Why Cognitive Test Alternatives Do Not Exist and

Why *g* Is More Valid Than We Thought

Jeffrey M. Cucina
&
Philip T. Walmsley

U.S. Customs and Border Protection

The views expressed are those of the author and do not necessarily reflect those of U.S. Customs and Border Protection or the U.S. Federal Government. Portions of this work have previously been presented at the he 2014 and 2015 meetings of the Society for Industrial-Organizational Psychology (SIOP), the January 2016 meeting of the Personnel Testing Council of Metropolitan Washington (PTC/MW), and a 2013 meeting at the George Washington University

Presented at the 2018 meeting of the International Personnel Assessment Council, Washington, DC

Overview

- Brief review of cognitive ability testing
- Will demonstrate that general mental ability is more valid than previously thought
- Will show that high-validity low adverse impact cognitive ability tests cannot exist
  - Provide mathematical proof
  - Based on past research
- Will discuss implications for personnel selection
- Will include high-level overview in each section

Cognitive Ability Testing

- One of the best predictors of training, job, and academic performance
- Many civil service tests are cognitive ability tests
- A concern with cognitive ability tests is adverse impact
  - Group differences (i.e., mean score test differences for different groups) is contributing factor
    - As well as cutoff score, recruitment practices, etc.
  - Here we use the general term "adverse impact" to refer to average differences between majority and minority groups
  - This has led to searches for alternatives for cognitive ability tests
- Scientifically speaking, cognitive ability is a construct(s), rather than a testing method
  - Can appear or be assessed in a variety of test types
  - Much of the variance, and validity, is due to general mental ability (*g*)

g as a latent variable

- Meta-analyses show that cognitive ability tests predict performance:
  - $r_{cognitive\ ability,\ job\ performance}$ = .51 (Schmidt & Hunter, 1998)
  - $r_{cognitive\ ability,\ job\ performance}$ = .66 (Hunter, Schmidt, & Le, 2006)
  - $r_{cognitive\ ability,\ training\ performance}$ = .63 (Schmidt & Hunter, 2004)
  - $r_{SAT,\ FGPA}$ = .59 (Berry & Sackett, 2009)
  - $r_{GRE,\ LSAT,\ GMAT,\ MAT,\ MCAT,\ PCAT,1sy\ Year\ GPA}$ = .41 to .59 (Kuncel & Hezlett, 2007)
- Even when combined with other predictors there is still a good amount of unexplained variance
- Applied perspective: focus on operational validity – obtain correlation between test and performance, then correct for range restriction and criterion unreliability

g as a latent variable

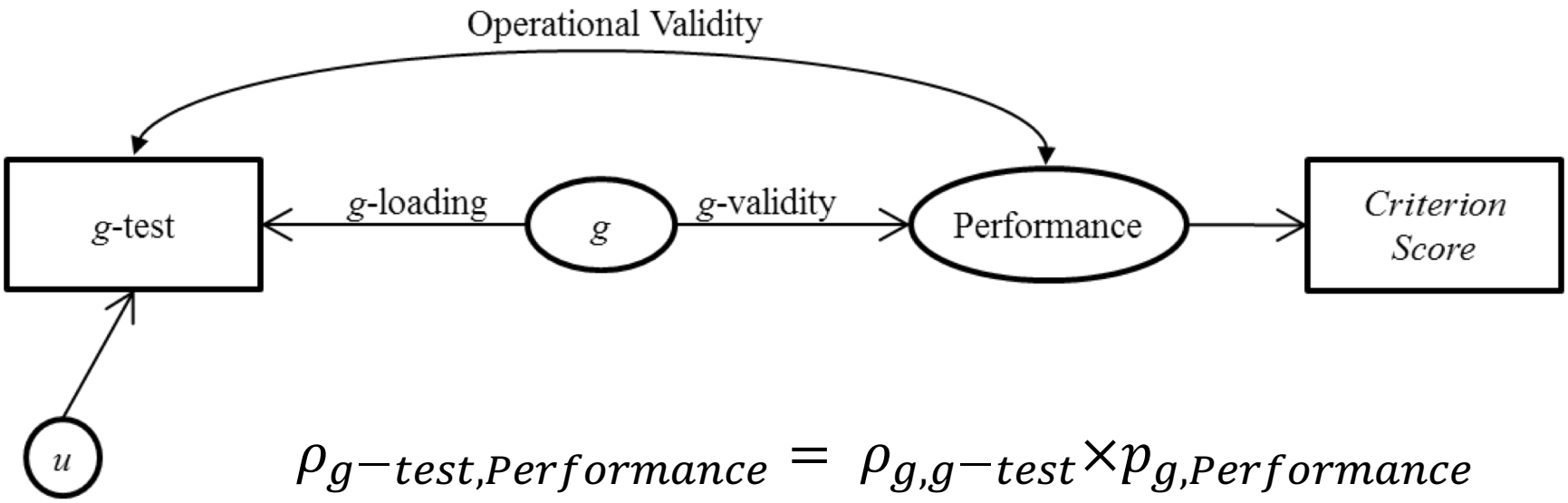
- Basic research perspective: make an additional correction for predictor unreliability → True score validity
  - Recall that cognitive ability test scores = *g* + broad factor(s) + narrow factor(s) + reliable specific variance + measurement error
  - Correcting for predictor reliability removes "measurement error"
  - True score = *g* + broad factor(s) + narrow factor(s) + reliable specific variance

- There is a problem here: "true scores" still include non-*g* variance (and we know that non-*g* variance rarely adds validity)
  - Jensen (1998): average *g*-loading for standardized tests is .87; *g*-loading of GATB G-score is .87 (note this served as foundation of Schmidt & Hunter's validity generalization work)
  - Reliabilities for standardized tests often in the .90s

g as a latent variable

- We used three approaches to estimate the validity of the latent variable *g*
  - Assumptions
    - *g*-loading is .87
    - Specific abilities have criterion-related validity of zero

1. Path Analysis Tracing Rule



$$\rho_{g-test,Performance} = \rho_{g,g-test} \times p_{g,Performance}$$

g as a latent variable

$$\rho_{g-test,Performance} = \rho_{g,g-test} \times p_{g,Performance}$$

$$p_{g,Performance} = \frac{\rho_{g-test,Performance}}{\rho_{g,g-test}}$$

$$g\text{−validity} = \frac{g-test\ validity}{g-loading}$$

$$g\text{−validity} = \frac{.51}{.87} = \mathbf{.59}$$

g as a latent variable

2. Correction for unreliability

- Reliability index: square root of reliability coefficient; gives the correlation between observed and true scores.
- If we define "reliability" of a g-loaded test as its "reliability" in measuring the latent true-score variable g, then:

$$g\text{-}validity = \frac{g\text{-}test\ validity}{\sqrt{reliability}} = \frac{g\text{-}test\ validity}{\sqrt{reliability\ index^2}}$$

$$= \frac{g\text{-}test\ validity}{reliability\ index}$$

$$= \frac{g\text{-}test\ validity}{g\text{-loading}} = \frac{.51}{.87} = \mathbf{.59}$$

g as a latent variable

3. Partial Correlation

- Compute partial correlation of *g*-test with performance after removing the effects of *u* (which we define as measurement error and reliable non-*g* specific variance).
- Nunnally & Bernstein (1994) formula for partial correlation:

$$r_{12\cdot3} = \frac{r_{12} - r_{23}r_{13}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

- Where: 1 = *g*-test scores
  2 = performance
  3 = *u* (=uniqueness)
  correlation between performance (2) and *u* (3) is zero

g as a latent variable

- Since 2 (performance) and *u* are uncorrelated:

$$r_{12\cdot3} = \frac{r_{12} - 0 \times r_{13}}{\sqrt{1-r_{13}^2}\sqrt{1-0}} = \frac{r_{12}}{\sqrt{1-r_{13}^2}}$$

- r_{13}^2 can be obtained using:
  - Path analysis with variance decomposition
  - Nunnally & Bernstein's (1994) formula for the multiple correlation for three uncorrelated predictors (5-21, p. 184)
  - Nunnally & Bernstein's formula (5-19, p. 183) for two predictors (assuming a correlation of zero between the two predictors)
  - All three methods give value of .49

$$r_{12\cdot3} = \frac{.51}{\sqrt{1-.49^2}} = \frac{.51}{\sqrt{1-.24}} = \frac{.51}{\sqrt{.76}} = \frac{.51}{.87} = \mathbf{.59}$$

1 = *g*-test scores
2 = performance
3 = *u* = meas. error + *s*

g as a latent variable: Results

Criterion	Source	*g*-test operational validity	Latent *g* validity	$r^2_{g\text{-}test}$	$r^2_{latent\ g}$
Training Level 3	SH04	.63	.72	40%	52%
Job Level 1	HSL06	.73	.84	53%	70%
Job Level 2	HSL06	.74	.85	55%	72%
Job Level 3	HSL06	.66	.76	44%	58%
Job Level 4	HSL06	.56	.64	31%	41%
Job Level 5	HSL06	.39	.45	15%	20%
Training Level 3	HH84	.57	.66	32%	43%
Job Level 1	HH84	.56	.64	31%	41%
Job Level 2	HH84	.58	.67	34%	44%
Job Level 3	HH84	.51	.59	26%	34%
Job Level 4	HH84	.40	.46	16%	21%
Job Level 5	HH84	.23	.26	5%	7%

Notes: MP13 = HH84=Hunter and Hunter (1984); = values used in Schmidt & Hunter (1998); BS09 = Berry and Sackett (2009); HSL06=Hunter, Schmidt, and Le (2006)

g as a latent variable: Results

Criterion	Source	*g*-test operational validity	Latent *g* validity	$r^2_{g\text{-}test}$	$r^2_{latent\ g}$
FGPA	MP13	.51	.59	26%	34%
FGPA	BS09 National	.59	.67	34%	45%
FGPA-ICGs	BS09 National	.71	.82	51%	67%
Cum. GPA	BS09 National	.52	.60	27%	36%
Cum. GPA ICGs	BS09 National	.69	.79	48%	63%
FGPA	BS09 College	.49	.57	24%	32%
FGPA-ICGs	BS09 College	.43	.50	19%	25%
Cum. GPA	BS09 College	.43	.50	19%	25%
Cum. GPA ICGs	BS09 College	.56	.64	31%	41%

Notes: MP13 = Mattern & Patterson (2013); BS09 = Berry & Sackett (2009); ICG = GPA estimated using Aggregate of Individual Course Grades; FGPA = Freshman/First-Year GPA; Cum. GPA = Cumulative GPA; National = corrections made using national applicant SDs, etc.; College = corrections made using individual school's SDs, etc.

g as a latent variable: Results

- High-level explanation
  - As mentioned, standardized ability tests measure general intelligence, narrow aspects of intelligence, and random error. The latter two do not correlate with performance.
  - Standardized ability tests are not pure measures of general intelligence, yet general intelligence is what predicts performance.
  - It is possible to estimate the correlation between general intelligence and performance.
  - For most jobs (i.e., those that are moderately complex), general intelligence has a validity of .76 for job performance.

Alternatives to *g*

- *Uniform Guidelines* call for a search for alternatives with equal validity and less adverse impact
- Search for alternatives to traditional *g*-tests with equal validity and lower group differences
  - "Despite many attempts, no one has been able to devise a mental test which can both eliminate the gap between races and meet the basic criteria necessary to validate a test." (Flynn, 1980, p.42)
  - "the search for the Holy Grail in personnel psychology." (Verive & McDaniel, 1996, p. 27)
  - "The 'holy grail' of American selection psychologists is an assessment of GMA that does not cause adverse impact." (Cook, 2009, p. 130)
  - "Attempts to develop new general measures that reduce group differences associated with *g* have largely been unsuccessful." (Tenopyr, 2002, p118)

Alternatives to *g*

- Addressed three questions
  1. Can a test measure *g* as well as a traditional *g*-test but without group differences?
  2. Can you measure *g* without group differences but with same criterion-related validity?
  3. Is it possible for another non-*g*-test related variable to have equal criterion-related validity as *g*-test?
- Used a variety of formulaic approaches (details are in the paper)
  - Path analysis
  - Variance decompositions of *g*-tests
  - Incremental validity
- Assumptions: *g*-tests are fair (under Cleary model), specific variance lacks validity, *g*-loading = .87, *d* = 1.0, etc.

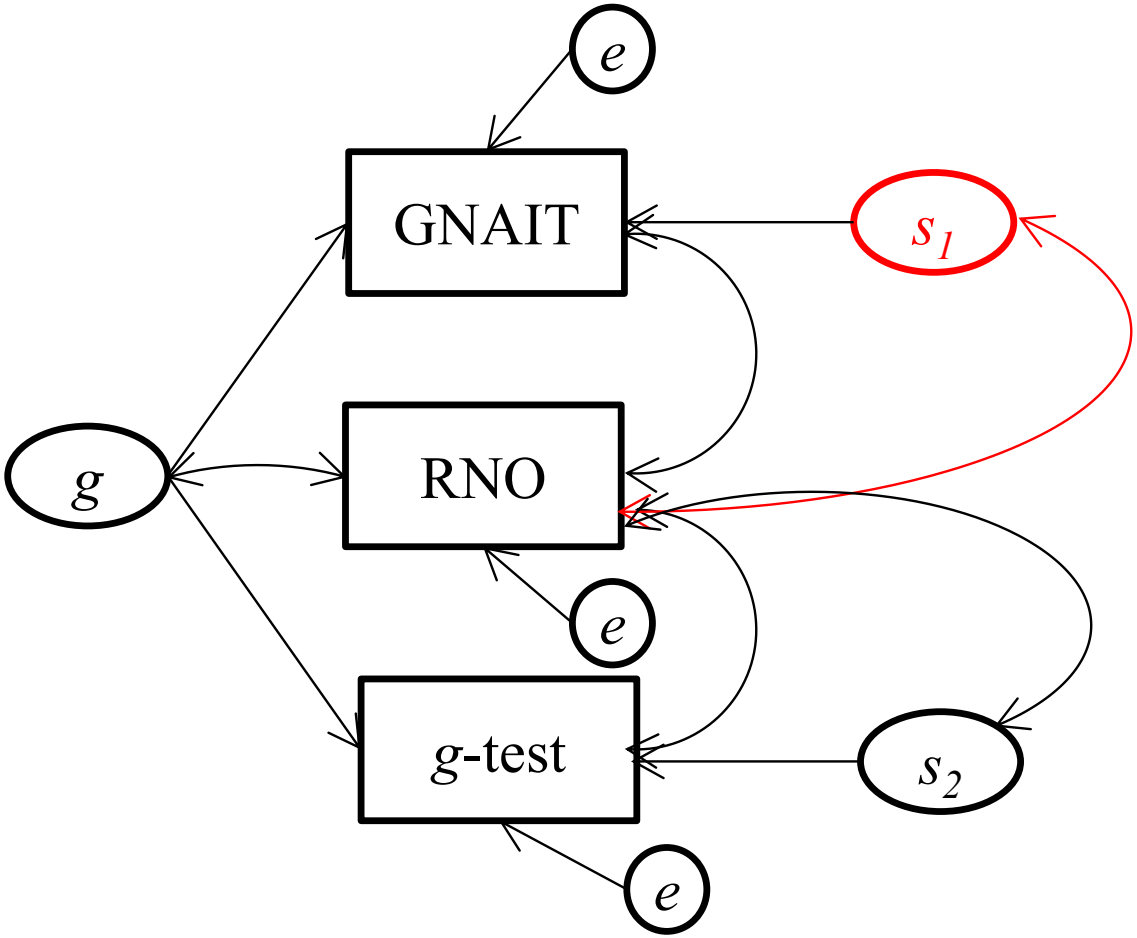
Alternatives to *g*

1. Can a test measure *g* as well as a traditional *g*-test but without group differences?

- No. If *g*-loading is .87, reliability is .90, etc. the specific variance must correlate -1.182 with RNO to give $d = 0$.
  - This is, of course, impossible
- We also tried partialling RNO from *g*-test; this requires specific variance to correlate -.950 with RNO
    - This is a very high correlation!
  - This is equivalent to within-group norming
  - We call this a GNAIT (*g* no adverse impact test)

Alternatives to *g*

- Explanation:
  - Consider the following model.



Alternatives to *g*

- We will assume the following values:
  - GNAIT and *g*-test have *g*-loadings of .87, reliabilities of .90.
    - With reliability of .90, path from *e* to test score is $\sqrt{1-.90} = .318$
  - $r_{g,s} = r_{s,e} = r_{g,e} = 0$
- We can decompose the variance as follows:

$$\sigma^2_{Test} = 1^2 = p^2_{g,Test} + p^2_{s,Test} + p^2_{e,Test}$$

$$1^2 = .87^2 + p^2_{s,Test} + .318^2$$

$$1 = .757 + p^2_{s,Test} + .10$$

$$1 - .757 - .10 = p^2_{s,Test} = .143$$

$$p_{s,Test} = \sqrt{p^2_{s,Test}} = \sqrt{.143} = .378$$

Alternatives to *g*

- Next, obtain observed correlation between RNO and *g*-test score.
  - Cohen (1998, formula 2.2.7)

$$r = \frac{d}{\sqrt{d^2 + \frac{1}{pq}}}$$

  - $d = 1.0, p = q = .50 \therefore r = .447$
- Now, obtain true score correlation between *g* and RNO

$$\rho_{RNO,g} = \frac{r_{RNO,g-test}}{\sqrt{reliability}} = \frac{.447}{\sqrt{.87^2}} = \frac{.447}{.87} = .514$$

Alternatives to *g*

- Finally, obtain correlation between GNAIT and RNO
  - Use path analysis tracing rule (Kenny, 1979/2004)

$$\rho_{RNO,GNAIT} = \rho_{RNO,g} \times p_{g,GNAIT} + \rho_{RNO,s1} \times p_{s1,GNAIT} + \rho_{RNO,e} \times p_{e,GNAIT}$$

$$0 = .514 \times .87 + \rho_{RNO,s1} \times .378 + 0 \times .316$$

$$0 = .447 + \rho_{RNO,s1} \times .378 + 0$$

$$-.447 = \rho_{RNO,s1} \times .378$$

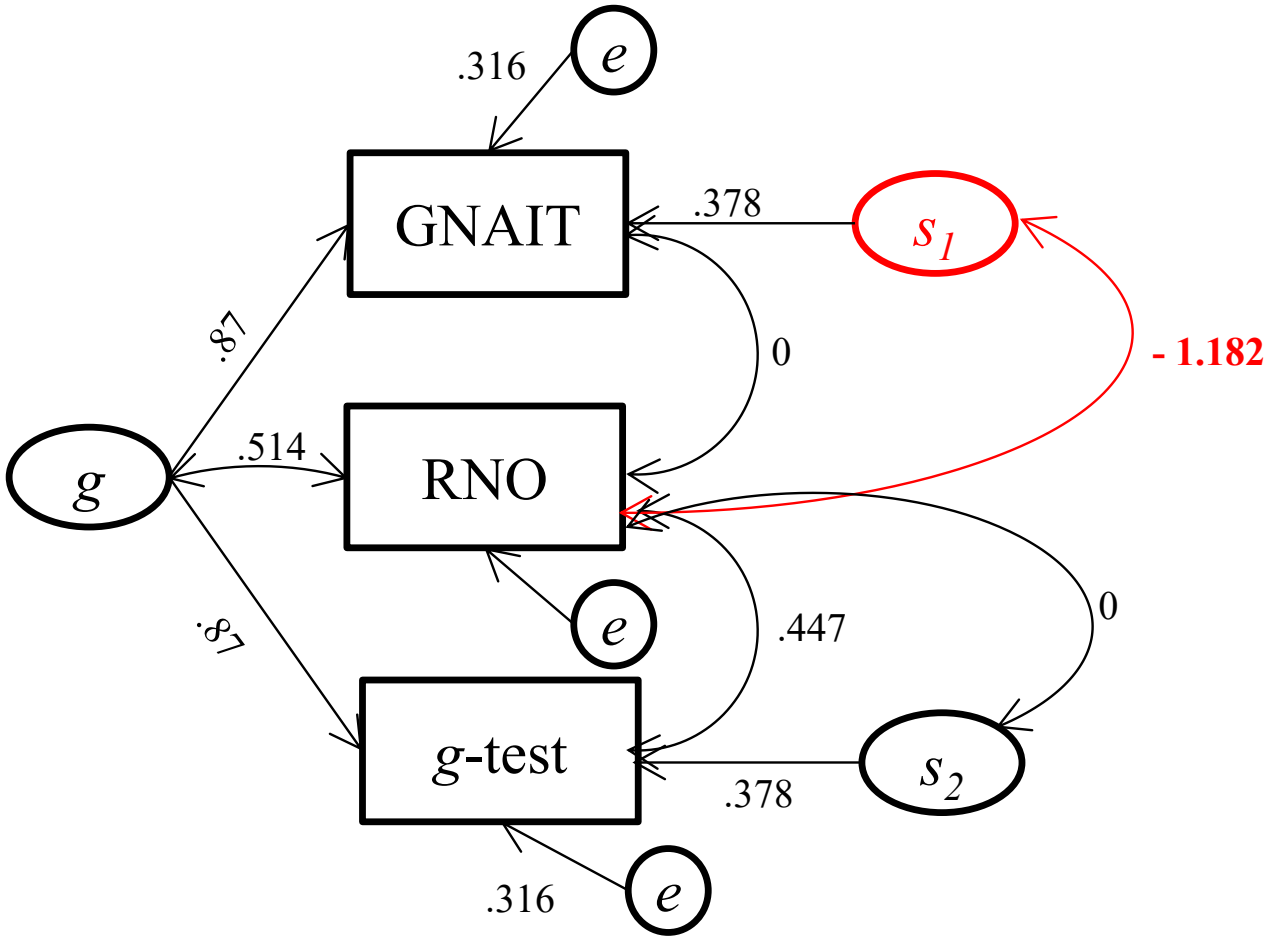
$$\frac{-.447}{.378} = \rho_{RNO,s1}$$

$$\rho_{RNO,s1} = -1.182$$

Note that *r* = -1.182 is an impossible value

Alternatives to *g*

- Finally, obtain correlation between GNAIT and RNO
  - Alternately, consider this model:



Alternatives to *g*

2. Can you measure *g* without group differences but with same criterion-related validity?
   - Not really.
   - Traditional *g*-test adds incremental validity over GNAIT.
   - Amount of incremental validity depends on percentage of minority and majority group members.
     - When *g*-test has validity of .51 and 50-50 split, GNAIT's validity is .47
     - When *g*-test has validity of .66 and 50-50 split, GNAIT's validity is .62
     - As you deviate from 50-50 split to 100-0 (or 0-100) split, GNAIT validity approaches *g*-test validity

Alternatives to *g*

- Explanation
  - Knowing that a GNAIT test with a *g*-loading of .87 and no group differences is impossible, what about the next best thing?
  - We can partial RNO from the *g*-test (equivalent to within-group norming)
  - Use formula for partial correlation (formula 5-14 from Nunnally & Bernstein, 1994, p. 176):

$$r_{12.3} = \frac{r_{12} - r_{23}r_{13}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Where: variable 1 is the test, variable 2 is latent variable *g*, variable 3 is RNO

$$r_{12.3} = \frac{.87 - .514 \times .447}{\sqrt{1 - .447^2}\sqrt{1 - .514^2}}$$

$$r_{12.3} = \frac{.87 - .230}{\sqrt{1 - .200}\sqrt{1 - .264}} = .834$$

- The *g*-loading has changed from .87 to .834

Alternatives to g

- Explanation
  - Using path analysis tracing rule, we can obtain a mathematically plausible value for correlation between s_1 and GNAIT.

$$\rho_{RNO,GNAIT} = \rho_{RNO,g} \times p_{g,GNAIT} + \rho_{RNO,s1} \times p_{s1,GNAIT} + \rho_{RNO,e} \times p_{e,GNAIT}$$

$$0 = .514 \times .834 + \rho_{RNO,s1} \times .452 + 0 \times .318$$

$$0 = .429 + \rho_{RNO,s1} \times .452 + 0$$

$$-.429 = \rho_{RNO,s1} \times .452$$

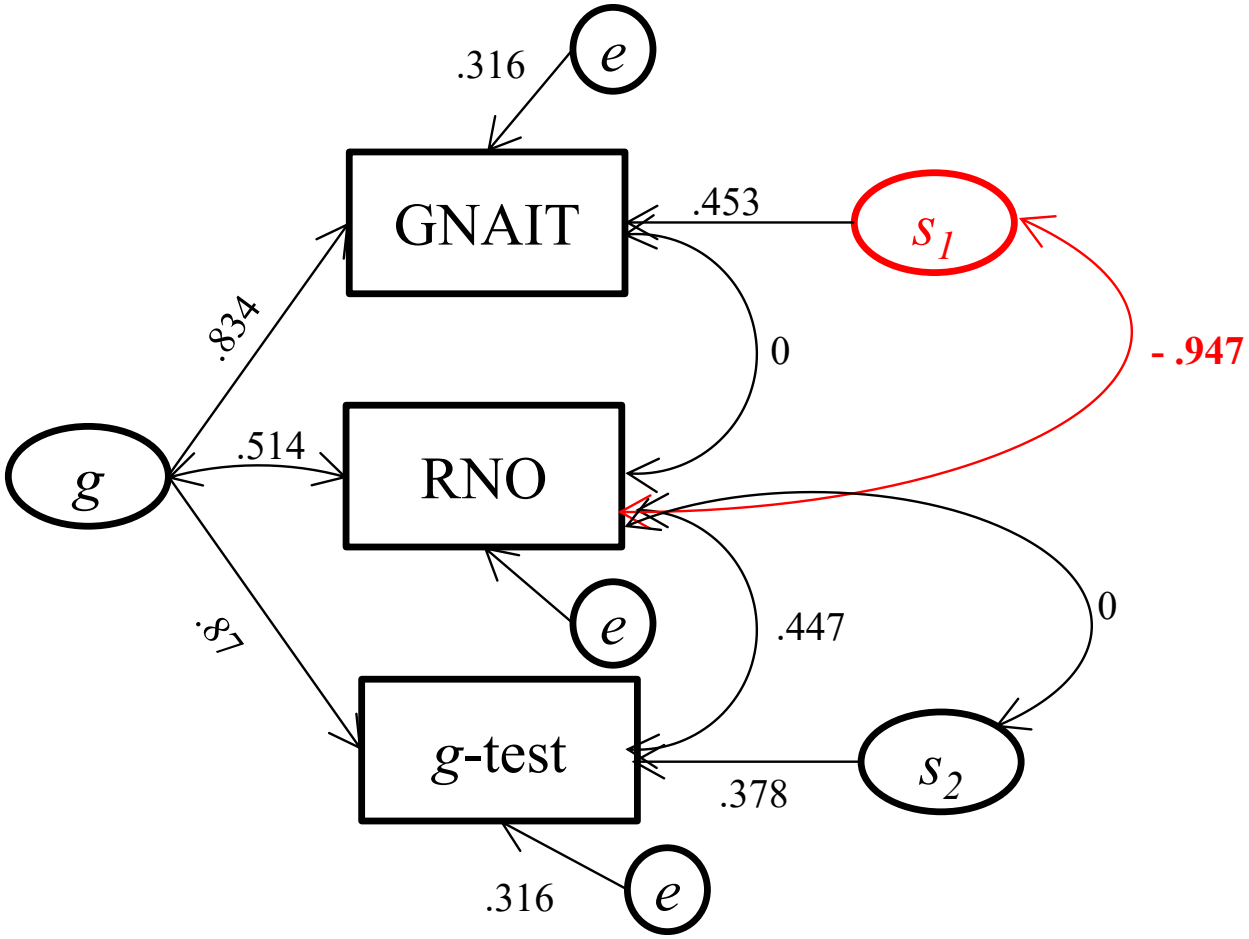
$$\frac{-.429}{.452} = \rho_{RNO,s1}$$

$$\rho_{RNO,s1} = -0.950$$

There is a practical issue of finding a latent variable that correlates -.950 with RNO

Alternatives to *g*

- Explanation
  - Here is the model:



Alternatives to *g*

- Explanation
- Partialling RNO out also impacts criterion-related validity.
  - The criterion-related validity of the GNAIT is lower than the *g*-test.
  - Validity drops from .510 to .469.
  - Using Schmidt's (2013) revised validity estimate of .660, the validity drops to .618.
  - The *g*-test always adds incremental validity over the GNAIT

- In these analyses we have assumed that $d = 1.0$, $p = q = .50$. The text of a SIOP poster provides the results for other values.
  - As the proportion of protected group members deviates from .50, the impact on validity is reduced.

Alternatives to *g*

- High-level explanation
  - Scores on standardized ability tests have three underlying sources: general intelligence, narrow aspects of intelligence, and some random error.
  - General intelligence is what predicts job and training performance (in most jobs). The narrow aspects of intelligence and random error do not predict performance.
  - Adverse impact is primarily due to general mental ability, not narrow aspects of intelligence, nor random error.
  - Standardized ability tests measure general intelligence much more so than the narrow aspects of intelligence or random error.

Alternatives to *g*

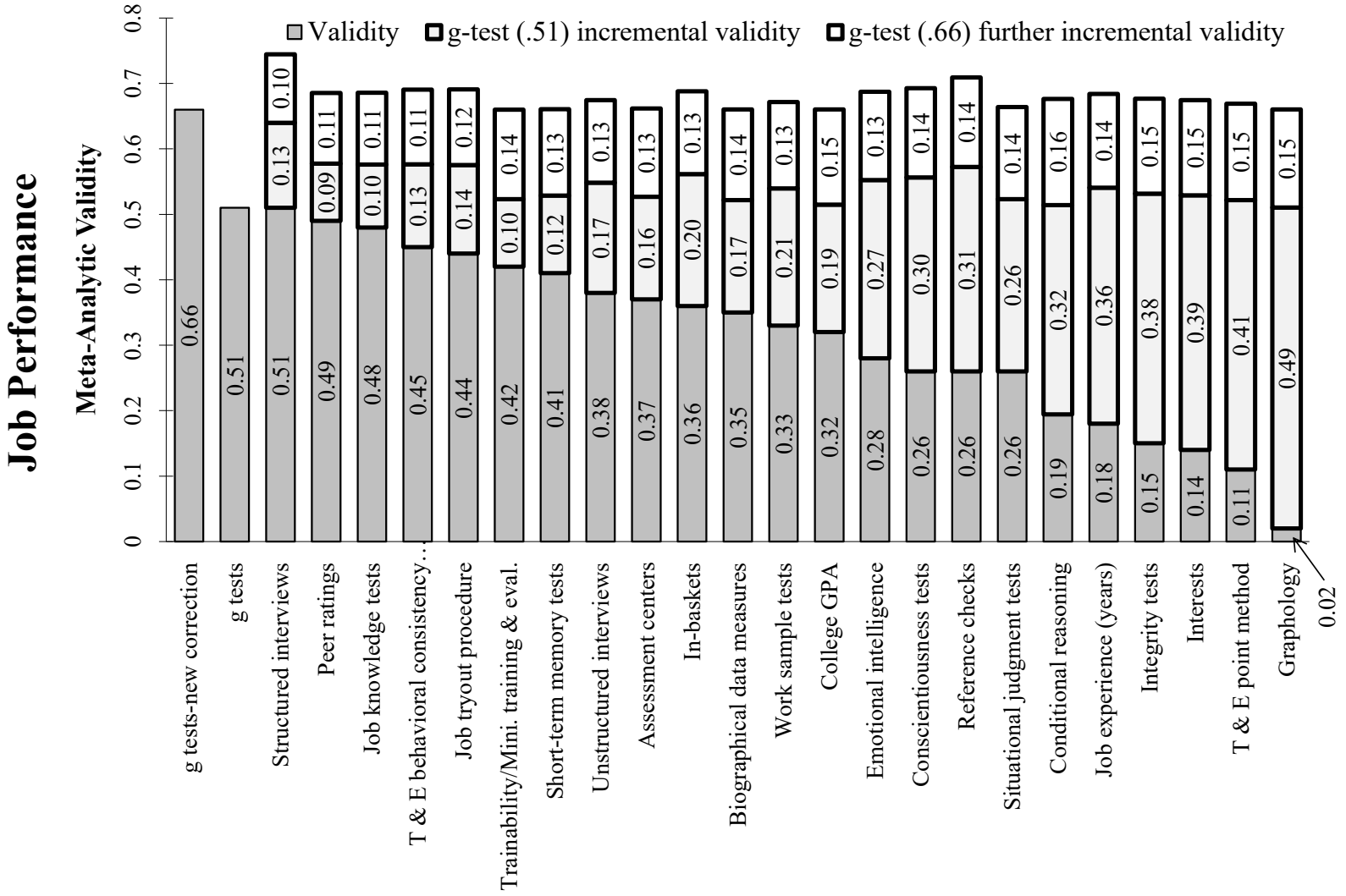
- High-level explanation
  - In order for a standardized ability test to not have adverse impact and retain its validity, the narrow aspects of intelligence would have to correlate with majority/minority group status in the opposite direction that general intelligence does.
    - However, these tests measure such a small amount of the narrow aspects that it is not possible to cancel out the adverse impact that is due to general intelligence.

Alternatives to *g*

3. Is it possible for another non-*g*-test related variable to have equal criterion-related validity as *g*-test?
   - Yes, but *g*-test always adds incremental validity
   - For *g*-test validity of .51, alternative could have validity as high as .86 (.75 for *g*-test validity of .66)
   - *g*-test adds incremental validity $\Delta R = .140$ (.249 for *g*-test validity of .66) and $\Delta R^2 = .260$ (.436 for *g*-test validity of .66)
   - Note that even a test with perfect validity ($r = 1.0$) has $d \geq .51$ (or .66) and violates 4/5ths ratio when less than ~75% applicants selected
     - Recall that Cleary-fair test with validity of .51 and $d = 1$ SD implies $d = .51$ on criterion
     - (see Sackett & Ellingson, 1997, Table 2, to convert *d* and selection ratio to 4/5 ratios)

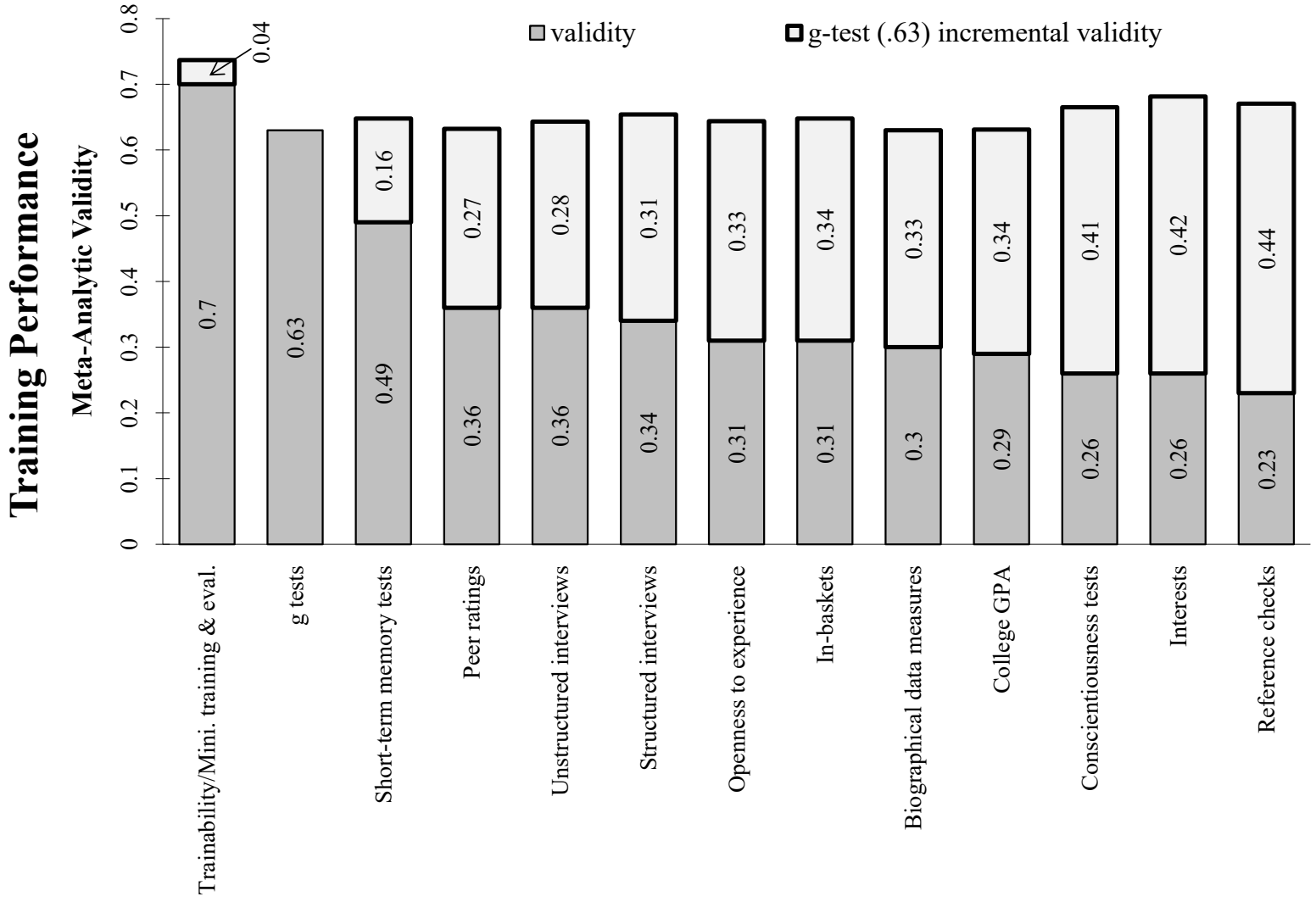
Alternatives to *g*

- *g*-tests add incremental validity over all known alternatives



Alternatives to *g*

- *g*-tests add incremental validity over all known alternatives



Alternatives to *g*

- Most known alternatives are semi-*g*-tests; partialling out *g* often lowers validity

Test	*g*-loading	Bivariate validity	Partial validity
Job knowledge tests	.55	.48	.11
Biodata	.57	.35	-.16
College GPA	.59	.32	-.24
Situational judgment tests	.33	.26	.01
Assessment centers	.57	.37	-.12
Work sample tests	.37	.33	.08
Structured interviews	.31	.51	.44
In-baskets	.30	.36	.21
Interests	.00	.14	.21
Conscientiousness tests	.09	.26	.30
Job tryout procedure	.44	.44	.19

Alternatives to *g*

- High-level explanation
  - It might be possible to develop a non-cognitive test that predicts performance as well as a standardized ability test but without adverse impact.
    - Non-cognitive tests include personality, interests, etc..
  - However, a standardized ability test would improve prediction of performance over and above this non-cognitive test. The resulting battery would then have adverse impact.
  - Even if you developed a test that perfectly predicted performance, it would also have adverse impact.
  - It is not possible to avoid adverse impact if you want to maximize the validity of your selection process.
  - Most personnel selection tests predict performance by virtue of measuring general intelligence.

Practical Implications

- Search for alternatives
  - Based on our analyses, there are no alternatives to cognitive ability tests that have equal validity and less adverse impact
- What should practitioners do?
  - Searching for alternatives is unlikely to resolve the validity-diversity dilemma
  - We suggest educating organizational leaders on tradeoffs of validity vs. diversity (e.g., expected job performance, ROI, adverse impact, diversity, etc. for different selection system options)
  - Be aware that many selection measures actually measure *g*
  - Be aware of basic research suggesting changes in test scores and group differences over time

Practical Implications

- Flynn (1984, 1987, 2007, 2012)
  - Reported that average IQ scores have been increasing about 1 *SD* each generation
  - Suggests that there are some environmental effects impacting *g*
- Dickens & Flynn (2006)
  - Reported some evidence that group differences are decreasing
  - However, this was not seen in Roth et al. (2001) meta-analysis
- There has been some debate in the intelligence literature on these findings
- However, it does suggest that one day group differences might disappear

Questions?
Comments?
Suggestions?