Poster
TITLE
How to Build a Cognitive Ability Test with Reduced Mean Group Differences

SHORTENED TITLE
How to Build a *g* Test with Reduced Mean Group Differences

ABSTRACT

Guided by psychometric principles, one can build cognitive ability tests with lower mean group differences by using items with low *g*-saturation and by reducing the reliability of the test. Such a diminished *g* test predicts *g*-related criteria worse than a *g*-test with high *g*-saturation and high reliability. Assertions about specific item types causing reduced mean differences are likely incorrect.

PRESS PARAGRAPH

Claims were made about building cognitive ability tests with lower than typical mean ethnic/racial differences suggesting that this was caused by the use of specific item types, but no empirical evidence was presented. The resulting test is proprietary and commercially distributed even though no peer reviewed research on it has documented the claims. Here, we show that one can use basic psychometric principles to build cognitive ability tests with lower mean group differences simply by using items with low *g*-saturation and by reducing the reliability of the test. Such diminished *g* tests predict a *g*-related criterion worse than *g*-tests with high *g*-saturation and high reliability. Assertions about specific proprietary item types causing reduced mean group differences are likely incorrect.

WORD COUNT
2846

REQUIRED STATEMENT OF FUNDING AGENCY

This research was funded by the U. S. Army Research Institute for the Behavioral and Social Sciences through Grant W911NF1310353 awarded to (name deleted to permit blind review). The views, opinions, and/or findings contained in this article are solely those of the authors and should not be construed as an official Department of the Army or Department of Defense position, policy, or decision, unless so designated by other documentation.

Revised February 12, 2018

General cognitive ability (*g*) tests typically show large mean White-Black score differences. Given this bleak situation, some researchers have sought to develop measures of general cognitive ability that have high validity for predicting job performance but result in low mean group differences. We refer to such measures as "alternative *g* tests."

The newest effort in alternative *g* tests is the Siena Reasoning Test (Goldstein, 2008). Although no publications or a test manual could be located, the measure has been offered as a *g* test that shows smaller mean racial differences than previous measures of *g*. Yusko, Goldstein, Oliver, and Hanges (2010) argued that the Siena Reasoning Test shows reduced mean racial differences because it seeks to reduce reliance on prior knowledge, reduce the use of language, and incorporates graphical stimuli. We are not aware of these authors providing any empirical support for their assertions concerning item characteristics and their influence on mean group differences.

Some research findings are contrary to the assertions concerning the Siena Reasoning Test. Other tests have also reduced the use of language. For example, the Davis-Eells games asked children to interpret events depicted in a series of cartoons. Jensen's (1980) review concluded that the test was "remarkably unsuccessful" (p. 643) at reducing White-Black mean score differences. Like the Siena Reasoning Test, other efforts have used graphical items. The Raven's Progressive Matrices test (Raven, Court & Raven, 1994) typically shows large White-Black mean differences even though the items do not rely on prior knowledge or language and are entirely graphical. These results are counter to the assertions about item characteristics in the Siena Reasoning Test being the cause of reductions in mean group differences.

In contrast to assertions about smaller mean group differences being due to special types of items, we argue that alternative *g* tests show smaller mean racial differences than traditional psychometric *g* tests because they have lower *g* saturation. That is, the tests have lower mean group differences because they measure *g* less well. We also argue that reliability of the test influences mean group differences. As an extreme example, a test score obtained by generating random numbers for each respondent will have a reliability of zero and no mean group differences, on average. Likewise, a cognitive ability test with a reliability of .70 will show lower mean group differences than a test with a reliability of .80, on average. Finally, *g*-tests with lower group differences should have lower predictive value for criteria that *g* predicts such as job performance and educational attainment. Lacking a job performance measure in the collected data set, we offer that *g* tests with lower magnitude mean group differences will have lower correlations with educational criteria. Thus, formal hypotheses are:

Hyp 1: Cognitive ability tests with lower *g*-saturation will have lower mean group differences than cognitive ability tests with higher g-saturation.

Hyp 2: Cognitive ability tests with lower reliability will have lower mean group differences than cognitive ability tests with higher reliability.

Hyp 3: Cognitive ability tests with lower *g*-saturation will have lower correlations with educational attainment than cognitive ability tests with higher *g*-saturation.

Method

Sample

Data were collected using Amazon Mechanical Turk. After data screening for inattentive responders, the analysis file consisted of 927 respondents. Of these, 209 were Asian (non-Hispanic), 236 were Black (non-Hispanic) and 246 were White (non-Hispanic). The remaining 236 respondents were Hispanic.

Measures

We developed or obtained 194 items grouped into 12 scales. As recommended by Major, Johnson, and Bouchard (2011), we used more than seven indicators to derive a *g* factor. Based on recommendations from Carroll (1993), we used a diverse set of item types. Also, following Ashton and Lee (2005) and Kvist and Gustafsson (2008), we used several types of fluid items because fluid items often have narrower bandwidth than measures of crystalized ability. Consistent with recommendations, we conducted a principal factor analysis (Major et al., 2011) of the 12 scales to derive a *g* factor.

After deletion of items showing differential item functioning (DIF), 159 items remained. The *g*-loading of an item was defined as the correlation of the item with the *g* factor. The items were divided into two sets based on the mean *g*-loading of the items. The set of items consisting of low-*g* items (items with *g*-loading below the mean) contained 79 items and the set of items consisting of high-*g* items (items with *g*-loading at or above the mean) consisted of 80 items.

One hundred low-*g* 30-item tests were created. Each test was created by randomly selecting 30 items from the low-*g* item set. One hundred low-*g* 40-item scales were then created by drawing 40 items for each test randomly from the low-*g* items set. The same process was used to create 100 30-item high-*g* tests which draw their items from the high-g item set and to create 100 40-item high-*g* tests. Finally, we created one more low-*g* test using all 79 items in the low-*g* item set and one more high-*g* test using all 80 items in the high-*g* item set. For each of these 402 tests, we calculated White-minority mean score differences expressed as a standardized mean difference. We also calculated the internal consistency reliability (alpha) of each of the 402 tests and the correlation of each test with educational attainment.

Results

Table 1 shows the 12 scales and the number of items in each (after removal of DIF items). Table 2 presents scale intercorrelations. The GATB Object matching scale is best classified as perceptual speed. The GATB Three-dimensional space scale measures spatial ability. Two logic-based measurement scales were developed to be comparable in logical structure of the items except one used all real words and one used some fake words (see Figures 1 and 2). These two scales are best classified as fluid *g* measures. Items with fake words have been offered as items less tied to culture and may be classified by some as alternative *g* items. Given that the items in these scales were built to have the same logical structure, one can compare the scale scores to determine if the distinction between real versus fake words have an impact on mean group differences.

Each item in the sentence revision scale presented a sentence and the respondent decided whether the sentence was grammatically and stylistically the best; or, they picked from alternative rephrasings of the sentence (see Figure 3). This scale is best classified as a crystallized *g* scale. With permission of Robert Sternberg, we include three scales from his Sternberg Triarchic Abilities Test (January 2001 version). One required respondents to infer the meaning of fake words from their context in a sentence (see Figure 4). The second presented respondents with new mathematical operators and the respondents attempted to solve mathematical arguments using mathematical operators (see Figure 5). The third is a traditional number series test. All three of the scales are likely best classified as fluid reasoning. Some might call the first two of the three scales alternative *g* measures given the use of fake words or novel items. The next scale presented eight objects that vary by size, shape and shading (see Figure 6) and respondents answered questions such as “Which three objects match exactly in size and shape, but differ in shading?”. These items are best classified as fluid intelligence. Because of the graphics, some might classify these as alternative *g* items. The next scale (see Figure 7) presented a table and asked respondents questions about the table contents. This scale might be classified as following directions. The last two scales provided information that compares four entities (e.g., cats, objects) and asks respondents questions. These items are best classified as fluid intelligence. The first set presented the entities as pictures (see Figure 8) and the second set presented the entities in words (see Figure 9). The items in each scale were designed to have the same logical structure. One can compare the scale scores to determine if the distinction between graphics versus words have an impact on mean group differences.

For the 159 items, the mean item *g*-loading was .30. To examine the effect of *g*-loading on mean group differences, we correlated the *g*-loading of the item with the mean group differences of the item. The mean group differences are expressed as standardized mean differences in which a positive *d* indicates that the White group had a larger mean score on an item than the minority group. The correlation of *g*-loading with mean White-Asian *d* was -.01. For White-Black *d*, the correlation was .60, and the correlation with White-Hispanic *d* was .35. Thus, on average, as item *g*-loading increases so the does the magnitude of White-minority item score differences.

One can use the item *g*-loading data to generate different *g* tests that vary in White-minority group differences. To demonstrate this, the 159 items were split at the mean of *g*-loading into a low-*g* group of 79 items and a high-*g* group of 80 items. From the low-*g* item group, we randomly selected 30 items and determined the mean White-minority differences, the internal consistency reliability (alpha) of the test, and the correlation of the test with educational attainment. We replicated this 30-item test construction process 100 times so that one can observe the range of possible group differences, reliabilities, and correlations with educational attainment for 30-item low-*g* scales. We repeated this process to obtain 100 40-item scales and their relevant statistics. Finally, we created a low-*g* scale with all 79 low-*g* items and a high-*g* scale with all 80 high-*g* items. Results are shown in Table 3 and support all three hypotheses.

We detail the results for the White-Asian analyses for the 30-item scales to explain the presented statistics. Mean score differences for Whites and Asians were expressed as standardized mean differences with a positive *d* indicating a score advantage for Whites and a negative *d* indicating a score advantage for Asians. Across the 100 low-*g* tests, each with 30 items, the mean White-Asian difference was 0.04. This is the mean *d* across 100 tests. For these 100 tests, the minimum *d* was -.09 and the maximum *d* was .16. The *d* of 0.04 indicates a very small score advantage for Whites over Asians. Now consider the 30 item scales derived from the high-*g* item set. The mean White-Asian *d* across the 100 tests, each with 30 items was -0.01. This indicates a very small mean score advantage for Asians over Whites. Across the 100 tests, the minimum White-Asian *d* was -.09 and the maximum was .08. For both the low-*g* tests and the high *g* tests, we concluded that the White-Asian mean score differences were very small. Now consider the reliability statistics. The 30-item low-*g* tests have a mean reliability of .60, with a minimum reliability of .54 and a maximum reliability of .68. The 100 high-*g* 30-item tests had a mean reliability of .84. The minimum reliability was .83 and the maximum reliability was .85. The mean correlations between the low-*g* scales and educational attainment was .08 (minimum = .01, maximum = .14), but the high-*g* scale was correlated .13 (minimum = .09, maximum = .17).

The reliability differences between the low-*g* and the high-*g* tests is due to the correlation between item *g*-loading and item variance (*r* = .56). Low-*g* items have less variance (mean variance of 79 items = .14) than high-*g* items (mean variance of 80 items = .18). Internal consistency reliability is a function, in part, of the intercorrelation among the items. Because a correlation is an indicator of shared variance, items with smaller variance will have smaller correlations with each other than items with larger variance, on average. This harms the reliability of low-*g* tests. Thus, for a test drawn from low-*g* items to have the same reliability as a test drawn from high-*g* items, it will need to have more items than the high-*g* test. Whereas many scholars consider a reliability of .80 to be the minimum reliability of a test to be used be used for making decisions (i.e., hiring) that affect people's lives, one should expect the need to make tests drawn from low *g*-item pools to be longer than high-*g* tests. As seen in the last section of table, the reliability was .80 for the low-*g* test with 79 items. In contrast, the average reliability of the test formed from all 80 high-*g* items was .93.

Now consider the White-Black comparisons on 30-item tests. For low-*g* tests, the mean White-Black *d* was .45 compared to the *d* of .57 for the high-*g* tests. One could alter the criterion for low-*g* items such that the tests have an even lower *g saturation* than in this analysis. This would result in even lower mean White-Black differences, on average, but would require many more items to achieve a reliability of .80.

For 30-item tests, the mean White-Hispanic *d* is .18 for the low-*g* tests and .25 for the high-*g* tests. Consistent with other U.S. samples, means group differences for Hispanics are smaller than group White-Black mean differences and larger than White-Asian mean differences.

Mean group difference results for the 40-item tests are not much different than the 30-item tests. White-Asian differences continue to be very small. White-Black and White-Hispanic differences increase slightly and this result is best attributed to the increased reliability of 40-item tests relative to the 30-item tests. Moving from 30 to 40 items, improves the reliability of the low-*g* tests from .60 to .66 and for the high-*g* tests from .84 to .88. In the last section of Table 3, we show the low and high-*g* tests composed of all available items. The magnitude of the group difference increase is due to the increase in reliability.

The correlation between the *g* scales and educational attainment covaries with *g*-loading and scale reliability. For example, the correlation is .09 for the low-*g* scale with 79 items and .14 for the high-*g* scale with 80 items. Thus, the correlation is 36% lower in the low-*g* scale than in the high-*g* scale.

Group difference comparisons for paired scales

We compared mean group differences for the two logic-based measurement scales with the same logical structure, one with real words and one with some fake words (see Figures 1 and 2). We also compared mean group difference for the comparison item scales with the same logical structure (see Figures 8 and 9). We used the mean of scored items as the scale scores.

Results for the paired scales are shown in Table 4. The statistical significance of the *d* between the two measures was determined by confidence interval overlap. The difference between the *d* of the real word version and the fake word version of the logic-based measurement scales was not statistically significant. Thus, using fake words in a logic-based measurement scale did not reduce mean group differences. The difference between the *d* for pictures (graphic comparison scale) versus words (written comparison scale) comparisons, was also not statistically significant. Thus, there appears to be no benefit in presenting items as pictures rather than words when comparing stimuli.

In summary, mean group differences in cognitive ability scales are driven by the *g*-saturation of the test and the reliability of the test. Reducing the *g*-saturation and reliability of a test also results in lower correlations with an external *g*-relevant criterion. Using fake words vs.

real words had no significant effect on mean group differences in logic-based measurement scales. In the analysis of scales that compare stimuli using graphics vs. words, mean group differences were not significantly different. Thus, this study casts additional doubt on the assertions concerning item types offered to explain why the Siena Reasoning Test has reduced mean group differences. We offer the alternative explanation that the test has lower mean group differences simply because the test has lower *g*-saturation. Consistent with this alternative explanation is the finding of Yusko, Goldstein, Scherbaum, and Hanges (2012) that the Siena Reasoning Test is correlated only .42, on average, with traditional cognitive ability tests. In conclusion, anyone with a pool of cognitive ability items can easily build a test with lower than typical mean group differences by using items that measure *g* less well. Such a test can be expected to have lower correlations with *g*-related criteria, such as job performance, than a test with high-*g* saturation.

References

Goldstein, H. (2008, November). *Building cognitive ability tests with reduced adverse impact*. Paper presented at the Mid-Atlantic Personnel Assessment Council.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.

Raven, J. C., Court, J. H., & Raven, J. (1994). *Advanced progressive matrices: Sets I and II. Manual for Raven's progressive matrices and vocabulary scales.* Oxford, England: Oxford Psychologists Press.

Yusko, K.P., Goldstein, H.W., Oliver, L.O. & Hanges, P.J. (2010). *Building cognitive ability tests with reduced adverse impact: Lowering reliance on prior knowledge*. Paper presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology. Atlanta.

Yusko, K.P., Goldstein, H.W., Scherbaum, C.A., & Hanges, P.J. (2012). *Siena Reasoning Test: Measuring intelligence with reduced adverse impact.* Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology. San Diego.

Table 1 Items by scale

Scales	Number of Items
GATB Object matching	39
GATB Three-dimensional space	12
Logic-based measurement (real words)	20
Logic-based measurement (fake words)	21
Sentence revision	11
STAT: Fake words	6
STAT: Unusual mathematical operators	5
STAT: Number series	5
Size, shape & shading	9
Table coding	6
Graphic comparison	13
Written comparison	12

Table 2. Scale correlation matrix

Scales	1	2	3	4	5	6	7	8	9	10	11	12
1. GATB Object matching												
2. GATB Three-dimensional space	.30											
3. Logic-based measurement (real words)	.21	.37										
4. Logic-based measurement (fake words)	.26	.35	.69									
5. Sentence revision	.23	.32	.48	.47								
6. STAT: Fake words	.27	.22	.39	.41	.48							
7. STAT: Unusual mathematical operators	.37	.37	.47	.44	.39	.31						
8. STAT: Number series	.26	.33	.37	.36	.34	.20	.56					
9. Size, shape & shading	.28	.38	.42	.40	.38	.36	.39	.33				
10. Table coding	.26	.20	.38	.37	.36	.31	.37	.26	.31			
11. Graphic comparison	.35	.32	.41	.37	.28	.30	.41	.33	.43	.23		
12. Written comparison	.33	.34	.47	.44	.44	.40	.46	.39	.44	.34	.52	

Table 3. Demonstration of how item *g*-loading and test length can be used to manipulate mean group differences, reliability, and correlations with an external variable (educational attainment)

30 item Scales						
	Low *g*	High *g*	Low *g* Reliability	High *g* Reliability	Low *g* Correlation with Education	High *g* Correlation with Education
White - Asian *d* (Min Max)	0.04 (-0.09,0.16)	-0.01 (-.09,0.08)	.60 (.54, .68)	.84 (.83, .85)	.08 (.01, .14)	.13 (.09, .17)
White – Black *d* (Min, Max)	0.45 (0.35,0.58)	0.57 (0.52,0.64)				
White – Hispanic *d* (Min, Max)	0.18 (0.08,0.28)	0.25 (0.17,0.33)				
40 items scales						
	Low *g*	High *g*	Low *g* Reliability	High *g* Reliability	Low *g* Correlation with Education	High *g* Correlation with Education
White - Asian *d* (Min Max)	0.04 (-0.06,0.14)	0.00 (0.07,0.07)	.66 (.62, .70)	.88 (.86,.89)	.09 (.05, .14)	.13 (.09, .16)
White – Black *d* (Min, Max)	0.47 (0.37,0.58)	0.58 (0.52,0.64)				
White – Hispanic *d* (Min, Max)	0.19 (0.12,0.29)	0.26 (0.20,0.33)				
All Items (79 items for Low *g* and 80 items for High *g*)						
	Low *g*	High *g*	Low *g* Reliability	High *g* Reliability	Low *g* Correlation with Education	High *g* Correlation with Education
White - Asian *d*	0.05	-0.01	.80	.93	.09	.14
White – Black *d*	0.52	0.60				
White – Hispanic *d*	0.21	0.27				

Table 4. Comparison of scales designed to have the same logical structure

	Logic-based measurement (real words)	Logic-based measurement (fake words	*d* difference significant?
White-Asian *d* (confidence interval)	0.11 (-0.08, 0 .29)	0.03 (-0.15, 0.22)	No
White-Black *d* (confidence interval	0.37 (0.19, 0.55)	0.38 (0.20, 0.56)	No
White-Hispanic *d* (confidence interval)	0.19 (0.01, 0.36)	0.11 (-0.07, 0.29)	No
	Picture (Graphic Comparison)	Words (Written Comparison)	*d* difference significant?
White-Asian *d* (confidence interval)	0.06 (-0.12, 0.25)	-0.11 (-0.29, 0.08)	No
White-Black *d* (confidence interval)	0.42 (0.24, 0.60)	0.34 (0.16, 0.52)	No
White-Hispanic *d* (confidence interval)	0.29 (0.03, 0.39)	0.33 (0.15, 0.51)	No

Figure 1. Instruction items from the logic-based measure scale using real words

LBM Verbal Reasoning
Instructions

The following passage describes a set of facts. The passage is followed by several conclusions. Read the passage and then decide whether each conclusion is:

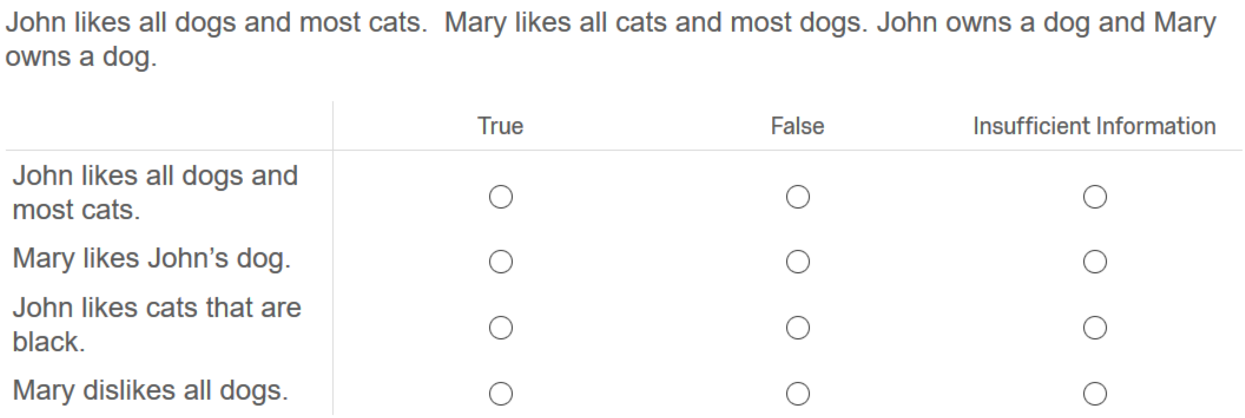
True, which means that the conclusion has to be true from the facts given; or

False, which means that the conclusion has to be false because it is contrary to the facts given;

Or, whether there is **Insufficient information to decide**, which means that there is insufficient information for you to determine whether the facts mean that the conclusion is true or is false.

Base your evaluation of the conclusions **SOLELY** on the information in the passage. Do **NOT** use any outside factual information to reach your conclusion. Work exclusively with the information provided.

Example:



John likes all dogs and most cats. Mary likes all cats and most dogs. John owns a dog and Mary owns a dog.

	True	False	Insufficient Information
John likes all dogs and most cats.	○	○	○
Mary likes John’s dog.	○	○	○
John likes cats that are black.	○	○	○
Mary dislikes all dogs.	○	○	○

Explanation:
1) **True** The first fact states that John likes all dogs. Thus, conclusion 1 is true.
2) **Insufficient information to decide**. The third fact indicates that John owns a dog. The second fact indicates that Mary likes most dogs. The facts do not indicate whether Mary likes John’s dog. Thus, there is insufficient information to determine whether conclusion 2 is true or false.
3) **Insufficient information to decide**. The first fact indicates that John likes most cats. One does not know if the cats that John likes include black cats. Thus, there is insufficient information to determine whether conclusion 2 is true or false.
4) **False**. The second fact indicates that Mary likes most dogs. Thus, the conclusion that Mary dislikes all dogs must be false.

Figure 2. Instruction items from the logic-based measure scale using fake words

LBM Verbal Reasoning With Unusual Words
Instructions

The following passage describes a set of facts. The passage contains some unusual words (e.g., *dosf*) that are presented in italics. You do not need to know the meaning of the words to answer the questions. The passage is followed by several conclusions. Read the passage and then decide whether each conclusion is:

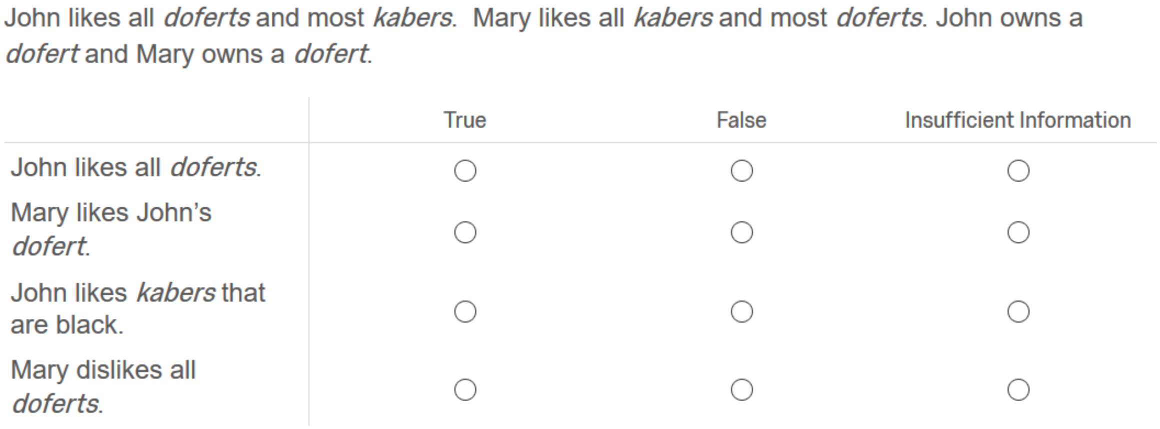
True, which means that the conclusion has to be true from the facts given; or

False, which means that the conclusion has to be false because it is contrary to the facts given;

Or, whether there is **Insufficient information to decide**, which means that there is insufficient information for you to determine whether the facts mean that the conclusion is true or is false.

Base your evaluation of the conclusions **SOLELY** on the information in the passage. Do **NOT** use any outside factual information to reach your conclusion. Work exclusively with the information provided.

Example:



John likes all *doferts* and most *kabers*. Mary likes all *kabers* and most *doferts*. John owns a *dofert* and Mary owns a *dofert*.

	True	False	Insufficient Information
John likes all *doferts*.	○	○	○
Mary likes John's *dofert*.	○	○	○
John likes *kabers* that are black.	○	○	○
Mary dislikes all *doferts*.	○	○	○

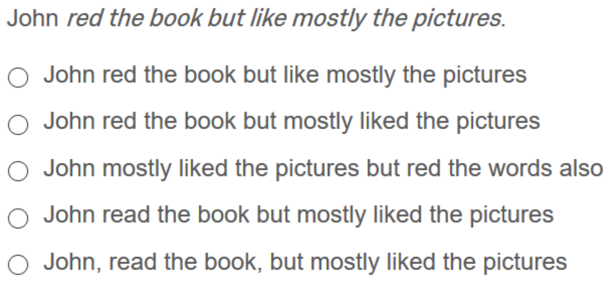
Explanation:
1) **True.** The first fact states that John likes all *doferts*. Thus, conclusion 1 is true.
2) **Insufficient information to decide**. The third fact indicates that John owns a *dofert*. The second fact indicates that Mary likes most *doferts*. The facts do not indicate whether Mary likes John's *dofert*. Thus, there is insufficient information to determine whether conclusion 2 is true or false.
3) **Insufficient information to decide**. The first fact indicates that John likes most *kabers*. One does not know if the *kabers* that John likes include black *kabers*. Thus, there is insufficient *information* to determine whether conclusion 3 is true or false.
4) **False**. The second fact indicates that Mary likes most *doferts*. Thus, the conclusion that Mary dislikes all *doferts* must be false.

Figure 3. Instruction item from the sentence revision scale

Sentence Revision
Instructions

Each item in this section presents one sentence. Part of the sentence or the entire sentence is underlined and may need to be revised. If the underlined section seems correct, choose the first response but read the other choices to make sure that the first response is the best choice. If the underlined section seems incorrect, choose the response from the remaining responses that makes the sentence correct. If none of the responses seem to make the sentence correct, pick the response that presents the sentence most clearly.

Example:



John *red the book but like mostly the pictures*.

○ John red the book but like mostly the pictures

○ John red the book but mostly liked the pictures

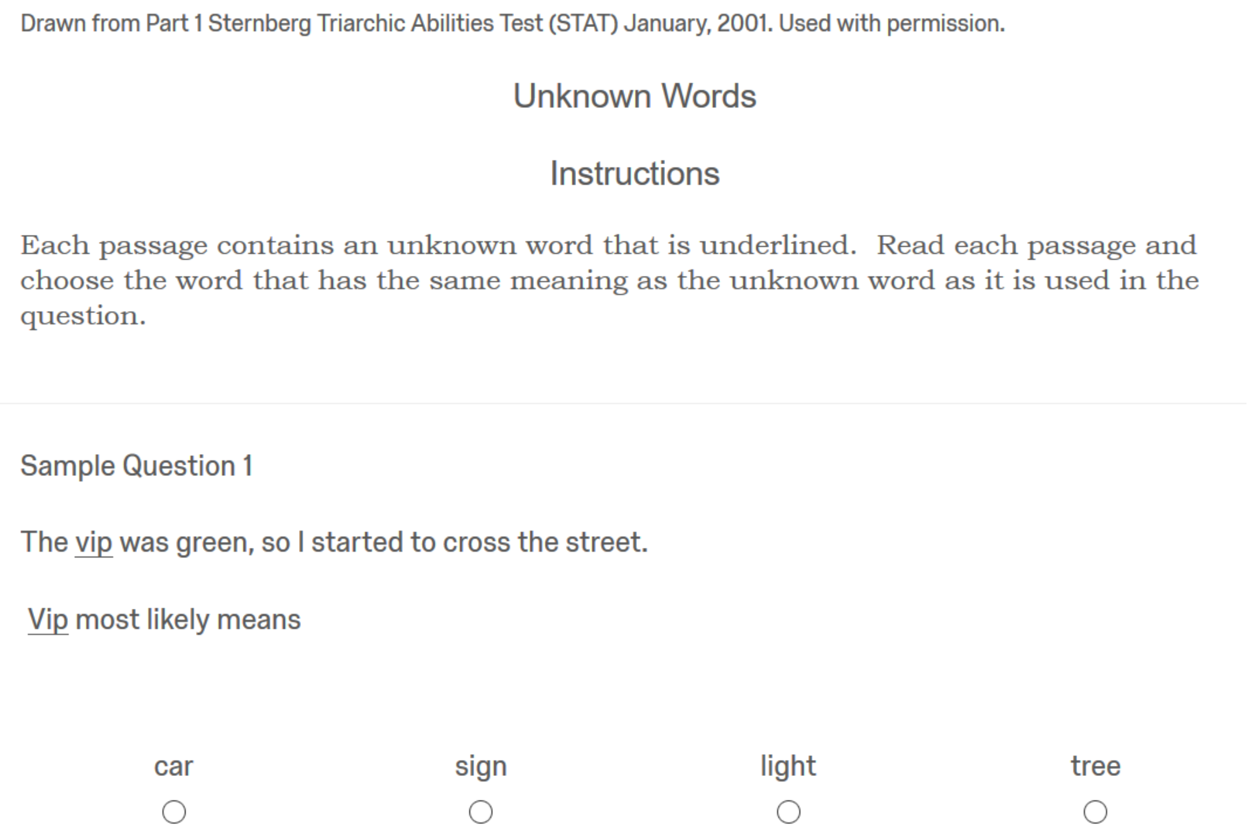
○ John mostly liked the pictures but red the words also

○ John read the book but mostly liked the pictures

○ John, read the book, but mostly liked the pictures

The best sentence was the fourth option: *John read the book but mostly liked the pictures*. Pick that as the answer.

Figure 4. Instruction item from the Sternberg Triarchic Abilities Test (January 2001 version) scale that uses fake words.



Drawn from Part 1 Sternberg Triarchic Abilities Test (STAT) January, 2001. Used with permission.

Unknown Words

Instructions

Each passage contains an unknown word that is underlined. Read each passage and choose the word that has the same meaning as the unknown word as it is used in the question.

Sample Question 1

The vip was green, so I started to cross the street.

Vip most likely means

car	sign	light	tree
○	○	○	○

Figure 5. Instruction item from the Sternberg Triarchic Abilities Test (January 2001 version) scale that uses new mathematical operators.

Drawn from Part 8 Sternberg Triarchic Abilities Test (STAT) January, 2001. Used with permission.

Mathematical Operations
Instructions

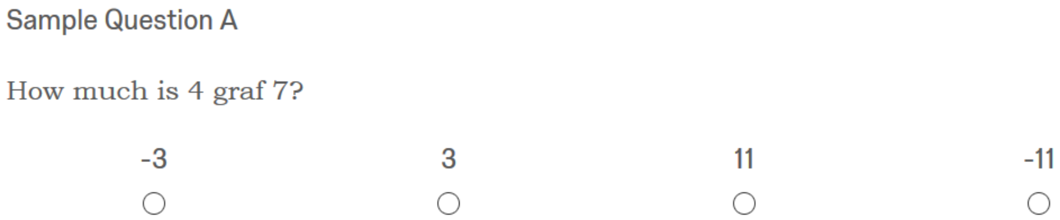
In each problem below, you will employ unusual mathematical operations in order to reach the solution. There are three unusual operations: graf, flix, and trup. First, read how the operations are defined. Then, decide what is the correct answer to the question.

There is a new mathematical operation called graf. It is defined as follows:

$$x \text{ graf } y = x + y, \text{ if } x < y$$

but $$x \text{ graf } y = x – y, \text{ if otherwise}$$

(*the instructions then continue, defining the operators flix and trup*)



Sample Question A

How much is 4 graf 7?

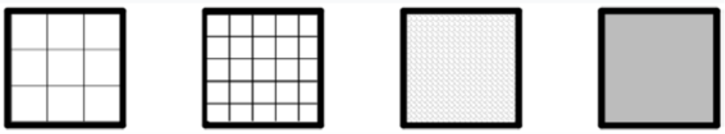
-3	3	11	-11
○	○	○	○

Figure 6. Instructions for the size, shape and shading items

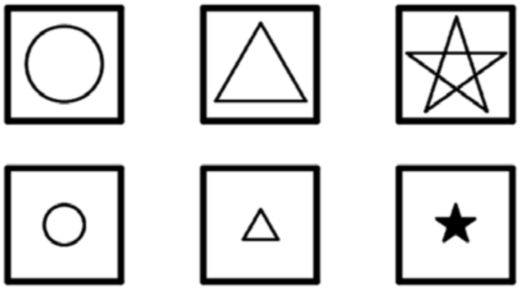
Size, Shape, and Shading
Instructions

The next type of question presents 8 boxes, labelled A through H. The boxes can differ from each other with respect to the shape inside the box, whether the shape is small or large, and the shading inside the box.

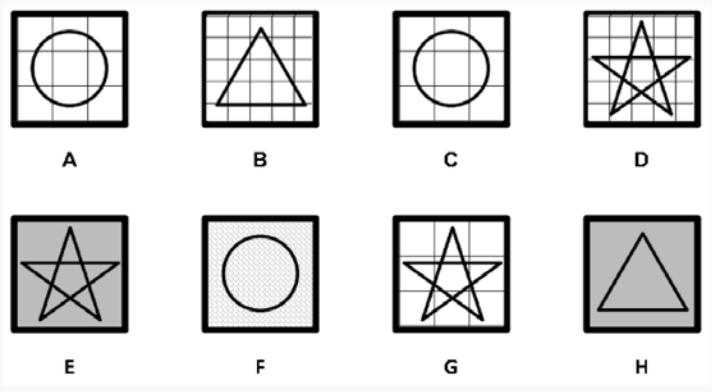
Below are boxes showing the four types of shading.

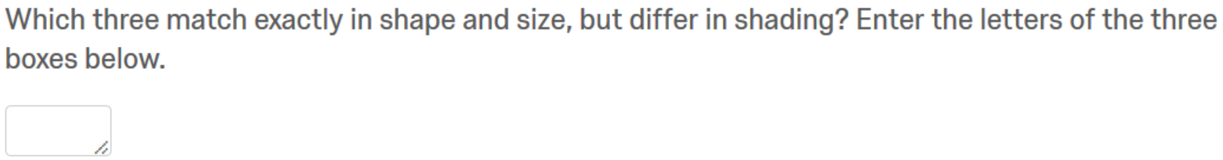


Below are the types of shapes. Note that some are large and some are small.



In the example question that follows, look at the boxes and answer the question about the boxes.





Which three match exactly in shape and size, but differ in shading? Enter the letters of the three boxes below.

Figure 7. Instructions for the table coding items

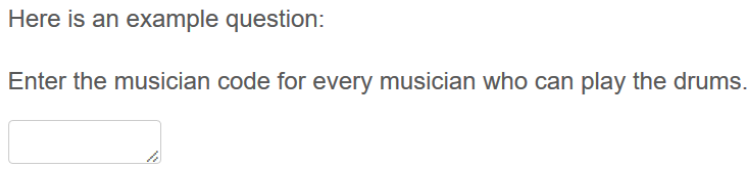
Table Questions

Instructions

In this section, a table is presented and then you are asked questions based on information in the table. Look at the table below.

A band is being formed. The table below shows information on musicians who might join the band. The first column shows person codes assigned to each musician. The last three columns indicate whether the musician can play the piano, the guitar, and drums. Some musicians can play one instrument, some can play two instruments and some can play three instruments.

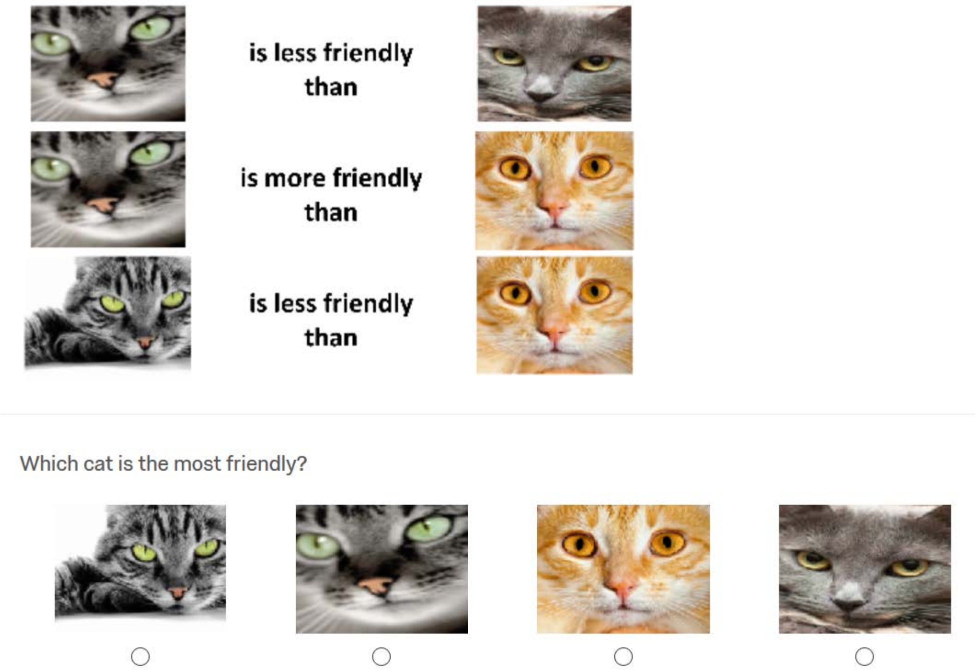
Musician Code	Plays Piano	Plays Guitar	Plays Drums
A	Yes	No	Yes
B	No	Yes	No
C	Yes	No	Yes
D	Yes	Yes	No
E	Yes	No	No
F	No	No	Yes
G	No	Yes	Yes
H	Yes	Yes	Yes



Here is an example question:

Enter the musician code for every musician who can play the drums.

Figure 8. Comparison item using pictures



is less friendly than

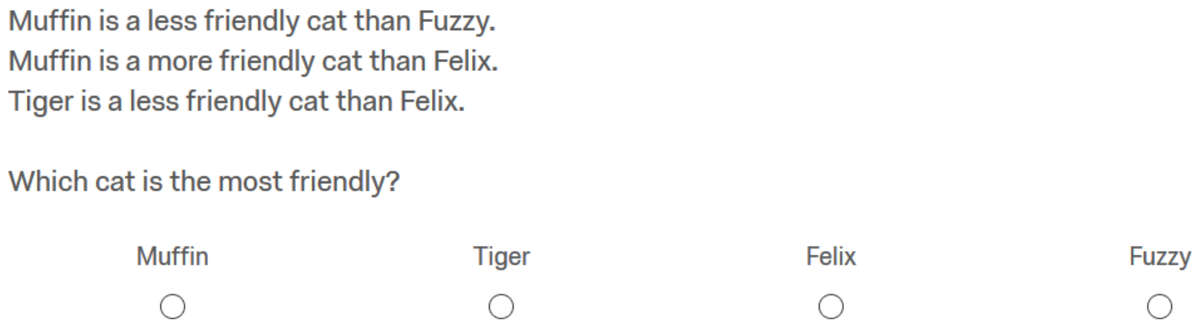
is more friendly than

is less friendly than

Which cat is the most friendly?

○ ○ ○ ○

Figure 9. Comparison item using words



Muffin is a less friendly cat than Fuzzy.
Muffin is a more friendly cat than Felix.
Tiger is a less friendly cat than Felix.

Which cat is the most friendly?

Muffin	Tiger	Felix	Fuzzy
○	○	○	○