	MPBCC	Var
A	0.543	0.58	0.553	0.465	0.359	0.242
B	0.422	0.63	0.39	0.385	0.28	0.211
C	0.395	0.488	0.411	0.373	0.257	0.247
D	0.487	0.35	0.452	0.436	0.34	0.197

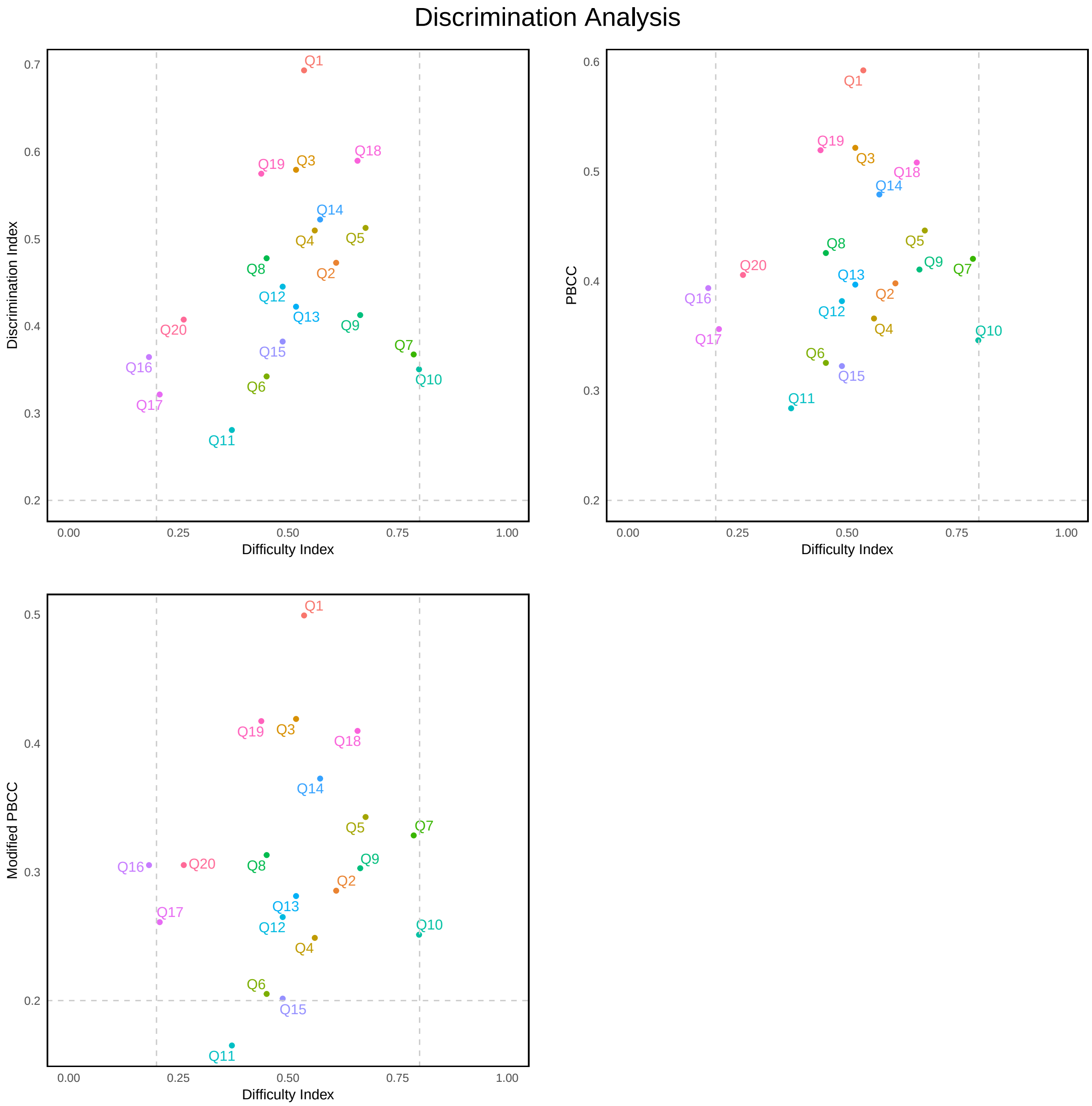
Test Summary by Item

Table 5: Classic Test Theory Summary by Item

Question	Title	Concept	Alpha WOI	Difficulty	Item Var	Discrimination	PBCC	MPBCC
Q1	Question 1	A	0.721	0.537	0.25	0.693	0.592	0.499
Q2	Question 2	A	0.739	0.61	0.239	0.473	0.398	0.285
Q3	Question 3	A	0.728	0.518	0.251	0.579	0.521	0.419
Q4	Question 4	A	0.742	0.561	0.248	0.51	0.366	0.249
Q5	Question 5	A	0.734	0.677	0.22	0.513	0.446	0.343
Q6	Question 6	B	0.745	0.451	0.249	0.342	0.325	0.205
Q7	Question 7	B	0.736	0.787	0.169	0.367	0.42	0.328
Q8	Question 8	B	0.737	0.451	0.249	0.478	0.425	0.313
Q9	Question 9	B	0.737	0.665	0.224	0.413	0.41	0.303
Q10	Question 10	B	0.741	0.799	0.162	0.35	0.346	0.251
Q11	Question 11	C	0.748	0.372	0.235	0.281	0.284	0.165
Q12	Question 12	C	0.741	0.488	0.251	0.445	0.382	0.265
Q13	Question 13	C	0.739	0.518	0.251	0.422	0.397	0.281
Q14	Question 14	C	0.732	0.573	0.246	0.522	0.479	0.373
Q15	Question 15	C	0.746	0.488	0.251	0.382	0.322	0.202
Q16	Question 16	D	0.738	0.183	0.15	0.364	0.393	0.305
Q17	Question 17	D	0.74	0.207	0.165	0.321	0.356	0.261
Q18	Question 18	D	0.729	0.659	0.226	0.59	0.508	0.41
Q19	Question 19	D	0.728	0.439	0.248	0.575	0.519	0.417
Q20	Question 20	D	0.737	0.262	0.195	0.407	0.405	0.305

Item Discrimination

The below scatter plots compare the three measures of item discrimination with the item difficulty. Dotted guidelines indicate the recommended ranges for each index. Note that the full difficulty index range is from 0 to 1 and the full range of the discrimination indices is from -1 to 1 — although a discrimination index (or PBCC or Modified PBCC) of less than 0.2 is not recommended.



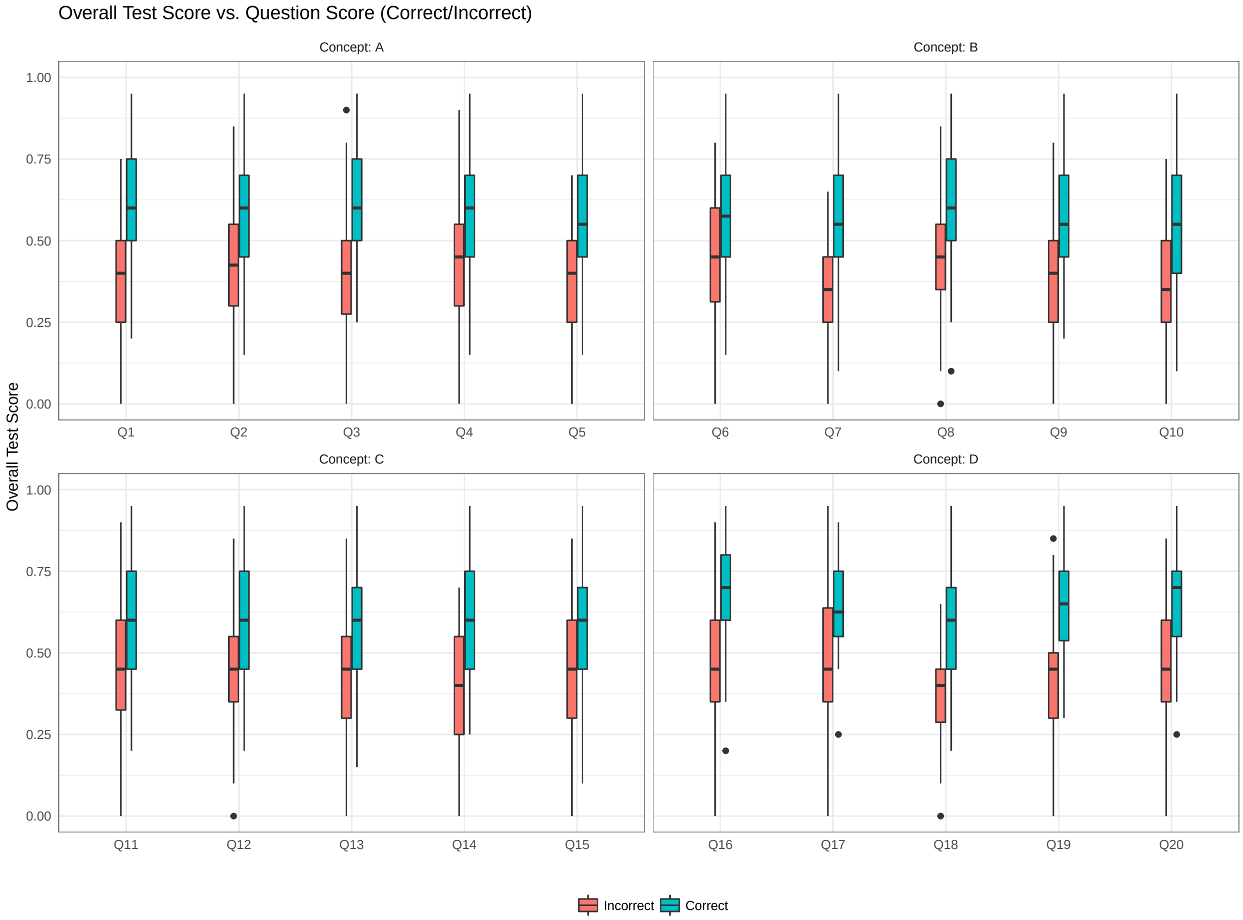
Overall Score vs. Question Score

The following plot compares the respondents' performance on a single item (correct or incorrect) to their overall score on the test. The plots are organized by concept group, and within in subplot, the boxplot displays the range of overall test scores among the respondents who correctly and incorrectly answered each question.

Intuitively, a question for which there is very little overlap between the boxplots of the correct and incorrect group is more discerning between the high and low performing students. Questions for which the box plots are mostly overlapping are not as good at differentiating between students.

Additionally, the range of each boxplot indicates whether the question is correctly (or incorrectly) answerd by students with a wide range of overall performance or more consistently by a students of a particular overal ability.

Generally, it is best for the boxplot of the correct group to be mostly above the boxplot of the incorrect group. Questions that have complete overlap between the two boxplots should be reviewed.



Distractor Analysis

The following plot and table compare the percentage of all respondents who select a given option for each item. These tables allow the test administrator to analize the performance of item options and to determine if the choice of distracting items reveals information about the misconceptions in students' knowledge. Repondents are grouped into the upper and lower 33rd percentiles by overall test score. For this report, there were 64 respondents in the upper 33rd percentile and 54 repondents in the lower 33rd percentile. Percentages are calculated relative to the total number of respondents, in this case 164 students.

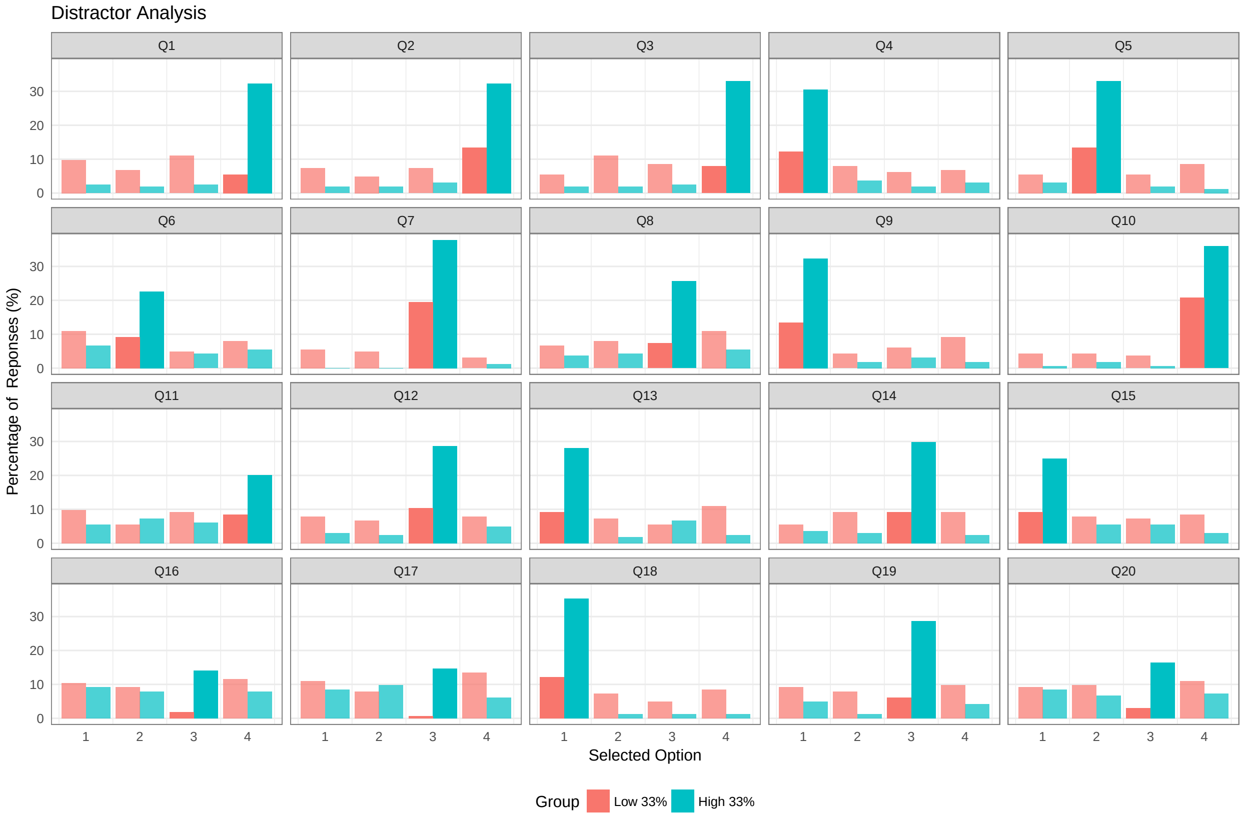


Table 6: Percentage of total respondents ($N = 164$) from upper (*High*, $N = 64$) and lower (*Low*, $N = 54$) 33rd percentiles having chosen each item option. The percentage of students choosing the correct option for each item are highlighted in bold.

Question	Title	1H	1L	2H	2L	3H	3L	4H	4L
Q1	Question 1	2.44	9.76	1.83	6.71	2.44	10.98	**32.32**	**5.49**
Q2	Question 2	1.83	7.32	1.83	4.88	3.05	7.32	**32.32**	**13.41**
Q3	Question 3	1.83	5.49	1.83	10.98	2.44	8.54	**32.93**	**7.93**
Q4	Question 4	**30.49**	**12.20**	3.66	7.93	1.83	6.10	3.05	6.71
Q5	Question 5	3.05	5.49	**32.93**	**13.41**	1.83	5.49	1.22	8.54
Q6	Question 6	6.71	10.98	**22.56**	**9.15**	4.27	4.88	5.49	7.93
Q7	Question 7	0.00	5.49	0.00	4.88	**37.80**	**19.51**	1.22	3.05
Q8	Question 8	3.66	6.71	4.27	7.93	**25.61**	**7.32**	5.49	10.98
Q9	Question 9	**32.32**	**13.41**	1.83	4.27	3.05	6.10	1.83	9.15
Q10	Question 10	0.61	4.27	1.83	4.27	0.61	3.66	**35.98**	**20.73**
Q11	Question 11	5.49	9.76	7.32	5.49	6.10	9.15	**20.12**	**8.54**
Q12	Question 12	3.05	7.93	2.44	6.71	**28.66**	**10.37**	4.88	7.93
Q13	Question 13	**28.05**	**9.15**	1.83	7.32	6.71	5.49	2.44	10.98
Q14	Question 14	3.66	5.49	3.05	9.15	**29.88**	**9.15**	2.44	9.15
Q15	Question 15	**25.00**	**9.15**	5.49	7.93	5.49	7.32	3.05	8.54
Q16	Question 16	9.15	10.37	7.93	9.15	**14.02**	**1.83**	7.93	11.59
Q17	Question 17	8.54	10.98	9.76	7.93	**14.63**	**0.61**	6.10	13.41
Q18	Question 18	**35.37**	**12.20**	1.22	7.32	1.22	4.88	1.22	8.54
Q19	Question 19	4.88	9.15	1.22	7.93	**28.66**	**6.10**	4.27	9.76
Q20	Question 20	8.54	9.15	6.71	9.76	**16.46**	**3.05**	7.32	10.98

Item Review Recommendations

Review Recommendations Criteria

Alpha If *Cronbach's Alpha* for the test with the item deleted is less than the alpha coefficient for the whole test then the recommendation is to **Keep** the item.

Jorion If the *Difficulty Index* is between 0.3 and 0.9, and the *Discrimination Index* is greater than 0.2, then the recommendation is to **Keep** the item.

Versatile This recommendation is based on the *Difficulty Index* and *PBCC* and provides a range of recommendations from **Remove** to **Review** through **Keep**, favoring positive PBCC values near to or greater than 0.3 and higher difficulty values. The criteria for this recommendation are based the criteria published by Sleeper (2011), reproduced below.

Stringent If the *Difficulty Index* is between 0.3 and 0.9, and the *Point-Biserial Correlation Coefficient* is greater than 0.3, then the recommendation is to **Keep** the item.

"Versatile" Recommendation Criteria

The *Versatile* recommendation criteria are based on criteria published by Sleeper (2011). The table below reproduces the source material that is unfortunately no longer available online.

Table 7: *Versatile* recommendation criteria from Sleeper (2011)

Difficulty Score (%)	PBCC $[0.3, 1.0]$	PBCC $[0.15, 0.3)$	PBCC $[0.0, 0.15)$	PBCC $[-1, 0)$
$[0, 30]$	Review	Review/Remove	Remove	Remove
$(30, 50]$	Keep (Tough)	Review	Review/Remove	Remove
$(50, 80]$	Keep	Keep	Review/Keep	Review
$(80, 100]$	Keep	Keep	Keep (Easy)	Review

Review Recommendations Table

Table 8: Recommendations for each test item based on the criteria described above.

Question	Title	Concept	Check Alpha	Check Jorion	Check Versatile	Check Stringent
Q1	Question 1	A	Keep	Keep	Keep	Keep
Q2	Question 2	A	Keep	Keep	Keep	Keep
Q3	Question 3	A	Keep	Keep	Keep	Keep
Q4	Question 4	A	Keep	Keep	Keep	Keep
Q5	Question 5	A	Keep	Keep	Keep	Keep
Q6	Question 6	B	Keep	Keep	Keep (Tough)	Keep
Q7	Question 7	B	Keep	Keep	Keep	Keep
Q8	Question 8	B	Keep	Keep	Keep (Tough)	Keep
Q9	Question 9	B	Keep	Keep	Keep	Keep
Q10	Question 10	B	Keep	Keep	Keep	Keep
Q11	Question 11	C	Remove	Keep	Review	Remove
Q12	Question 12	C	Keep	Keep	Keep (Tough)	Keep
Q13	Question 13	C	Keep	Keep	Keep	Keep
Q14	Question 14	C	Keep	Keep	Keep	Keep
Q15	Question 15	C	Keep	Keep	Keep (Tough)	Keep
Q16	Question 16	D	Keep	Remove	Review	Remove
Q17	Question 17	D	Keep	Remove	Review	Remove
Q18	Question 18	D	Keep	Keep	Keep	Keep
Q19	Question 19	D	Keep	Keep	Keep (Tough)	Keep
Q20	Question 20	D	Keep	Remove	Review	Remove

Item Response Theory

Model Summary

The model selected by the user for this analysis was the one-factor logistic model, which had an AIC of 3986.9.

Model Parameters

Difficulty The difficulty parameter, β, sometimes called the threshold parameter, describes the difficulty of a given item. It is the only parameter estimated in the 1-PL (Rasch) model.

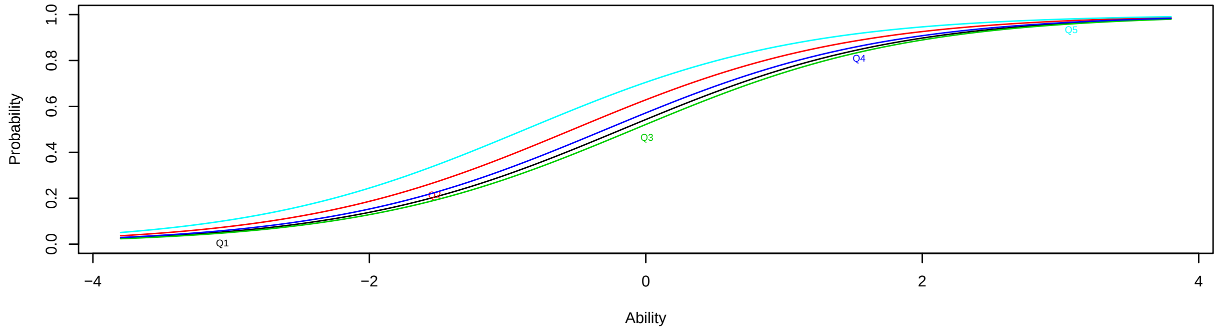
Discrimination In the 1-PL Rasch model, the discrimination parameter is assumed to be equivalent across all items. This assumption leads to consistent ICC curves where more difficult questions are always less easy for all students. When the discrimination parameter is allowed to vary, for two items of similar difficulty one item can be both easier for low-performing students and harder for high-performing students when compared with the second item (or vice-versa).

Prob. The probability column gives the probability that an average student will correctly answer the item, i.e. $P(x_i = 1|z = 0)$.

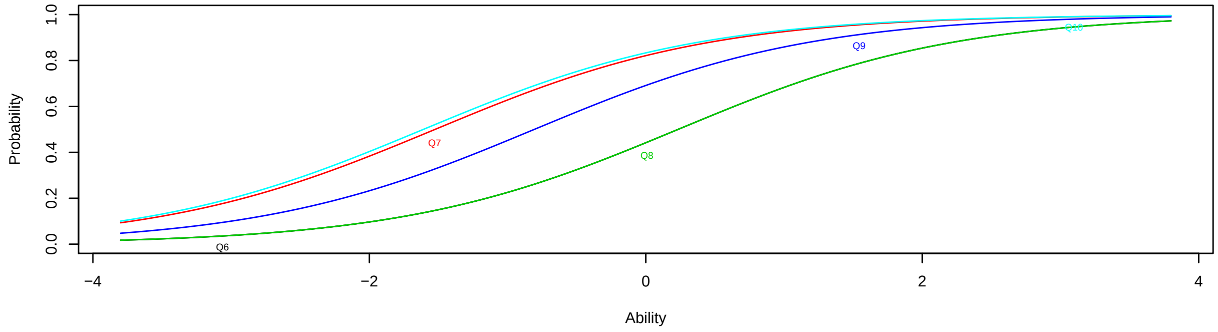
Question	Difficulty	Discrimination	$P(x_i = 1\|z = 0)$
Q1	-0.1725	1	0.543
Q2	-0.5269	1	0.6288
Q3	-0.0857	1	0.5214
Q4	-0.2892	1	0.5718
Q5	-0.8711	1	0.705
Q6	0.2334	1	0.4419
Q7	-1.525	1	0.8212
Q8	0.2334	1	0.4419
Q9	-0.8067	1	0.6914
Q10	-1.609	1	0.8333
Q11	0.6211	1	0.3495
Q12	0.0593	1	0.4852
Q13	-0.0855	1	0.5214
Q14	-0.348	1	0.5861
Q15	0.0594	1	0.4851
Q16	1.745	1	0.1487
Q17	1.569	1	0.1724
Q18	-0.7746	1	0.6845
Q19	0.2919	1	0.4275
Q20	1.217	1	0.2284

Item Characteristic Curves

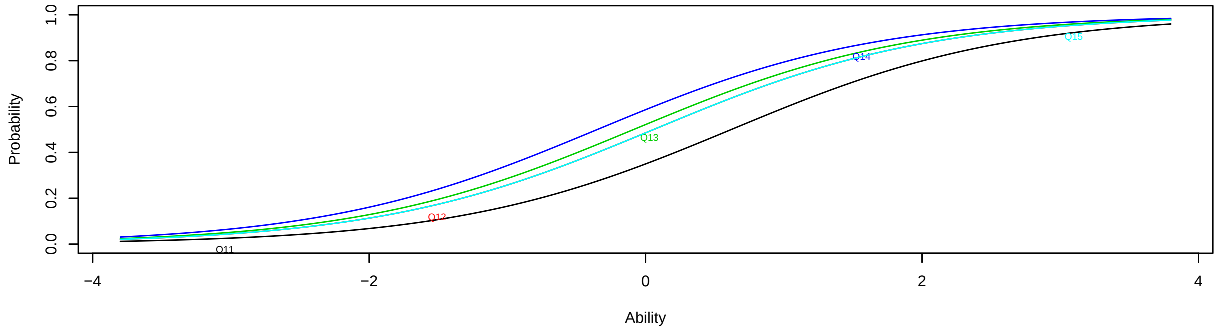
Item Characteristic Curves: Concept "A"



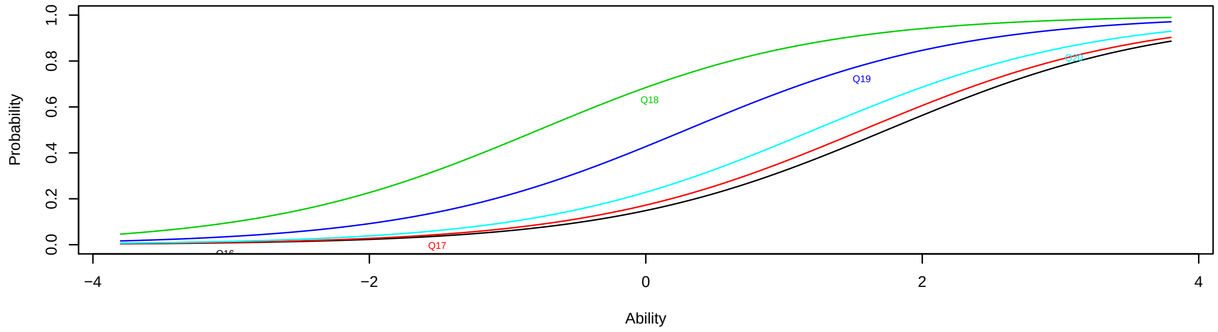
Item Characteristic Curves: Concept "B"



Item Characteristic Curves: Concept "C"



Item Characteristic Curves: Concept "D"

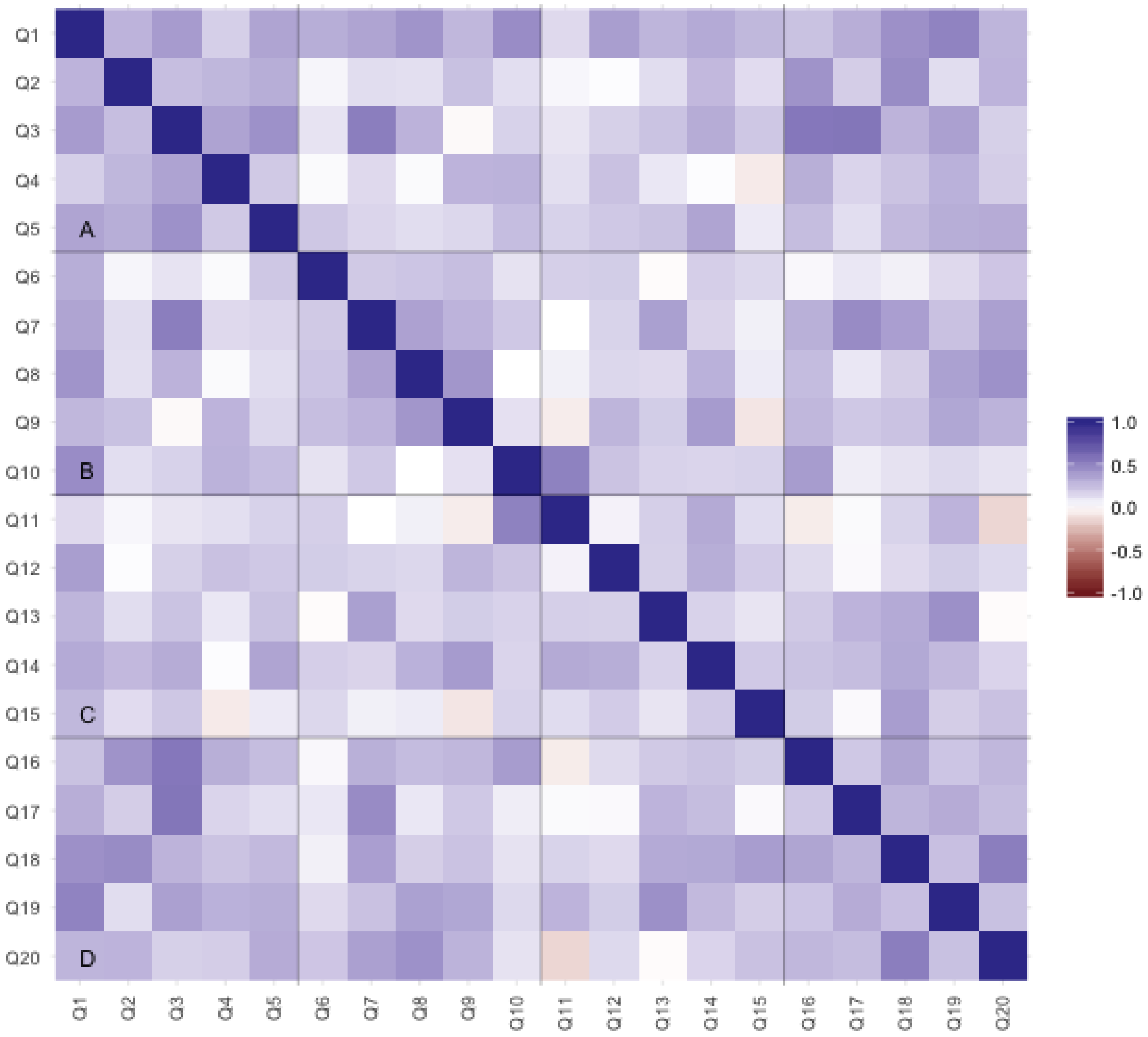


Introductory Factor Analysis

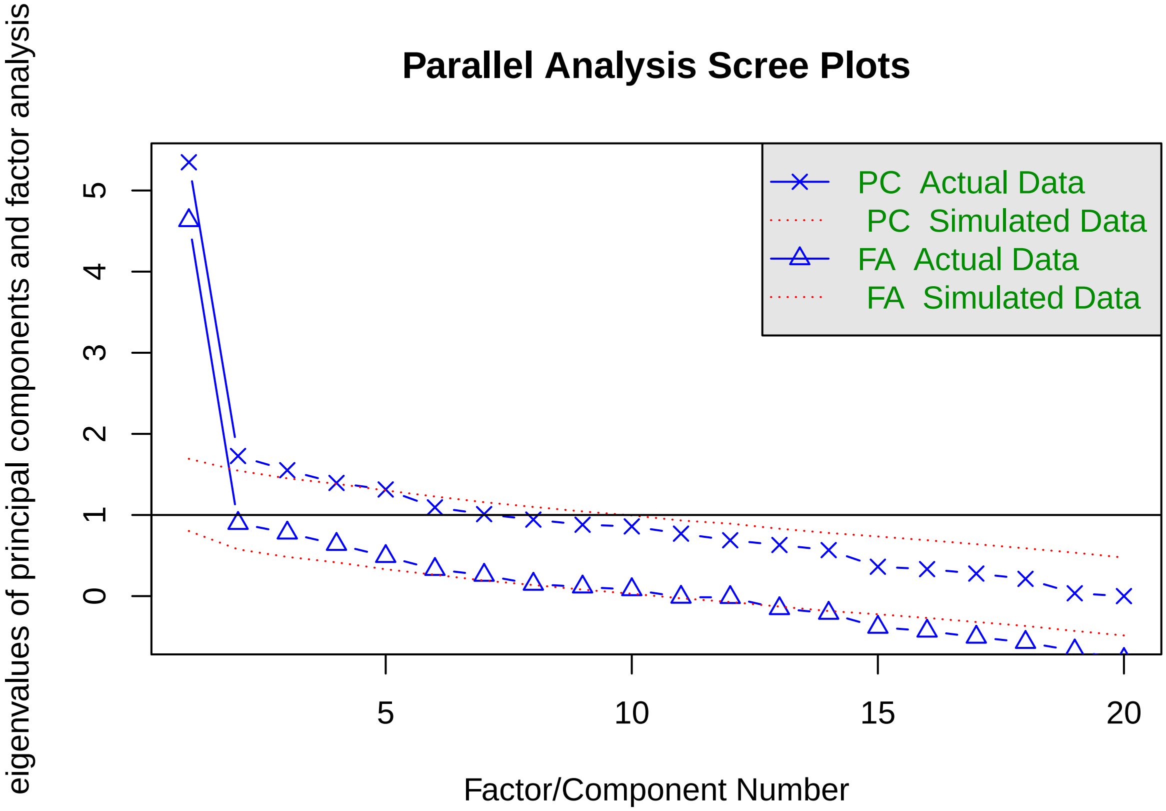
Tetrachoric Plot

The following plot shows the item-by-item *tetrachoric correlation* for all questions in the test. The tetrachoric correlation estimates the correlation between two variables whose measurement is artificially dichotomized but whose underlying joint ditribution is a bivariate normal distribution. In the case of Item Response Theory, the tetrachoric correlation is seen as the correlation between the response to two items when "each item is assumed to represent an underlying ability which is reflected as a probability of responding correctly to the item and the items are coded as correct or incorrect" (Revelle, 2017).

Considering that the tetrachoric correlation matrix represents item correlations for all test items, then the structure of this matrix directly correpsonds to the structure of the underlying latent variables measured by the test. This is — at a very high level — the goal of factor analysis. For more information, the Personality Project webpage provides excellent resources. The tetrachoric correlation plot is included here for visual inspection of the underlying structure, as this matrix will be used in the factor analysis that follows.


Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q18
Q19
Q20
A
B
C
D
1.0
0.5
0.0
-0.5
-1.0

Scree Plot



A method for determining the number of factors or components in the tetrachoric correlation matrix of the test responses is to examine the scree plot of the eigenvalues of the correlation matrix. Typically, when using a scree plot, the analyst is looking for a sharp break in the slope of the line between the eigenvalues of the correlation matrix. In parallel analysis, the scree of factors from the observed data is compared to that of a random data matrix of the same size as the observed. Parallel analysis suggests a number of factors/components by comparing the eigenvalues of the factors/components of the observed data to the random data and keeping those that are greater than the random data.

Parallel analysis for the test results in this report suggest that the number of factors is 12 and the number of components is 5.

Exploratory Factor Analysis

The table below presents the factor loadings, where 12 were explored, using the `fa()` function from the `psych` package (see Revelle (2016) for more information on the options available for this function). In this report, the EFA used the `'varimax'` rotation method and the `'minres'` factoring method. Factors with absolute value loadings less than 0.3 were suppressed.

Table 10: Exploratory Factor Analysis with 12 factors using `'varimax'` rotation and `'minres'` factoring.

Question	Concept	MR1	MR10	MR11	MR12	MR2	MR3	MR4	MR5	MR6	MR7	MR8	MR9
Q1	A	-	-	0.462	-	-	-	-	-	-	-	-	-
Q2	A	-	-	-	0.309	-	-	-	-	0.430	-	-	-
Q3	A	0.742	-	-	0.432	-	-	-	-	-	-	-	-
Q4	A	-	-	-	-	-	-	-	-	-	-	-	0.955
Q5	A	-	-	-	-	-	-	-	-	-	0.946	-	-
Q6	B	-	-	-	-	-	0.965	-	-	-	-	-	-
Q7	B	0.566	-	-	-	-	-	-	-	-	-	-	-
Q8	B	-	-	-	-	-	-	-	-	-	-	0.953	-
Q9	B	-	-	-	-	-	-	-	0.816	-	-	-	-
Q10	B	-	-	-	-	0.920	-	-	-	-	-	-	-
Q11	C	-	0.945	-	-	-	-	-	-	-	-	-	-
Q12	C	-	-	0.582	-	-	-	-	-	-	-	-	-
Q13	C	-	-	-	-	-	-	0.953	-	-	-	-	-
Q14	C	-	0.406	0.421	-	-	-	-	0.366	-	-	-	-
Q15	C	-	-	0.387	-	-	-	-	-	0.384	-	-	-
Q16	D	-	-	-	0.880	-	-	-	-	-	-	-	-
Q17	D	0.757	-	-	-	-	-	-	-	-	-	-	-
Q18	D	-	-	-	-	-	-	-	-	0.865	-	-	-
Q19	D	-	-	-	-	-	-	0.324	-	-	-	-	-
Q20	D	-	-	-	-	-	-	-	-	0.583	-	0.339	-

References

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment; Evaluation. Retrieved from http://echo.edres.org:8080/irt/baker/

Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences* (1st ed.). Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, *39*(1), 62–79. https://doi.org/10.1177/0146621614561315

Fletcher, T. D. (2010). *Psychometric: Applied psychometric theory*. Retrieved from https://CRAN.R-project.org/package=psychometric

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, *2*(2), 61–103. https://doi.org/10.1207/s15366359mea0202/_1

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*(4), 454–496. https://doi.org/10.1002/jee.20104

Revelle, W. (2016). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University; https://CRAN.R-project.org/package=psych. Retrieved from https://CRAN.R-project.org/package=psych

Revelle, W. (2017). Northwestern University; http://personality-project.org/r/book/.

Rizopoulos, D. (2006). Ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05/

Sleeper, R. (2011). Keep, toss or revise? Tips for post-exam item analysis. http://www.ttuhsc.edu/sop/administration/enhancement/documents/Sleeper_Handout.ppt (URL no longer valid).