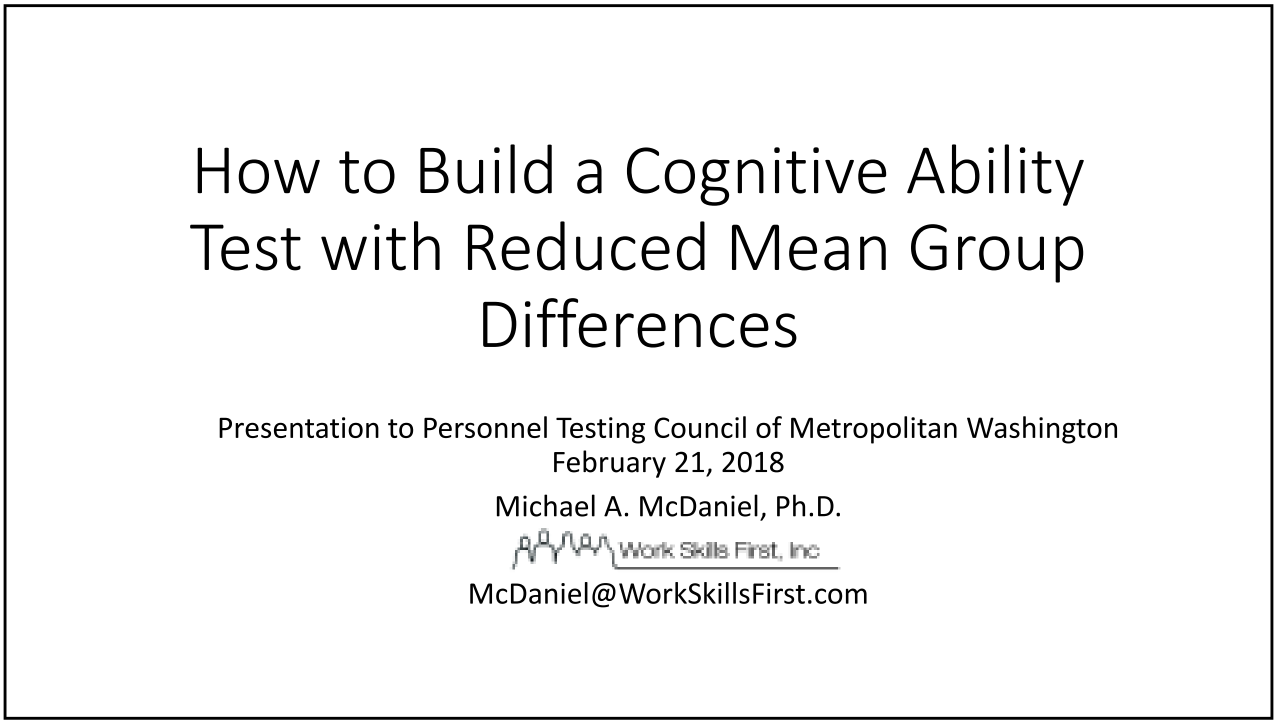


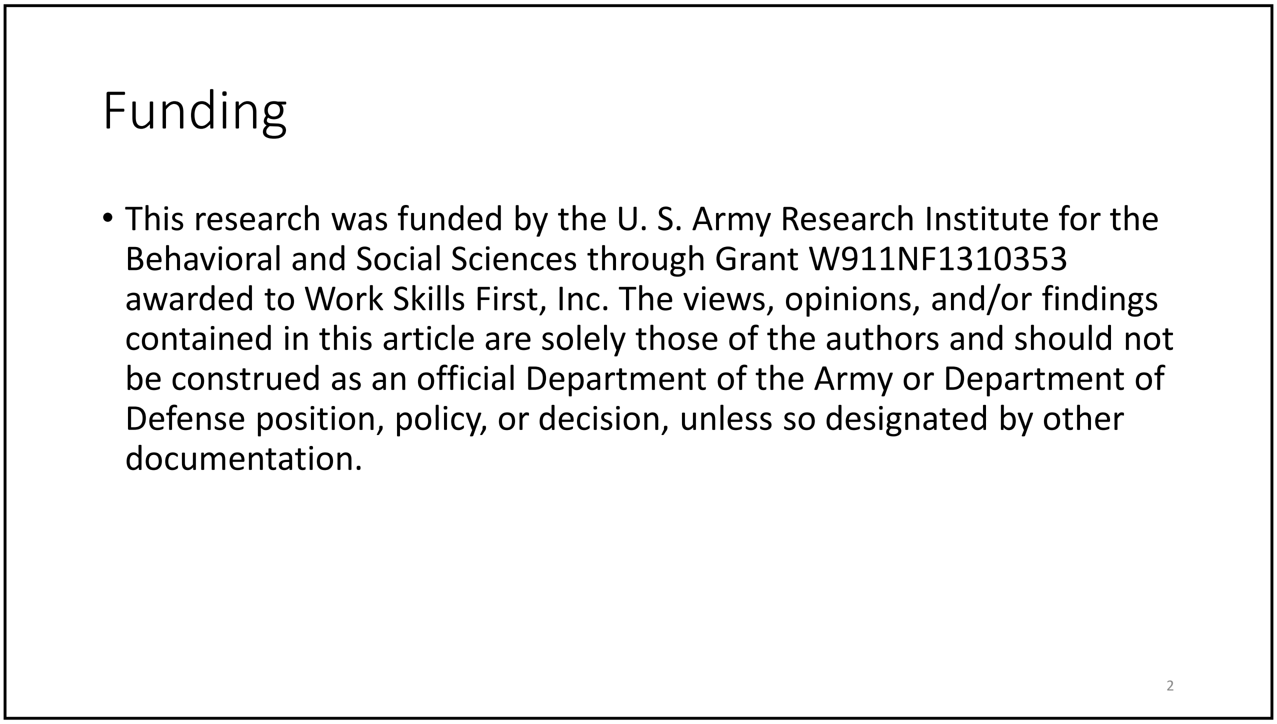
# How to Build a Cognitive Ability Test with Reduced Mean Group Differences

Presentation to Personnel Testing Council of Metropolitan Washington
February 21, 2018

Michael A. McDaniel, Ph.D.

Work Skills First, Inc

McDaniel@WorkSkillsFirst.com

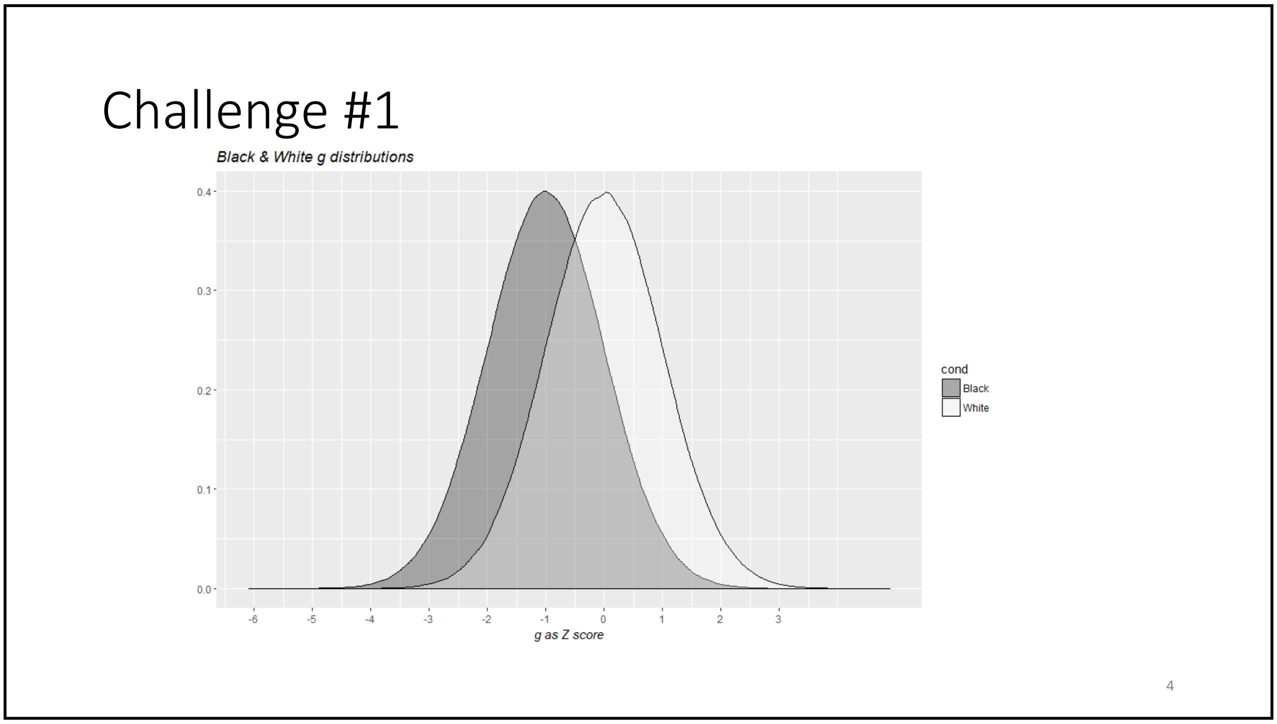


# Funding

- This research was funded by the U. S. Army Research Institute for the Behavioral and Social Sciences through Grant W911NF1310353 awarded to Work Skills First, Inc. 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.

# Overview

- Two challenges in building a general cognitive ability (*g)* test with low mean group differences
- Ways to build a *g* test with low mean group differences
- Brief summary of grant research approach and findings
- Summary



## Challenge #1

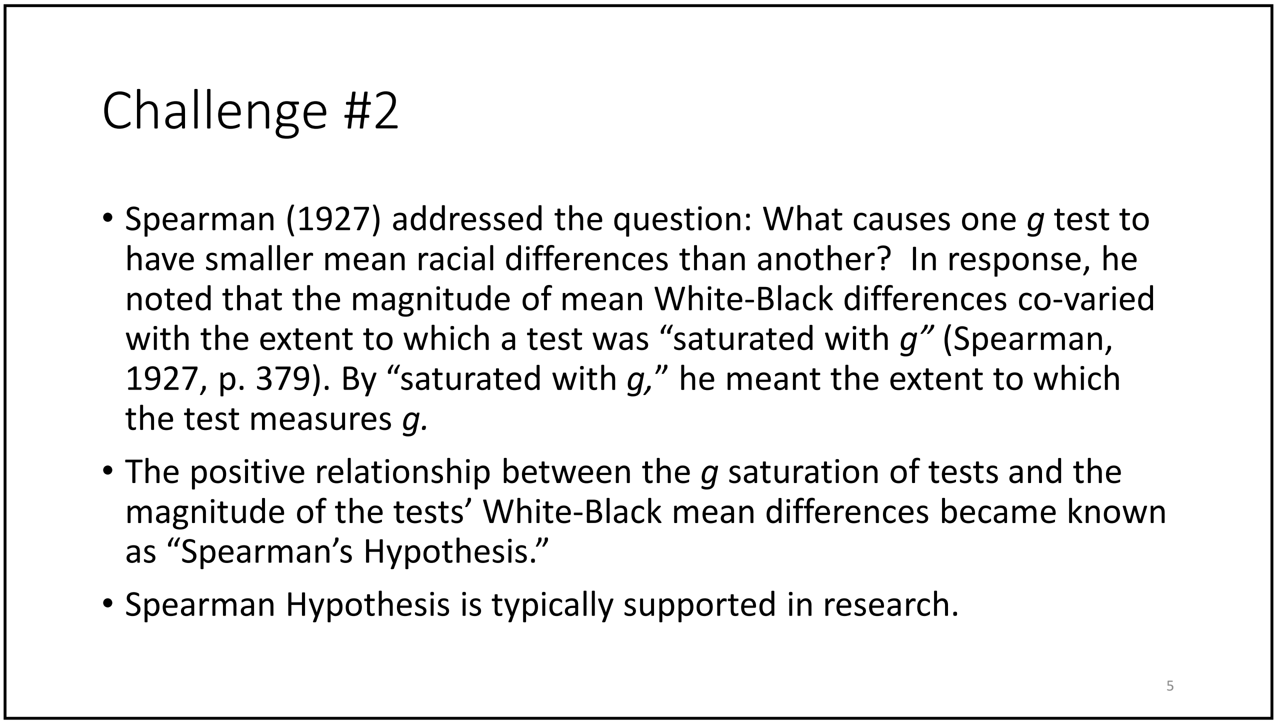
Challenge #1 is that there are large Black-White mean differences in *g*.

The graph shows two density plots. This simulates the distribution of Whites and Blacks mean cognitive ability (*g*) in a *z* score metric (0,1). Whites have a mean of zero, corresponding to an IQ of 100 and Blacks have a mean of -1 corresponding to an IQ of 85.

This is an accurate but terrible situation.

If we added Hispanic and Asian, the Hispanic mean would be between the Black mean and the White mean. The Asian mean would be slightly higher than the White mean.

It appears that if one wants to measure *g* well yet have smaller mean group differences, one might need to build the test with easy items. Such a test would differentiate respondents well at low levels of *g* but differentiate respondents poorly near the mean or at higher scores. In most personnel selection screening scenarios, one wants to differentiate respondents at or above the mean.



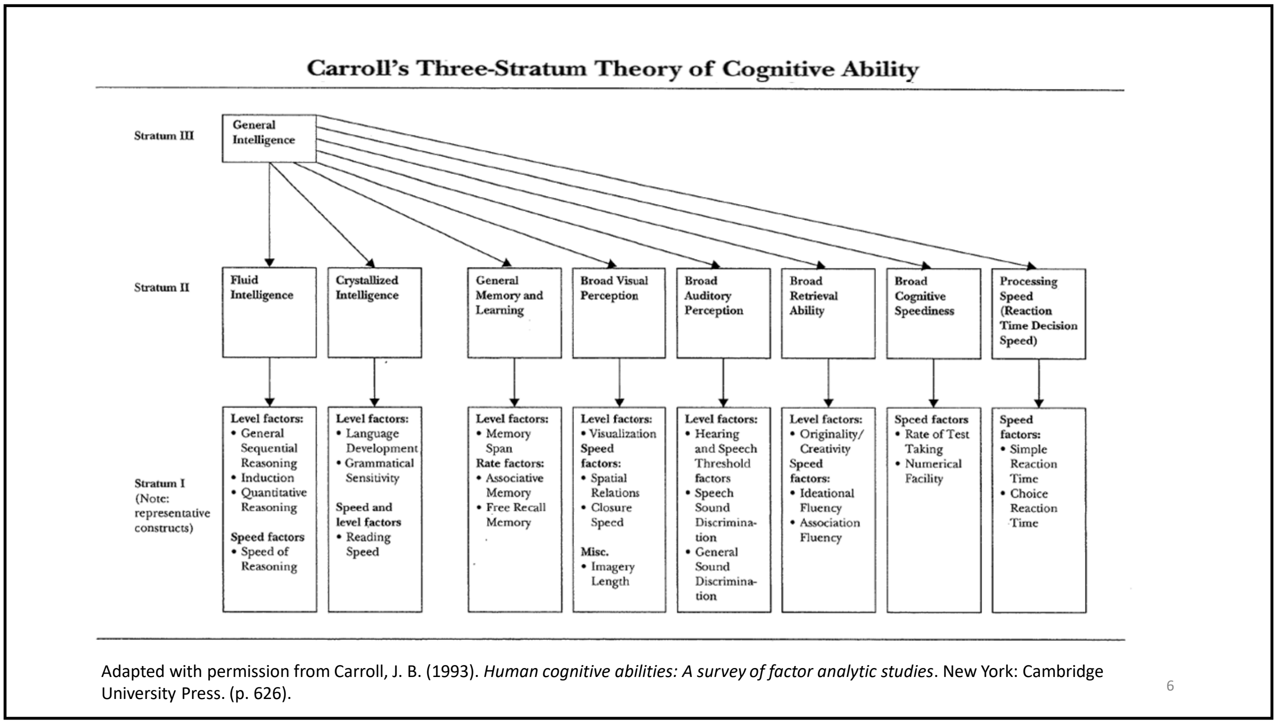
# Challenge #2

- Spearman (1927) addressed the question: What causes one *g* test to have smaller mean racial differences than another? In response, he noted that the magnitude of mean White-Black differences co-varied with the extent to which a test was “saturated with *g”* (Spearman, 1927, p. 379). By “saturated with *g,*” he meant the extent to which the test measures *g.*
- The positive relationship between the *g* saturation of tests and the magnitude of the tests’ White-Black mean differences became known as “Spearman’s Hypothesis.”
- Spearman Hypothesis is typically supported in research.

Challenge #2 is that if Spearman’s Hypothesis is correct, then one can’t build a *g* test that measures *g* well (has high *g* saturation) and has low mean group differences.

However, one can build a *g* test that does not measure *g* well (has low *g* saturation) that has smaller mean group differences.

So what *g*-ish measures assess *g* poorly? Carroll’s three stratum model suggests constructs that are related to *g* but which have lower *g* saturation. Interested parties can consult Chapters 5 through 15 of Carroll (1993) for references to studies using less *g*-ish measures. Google searches also generate helpful information.



## Carroll's Three-Stratum Theory of Cognitive Ability

Stratum III: General Intelligence

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Stratum II | Fluid Intelligence | Crystallized Intelligence | General Memory and Learning | Broad Visual Perception | Broad Auditory Perception | Broad Retrieval Ability | Broad Cognitive Speediness | Processing Speed (Reaction Time Decision Speed) |
| Stratum I (Note: representative constructs) | **Level factors:**<br>• General Sequential Reasoning<br>• Induction<br>• Quantitative Reasoning<br>**Speed factors**<br>• Speed of Reasoning | **Level factors:**<br>• Language Development<br>• Grammatical Sensitivity<br>**Speed and level factors**<br>• Reading Speed | **Level factors:**<br>• Memory Span<br>**Rate factors:**<br>• Associative Memory<br>• Free Recall Memory | **Level factors:**<br>• Visualization<br>**Speed factors:**<br>• Spatial Relations<br>• Closure Speed<br>**Misc.**<br>• Imagery Length | **Level factors:**<br>• Hearing and Speech Threshold factors<br>• Speech Sound Discrimination<br>• General Sound Discrimination | **Level factors:**<br>• Originality/ Creativity<br>**Speed factors:**<br>• Ideational Fluency<br>• Association Fluency | **Speed factors**<br>• Rate of Test Taking<br>• Numerical Facility | **Speed factors:**<br>• Simple Reaction Time<br>• Choice Reaction Time |

Adapted with permission from Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press. (p. 626).

In the graphic, the stratum 2 abilities that are most *g* saturated are to the left and the *g* saturation of abilities moving right across the graphic have less and less *g* saturation. Thus, fluid ability has the most *g* saturation, crystallized ability has the second most *g* saturation, and processing speed has the least *g* saturation.

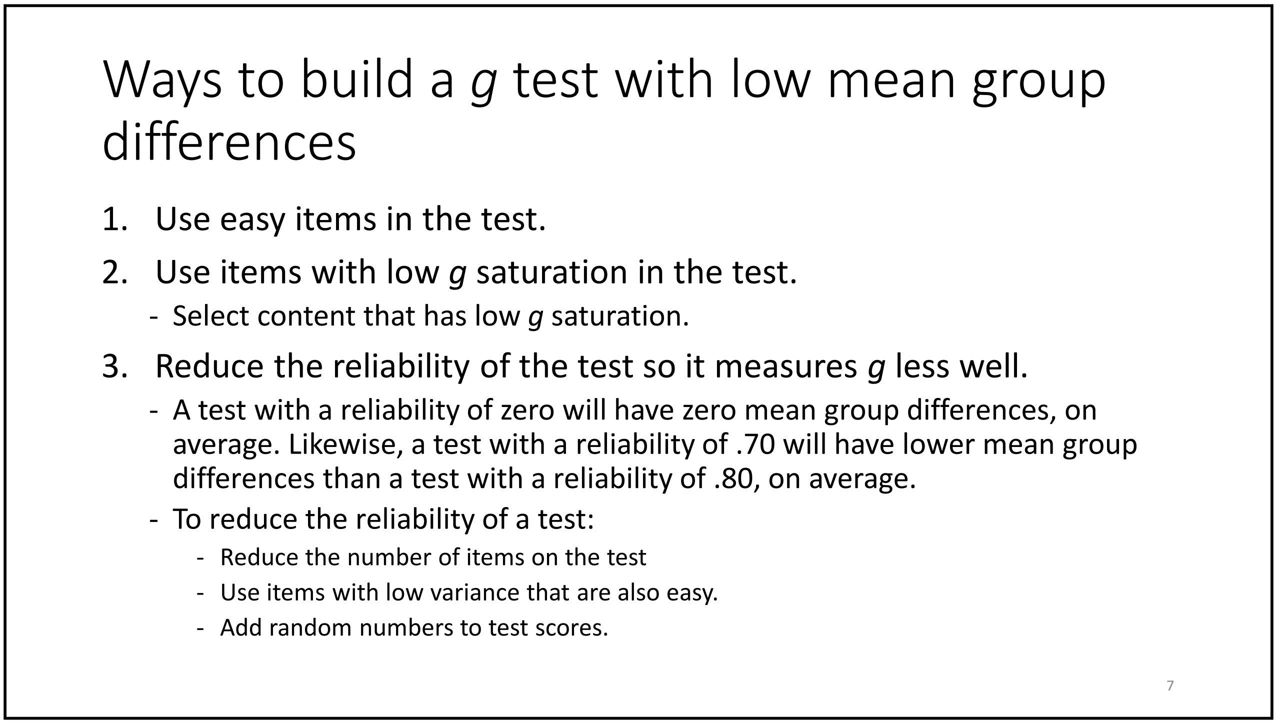
To locate less *g*-saturated measures to be used in building a low *g* test, one can look in Carroll (1993) book to find references for studies that have used the measures.

Additional sources for information and samples tests are:

Ekstrom, R.B., French, J.W., Harmen, H.H., & Dermen, D. (1976a). *Kit of factor-referenced cognitive tests*. ETS: Princeton, NJ.

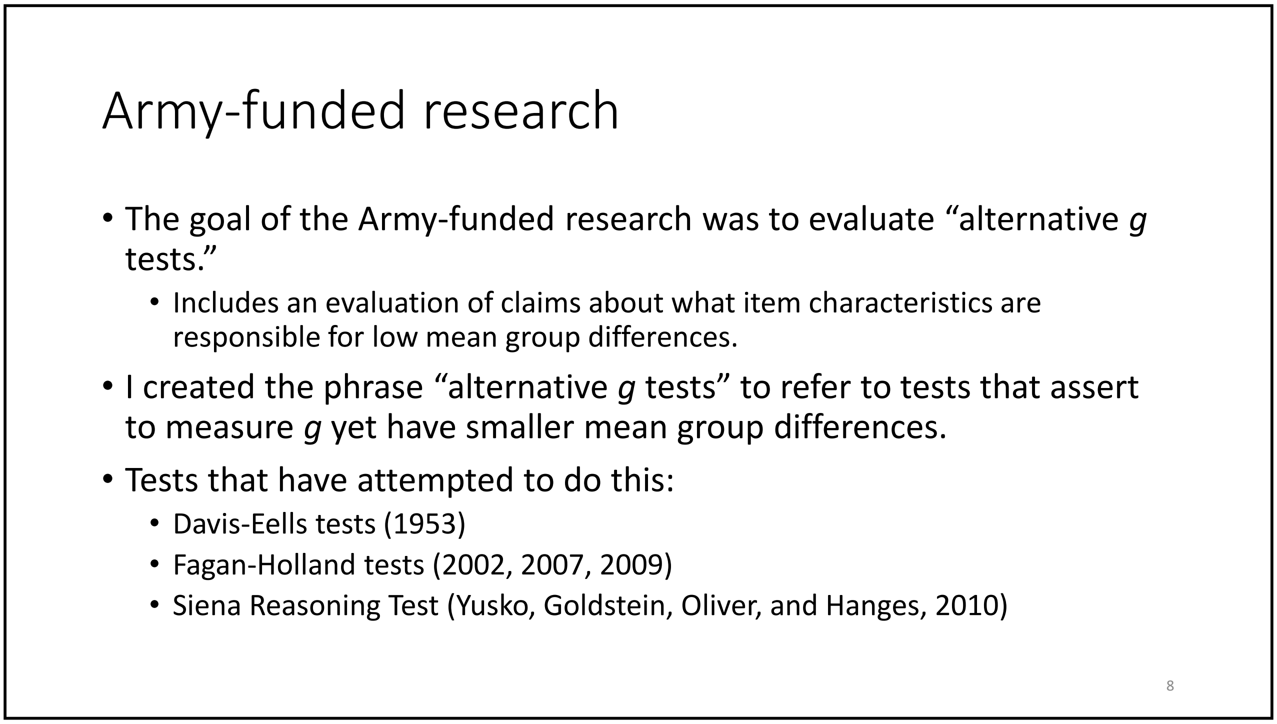
Ekstrom, R.B., French, J.W., Harmen, H.H., & Dermen, D. (1976b). *Manual for kit of factor-referenced cognitive tests*. ETS: Princeton, NJ.

These reports can be downloaded for free from ETS. See the references slide for links to the reports.



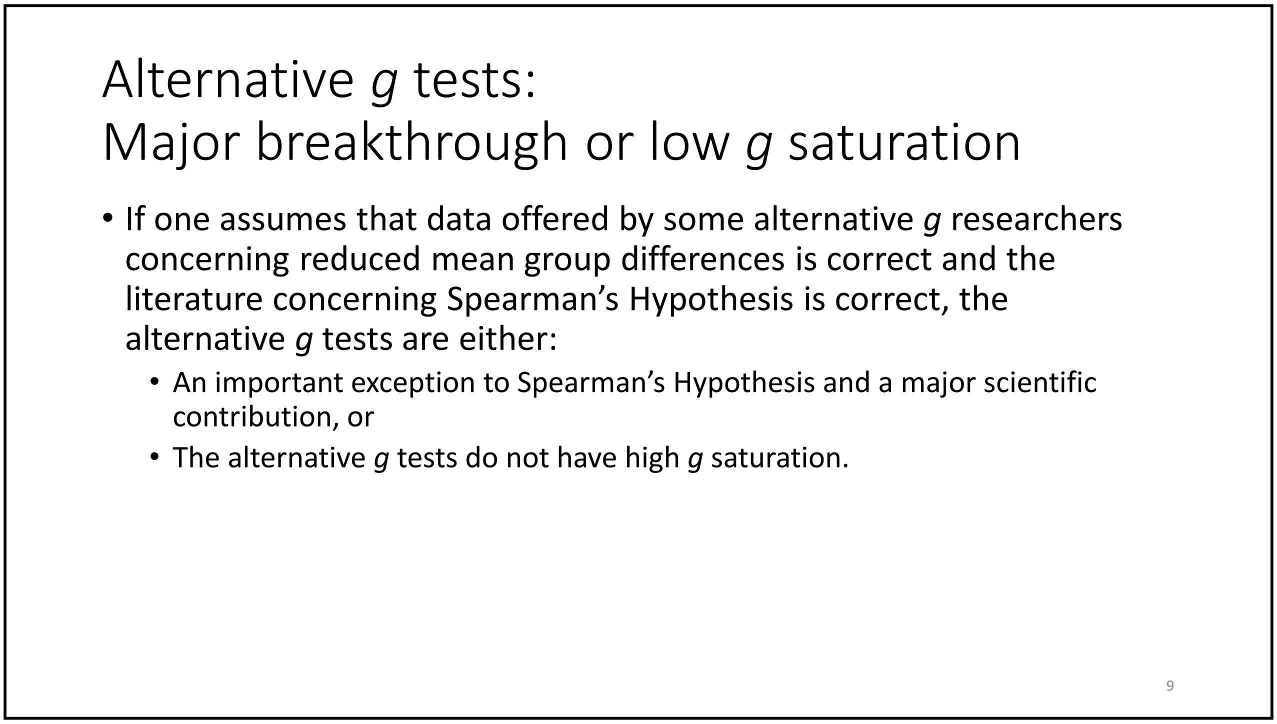
# Ways to build a *g* test with low mean group differences

1. Use easy items in the test.
2. Use items with low *g* saturation in the test.
   - Select content that has low *g* saturation.
3. Reduce the reliability of the test so it measures *g* less well.
   - A test with a reliability of zero will have zero mean group differences, on average. Likewise, a test with a reliability of .70 will have lower mean group differences than a test with a reliability of .80, on average.
   - To reduce the reliability of a test:
     - Reduce the number of items on the test
     - Use items with low variance that are also easy.
     - Add random numbers to test scores.



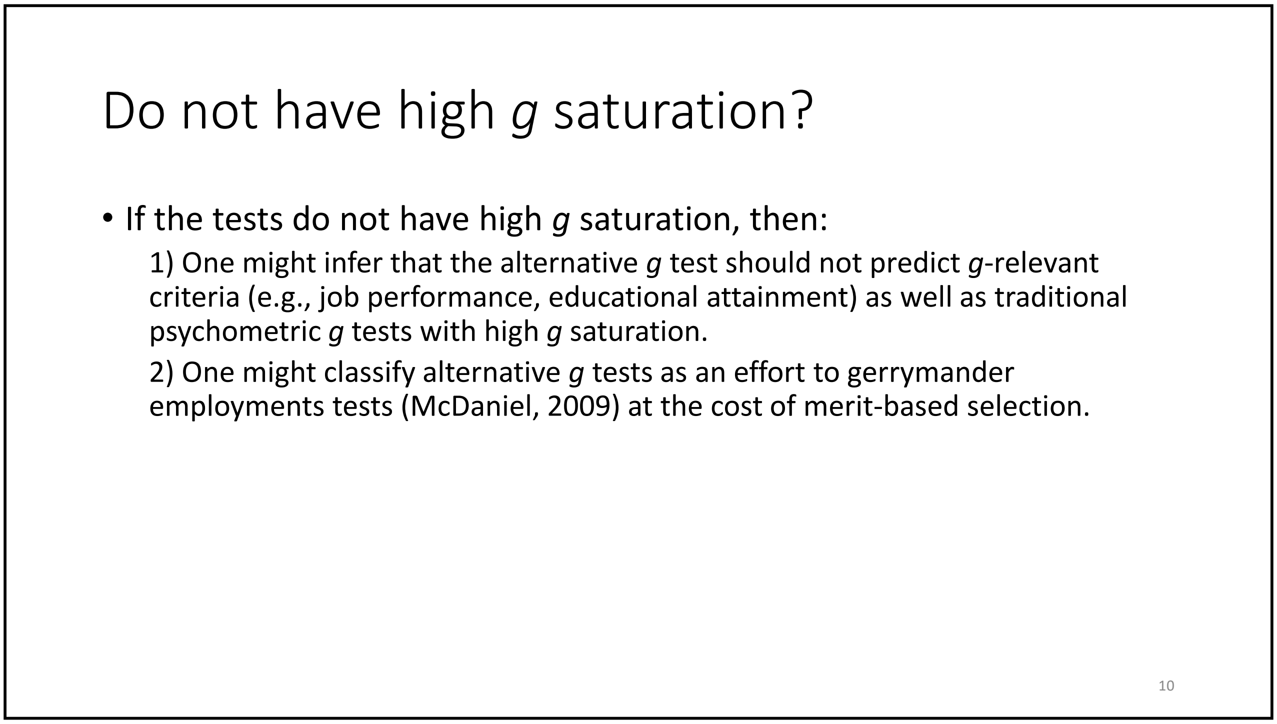
# Army-funded research

- The goal of the Army-funded research was to evaluate "alternative *g* tests."
    - Includes an evaluation of claims about what item characteristics are responsible for low mean group differences.
- I created the phrase "alternative *g* tests" to refer to tests that assert to measure *g* yet have smaller mean group differences.
- Tests that have attempted to do this:
    - Davis-Eells tests (1953)
    - Fagan-Holland tests (2002, 2007, 2009)
    - Siena Reasoning Test (Yusko, Goldstein, Oliver, and Hanges, 2010)



# Alternative *g* tests:<br>Major breakthrough or low *g* saturation

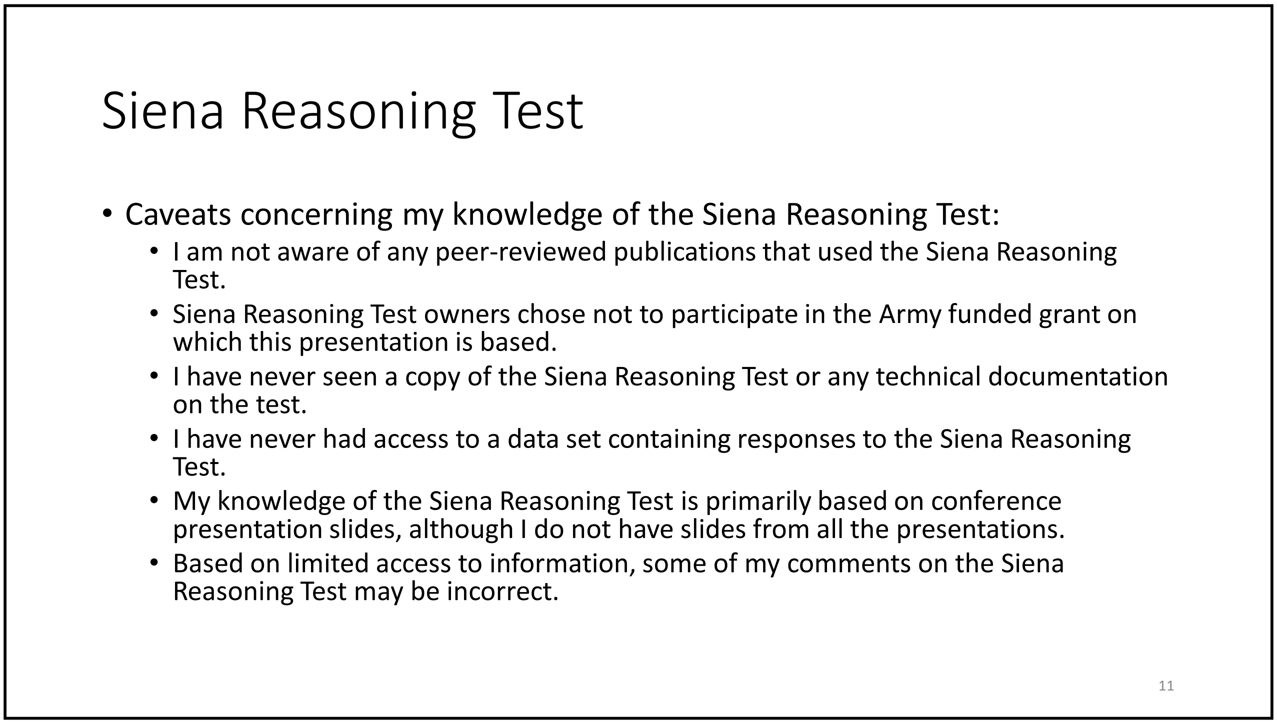
- If one assumes that data offered by some alternative *g* researchers concerning reduced mean group differences is correct and the literature concerning Spearman’s Hypothesis is correct, the alternative *g* tests are either:
  - An important exception to Spearman’s Hypothesis and a major scientific contribution, or
  - The alternative *g* tests do not have high *g* saturation.



# Do not have high *g* saturation?

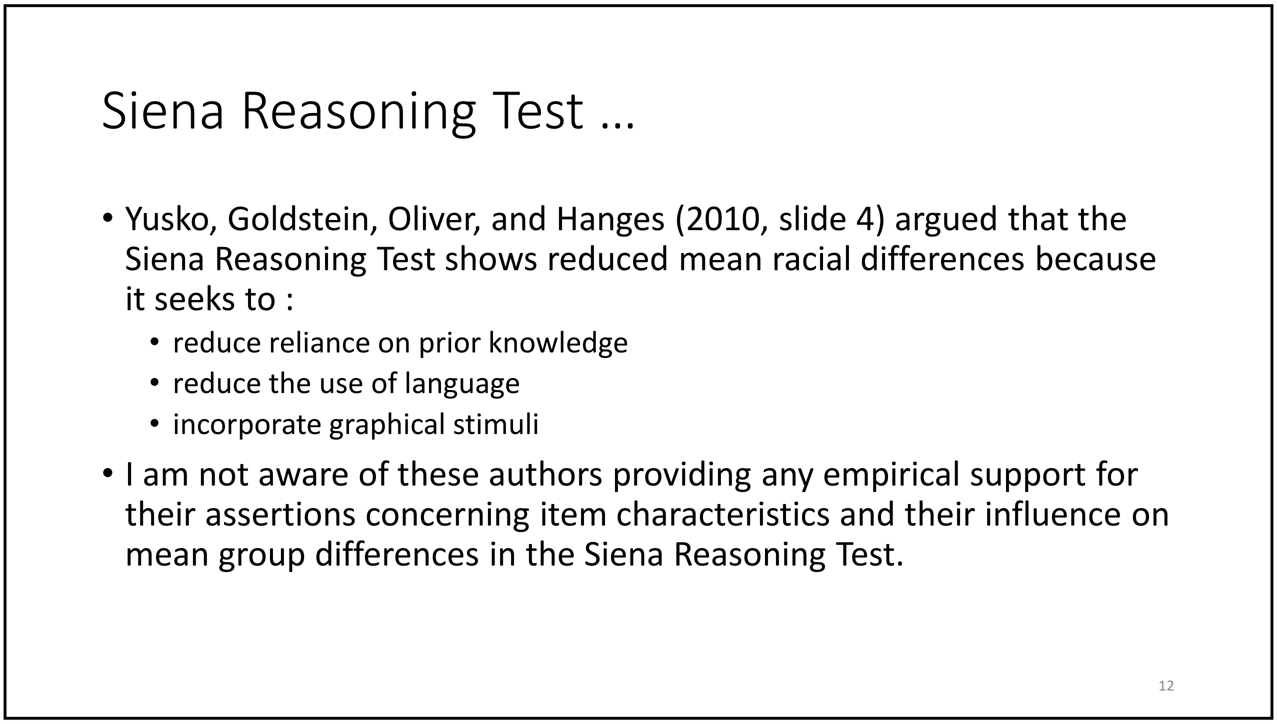
- If the tests do not have high *g* saturation, then:
  1) One might infer that the alternative *g* test should not predict *g*-relevant criteria (e.g., job performance, educational attainment) as well as traditional psychometric *g* tests with high *g* saturation.

  2) One might classify alternative *g* tests as an effort to gerrymander employments tests (McDaniel, 2009) at the cost of merit-based selection.



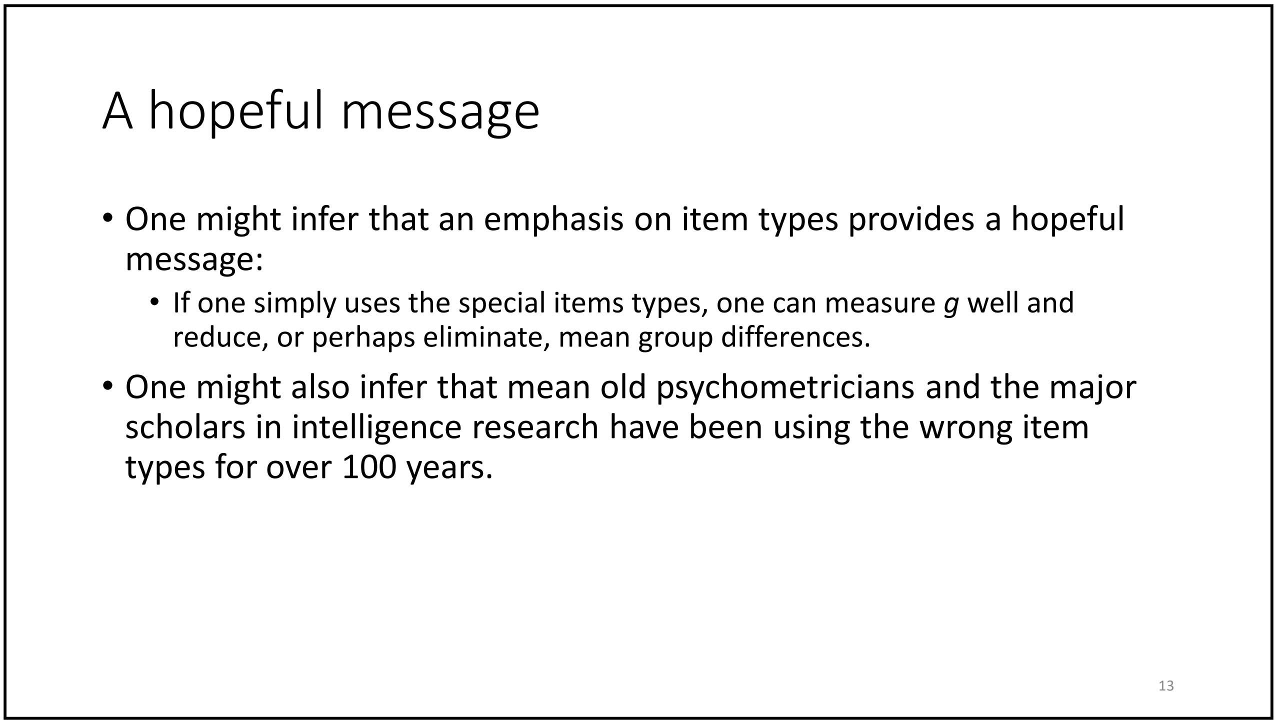
# Siena Reasoning Test

- Caveats concerning my knowledge of the Siena Reasoning Test:
    - I am not aware of any peer-reviewed publications that used the Siena Reasoning Test.
    - Siena Reasoning Test owners chose not to participate in the Army funded grant on which this presentation is based.
    - I have never seen a copy of the Siena Reasoning Test or any technical documentation on the test.
    - I have never had access to a data set containing responses to the Siena Reasoning Test.
    - My knowledge of the Siena Reasoning Test is primarily based on conference presentation slides, although I do not have slides from all the presentations.
    - Based on limited access to information, some of my comments on the Siena Reasoning Test may be incorrect.



# Siena Reasoning Test …

- Yusko, Goldstein, Oliver, and Hanges (2010, slide 4) 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
  - incorporate graphical stimuli
- I am not aware of these authors providing any empirical support for their assertions concerning item characteristics and their influence on mean group differences in the Siena Reasoning Test.

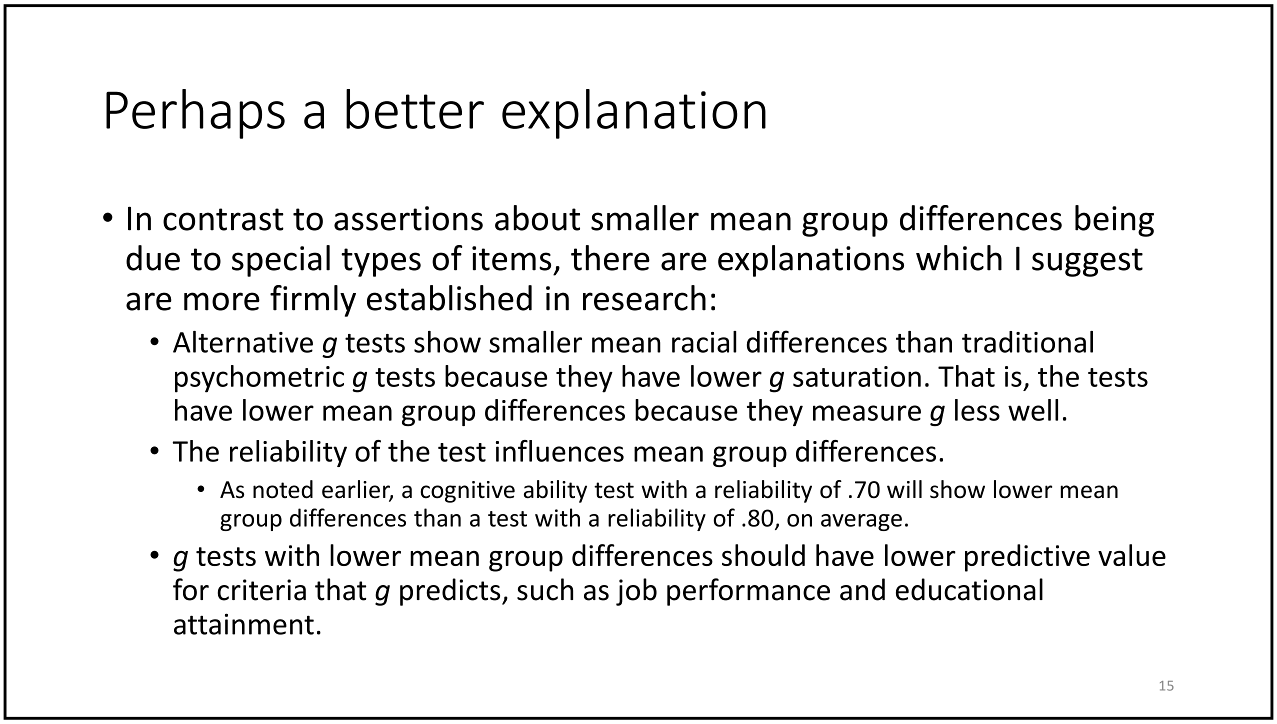


# A hopeful message

- One might infer that an emphasis on item types provides a hopeful message:
  - If one simply uses the special items types, one can measure *g* well and reduce, or perhaps eliminate, mean group differences.
- One might also infer that mean old psychometricians and the major scholars in intelligence research have been using the wrong item types for over 100 years.

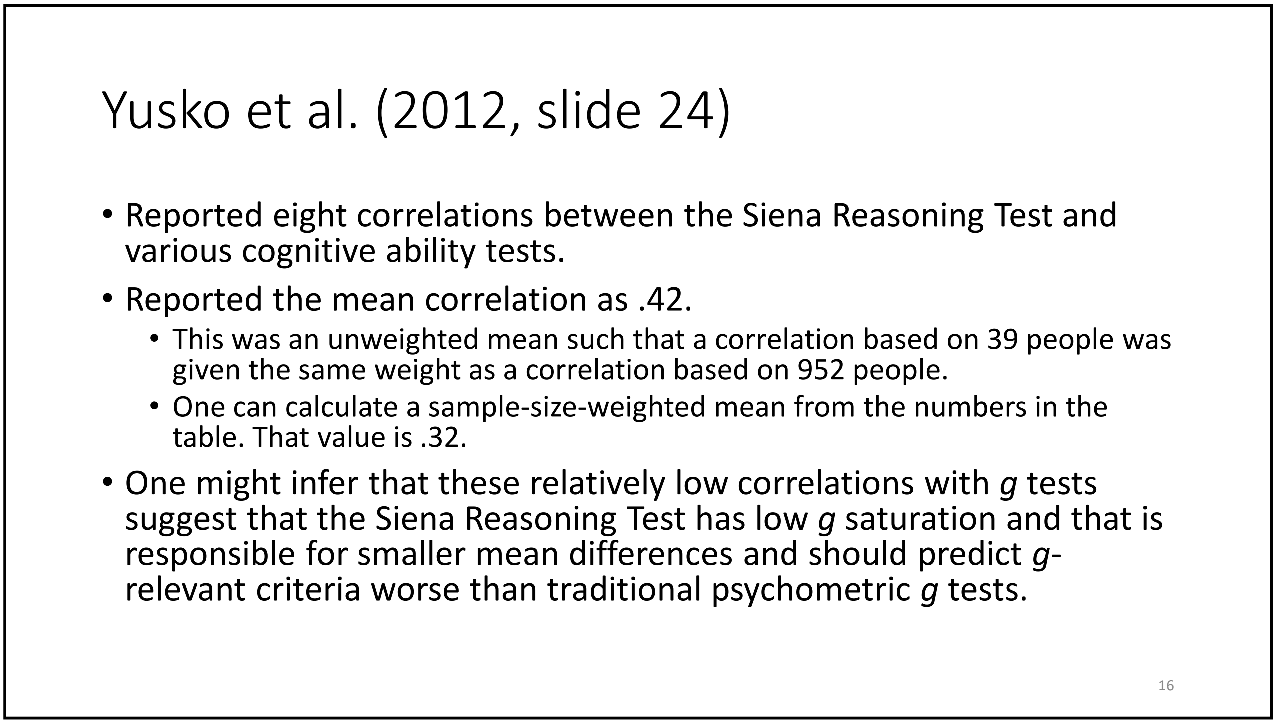
# Does research support the assertion of item types that can minimize mean group differences?

- No.
- Past research does not support the assertions about the item types being responsible for reduced mean group differences:
  - The Davis-Eells tests sought to limit verbal content and used graphical items but did not substantially reduce White-Black mean racial differences (Jenson, 1980)
  - The Raven's Progressive Matrices and the Advanced Raven's Progressive Matrices (Raven, Court & Raven, 1994; Raven, Raven and Court, 1998) do not rely on prior knowledge or language and their items are entirely graphical. The Raven's tests typically show large White-Black mean differences.



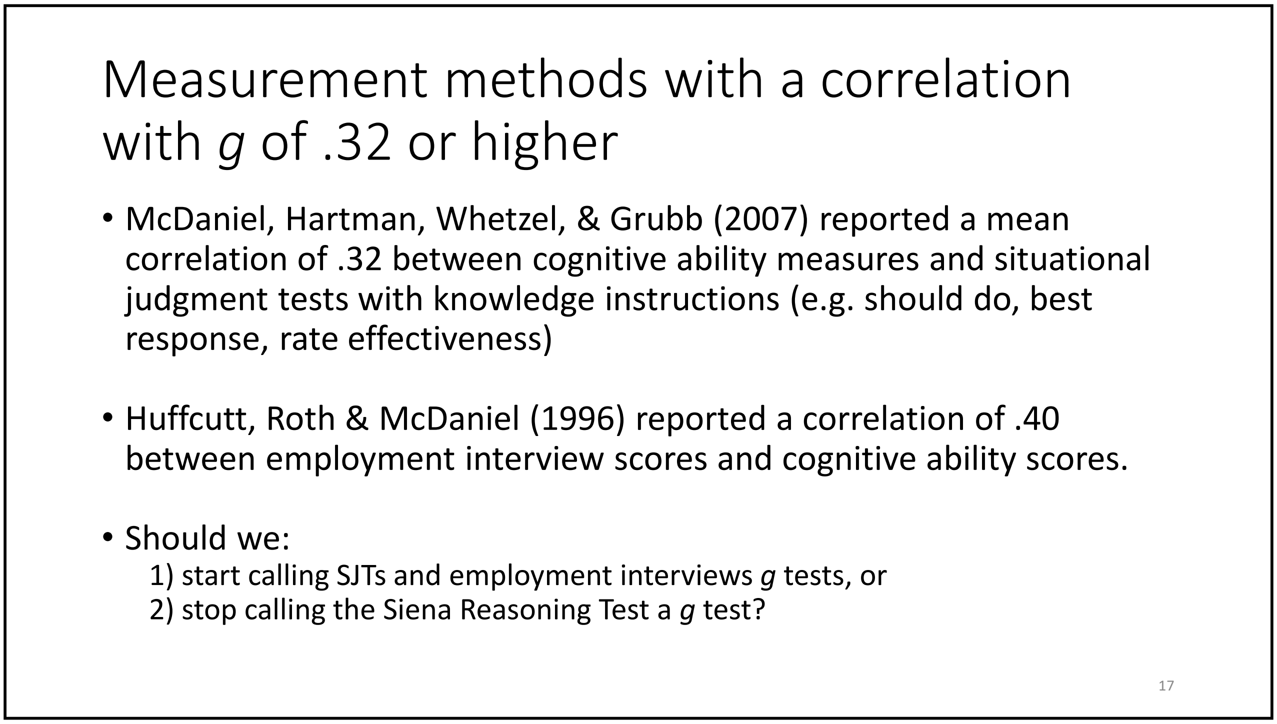
# Perhaps a better explanation

- In contrast to assertions about smaller mean group differences being due to special types of items, there are explanations which I suggest are more firmly established in research:
  - 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.
  - The reliability of the test influences mean group differences.
    - As noted earlier, 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.
  - *g* tests with lower mean group differences should have lower predictive value for criteria that *g* predicts, such as job performance and educational attainment.



# Yusko et al. (2012, slide 24)

- Reported eight correlations between the Siena Reasoning Test and various cognitive ability tests.
- Reported the mean correlation as .42.
  - This was an unweighted mean such that a correlation based on 39 people was given the same weight as a correlation based on 952 people.
  - One can calculate a sample-size-weighted mean from the numbers in the table. That value is .32.
- One might infer that these relatively low correlations with *g* tests suggest that the Siena Reasoning Test has low *g* saturation and that is responsible for smaller mean differences and should predict *g*-relevant criteria worse than traditional psychometric *g* tests.



# Measurement methods with a correlation with *g* of .32 or higher

- McDaniel, Hartman, Whetzel, & Grubb (2007) reported a mean correlation of .32 between cognitive ability measures and situational judgment tests with knowledge instructions (e.g. should do, best response, rate effectiveness)

- Huffcutt, Roth & McDaniel (1996) reported a correlation of .40 between employment interview scores and cognitive ability scores.

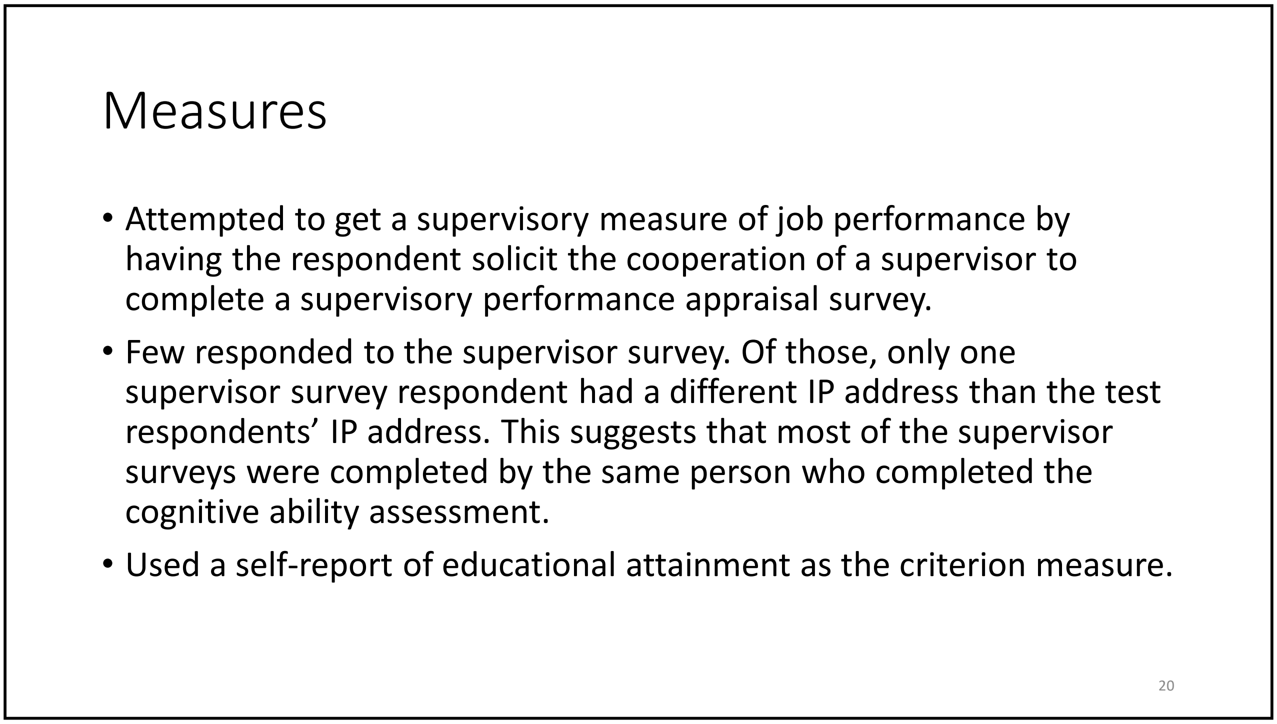
- Should we:
  1) start calling SJTs and employment interviews *g* tests, or
  2) stop calling the Siena Reasoning Test a *g* test?

# Brief summary of grant research

18

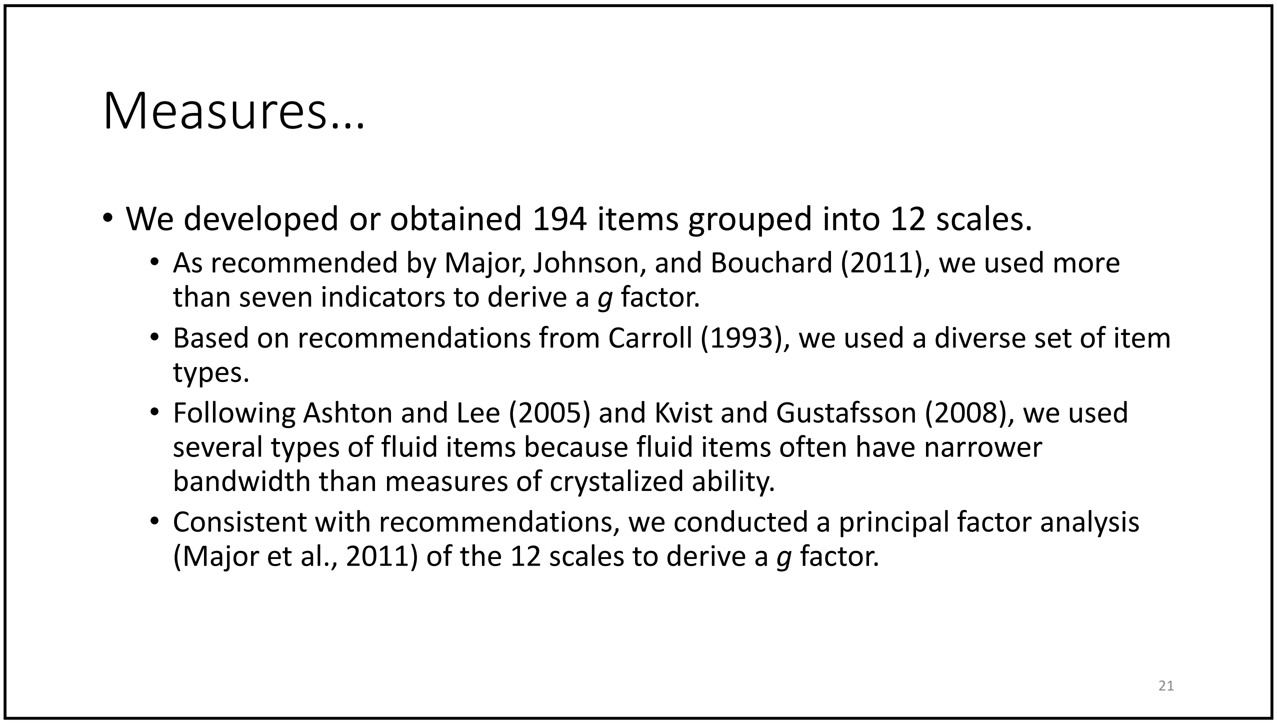
# Army research sample

- Data were collected using Amazon Mechanical Turk (AMT). 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.



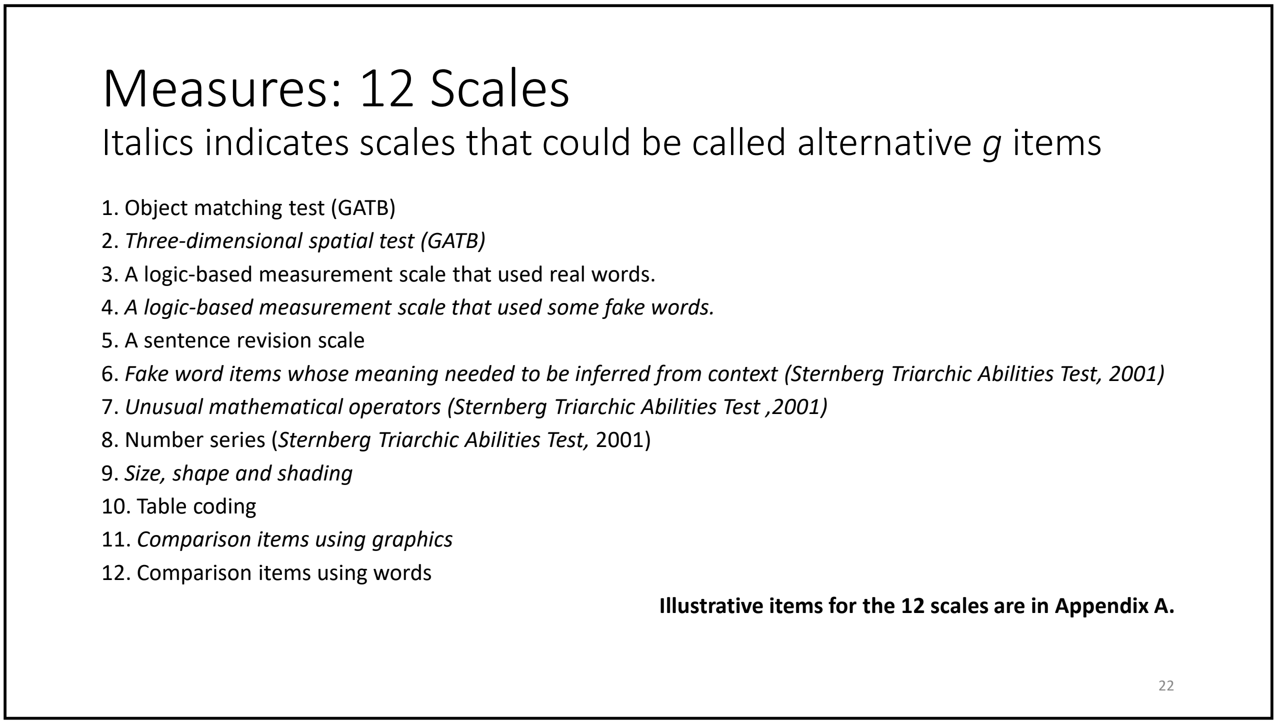
## Measures

- Attempted to get a supervisory measure of job performance by having the respondent solicit the cooperation of a supervisor to complete a supervisory performance appraisal survey.
- Few responded to the supervisor survey. Of those, only one supervisor survey respondent had a different IP address than the test respondents’ IP address. This suggests that most of the supervisor surveys were completed by the same person who completed the cognitive ability assessment.
- Used a self-report of educational attainment as the criterion measure.



## 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.
  - 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.



# Measures: 12 Scales

## Italics indicates scales that could be called alternative *g* items

1. Object matching test (GATB)
2. *Three-dimensional spatial test (GATB)*
3. A logic-based measurement scale that used real words.
4. *A logic-based measurement scale that used some fake words.*
5. A sentence revision scale
6. *Fake word items whose meaning needed to be inferred from context (Sternberg Triarchic Abilities Test, 2001)*
7. *Unusual mathematical operators (Sternberg Triarchic Abilities Test ,2001)*
8. Number series (*Sternberg Triarchic Abilities Test,* 2001)
9. *Size, shape and shading*
10. Table coding
11. *Comparison items using graphics*
12. Comparison items using words

**Illustrative items for the 12 scales are in Appendix A.**

# Measures that might be classified as alternative *g* items

2. Three-dimensional spatial test:
   - Graphical, non-verbal and do not rely on prior knowledge

4. A logic-based measurement scale that used some fake words:
   - The meaning of fake words need to be inferred and thus have little reliance on prior knowledge.

6. Fake word items whose meaning needed to be inferred from context (Sternberg Triarchic Abilities Test, 2001):
   - Little reliance on prior knowledge.

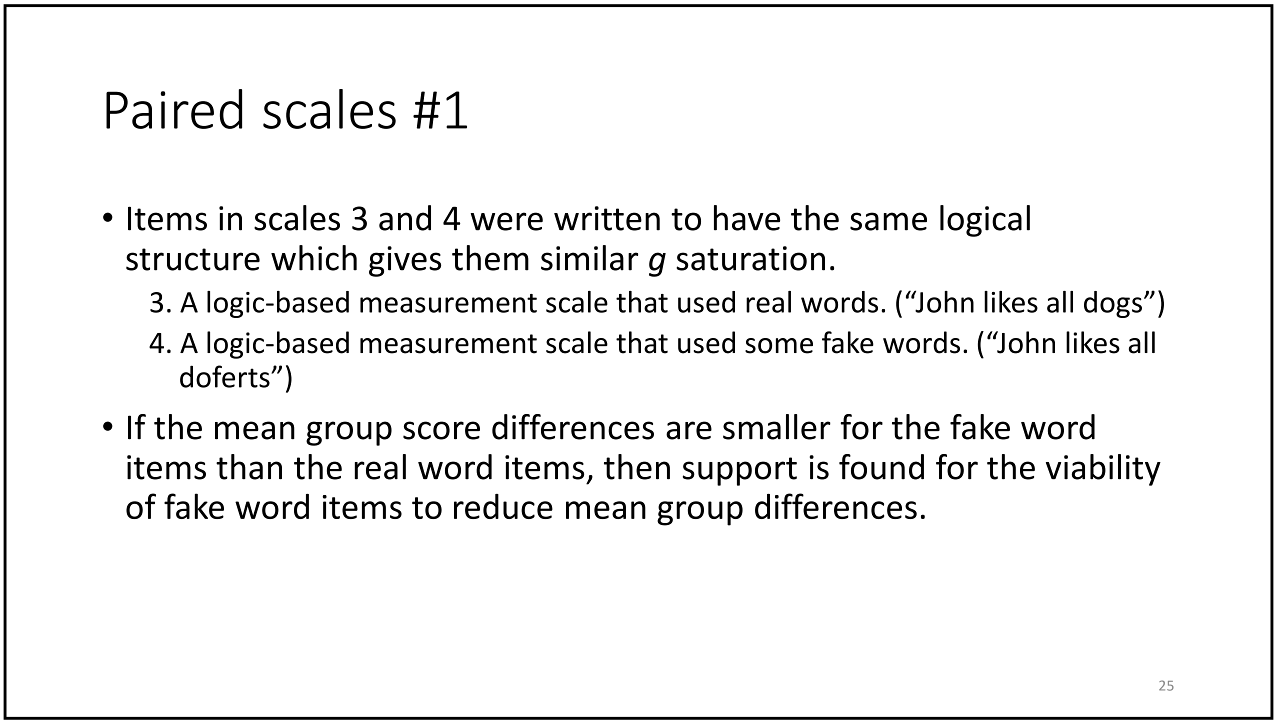
7. Unusual mathematical operators (Sternberg Triarchic Abilities Test, 2001):
   - Little reliance on prior knowledge.

9. Size, shape and shading:
   - Graphic items with little reliance on prior knowledge

11. Comparison items using graphics
    - Graphic items with little reliance on prior knowledge

# Comparing items types on mean group differences

- With respect to mean group differences, comparisons of item types can confound effects due to item types with effects due to *g* saturation.
- Vocabulary items tend to have large mean group differences because they have substantial *g* saturation.
- Single digit addition items (e.g., 2 + 3), tend to have small mean group differences because they have minimal *g* saturation.
- To meaningfully compare items types, one should control the *g* saturation by approaches such as items having the same logical structure.



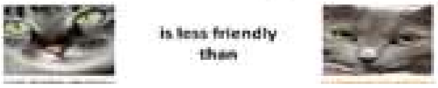
## Paired scales #1

- Items in scales 3 and 4 were written to have the same logical structure which gives them similar *g* saturation.
    3. A logic-based measurement scale that used real words. (“John likes all dogs”)
    4. A logic-based measurement scale that used some fake words. (“John likes all doferts”)
- If the mean group score differences are smaller for the fake word items than the real word items, then support is found for the viability of fake word items to reduce mean group differences.

# Paired scales #2

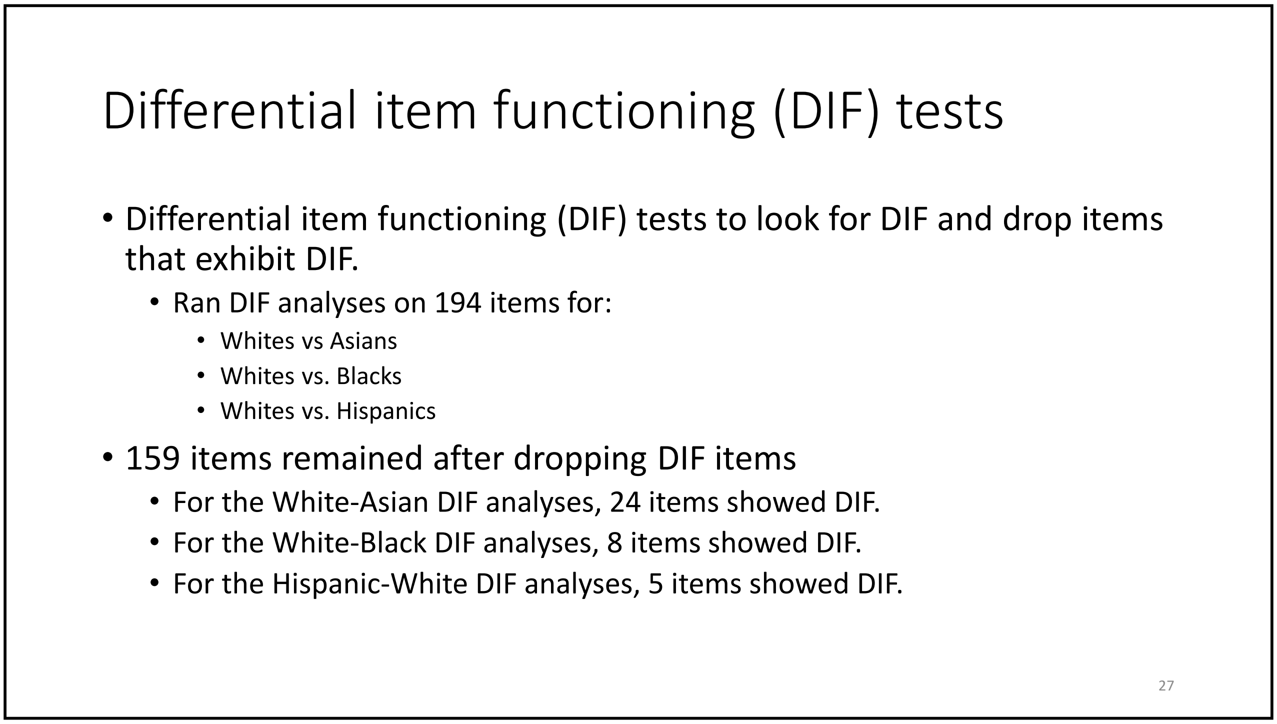
- Items in scales 11 and 12 were written to have same logical structure

  11. Comparison items using graphics:



  12. Comparison items using words: Muffin is a less friendly cat than Fuzzy.

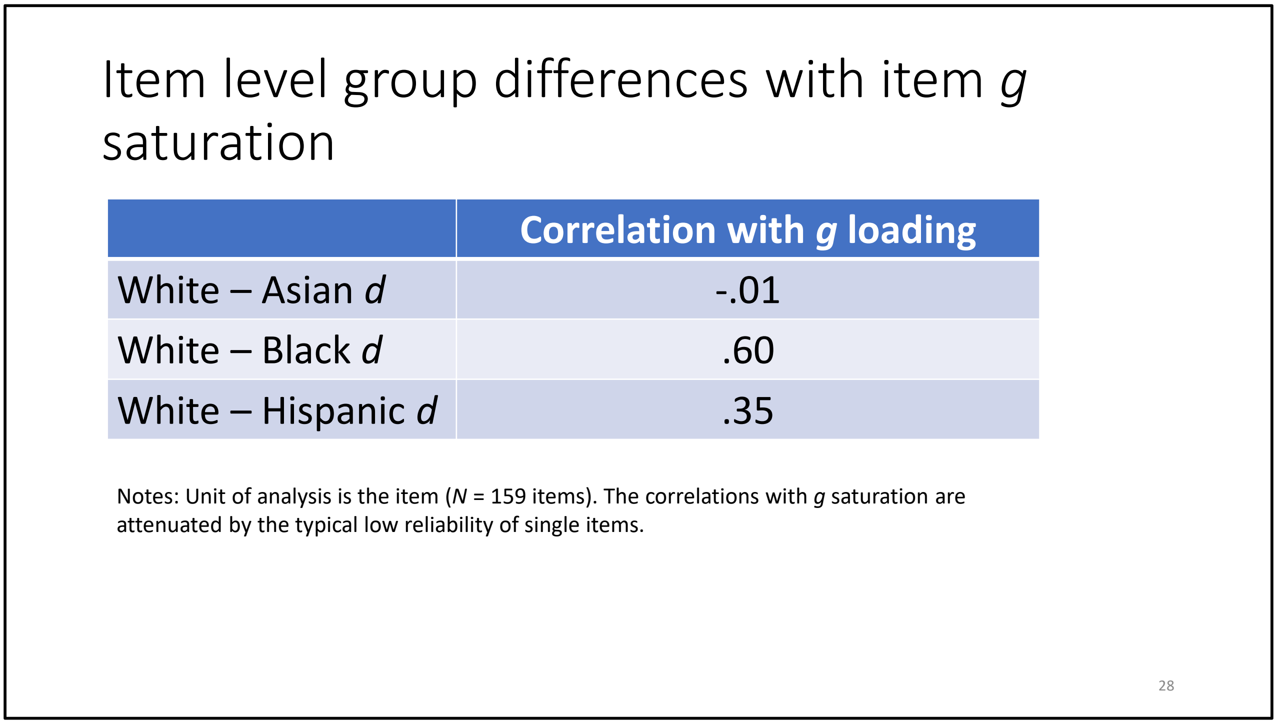
- If the mean group score differences are smaller for the graphic comparison items then the word comparison items, then support is found for the viability of graphic items to reduce mean group differences.



## Differential item functioning (DIF) tests

- Differential item functioning (DIF) tests to look for DIF and drop items that exhibit DIF.
  - Ran DIF analyses on 194 items for:
    - Whites vs Asians
    - Whites vs. Blacks
    - Whites vs. Hispanics
- 159 items remained after dropping DIF items
  - For the White-Asian DIF analyses, 24 items showed DIF.
  - For the White-Black DIF analyses, 8 items showed DIF.
  - For the Hispanic-White DIF analyses, 5 items showed DIF.

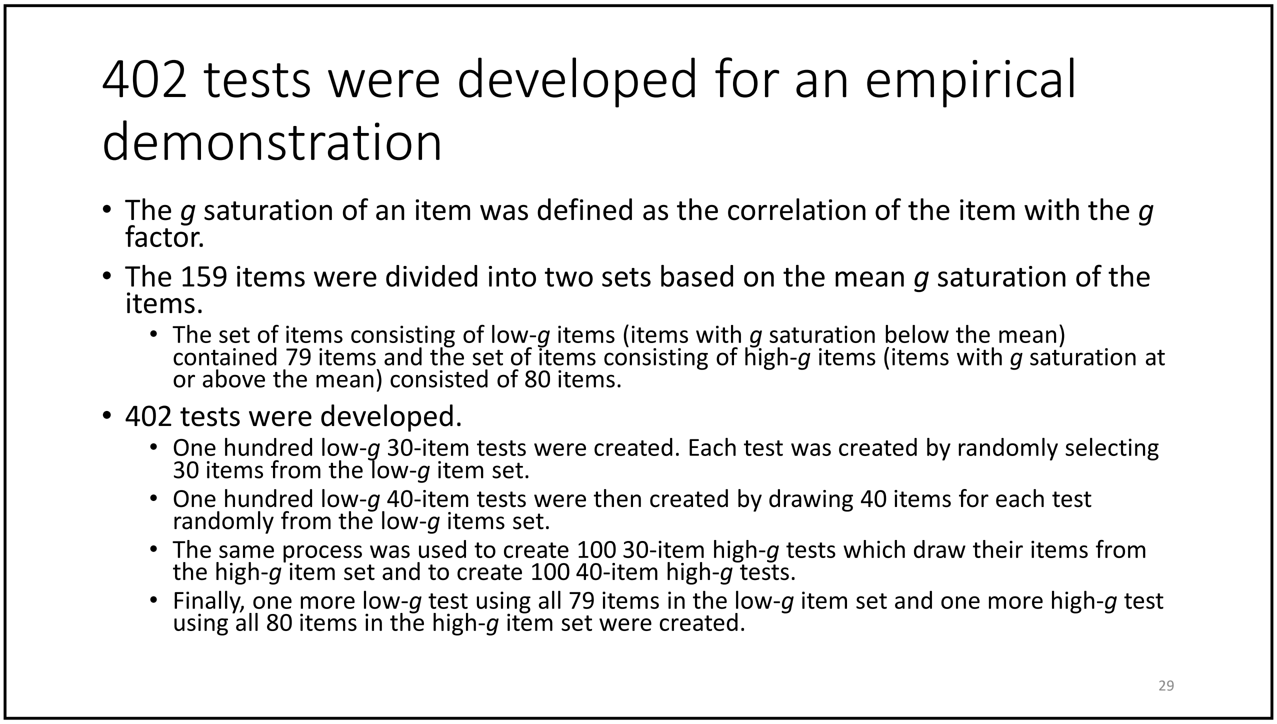
If someone notices the numbers sum to higher than 159: Of these 37 items with DIF, one item was indicating DIF for both White-Black and White-Asian (37 -1 = 36 DIF items) and one was indicating DIF for both White-Black and White-Hispanic analyses (36 -1 = 35). Thus, 35 of 194 items showed DIF and 159 of 194 items did not show DIF.



# Item level group differences with item *g* saturation

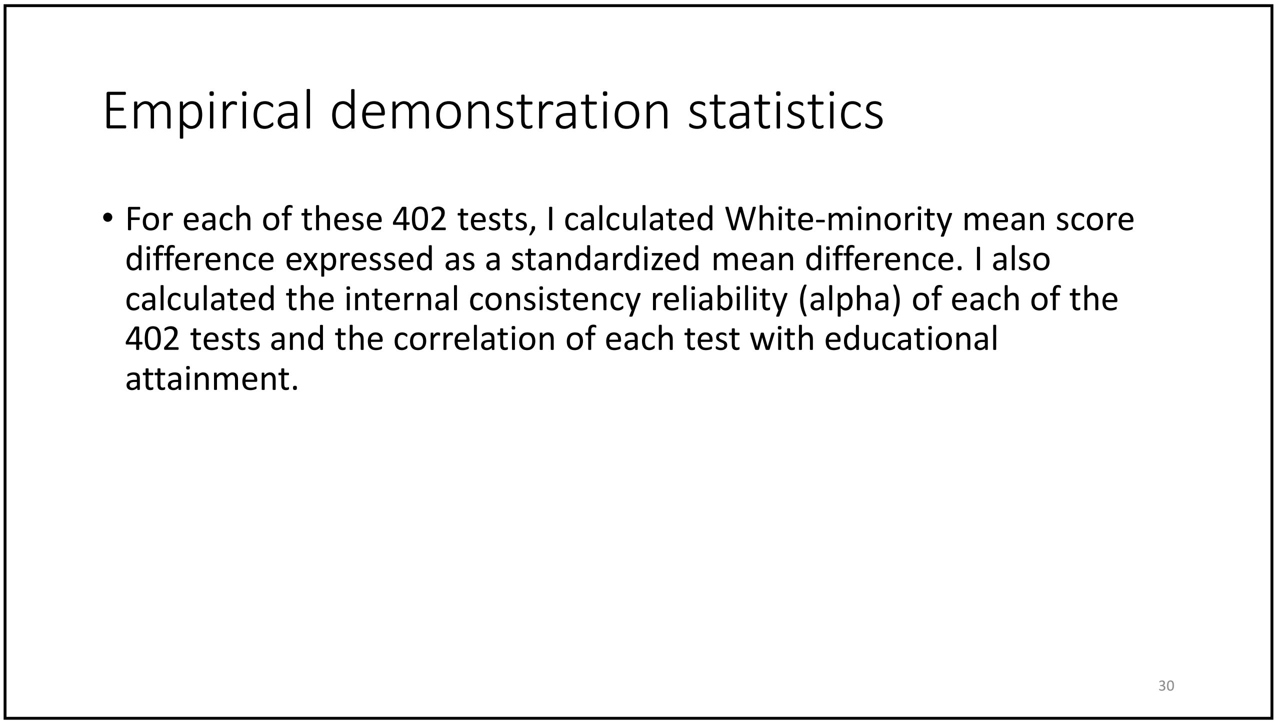
| | **Correlation with *g* loading** |
|---|---|
| White – Asian *d* | -.01 |
| White – Black *d* | .60 |
| White – Hispanic *d* | .35 |

Notes: Unit of analysis is the item (*N* = 159 items). The correlations with *g* saturation are attenuated by the typical low reliability of single items.



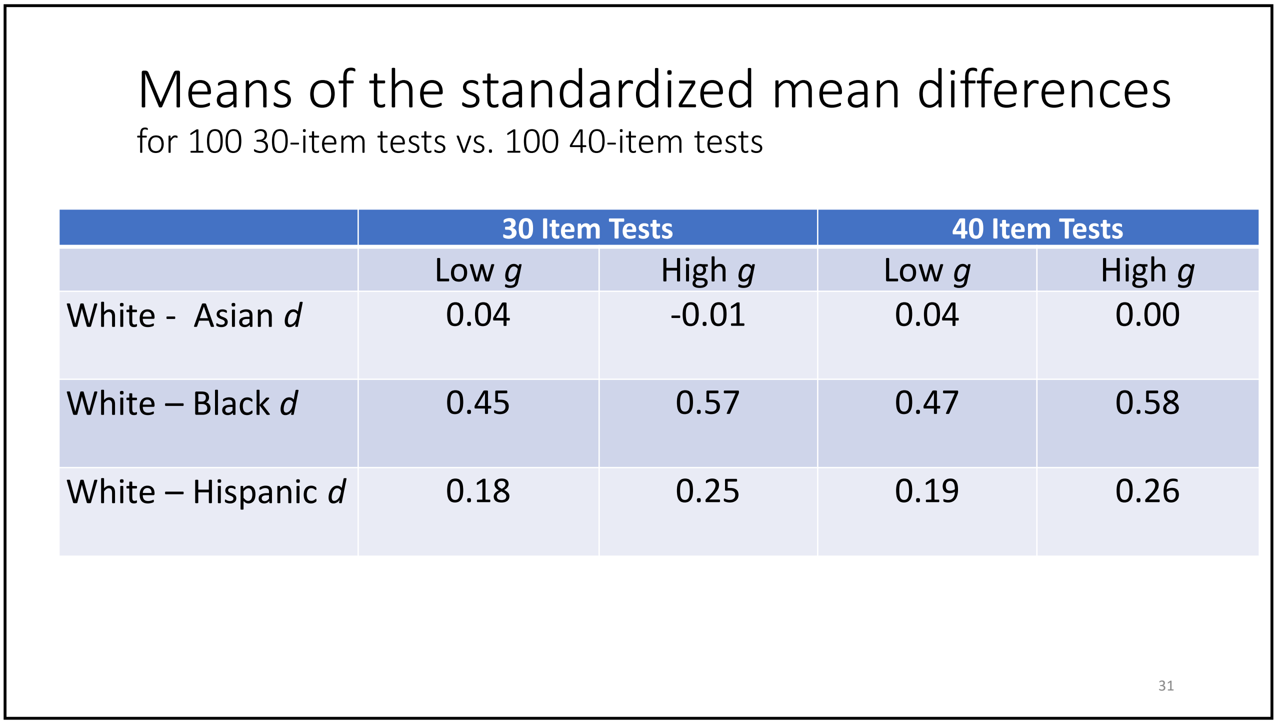
# 402 tests were developed for an empirical demonstration

- The *g* saturation of an item was defined as the correlation of the item with the *g* factor.
- The 159 items were divided into two sets based on the mean *g* saturation of the items.
    - The set of items consisting of low-*g* items (items with *g* saturation below the mean) contained 79 items and the set of items consisting of high-*g* items (items with *g* saturation at or above the mean) consisted of 80 items.
- 402 tests were developed.
    - 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 tests 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, 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 were created.



# Empirical demonstration statistics

- For each of these 402 tests, I calculated White-minority mean score difference expressed as a standardized mean difference. I also calculated the internal consistency reliability (alpha) of each of the 402 tests and the correlation of each test with educational attainment.



# Means of the standardized mean differences

for 100 30-item tests vs. 100 40-item tests

| | 30 Item Tests | | 40 Item Tests | |
|---|---|---|---|---|
| | Low *g* | High *g* | Low *g* | High *g* |
| White - Asian *d* | 0.04 | -0.01 | 0.04 | 0.00 |
| White – Black *d* | 0.45 | 0.57 | 0.47 | 0.58 |
| White – Hispanic *d* | 0.18 | 0.25 | 0.19 | 0.26 |

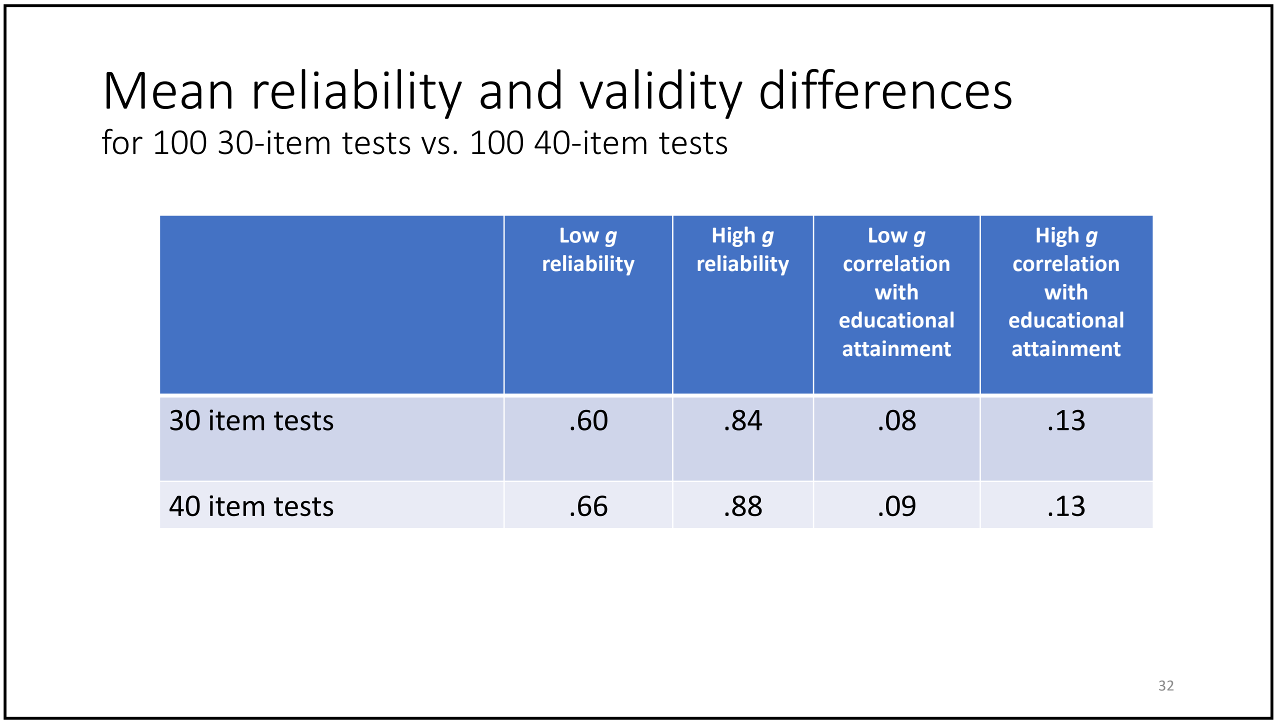
The numbers displayed are standardized mean differences (Cohen's *d*) which have a mean of zero and a standard deviation of 1. A *d* of zero indicates no mean group difference and a *d* of 1 indicates a one standard deviation difference.

A positive *d* indicates that the mean difference favors Whites and a negative *d* indicates that the mean difference favors the minority group.

The mean group differences for White vs Asians are trivial in magnitude. The statistics for Blacks and Hispanics are not trivial.

30 item tests have slightly lower mean group differences than the 40 item tests

The low *g* tests, the tests with lower *g* saturation, have lower mean group differences that the high *g* tests.



# Mean reliability and validity differences

for 100 30-item tests vs. 100 40-item tests

| | Low *g* reliability | High *g* reliability | Low *g* correlation with educational attainment | High *g* correlation with educational attainment |
|---|---|---|---|---|
| 30 item tests | .60 | .84 | .08 | .13 |
| 40 item tests | .66 | .88 | .09 | .13 |

32

Low *g* tests have lower reliability than high *g* scales of the same length.
Low *g* test predict educational attainment worse than high *g* tests

If anyone asks, the differences in reliability between the low-*g* tests and the high-*g* tests are due to the strong correlation between item *g* saturation and item variance (*r* = .56).

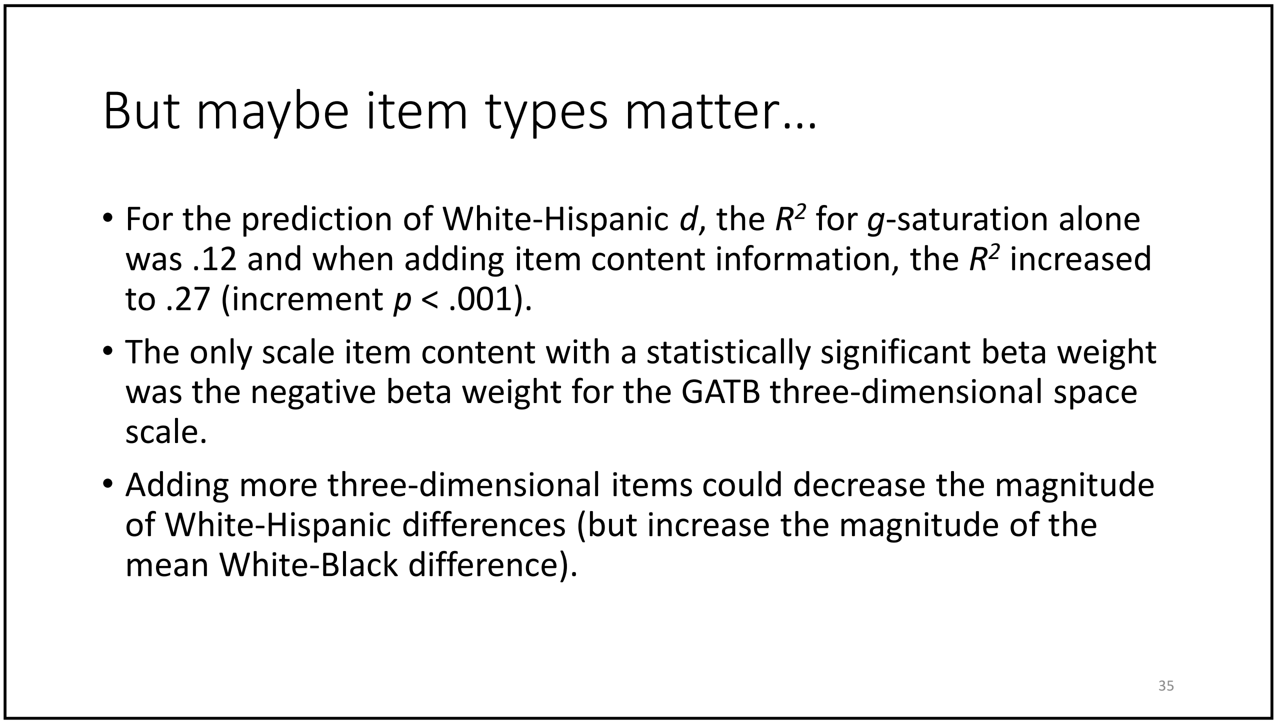
> Low *g* items have less variance (mean variance of 79 items = .14) than high-*g* items (mean variance of 80 items = .18).
>
> Internal consistency reliability is a function, in part, of the intercorrelation among the items.
>
> Because a correlation is an indicator of shared variance, items with smaller variance will have smaller correlations with each other than items with larger variance. This harms the reliability of low-*g* tests.

# But maybe item types might matter

- Created dummy variable for the 12 item types.
- Ran a two-step hierarchical regression
  - Item is the unit of analysis (*N* = 159)
  - Dependent variable is the White-Minority *d* (e.g., group mean difference)
  - First step: Independent variable was *g*-saturation
  - Second step: 11 item type dummy variables (The STAT Fake Words scale was excluded to prevent the matrix from being singular. )
- Given time constraints, I restrict results presented to White-Black and White-Hispanic.

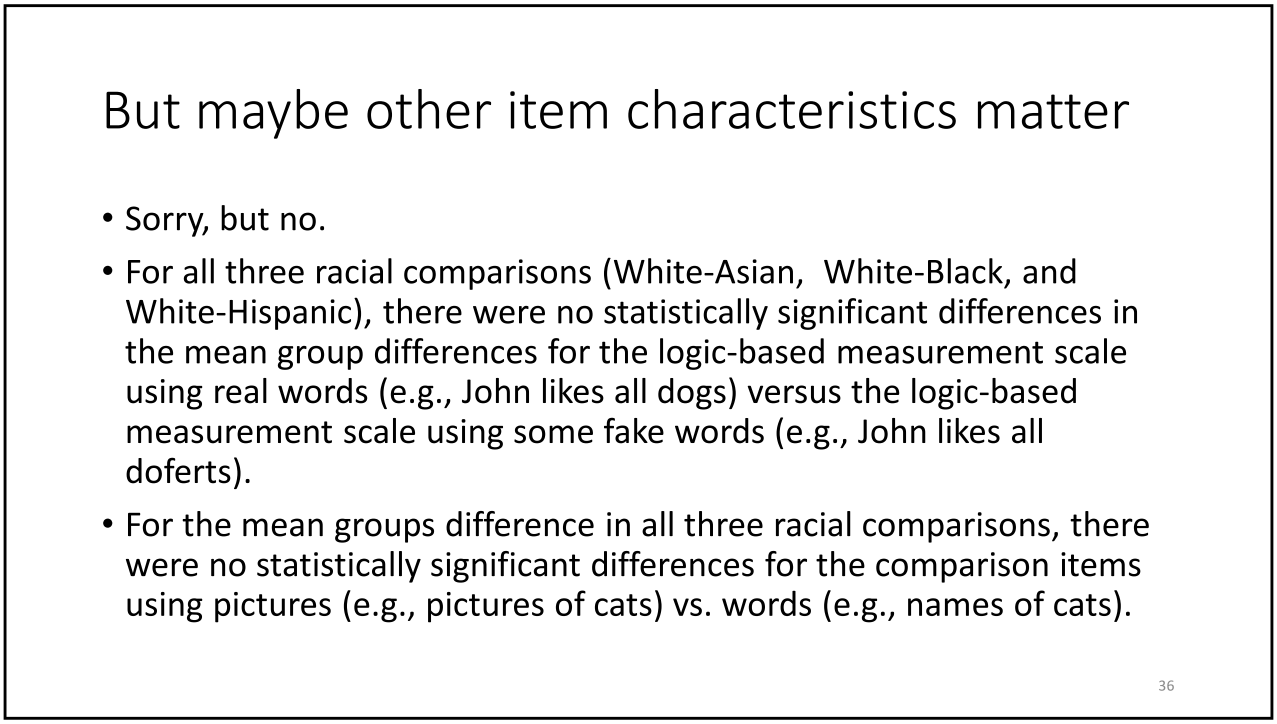
# But maybe item types matter…

- For the prediction of White-Black *d*, the $R^2$ was .36 for *g*-saturation alone and when adding in the item content information, the $R^2$ rose to .52 (increment $p < .001$).
- There were no statistically significant negative beta weights meaning that no content area could be added to reduce the White-Black *d*.
- However, there were two statistically significant positive beta-weights: GATB object matching and GATB three-dimensional space. One could decrease the White-Black mean *d* by reducing the number of these items in the test.



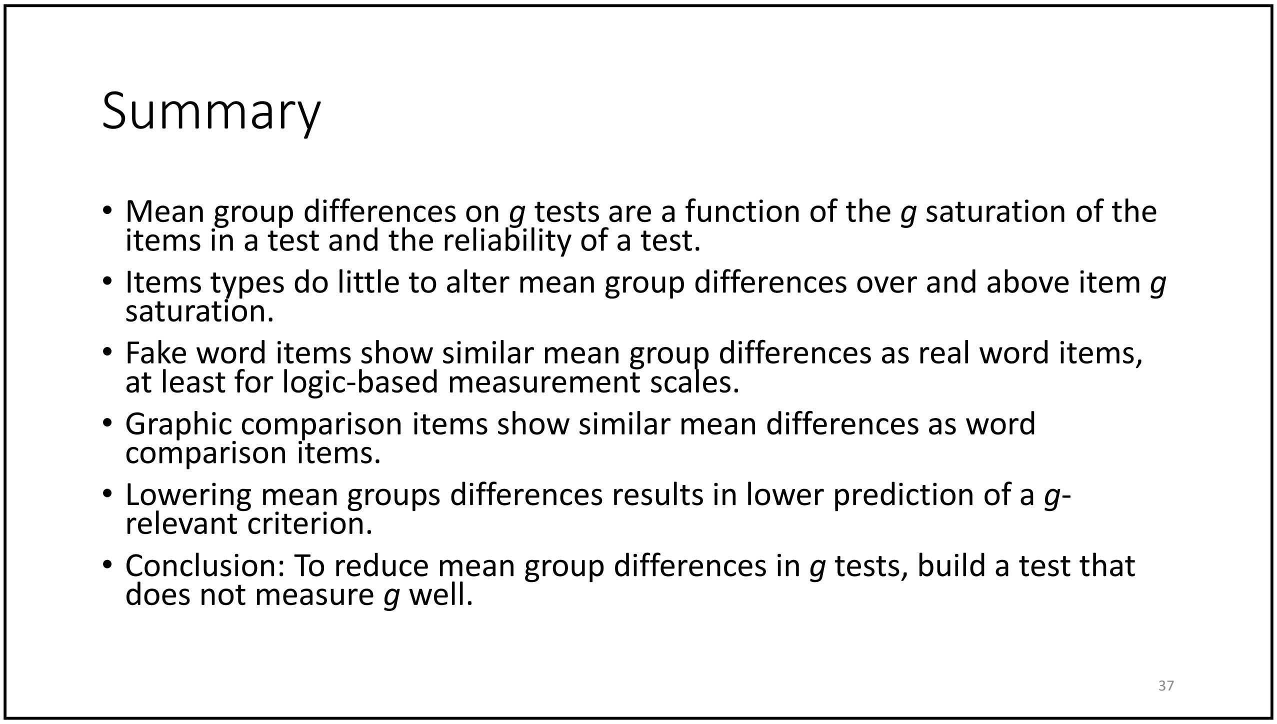
# But maybe item types matter…

- For the prediction of White-Hispanic *d*, the $R^2$ for *g*-saturation alone was .12 and when adding item content information, the $R^2$ increased to .27 (increment $p < .001$).
- The only scale item content with a statistically significant beta weight was the negative beta weight for the GATB three-dimensional space scale.
- Adding more three-dimensional items could decrease the magnitude of White-Hispanic differences (but increase the magnitude of the mean White-Black difference).



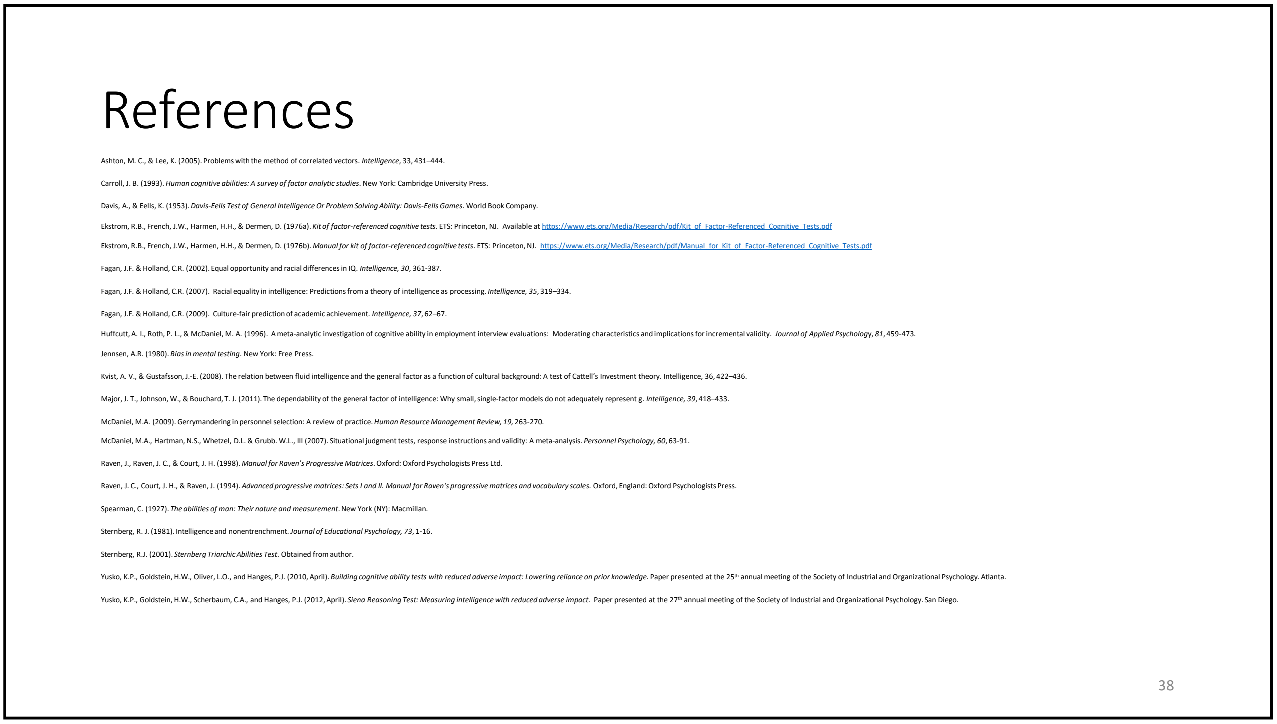
# But maybe other item characteristics matter

- Sorry, but no.
- For all three racial comparisons (White-Asian, White-Black, and White-Hispanic), there were no statistically significant differences in the mean group differences for the logic-based measurement scale using real words (e.g., John likes all dogs) versus the logic-based measurement scale using some fake words (e.g., John likes all doferts).
- For the mean groups difference in all three racial comparisons, there were no statistically significant differences for the comparison items using pictures (e.g., pictures of cats) vs. words (e.g., names of cats).



# Summary

- Mean group differences on *g* tests are a function of the *g* saturation of the items in a test and the reliability of a test.
- Items types do little to alter mean group differences over and above item *g* saturation.
- Fake word items show similar mean group differences as real word items, at least for logic-based measurement scales.
- Graphic comparison items show similar mean differences as word comparison items.
- Lowering mean groups differences results in lower prediction of a *g*-relevant criterion.
- Conclusion: To reduce mean group differences in *g* tests, build a test that does not measure *g* well.



# References

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, 33, 431–444.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Davis, A., & Eells, K. (1953). *Davis-Eells Test of General Intelligence Or Problem Solving Ability: Davis-Eells Games*. World Book Company.

Ekstrom, R.B., French, J.W., Harmen, H.H., & Dermen, D. (1976a). *Kit of factor-referenced cognitive tests*. ETS: Princeton, NJ. Available at https://www.ets.org/Media/Research/pdf/Kit_of_Factor-Referenced_Cognitive_Tests.pdf

Ekstrom, R.B., French, J.W., Harmen, H.H., & Dermen, D. (1976b). *Manual for kit of factor-referenced cognitive tests*. ETS: Princeton, NJ. https://www.ets.org/Media/Research/pdf/Manual_for_Kit_of_Factor-Referenced_Cognitive_Tests.pdf

Fagan, J.F. & Holland, C.R. (2002). Equal opportunity and racial differences in IQ. *Intelligence, 30*, 361-387.

Fagan, J.F. & Holland, C.R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence, 35*, 319–334.

Fagan, J.F. & Holland, C.R. (2009). Culture-fair prediction of academic achievement. *Intelligence, 37*, 62–67.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, *81*, 459-473.

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

Kvist, A. V., & Gustafsson, J.-E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell’s Investment theory. Intelligence, 36, 422–436.

Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence, 39*, 418–433.

McDaniel, M.A. (2009). Gerrymandering in personnel selection: A review of practice. *Human Resource Management Review, 19,* 263-270.

McDaniel, M.A., Hartman, N.S., Whetzel, D.L. & Grubb. W.L., III (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices*. Oxford: Oxford Psychologists Press Ltd.

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.

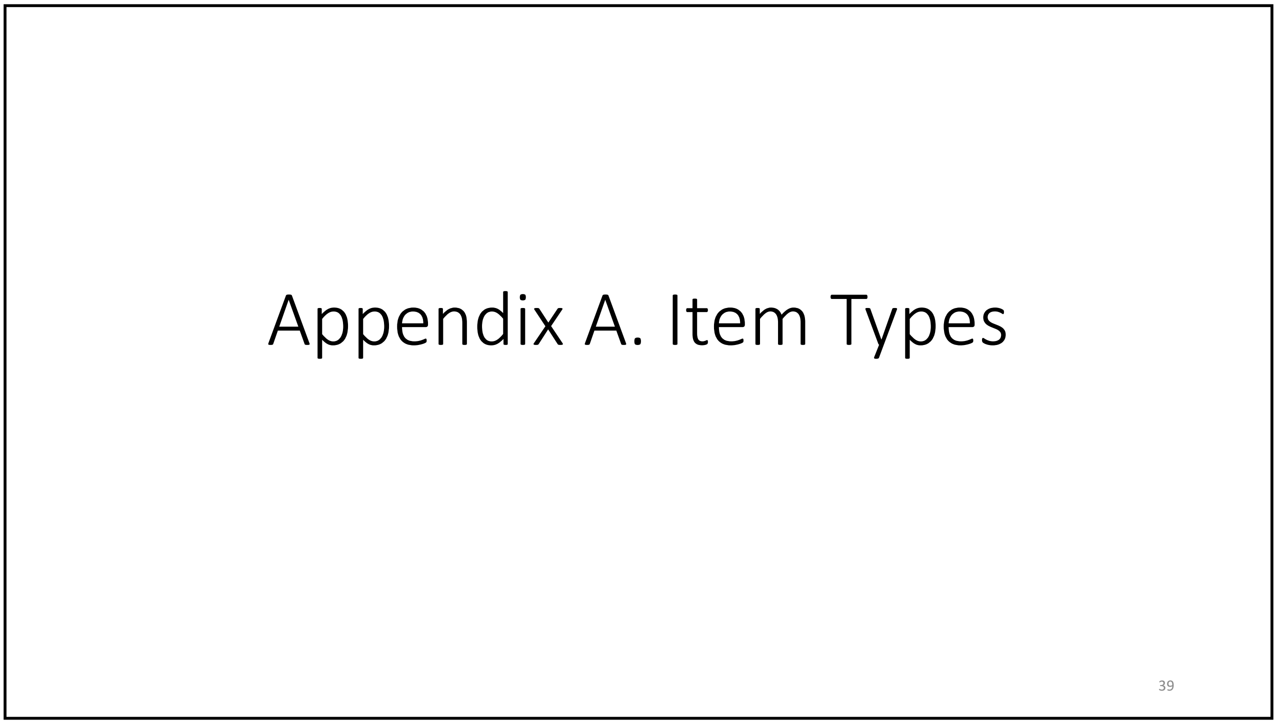
Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York (NY): Macmillan.

Sternberg, R. J. (1981). Intelligence and nonentrenchment. *Journal of Educational Psychology, 73*, 1-16.

Sternberg, R.J. (2001). *Sternberg Triarchic Abilities Test*. Obtained from author.

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

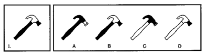
Yusko, K.P., Goldstein, H.W., Scherbaum, C.A., and Hanges, P.J. (2012, April). *Siena Reasoning Test: Measuring intelligence with reduced adverse impact*. Paper presented at the 27th annual meeting of the Society of Industrial and Organizational Psychology. San Diego.



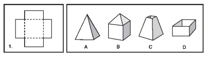
# Appendix A. Item Types

# 12 scales

1. The GATB object matching test, best classified as perceptual speed



2. The GATB three-dimensional spatial test, best classified as spatial ability



# 12 scales …

3. A logic-based measurement scale that used real words. Best classified as fluid intelligence.

Given a set of facts:

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

Respondents indicate whether a statement (e.g., Mary likes John’s dog) is true, false, or there is insufficient information to decide.

# 12 Scales …

4. A logic-based measurement scale that used some fake words. Best classified as fluid intelligence.

Given a set of facts:

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

Respondents indicate whether a statement (e.g., Mary likes John's *dofert*) is true, false, or there is insufficient information to decide.

Items with fake words could be considered non-entrenched (Sternberg, 1981) with few ties to culture and past knowledge.

# 12 Scales …

5. A sentence revision scale, best classified as crystalized intelligence.

Behind one of the two houses was a swimming pool, and the other has a garden behind it.

A) and the other has a garden behind it

B) and behind the other is a garden

C) and behind the other was a garden

D) and a garden was behind the other house

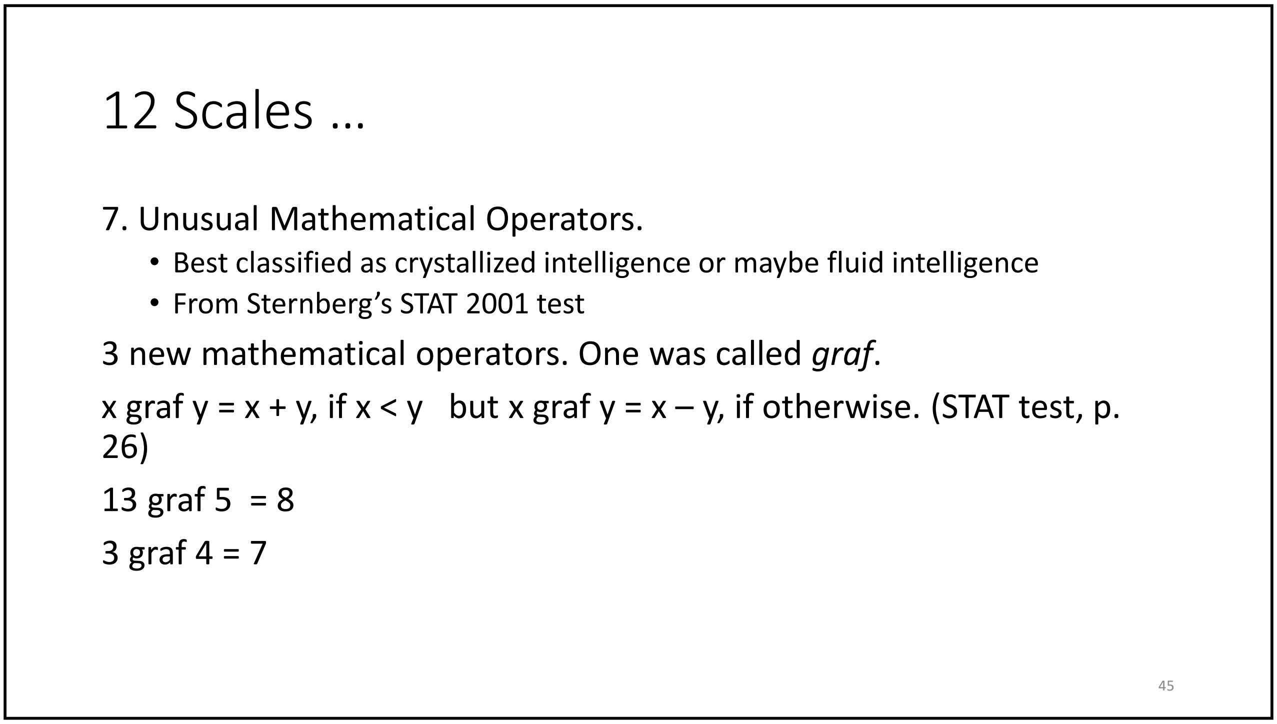
E) and the other house has a garden behind it

# 12 Scales …

6. Fake word items in which the meaning needed to be inferred from context.

- Possibly fluid intelligence
- From Sternbeg’s STAT 2001 test

The vip was green, so I started to cross the street. (STAT test, p. 1)



# 12 Scales …

7. Unusual Mathematical Operators.

- Best classified as crystallized intelligence or maybe fluid intelligence
- From Sternberg’s STAT 2001 test

3 new mathematical operators. One was called *graf*.

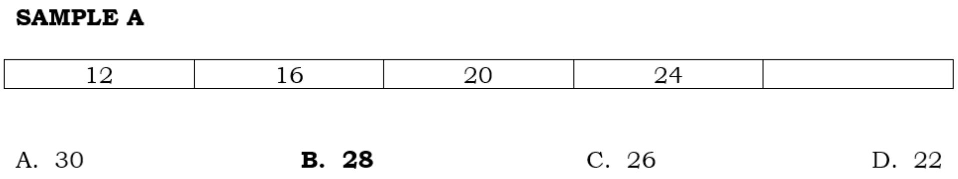
x graf y = x + y, if $x < y$ but x graf y = x – y, if otherwise. (STAT test, p. 26)

13 graf 5 = 8

3 graf 4 = 7

# 12 Scales …

8. Number series
    - Best classified as fluid intelligence
    - From Sternberg’s STAT 2001 test (page 4)
    - Traditional number series scale



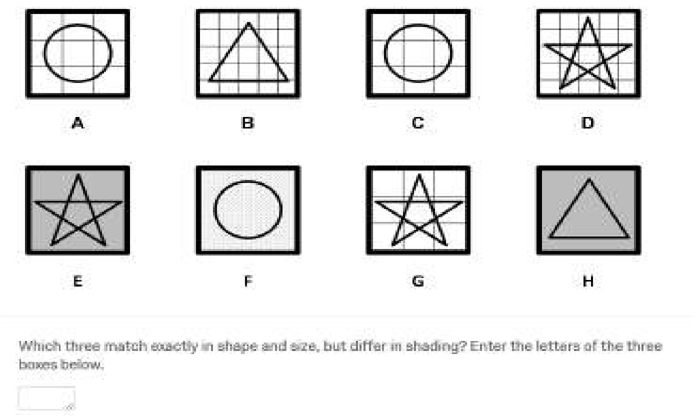
**SAMPLE A**

|  |  |  |  |  |
|---|---|---|---|---|
| 12 | 16 | 20 | 24 |  |

A. 30 **B. 28** C. 26 D. 22

# 12 Scales …

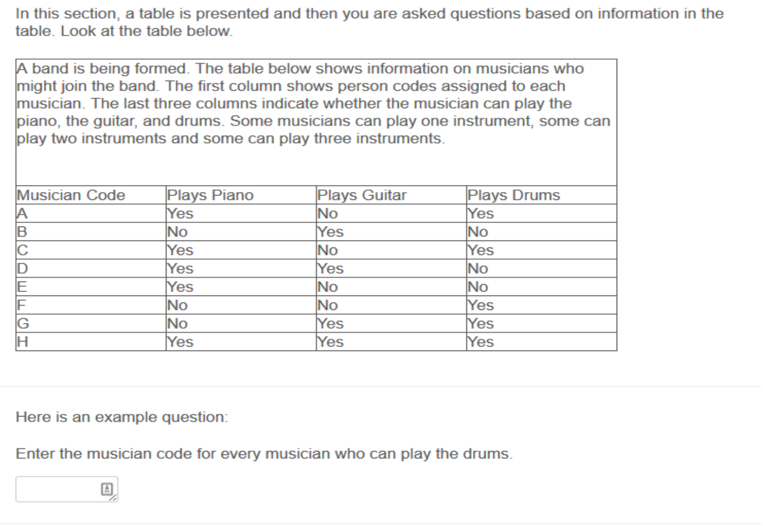
9. Size, shape and shading



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

# 12 Scales …

10.Table coding



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.

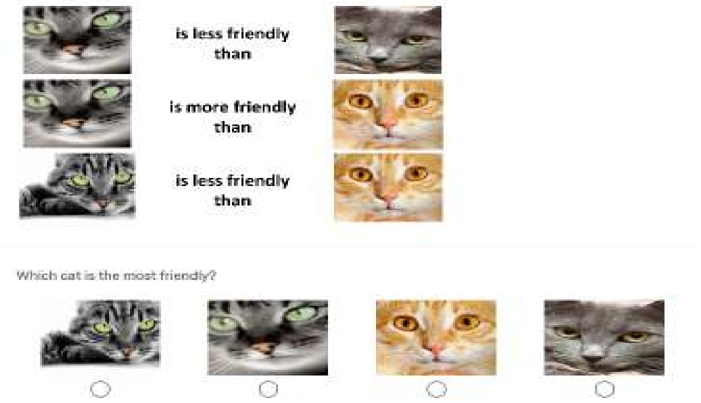
| Musician Code | Plays Piano | Plays Guitar | Plays Drums |
|---|---|---|---|
| A | Yes | No | Yes |
| B | No | Yes | No |
| C | Yes | No | Yes |
| D | Yes | Yes | No |
| E | Yes | No | No |
| F | No | No | Yes |
| G | No | Yes | Yes |
| H | Yes | Yes | Yes |

Here is an example question:

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

# 12 Scales …

## 11. Comparison items using graphics



is less friendly than

is more friendly than

is less friendly than

Which cat is the most friendly?

○ ○ ○ ○

# 12 Scales …

12. Comparison items using words

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

Which cat is the most friendly?