Online Gibson Cognitive Skills Test

Phase 1 Report

Item Analysis, Test Reliability, and Effects of Early Termination Rule

March 10, 2008

Vonda Kiplinger, Ph. D.
WindWalker Consulting
Lafayette, Colorado

Background

This report describes the results of Phase I of the evaluation of the online Gibson Cognitive Skills Test. Examined are item characteristics, test reliability and impact of the early termination rule on examinee performance.

The online Gibson Cognitive Skills Test (GCST) is designed as a comprehensive online assessment of individuals' level of cognitive skills development. This assessment is based on the cognitive skills training and assessments developed by Dr. Ken Gibson of LearningRx. The cognitive skills subtests of the GCST administered in the online assessment and included in this evaluation are:

- Processing Speed
- Working Memory
- Long-Term Memory
- Word Attack
- Visual Processing
- Auditory Analysis – Segmenting and Drop
- Logic and Reasoning

Item Response Theory (IRT) is used to analyze the results of the online Gibson Test and was implemented with the BILOG-MG software.

The primary techniques for investigating item adequacy and empirical test reliability are:

- Test reliability coefficient
- Test-retest reliability
- Item difficulty
- Item-test correlation
- Item discrimination
- Differential Item Functioning (DIF)
- Impact of the early termination rule

Definition of Terms:

- *The test reliability coefficient*, in this case, is an *internal consistency* measure across all items within a test. When examinees perform consistently across items within a test, this is taken as an indication that the items are measuring the same type of performance or represent the same cognitive domain. In addition, higher reliability coefficients mean lower measurement error. The highest values reported for commercially available achievement tests are in the 0.70s, 0.80s and 0.90s.

- *Test-retest reliability* indicates how consistently a test measures examinee performance when the same examinees are administered the same test on two different measurement occasions. Scores from the two testing occasions are correlated to produce a *coefficient of stability*. Few, if any, standards exist for judging the minimum acceptable value for a test-retest reliability estimate. The highest values reported for commercially available achievement tests are in the 0.70s, 0.80s and 0.90s. In order to estimate test-retest reliability, a sample of 51 examinees was re-administered each subtest approximately

seven weeks after the initial administration. No online cognitive training occurred during the intervening period.

- *Item-test correlation* indicates the strength of the relationship between examinee response to a particular item and examinee total test score. The higher the correlation, the stronger the relationship (maximum value 1.0).

- *Item discrimination* describes how well an item discriminates among examinees of lower ability and those of higher ability. Lower values indicate less item discrimination, while higher values indicate greater discrimination. (High item discrimination is a good thing.)

- *Differential Item Functioning (DIF)* occurs when a test item "functions" differently for examinees in different groups (e.g., males vs. females). For example, DIF is indicated if males and females are at the same ability level, but one gender typically answers the item incorrectly, while the other typically responds correctly. It this case, there is some type of item bias related to gender.

- *Impact of the early termination rule*. All but two subtests, Working Memory and Long-Term Memory, are terminated early if the examinee responds incorrectly to three consecutive items. In order to evaluate the impact of early termination, a sample of 153 examinees was allowed to complete the subtests with the early termination rule "turned off." The sample was then scored (1) as if the early termination rule was in effect[1] and (2) with it turned off[2]. The results of these two "scorings" are compared in order to evaluate the impact of this rule.

Results

The results of analyses are discussed below. A summary and implications for test construction also are provided. *Test Construction Implications* for each subtest is highlighted in a text box at the end of the discussion of each subtest.

Processing Speed

Reliability

Analysis indicates that the Processing Speed subtest is highly reliable. The *internal consistency reliability coefficient* is 0.95 (1.00 is maximum value). Tests whose reliability coefficients are in the 0.8 and 0.9 ranges are considered highly reliable.

The *test-retest reliability (stability coefficient)* for the Processing Speed subtest is moderate in value, 0.57. Overall, examinees performed significantly better on the retest than on the initial test. For discussion purposes, the score metric is terms of percent correct. The average score on the retest is 77, as compared to an initial test average of 74. Approximately seven weeks elapsed between initial and second administration of the test. Therefore, it is unlikely that the difference is the result of a "practice effect."

[1] Scores were computed based on responses given prior to the examinee missing three consecutive items. All responses after the three consecutive misses were discounted.
[2] All examinee responses were included in the scoring.

Item Difficulty and Item-Test Correlation

The item-test correlations for the Processing Speed subtest are all relatively low (<0.68), but that might not be unexpected given the nature of the subtest. This is a speed test in which initial items are required to be easy. The first 24 items out of a total of 45 items were answered correctly by 90-99 percent of examinees. Fifty percent or more examinees answered the first 36 items correctly.

Four of the items exhibit item-test correlations < 0.2 and also are among the easiest of items. These items are #s 1, 2, 4, and 7. Table A-1 in the Appendix provides the results from the item analyses produced by the IRT program.

Item Discrimination

The same four items also exhibit the lowest item discrimination. That is, they do not adequately discriminate among examinees of lower and higher speed processing abilities.

DIF

None of the items exhibit significant gender bias, or differential item functioning (DIF).

Impact of Early Termination Rule

With the early termination rule "turned off," 42 of the 153 examinees are able to correctly answer from 1 to 30 additional items. The average subtest percent correct when the early termination rule is in effect is 64, compared to 65 without early termination.

Implications

Test Construction Implications for the Processing Speed Subtest

Evaluation of this test strongly suggests that 4 items could be dropped from the test (with no adverse effects on reliability or amount of test information provided). Items 1, 2, 4 and 7 should be dropped for the following reasons:

- These four items appear unrelated to total test score (correlation coefficients < 0.2).
- The same four items do not adequately discriminate among examinees of lower and higher abilities related to processing speed.

This is a rather lengthy test (45 items), and fatigue may be a factor in end-of-test performance. Dropping items 1, 2, 4 and 7 should not adversely affect test performance and could very likely enhance performance, as well as the reliability and validity of the test.

Early termination appears to have a slight negative effect on examinee performance.

Working Memory

Reliability

The *reliability (internal consistency)* of the Working Memory subtest is somewhat lower (0.73) than that of Processing Speed (0.95), which is not unexpected given the shorter length of the subtest (20 items).

The *test-retest reliability (stability)* is relatively high, 0.66, and while examinees appeared to perform somewhat better on the retest (65 vs. 63 percent correct), the difference is not statistically significant.

Item Difficulty, Item-Test Correlation, Item Discrimination, and DIF

Item difficulties range from 22 percent correct response to item 14 to 96 percent correct for item 1. Table A-2 provides the results from the item analyses produced by the IRT program.

The item-test correlations for the Working Memory subtest are all relatively low, ranging from 0.06 to 0.46. The results of the item difficulty, item- test correlation, item discrimination, and DIF analyses are summarized below.

The Working Memory subtest consists of only 20 items and their "behavior" is somewhat erratic. Of the 20 items:

- Three items are extremely easy, with correct responses from approximately 90 – 97 % of examinees (in order of difficulty, items 6, 5, 1).
- Seven appear unrelated to total subtest score (correlation coefficients < 0.2), and all items are less than 0.45 (1.0 is the maximum value).
- Ten do not adequately discriminate among examinees with lower and higher levels of working memory.
- Four demonstrate significant differential item functioning (DIF) with regard to gender, as shown in Figures 1-5.

There is a high degree of overlap among these items with regard to low item-test correlations, low discrimination and high DIF, as illustrated in Table 1.

Table 1. Summary of Item Analysis Findings for the Working Memory Subtest

Item	Low Item-Test Correlation	Low Discrimination	High DIF
1	√	√[a]	
3		√	√
5	√	√	
6	√	√	√
8	√	√	
12			√
13		√	
14	√	√	
15	√	√	
16		√	
17	√	√	
19			√

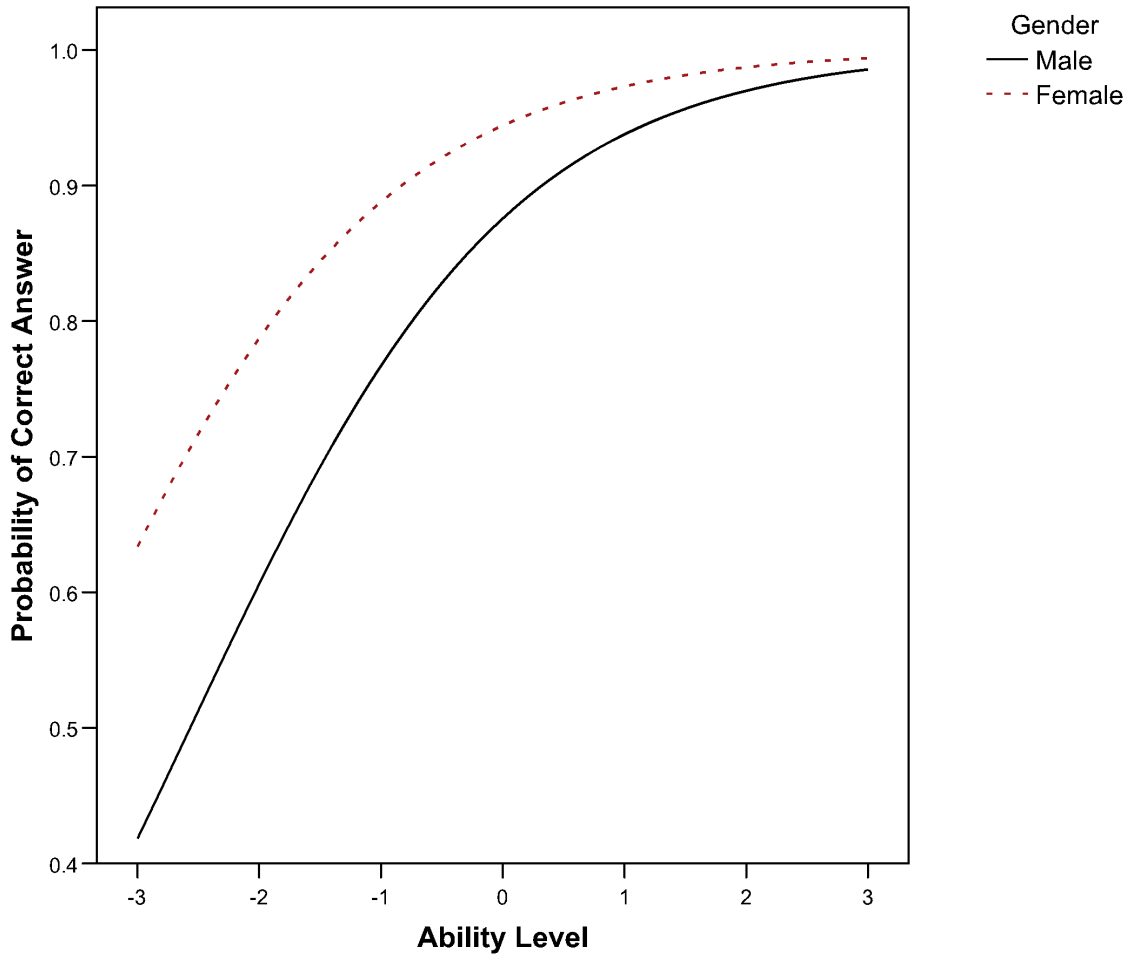
[a] This is the easiest item on the subtest. 97% responded correctly; therefore, there is no discrimination between students of high and low ability.

Items not included in the above table do not demonstrate low item-test correlations, low discrimination, or high DIF. These are Items 2, 4, 7, 10, 11, 18, and 20.

The nature of the item bias, as indicated by the DIF analysis, is discussed following each figure presented below.

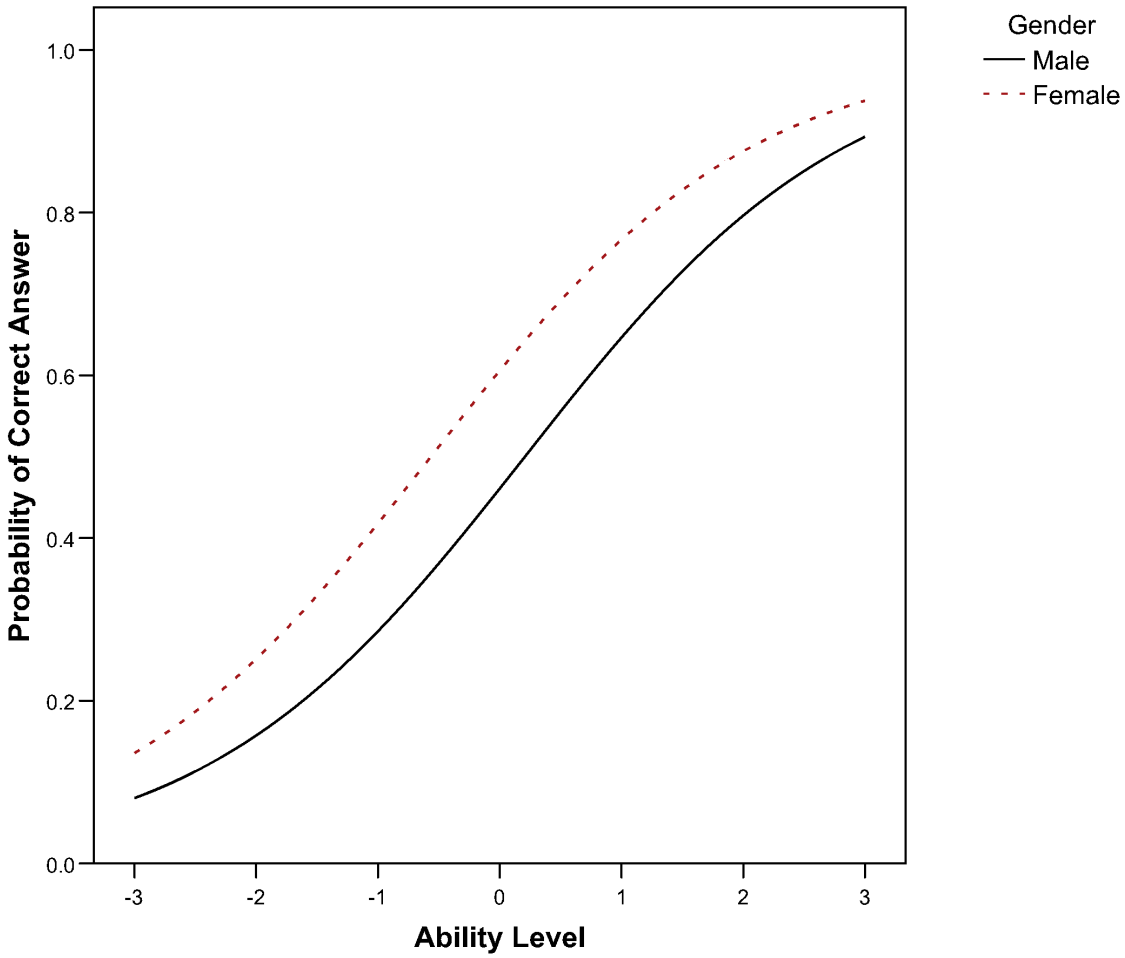
For each figure, the X (horizontal) axis represents values of the "ability" scale for the trait (e.g., working memory) being measured. In the Item Response Theory metric, a value of zero (0) denotes average ability, negative values indicate below average ability, and positive values indicate above average ability. The more negative or more positive the value, the lower or higher the ability level, respectively. The Y (vertical) axis indicates the probability that an examinee, at any given ability level, will answer the item correctly.

Figure 1. DIF Curves for Item 6, Working Memory Subtest



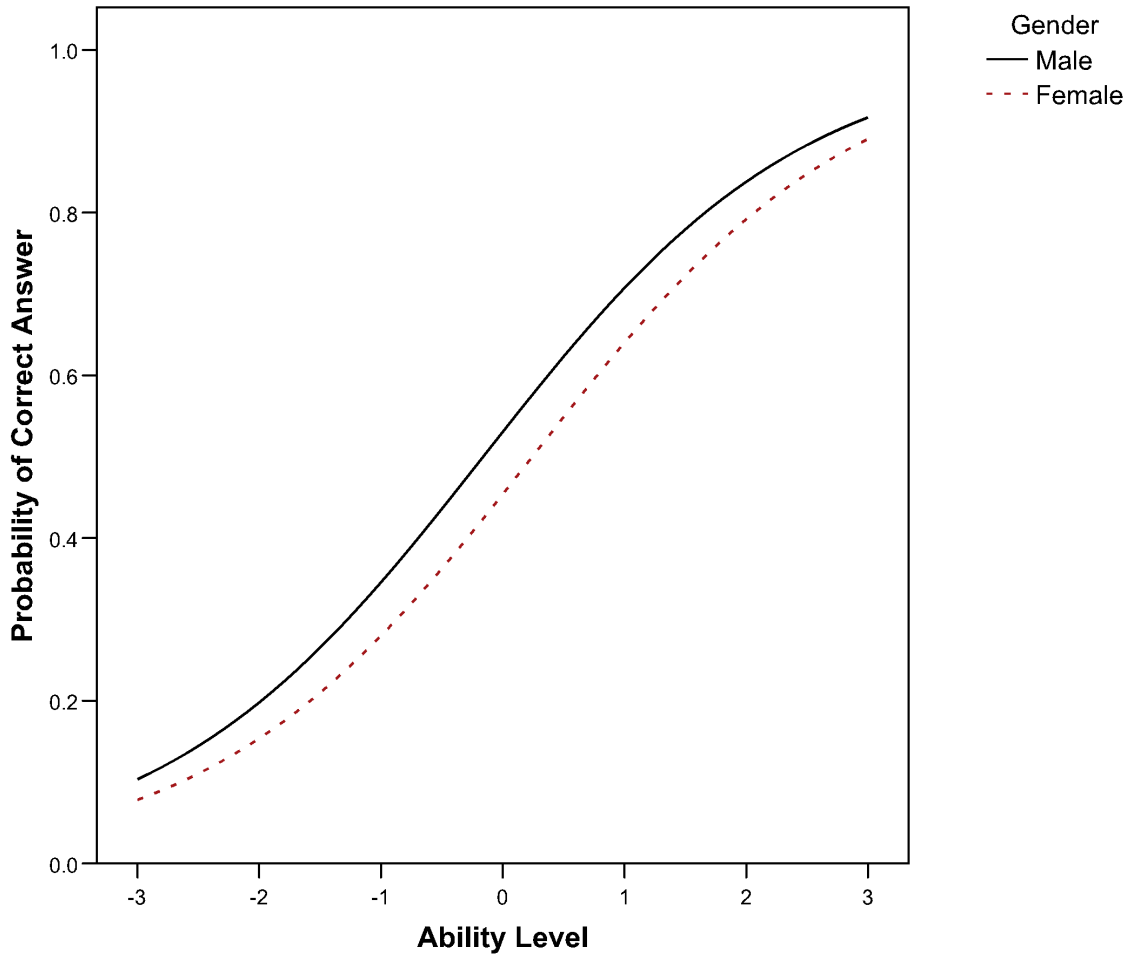
The DIF analysis indicates that item 6 is biased, as demonstrated by the separation of the "male" and "female" lines in Figure 1. The bias is in favor of females, as indicated by the red dashed "female" line appearing above the black "male" line. This means that when a male and a female are at the same ability level, the female is significantly more likely to select the correct answer than is the male. The amount of bias is greatest for examinees from the lower to above-average range in ability, but attenuates somewhat for those demonstrating the very highest level of working memory. Item 6 demonstrates the greatest amount of item bias in the Working Memory subtest.

Figure 2. DIF Curves for Item 19, Working Memory Subtest



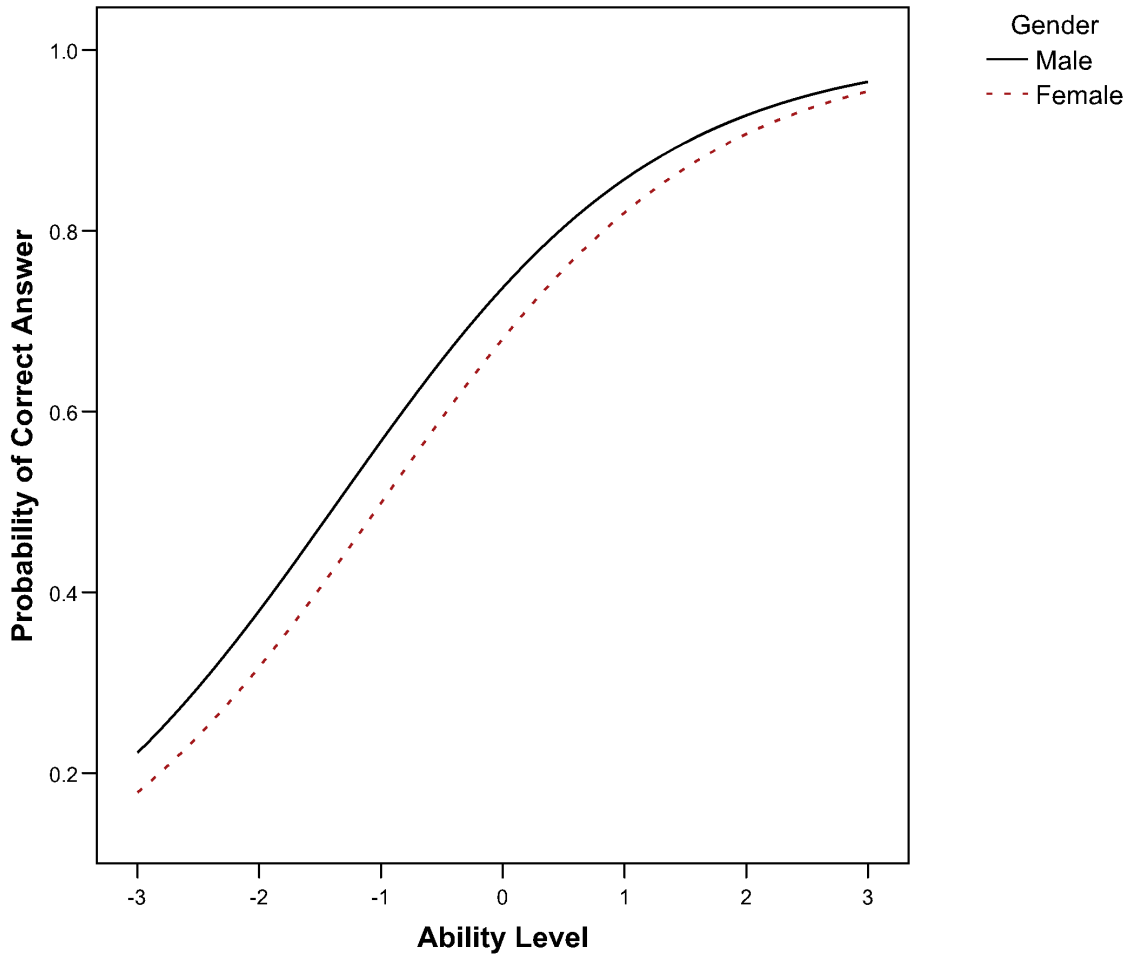
Item 19 demonstrates significant bias in favor of females at all levels of working memory.

Figure 3. DIF Curves for Item 12, Working Memory Subtest



Item 12 demonstrates significant bias in favor of males at all levels of working memory.

Figure 4. DIF Curves for Item 3, Working Memory Subtest



Item 3 demonstrates significant bias in favor of males at all levels of working memory, with attenuation at the highest ability levels.

Impact of Early Termination Rule

There is no early termination in the Working Memory subtest

Implications

Test Construction Implications for the Working Memory Subtest

The Working Memory subtest does not demonstrate a high degree of internal consistency:

- 7 items appear unrelated to total test score (correlation coefficients < 0.20),
- 10 do not adequately discriminate between examinees with lower and higher levels of working memory, and
- 4 items demonstrate significant differential item functioning (DIF).

Specifically:

- Items 1, 5, 6 (the three easiest items in the subtest), 8, 14,15 and 17 demonstrate low item-test correlation and low discrimination.
- In addition, item 6 also appears significantly biased against males.
- 3 other items, 3, 12, and 19 exhibit significant DIF. Other than the DIF, items 12 and 19 appear satisfactory in terms difficulty, item-test correlation and discrimination.
- Item 16 appears problematic in terms of low item discrimination.

In particular, item 6 appears (statistically) to be an excellent candidate for elimination or revision. Items, 1, 3, 5, 8, 9, 14, 16 and 17 also are strong candidates

The findings of the item and DIF analyses suggest that the indicated items bear closer scrutiny and that the Working Memory test should be re-evaluated substantively.

Long-Term Memory

Reliability

The 10 long-term memory items that comprise the "Long-Term Memory" subtest are a subset of the 20-item Working Memory subtest discussed above. As expected, due to the limited number of items, the reliability of this shorter Long-Term Memory subtest is lower than that of the Working Memory subtest. The internal consistency *reliability coefficient* is 0.65 (as compared with 0.73 for the longer subtest), and indicates relatively low internal consistency.

The *test-retest reliability coefficient* for the Long-Term Memory subtest is moderately high, 0.64. On average, examinees performed significantly better on the retest that on the initial testing (58 vs. 51 percent correct). It is unlikely that this difference is the result of a "practice effect" since approximately seven weeks elapsed between initial and second administrations of the subtest and no cognitive training took place during the period.

Item Difficulty, Item-Test Correlation, Item Discrimination, and DIF

The long-term memory item subset represents the most difficult items in the full Working Memory subtest. Item difficulties for the long-term memory subset range from 25 to 84 percent correct response. Table A-3 provides the results from the item analyses produced by the IRT program.

The item-test correlations for the Long-Term Memory subtest are all relatively low, ranging from 0.13 to 0.43. The results of the item difficulty, item-test correlation, item discrimination, and DIF analyses are summarized below.

The subset of long-term memory items is characterized by the same attributes shown by those same items in the Working Memory analyses. For convenience, results for the specific Long-Term Memory items shown in Table 1 are reproduced below in Table 2. In addition, the last column of Table 2 provides the relevant Figure numbers for location of the DIF curves in the preceding section.

Two of the four high-DIF items in the Working Memory subtest are also in the Long-Term Memory subtest (items 12, and 19).

Table 2. Summary of Item Analysis Findings for the Long-Term Memory Subtest

Item	Low Item-Test Correlation	Low Discrimination	High DIF	DIF Figure Number
12			√	3
15	√	√		
16		√		
17	√	√		
19			√	2

Long-term memory items not included in the above table do not demonstrate low item-test correlations, low discrimination, or high DIF. These are Items 7, 9, 11, 18, and 20.

Impact of Early Termination Rule

There is no early termination in the Long-Term Memory Subtest

Implications

Test Construction Implications for the Long-Term Memory Subtest

The Long-Term Memory subtest does not demonstrate a high degree of internal consistency. Of the 10 items:

- 2 appear unrelated to total test score,
- 3 do not adequately discriminate among examinees with lower and higher levels working memory, and
- 2 items demonstrate significant differential item functioning (DIF).

Specifically,

- Items 15 and 17 demonstrate low item-test correlation and low discrimination.
- Item 16 exhibits low discrimination only.
- Items 12 and 19 exhibit high DIF.

Word Attack

Reliability
Analysis indicates that the Word Attack subtest is internally consistent, with an internal consistency *reliability coefficient* of 0.83.

However, the *test-retest reliability coefficient* is low-moderate, 0.52. Differences in performance on the initial and retesting (82 percent and 84 percent correct, respectively) are not statistically significant.

The test-retest reliability coefficient indicates the degree to which examinees' total scores on the initial test are similar to their total scores on the second administration of the same test; thus, it is a stability coefficient. Examinee performance on the Word Attack subtest exhibits only modest temporal stability.

When a low or moderate stability coefficient is obtained, the following question must be asked:

> Does the low stability coefficient indicate that the test provides unreliable measures of the trait, or does it imply that the trait itself is unstable?

If the level of the trait being measured *will* change over time, then the obtained test-retest stability coefficient is not an appropriate estimate of test score reliability. Regarding measurement error and external threats to validity, one must consider whether an examinees' performance is altered by the first test administration so that the second test score will reflect effects of memory, practice, learning, or any other consequences of the first administration. In this case, it is difficult to think that any of these potential threats to reliability (and validity) are operating because no cognitive training occurred during the test-retest period and only about seven weeks elapsed between the administrations – probably insufficient time for either inherent cognitive development (e.g., maturation) or schooling-induced cognitive development to enhance performance on the retest.

Item Difficulty and Item-Test Correlation

Item difficulties for the 23 items of the Word Attack subtest range from 41 percent correct response for item 21 to 98 percent correct for item 2. One-quarter of the items received correct responses from 85 to 98 percent of examinees. Item statistics from the IRT analyses for the Word Attack subtest are provided in the Appendix, Table A-4.

Items in this subtest exhibit moderate to high correlation with total test score. Correlation coefficients range from 0.27 to 0.79.

Item Discrimination and DIF

Initial analyses indicate that all items in the Word Attack subtest are equally discriminating among examinees of low and high cognitive ability in this area. Furthermore, none of the items exhibit any *DIF*, or item bias, between male and female examinees.

Impact of Early Termination Rule

Forty-five examinees are able to respond correctly to a significant number of additional items when the Work Attack subtest is completed (between 1 and 11 additional items). Examinees

respond correctly to an average of 73 percent of the items, as compared to 67 percent with the early termination rule imposed.

Implications

Test Construction Implications for the Word Attack Subtest

The Work Attack subtest is internally consistent, as demonstrated by the high test reliability coefficient and high item-test correlations. The test-retest reliability coefficient (a measure of stability), on the other hand, is low-moderate (0.52). However, it is unlikely that administration of the initial test or intervening factors such as maturation, learning, practice, etc., significantly affects performance on the retest.

Based on these analyses, no items appear to be candidates for elimination or revision:

- All items in the Word Attack subtest are equally discriminating among examinees of low and high cognitive ability, and no DIF effects are observed.
- None of the items exhibit any gender DIF.

Early termination has a tremendous negative impact on examinee performance, and thus, on measurement of this cognitive domain. Therefore the practice should be discontinued.

Visual Processing

Reliability

The Visual Processing subtest is a highly reliable measure of the type of spatial relations ability required for solving picture puzzles. The internal consistency *reliability coefficient* is extremely high, 0.97.

The *test-retest reliability (stability coefficient)* is moderate, 0.58. Performance on the retest is substantially higher than on the initial test, 50 percent correct vs. 45 percent correct, respectively. Although it is impossible to ascertain the cause(s) of this difference, it is unlikely that extraneous factors such as examinee maturation, practice, learning, or schooling-induced cognitive development during the 7-week testing interval could have enhanced performance on the retest.

Item Difficulty

This subtest is the most difficult of all the subtests. The average percent correct, overall, is only 48. Males typically out-perform females, 49 percent to 47 percent.

The percent correct response to the 56 items in this subtest range from 0 percent for item 56 to 99 percent for item 1. Disregarding these two items, item difficulties range from 2 percent correct for items 55 and 54 to 95 percent for item 2. Item statistics from the IRT analyses are provided in Appendix Table A-6.

Item-Test Correlation

Ignoring item 56, item 1 exhibits the lowest item-test correlation (0.19). Items 2 – 5 also demonstrate relatively low correlations, with values less than 0.30. It seems likely that a fair

amount of guessing may be occurring, and results are also affected by the early termination rule (see discussion below), which would also adversely affect the item-test correlation.

Item Discrimination

Five items exhibit very low discrimination among examinees of lower and higher ability in the visual processing cognitive domain. These items are: 2, 3, 4, 5, and 13. As noted above, items 2 – 5 are also rather weakly linked to total score on this subtest.

DIF

The Visual Processing subtest consists of, effectively, spatial relations tasks – tasks at which males typically out-perform females. Accordingly, males significantly out-performed females on each and every task in this subtest. In particular, item 1 is extremely problematic, and should be dropped from the subtest.

DIF curves are not shown because all items are significantly biased against females.

Impact of Early Termination Rule

Imposition of the early termination rule has an even more damaging impact on examinee performance on the Visual Processing subtest than it has on the Word Attack subtest. One hundred and nineteen examinees respond correctly to additional items after completing the test. Between 1 and 24 additional items re answered correctly. Without early termination, examinees responded correctly to an average of 52 percent of the items, as compared to 40 percent with the early termination rule imposed.

Implications

Test Construction Implications for the Visual Processing Subtest

The Visual Processing subtest is a highly reliable measure of the type of spatial relations ability required for solving picture puzzles. The internal consistency *reliability coefficient* is extremely high, 0.97. However, the test-retest reliability coefficient, a measure of stability, is low-moderate (0.58). It is unlikely that administration of the initial test or intervening factors such as maturation, learning, practice, etc., would significantly affect performance on the retest.

Several items appear problematic:

- Item 1 is correctly answered by 99% of examinees, while none answered item 56 correctly. These 2 items should be dropped.
- Items 2-5 demonstrate low discrimination, are among the easiest items, and are rather weakly linked to total test score. These also are candidates for dropping.
- Item 13 does not discriminate among those with higher vs. lower visual processing abilities.

The early termination rule has a very serious negative impact on examinee performance, and hence on measurement of the visual processing cognitive domain. Therefore, early termination should be discontinued.

Auditory Analysis

The Auditory Analysis subtest consists of two subscales, Segmenting and Drop. Initial analyses were conducted with all items as a composite Auditory Analysis subtest. The Segmenting and Drop subscales were also analyzed separately.

Reliability

The total Auditory Analysis subtest possesses a very high degree of internal consistency: *test reliability* = 0.90. The rest reliability coefficients for the two subscales are very high, particularly for the Drop subscale. The subscale reliability coefficients are 0.79 for Segmenting and 0.87 for Drop.

The *test-retest reliability coefficients* for the total subtest and the two subscales also are very high: 0.83, 0.73, and 0.79, respectively. Examinee performance is essentially equal on the initial and second test administrations.

Item Difficulty and Item-Test Correlation

Item difficulties for the total Auditory Analysis subtest range from 5.2 percent correct for item 28 (item 13 of the Auditory Drop subtest)[3] to 88 percent for items 2 and 5. Item 1 also is quite easy, with 87 percent of examinees responding correctly.

Of the two subscales, Drop is much more difficult. Subscale scores are 67 percent correct for Segmenting, compared to 56 for Drop. Item statistics for the composite subtest and for the two subscales are provided in the Appendix, Tables A-6a – A-6c.

Tables A6a – A6c also indicate that items in the Drop subscale may be "tighter" measures of their cognitive domain than items in the Segmenting subscale, as implied by the higher item-test correlations. It is interesting to note that the two Segmenting items with the lowest item-test correlations are also two of the easiest items in the entire subtest (items 1 and 5, difficulties 87% and 88%), while the item with the lowest item-test correlation in the Drop subscale is the most difficult item in the entire subtest (item 28 in the total test, item 13 in the Drop subscale, difficulty = 5% correct).

Item Discrimination

Six items do not adequately discriminate among examinees possessing higher or lower auditory analysis abilities. These items are 1, 3, 4, 5, 9, and 28 (item 13 in the Drop subscale). There is some overlap among these items with regard to item difficulty, low item-test correlations, and low discrimination. Results are summarized in Table 3. Placement of items in the Segmenting or Drop subscale is noted.

[3] For the composite Auditory Analysis, the Segmenting and Drop subtests are combined and numbered consecutively, 1-15 for the segmenting subscale and 16-28 for the drop subscale. Thus, the same item is numbered 28 in the composite Auditory Analysis subtest and 13 in the original Drop test.

Table 3. Summary of Item Analysis Findings for the Auditory Analysis Subtest

Item	Subscale	Easiest Items	Hardest Item	Low Item-Test Correlation	Low Discrimination
1	Segmenting	√		√	√
2	Segmenting	√			
3	Segmenting				√
4	Segmenting				√
5	Segmenting	√		√	√
9	Segmenting				√
28(13)[a]	Drop		√	√	√

[a] For the composite Auditory Analysis, the Segmenting and Drop subtests were combined and numbered consecutively, 1-15 for the segmenting subscale and 16-28 for the drop subscale. Thus, the same item is numbered 28 in the composite Auditory Analysis subtest, but was numbered 13 in the original Drop test.

DIF

None of the items exhibit any *DIF*, or item bias, between male and female examinees.

Impact of Early Termination Rule

Overall, early termination has a significant negative effect on the Auditory Analysis subtest and both subscales (Segmenting and Drop), although the effect is more pronounced in the Drop subscale. Ninety-eight examinees are able to respond correctly to between 1and 12 addition items on either or both the Segmenting and Drop subscales. The total percent correct for Auditory Analysis without early termination is 57, compared with 49 percent when the rule is imposed. The impact is worse for "Drop" than it is for "Segmenting".

Implications

Test Construction Implications for the Auditory Analysis Subtest

The internal consistency for the total Auditory Analysis subtest is very high, as are the *test reliability* coefficients for the Segmenting and Drop subscales: 0.90, 0.79, and 0.87, respectively. The *test-retest reliability coefficients* for the total subtest and the two subscales also are very high: 0.83, 0.73, and 0.79, respectively. Examinee performance was essentially equal on the initial and second test administrations.

Of the 30 items:

- 3 appear unrelated to total test score (items 1, 5, and 28(13)).
- 6 do not adequately discriminate among examinees with lower and higher levels auditory analysis (items 1, 3, 4, 5, 9, and 28(13)).

No gender DIF was detected.

The early termination rule has a serious negative impact on measurement of the auditory analysis cognitive domain and should be discontinued.

Logic and Reasoning

Reliability

The Logic and Reasoning subtest appear reliable, with an *internal consistency* reliability coefficient of 0.79. The *test-retest stability* coefficient also is quite high, 0.70.

Item Difficulty and Item-Test Correlation

The Logic and Reasoning subtest consists of 16 items. All appear related to total test score (all correlations > 0.20). Items demonstrating the lowest item-test correlations are also the easiest items (items 1, 2, and 3). Item difficulties range from 3.1 percent correct for item10 to 92 percent correct for item 2. Item statistics are provided in Appendix Table A-7.

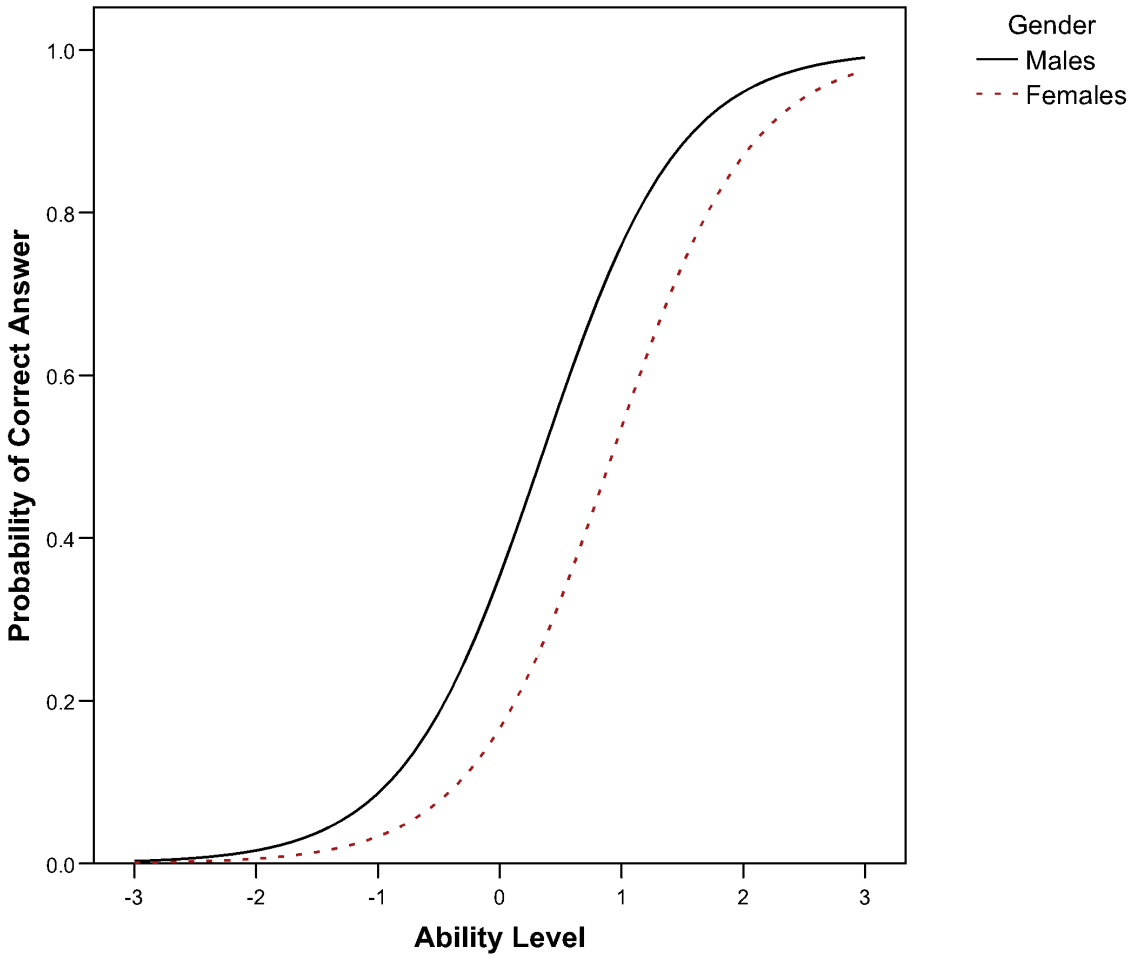
Item Discrimination

Items in the Logic and Reasoning subtest demonstrate a high degree of discrimination overall. Two items appear lower in discriminatory power (items 1 and 4), but the effect, in itself, is not sufficient cause for elimination.

DIF

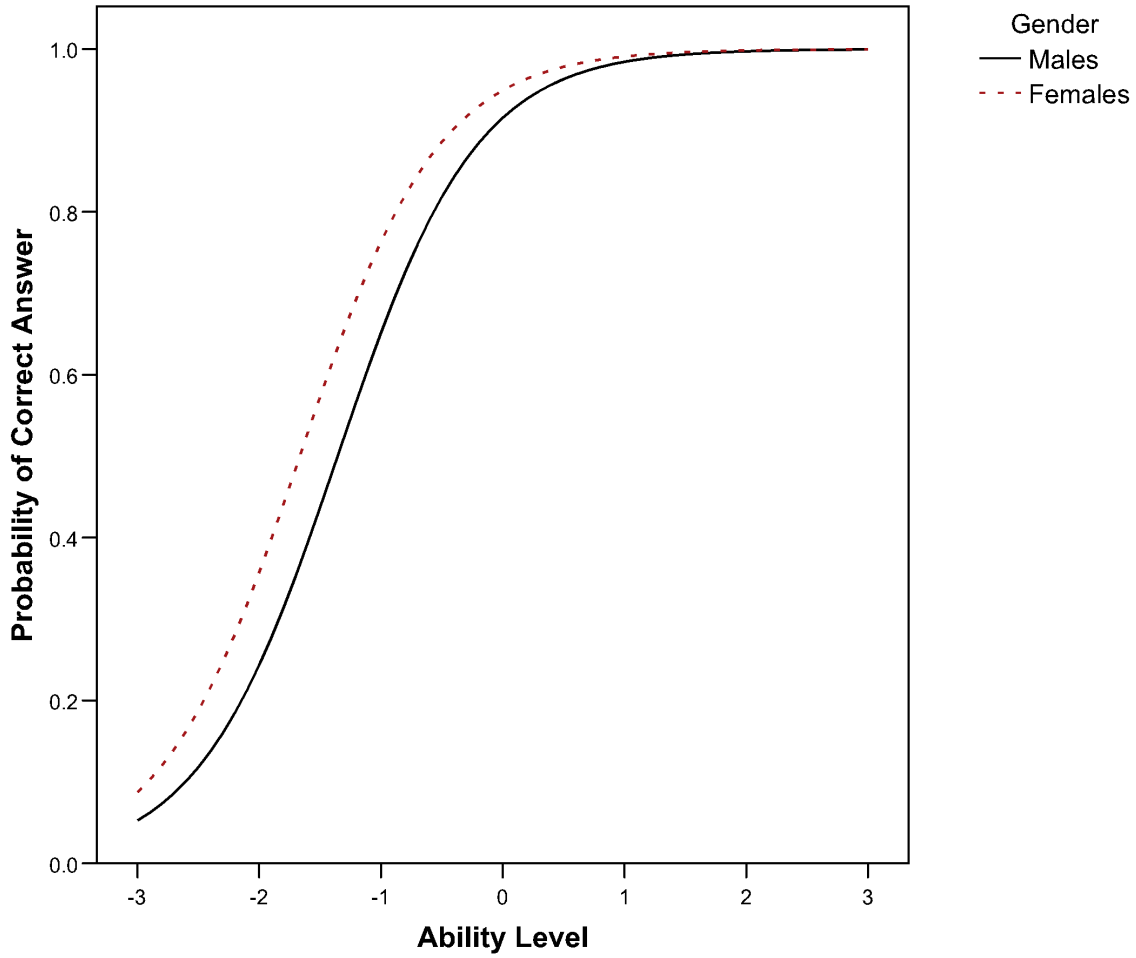
The two least discriminating items discussed above (1 and 4) also exhibit significant differential items function with regard to gender. The DIF curves for these two items are shown below.

Figure 5. DIF Curves for Item 4, Logic and Reasoning Subtest



Item 4 demonstrates significant bias in favor of males at all ability levels of logic and reasoning, although the affect is attenuated at the very lowest and highest ends of the ability distribution.

Figure 6. DIF Curves for Item 1, Logic and Reasoning Subtest



Item 1 exhibits significant bias toward females in the average to below average ability range.

Impact of Early Termination Rule

One hundred and twelve examinees were able to respond correctly to additional items when the Logic and Reasoning subtest was completed without early termination. Between 1 and 7 additional items were answered correctly. The total percent correct for the Logic and Reasoning subtest without early termination is 32, compared with 24 percent when the rule is imposed.

Implications

Test Construction Implications for the Logic and Reasoning Subtest

With the exception of items 1 and 4, the Logic and Reasoning subtest appears very well constructed. The subtest possesses high reliability, good item-test correlation, good item discrimination, and little DIF.

Items 1 and 4 may be candidates for revision or substitution with other, more suitable items due to lower discrimination and some gender DIF. This is a substantive decision.

Early termination has a significant negative impact on measurement of the logic and reasoning cognitive domain and should be discontinued.

APPENDIX

ITEM STATISTICS
FOR EACH SUBTEST OF THE GIBSON COGNITIVE SKILLS TEST

Processing Speed
Working Memory
Long-Term Memory
Word Attack
Visual Processing
Auditory Analysis – Segmenting and Drop Subscales
Logic and Reasoning

Interpretation of Tables

Tables in this Appendix provide the results from the item analyses produced by the IRT program.

- The ITEM and NAME columns indicate the item numbers.
- #TESTED is the number of examinees and # RIGHT is the number of examinees who responded correctly to the corresponding item number.
- PCT indicates the percentage of examinees who responded correctly, and is the measure of item difficulty (or more logically, item facility).
- LOGIT/1.7 is a scaling factor and can be ignored.
- ITEM-TEST CORRELATION is the correlation between examinee responses to the item and examinees' total test scores.

Items highlighted in yellow have little to no relationship to total test score.

Table A-1. Item Statistics for the Processing Speed Subtest

ITEM	NAME	#TESTED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	909.0	880.0	96.8	-2.01	0.174
2	ITEM0002	909.0	883.0	97.1	-2.07	0.190
3	ITEM0003	909.0	898.0	98.8	-2.59	0.266
4	ITEM0004	909.0	877.0	96.5	-1.95	0.171
5	ITEM0005	909.0	893.0	98.2	-2.37	0.315
6	ITEM0006	909.0	901.0	99.1	-2.78	0.256
7	ITEM0007	909.0	891.0	98.0	-2.30	0.172
8	ITEM0008	909.0	896.0	98.6	-2.49	0.344
9	ITEM0009	909.0	869.0	95.6	-1.81	0.257
10	ITEM0010	909.0	890.0	97.9	-2.26	0.338
11	ITEM0011	909.0	894.0	98.3	-2.40	0.354
12	ITEM0012	909.0	884.0	97.2	-2.10	0.338
13	ITEM0013	909.0	891.0	98.0	-2.30	0.400
14	ITEM0014	909.0	833.0	91.6	-1.41	0.227
15	ITEM0015	909.0	851.0	93.6	-1.58	0.312
16	ITEM0016	909.0	847.0	93.2	-1.54	0.316
17	ITEM0017	909.0	846.0	93.1	-1.53	0.368
18	ITEM0018	909.0	857.0	94.3	-1.65	0.413
19	ITEM0019	909.0	825.0	90.8	-1.34	0.350
20	ITEM0020	909.0	853.0	93.8	-1.60	0.417
21	ITEM0021	909.0	852.0	93.7	-1.59	0.501
22	ITEM0022	909.0	826.0	90.9	-1.35	0.528
23	ITEM0023	909.0	822.0	90.4	-1.32	0.535
24	ITEM0024	909.0	825.0	90.8	-1.34	0.614
25	ITEM0025	909.0	805.0	88.6	-1.20	0.605
26	ITEM0026	909.0	804.0	88.4	-1.20	0.639
27	ITEM0027	909.0	815.0	89.7	-1.27	0.672
28	ITEM0028	909.0	779.0	85.7	-1.05	0.647
29	ITEM0029	909.0	751.0	82.6	-0.92	0.643
30	ITEM0030	909.0	722.0	79.4	-0.79	0.675
31	ITEM0031	909.0	678.0	74.6	-0.63	0.680
32	ITEM0032	909.0	632.0	69.5	-0.49	0.669
33	ITEM0033	909.0	593.0	65.2	-0.37	0.651
34	ITEM0034	909.0	503.0	55.3	-0.13	0.606
35	ITEM0035	909.0	424.0	46.6	0.08	0.619
36	ITEM0036	909.0	450.0	49.5	0.01	0.680
37	ITEM0037	909.0	356.0	39.2	0.26	0.647
38	ITEM0038	909.0	314.0	34.5	0.38	0.640
39	ITEM0039	909.0	291.0	32.0	0.44	0.650
40	ITEM0040	909.0	258.0	28.4	0.54	0.629
41	ITEM0041	909.0	214.0	23.5	0.69	0.578
42	ITEM0042	909.0	191.0	21.0	0.78	0.565
43	ITEM0043	909.0	154.0	16.9	0.94	0.518
44	ITEM0044	909.0	125.0	13.8	1.08	0.476
45	ITEM0045	909.0	94.0	10.3	1.27	0.412

Table A-2. Item Statistics for the Working Memory Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	879.0	96.8	-2.01	0.066
2	ITEM0002	908.0	710.0	78.2	-0.75	0.448
3	ITEM0003	908.0	624.0	68.7	-0.46	0.220
4	ITEM0004	908.0	834.0	91.9	-1.42	0.262
5	ITEM0005	908.0	814.0	89.6	-1.27	0.162
6	ITEM0006	908.0	808.0	89.0	-1.23	0.166
7	ITEM0007	908.0	516.0	56.8	-0.16	0.415
8	ITEM0008	908.0	356.0	39.2	0.26	0.062
9	ITEM0009	908.0	763.0	84.0	-0.98	0.327
10	ITEM0010	908.0	731.0	80.5	-0.83	0.310
11	ITEM0011	908.0	495.0	54.5	-0.11	0.273
12	ITEM0012	908.0	449.0	49.4	0.01	0.299
13	ITEM0013	908.0	717.0	79.0	-0.78	0.198
14	ITEM0014	908.0	201.0	22.1	0.74	0.097
15	ITEM0015	908.0	530.0	58.4	-0.20	0.110
16	ITEM0016	908.0	224.0	24.7	0.66	0.253
17	ITEM0017	908.0	566.0	62.3	-0.30	0.166
18	ITEM0018	908.0	580.0	63.9	-0.34	0.289
19	ITEM0019	908.0	481.0	53.0	-0.07	0.463
20	ITEM0020	908.0	520.0	57.3	-0.17	0.419

Table A-3. Item Statistics for the Long-Term Memory Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	516.0	56.8	-0.16	0.385
2	ITEM0002	908.0	763.0	84.0	-0.98	0.291
3	ITEM0003	908.0	495.0	54.5	-0.11	0.257
4	ITEM0004	908.0	449.0	49.4	0.01	0.268
5	ITEM0005	908.0	530.0	58.4	-0.20	0.076
6	ITEM0006	908.0	224.0	24.7	0.66	0.244
7	ITEM0007	908.0	566.0	62.3	-0.30	0.130
8	ITEM0008	908.0	580.0	63.9	-0.34	0.286
9	ITEM0009	908.0	481.0	53.0	-0.07	0.428
10	ITEM0010	908.0	520.0	57.3	-0.17	0.402

Table A-4. Item Statistics for the Word Attack Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	815.0	89.8	-1.28	0.357
2	ITEM0002	908.0	891.0	98.1	-2.33	0.271
3	ITEM0003	908.0	626.0	68.9	-0.47	0.362
4	ITEM0004	908.0	835.0	92.0	-1.43	0.399
5	ITEM0005	908.0	865.0	95.3	-1.77	0.379
6	ITEM0006	908.0	867.0	95.5	-1.79	0.513
7	ITEM0007	908.0	831.0	91.5	-1.40	0.397
8	ITEM0008	908.0	794.0	87.4	-1.14	0.309
9	ITEM0009	908.0	864.0	95.2	-1.75	0.488
10	ITEM0010	908.0	714.0	78.6	-0.77	0.432
11	ITEM0011	908.0	817.0	90.0	-1.29	0.556
12	ITEM0012	908.0	610.0	67.2	-0.42	0.430
13	ITEM0013	908.0	732.0	80.6	-0.84	0.602
14	ITEM0014	908.0	602.0	66.3	-0.40	0.467
15	ITEM0015	908.0	786.0	86.6	-1.10	0.679
16	ITEM0016	908.0	541.0	59.6	-0.23	0.511
17	ITEM0017	908.0	752.0	82.8	-0.93	0.667
18	ITEM0018	908.0	725.0	79.8	-0.81	0.654
19	ITEM0019	908.0	794.0	87.4	-1.14	0.791
20	ITEM0020	908.0	776.0	85.5	-1.04	0.748
21	ITEM0021	908.0	370.0	40.7	0.22	0.377
22	ITEM0022	908.0	663.0	73.0	-0.59	0.608
23	ITEM0023	908.0	512.0	56.4	-0.15	0.503

Table A-5. Item Statistics for the Visual Processing Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	894.0	98.5	-2.45	0.195
2	ITEM0002	908.0	866.0	95.4	-1.78	0.223
3	ITEM0003	908.0	810.0	89.2	-1.24	0.227
4	ITEM0004	908.0	826.0	91.0	-1.36	0.284
5	ITEM0005	908.0	848.0	93.4	-1.56	0.270
6	ITEM0006	908.0	834.0	91.9	-1.42	0.347
7	ITEM0007	908.0	787.0	86.7	-1.10	0.342
8	ITEM0008	908.0	820.0	90.3	-1.31	0.340
9	ITEM0009	908.0	821.0	90.4	-1.32	0.434
10	ITEM0010	908.0	825.0	90.9	-1.35	0.402
11	ITEM0011	908.0	820.0	90.3	-1.31	0.446
12	ITEM0012	908.0	809.0	89.1	-1.24	0.447
13	ITEM0013	908.0	522.0	57.5	-0.18	0.379
14	ITEM0014	908.0	583.0	64.2	-0.34	0.465
15	ITEM0015	908.0	658.0	72.5	-0.57	0.487
16	ITEM0016	908.0	624.0	68.7	-0.46	0.540
17	ITEM0017	908.0	741.0	81.6	-0.88	0.629
18	ITEM0018	908.0	372.0	41.0	0.21	0.611
19	ITEM0019	908.0	546.0	60.1	-0.24	0.696
20	ITEM0020	908.0	426.0	46.9	0.07	0.651
21	ITEM0021	908.0	525.0	57.8	-0.19	0.792
22	ITEM0022	908.0	529.0	58.3	-0.20	0.787
23	ITEM0023	908.0	507.0	55.8	-0.14	0.784
24	ITEM0024	908.0	537.0	59.1	-0.22	0.821
25	ITEM0025	908.0	502.0	55.3	-0.12	0.787
26	ITEM0026	908.0	538.0	59.3	-0.22	0.824
27	ITEM0027	908.0	503.0	55.4	-0.13	0.808
28	ITEM0028	908.0	527.0	58.0	-0.19	0.819
29	ITEM0029	908.0	535.0	58.9	-0.21	0.834
30	ITEM0030	908.0	468.0	51.5	-0.04	0.777
31	ITEM0031	908.0	500.0	55.1	-0.12	0.778
32	ITEM0032	908.0	388.0	42.7	0.17	0.745
33	ITEM0033	908.0	100.0	11.0	1.23	0.323
34	ITEM0034	908.0	235.0	25.9	0.62	0.549
35	ITEM0035	908.0	341.0	37.6	0.30	0.755
36	ITEM0036	908.0	181.0	19.9	0.82	0.517
37	ITEM0037	908.0	366.0	40.3	0.23	0.793
38	ITEM0038	908.0	323.0	35.6	0.35	0.737
39	ITEM0039	908.0	275.0	30.3	0.49	0.685
40	ITEM0040	908.0	235.0	25.9	0.62	0.656
41	ITEM0041	908.0	260.0	28.6	0.54	0.681
42	ITEM0042	908.0	125.0	13.8	1.08	0.502
43	ITEM0043	908.0	155.0	17.1	0.93	0.562
44	ITEM0044	908.0	199.0	21.9	0.75	0.630
45	ITEM0045	908.0	123.0	13.5	1.09	0.510
46	ITEM0046	908.0	89.0	9.8	1.31	0.447
47	ITEM0047	908.0	47.0	5.2	1.71	0.347
48	ITEM0048	908.0	57.0	6.3	1.59	0.372
49	ITEM0049	908.0	32.0	3.5	1.95	0.298
50	ITEM0050	908.0	48.0	5.3	1.70	0.371

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
51	ITEM0051	908.0	21.0	2.3	2.20	0.249
52	ITEM0052	908.0	24.0	2.6	2.12	0.280
53	ITEM0053	908.0	23.0	2.5	2.15	0.276
54	ITEM0054	908.0	21.0	2.3	2.20	0.265
55	ITEM0055	908.0	19.0	2.1	2.26	0.250
56	ITEM0056	908.0	0.0	0.0	99.99	0.000

Table A-6a. Item Statistics for the Auditory Analysis Subtest (Composite)

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	793.0	87.3	-1.14	0.198
2	ITEM0002	908.0	800.0	88.1	-1.18	0.223
3	ITEM0003	908.0	763.0	84.0	-0.98	0.202
4	ITEM0004	908.0	690.0	76.0	-0.68	0.244
5	ITEM0005	908.0	803.0	88.4	-1.20	0.199
6	ITEM0006	908.0	585.0	64.4	-0.35	0.287
7	ITEM0007	908.0	784.0	86.3	-1.08	0.407
8	ITEM0008	908.0	613.0	67.5	-0.43	0.428
9	ITEM0009	908.0	302.0	33.3	0.41	0.210
10	ITEM0010	908.0	490.0	54.0	-0.09	0.429
11	ITEM0011	908.0	668.0	73.6	-0.60	0.530
12	ITEM0012	908.0	551.0	60.7	-0.26	0.550
13	ITEM0013	908.0	535.0	58.9	-0.21	0.554
14	ITEM0014	908.0	423.0	46.6	0.08	0.407
15	ITEM0015	908.0	382.0	42.1	0.19	0.463
16	ITEM0016	908.0	769.0	84.7	-1.01	0.374
17	ITEM0017	908.0	745.0	82.0	-0.89	0.401
18	ITEM0018	908.0	563.0	62.0	-0.29	0.356
19	ITEM0019	908.0	769.0	84.7	-1.01	0.469
20	ITEM0020	908.0	486.0	53.5	-0.08	0.530
21	ITEM0021	908.0	724.0	79.7	-0.81	0.522
22	ITEM0022	908.0	449.0	49.4	0.01	0.543
23	ITEM0023	908.0	529.0	58.3	-0.20	0.641
24	ITEM0024	908.0	335.0	36.9	0.32	0.471
25	ITEM0025	908.0	486.0	53.5	-0.08	0.627
26	ITEM0026	908.0	522.0	57.5	-0.18	0.719
27	ITEM0027	908.0	482.0	53.1	-0.07	0.647
28	ITEM0028	908.0	47.0	5.2	1.71	0.080
29	ITEM0029	908.0	378.0	41.6	0.20	0.626
30	ITEM0030	908.0	369.0	40.6	0.22	0.623

Table A-6b. Item Statistics for the Auditory Analysis – Segmenting Subscale of the Auditory Analysis Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	793.0	87.3	-1.14	0.170
2	ITEM0002	908.0	800.0	88.1	-1.18	0.226
3	ITEM0003	908.0	763.0	84.0	-0.98	0.217
4	ITEM0004	908.0	690.0	76.0	-0.68	0.290
5	ITEM0005	908.0	803.0	88.4	-1.20	0.269
6	ITEM0006	908.0	585.0	64.4	-0.35	0.301
7	ITEM0007	908.0	784.0	86.3	-1.08	0.423
8	ITEM0008	908.0	613.0	67.5	-0.43	0.479
9	ITEM0009	908.0	302.0	33.3	0.41	0.321
10	ITEM0010	908.0	490.0	54.0	-0.09	0.466
11	ITEM0011	908.0	668.0	73.6	-0.60	0.655
12	ITEM0012	908.0	551.0	60.7	-0.26	0.573
13	ITEM0013	908.0	535.0	58.9	-0.21	0.628
14	ITEM0014	908.0	423.0	46.6	0.08	0.488
15	ITEM0015	908.0	382.0	42.1	0.19	0.501

Table A-6c. Item Statistics for the Auditory Analysis – Drop Subscale of the Auditory Analysis Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0016	908.0	769.0	84.7	-1.01	0.441
2	ITEM0017	908.0	745.0	82.0	-0.89	0.423
3	ITEM0018	908.0	563.0	62.0	-0.29	0.413
4	ITEM0019	908.0	769.0	84.7	-1.01	0.514
5	ITEM0020	908.0	486.0	53.5	-0.08	0.570
6	ITEM0021	908.0	724.0	79.7	-0.81	0.585
7	ITEM0022	908.0	449.0	49.4	0.01	0.568
8	ITEM0023	908.0	529.0	58.3	-0.20	0.695
9	ITEM0024	908.0	335.0	36.9	0.32	0.493
10	ITEM0025	908.0	486.0	53.5	-0.08	0.700
11	ITEM0026	908.0	522.0	57.5	-0.18	0.787
12	ITEM0027	908.0	482.0	53.1	-0.07	0.723
13	ITEM0028	908.0	47.0	5.2	1.71	0.118
14	ITEM0029	908.0	378.0	41.6	0.20	0.661
15	ITEM0030	908.0	369.0	40.6	0.22	0.677

Table A-7. Item Statistics for the Logic and Reasoning Subtest

ITEM	NAME	#TRIED	#RIGHT	PCT	LOGIT/1.7	ITEM-TEST CORRELATION
1	ITEM0001	908.0	781.0	86.0	-1.07	0.230
2	ITEM0002	908.0	833.0	91.7	-1.42	0.233
3	ITEM0003	908.0	823.0	90.6	-1.34	0.222
4	ITEM0004	908.0	293.0	32.3	0.44	0.323
5	ITEM0005	908.0	548.0	60.4	-0.25	0.475
6	ITEM0006	908.0	229.0	25.2	0.64	0.398
7	ITEM0007	908.0	234.0	25.8	0.62	0.376
8	ITEM0008	908.0	227.0	25.0	0.65	0.614
9	ITEM0009	908.0	105.0	11.6	1.20	0.509
10	ITEM0010	908.0	28.0	3.1	2.03	0.211
11	ITEM0011	908.0	149.0	16.4	0.96	0.660
12	ITEM0012	908.0	105.0	11.6	1.20	0.611
13	ITEM0013	908.0	98.0	10.8	1.24	0.595
14	ITEM0014	908.0	37.0	4.1	1.86	0.346
15	ITEM0015	908.0	61.0	6.7	1.55	0.466
16	ITEM0016	908.0	36.0	4.0	1.87	0.399