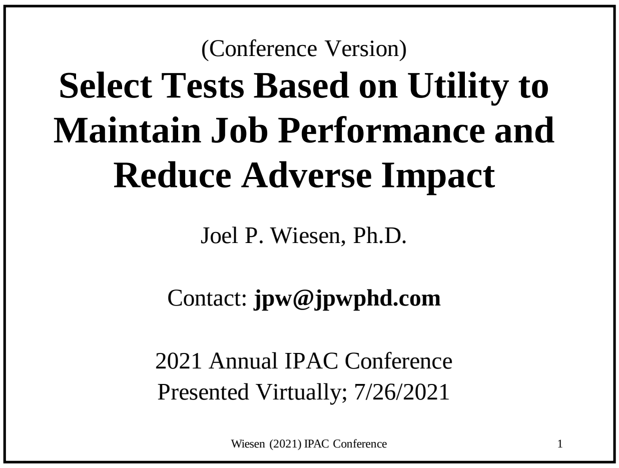
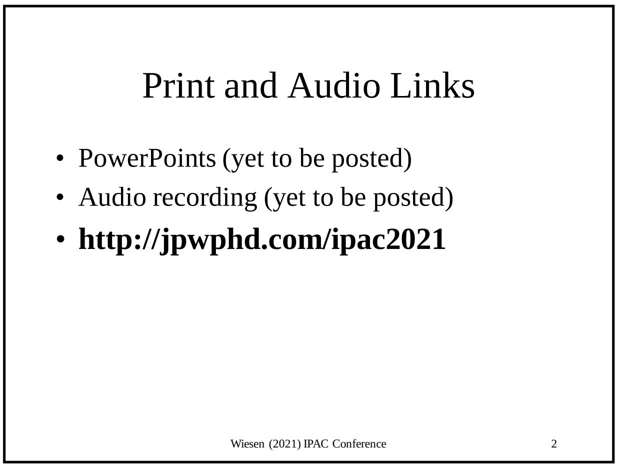

(Conference Version)

Select Tests Based on Utility to Maintain Job Performance and Reduce Adverse Impact

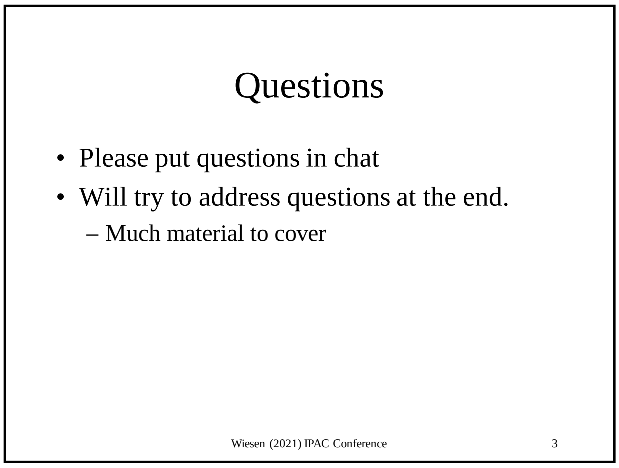
Joel P. Wiesen, Ph.D.

Contact: **jpw@jpwphd.com**

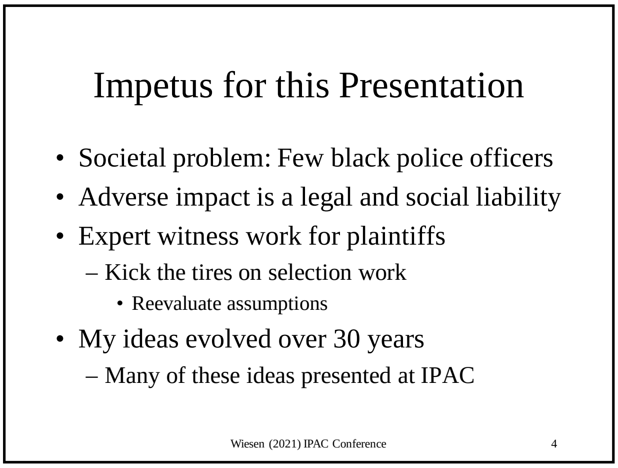
2021 Annual IPAC Conference
Presented Virtually; 7/26/2021


Print and Audio Links

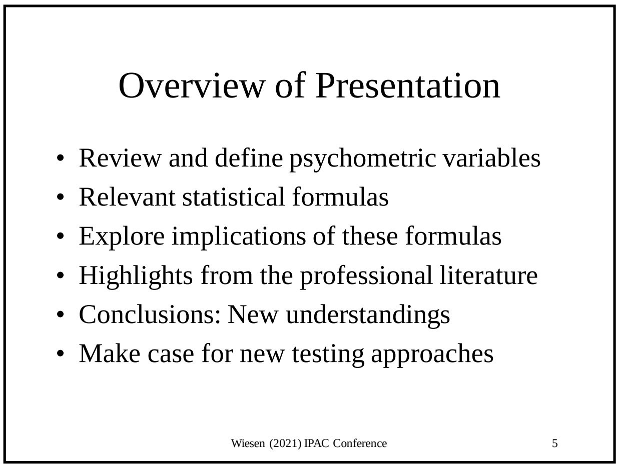
- PowerPoints (yet to be posted)
- Audio recording (yet to be posted)
- **http://jpwphd.com/ipac2021**


Questions

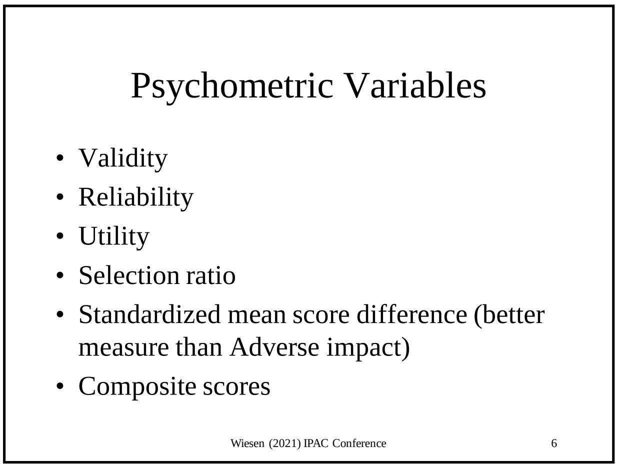
- Please put questions in chat
- Will try to address questions at the end.
  - Much material to cover


Impetus for this Presentation

- Societal problem: Few black police officers
- Adverse impact is a legal and social liability
- Expert witness work for plaintiffs
  - Kick the tires on selection work
    - Reevaluate assumptions
- My ideas evolved over 30 years
  - Many of these ideas presented at IPAC


Overview of Presentation

- Review and define psychometric variables
- Relevant statistical formulas
- Explore implications of these formulas
- Highlights from the professional literature
- Conclusions: New understandings
- Make case for new testing approaches


Psychometric Variables

- Validity
- Reliability
- Utility
- Selection ratio
- Standardized mean score difference (better measure than Adverse impact)
- Composite scores

Validity

- Joint Standards:
  "The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test." (glossary)
- Usually denoted as r
  - r can vary from -1 to 1
  - r usually is a Pearson correlation coefficient

Reliability

- SIOP:
  "The degree to which scores for a group of assessees are consistent over one or more potential sources of error (e.g. time, raters, items, conditions of measurement) in the application of a measurement procedure"

Validity Reliability Relationship

- Validity is limited by reliability
- Formula for theoretical validity in terms of observed validity and reliability of the two measures:

$$r_{\bar{x}\bar{y}} = \frac{r_{xy}}{\sqrt{(r_{xx} r_{yy})}}$$

- Can use this to correct r Guion (2011, pg 163)

Validity/Reliability Implications

- If the reliability of your test score is .6, the validity can be no higher than .77
- If the reliability of your job performance measure also is .6, the validity of the test can be no higher than .6.

Practical Observation

- Content validity ratings may ignore this relationship between validity and reliability.
  - SMEs assume we have reliable measures of the KSAPs they rate

Validity - Job Performance

- Test users often assume that high validity and many applicants result in high job performance.
  - **This is often not so!**
- Utility tells us about job performance level
- Validity is only one factor of Utility

Utility

- SIOP:
  "Projected productivity gains or utility estimates for each employee and the organization due to use of the selection procedure" (SIOP, 2017, page 46)
- We will focus here on **job performance**
- Can consider diversity in evaluating utility
  (Cascio & Aguinis, 2011, page 331)

What Drives Utility?

- Quality of applicants (Q)
  - Proportion of applicants who can do the job
- Selection ratio (SR)
  - Ratio of openings to applicants
- Validity (r)

(Cascio & Aguinis, 2011, pg 328)

Practical Implications of Q

- Can only select from among applicants
- If no good applicants, cannot hire superstars
- If all applicants great, all hires will be great
  - Random hiring will yield superstars

  NOTE: The above does not depend on r
- Must pay attention to recruitment
- Cannot recruit more after we see test scores

Practical Implications of SR

- A lower SR results in:
  - More disappointed applicants
  - Higher expected job performance
  - More false negatives (can do job but not hired)
  - Fewer false positives (hired but cannot do job)
  - More severe adverse impact when *d* > 0
    (*d* is the standardized mean score group difference)

Formula for *d*

- Standardized mean score difference

```
        M1 – M2
d = ---------------
          Sp
```

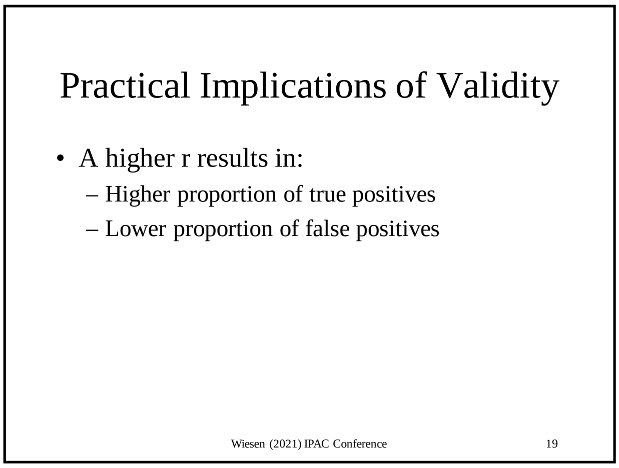
- Where M1 and M2 are group means and Sp is the pooled estimated population standard deviation

Formula for Sp

- Pooled estimated population standard

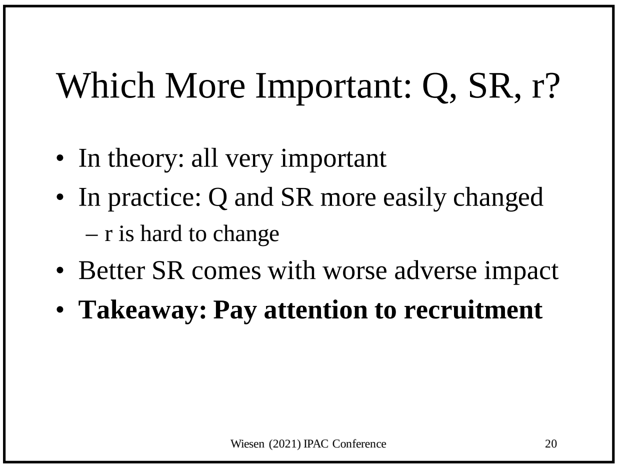
$$S_p = \sqrt{\frac{(N_1 - 1) \cdot S_1^2 + (N_2 - 1) \cdot S_2^2}{N_1 + N_2 - 2}}$$

(Source SPSS website:
https://www.spss-tutorials.com/cohens-d/)



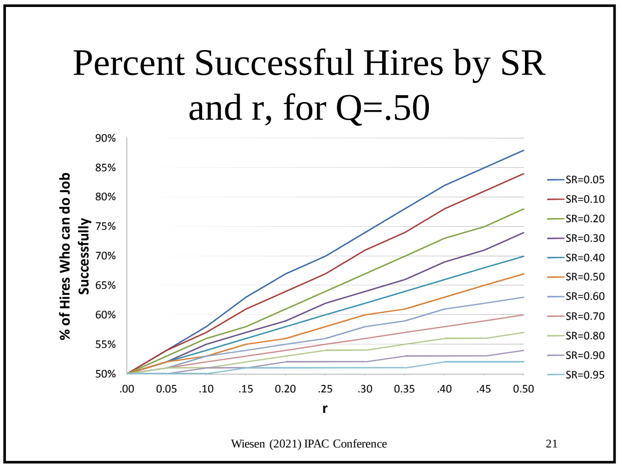
Practical Implications of Validity

- A higher r results in:
  - Higher proportion of true positives
  - Lower proportion of false positives

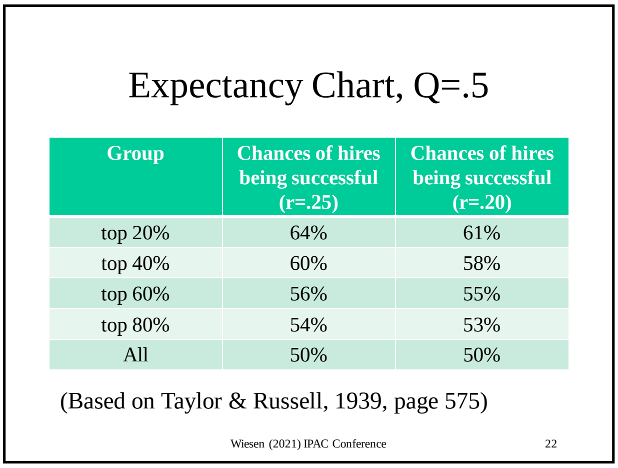


Which More Important: Q, SR, r?

- In theory: all very important
- In practice: Q and SR more easily changed
  - r is hard to change
- Better SR comes with worse adverse impact
- **Takeaway: Pay attention to recruitment**



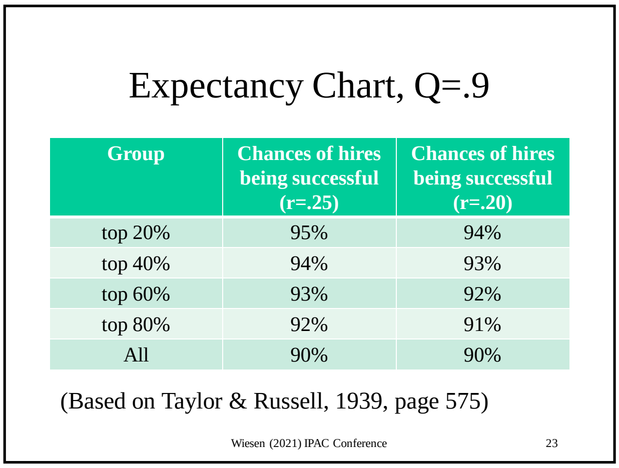
Percent Successful Hires by SR and r, for Q=.50



Expectancy Chart, Q=.5

Group	Chances of hires being successful (r=.25)	Chances of hires being successful (r=.20)
top 20%	64%	61%
top 40%	60%	58%
top 60%	56%	55%
top 80%	54%	53%
All	50%	50%

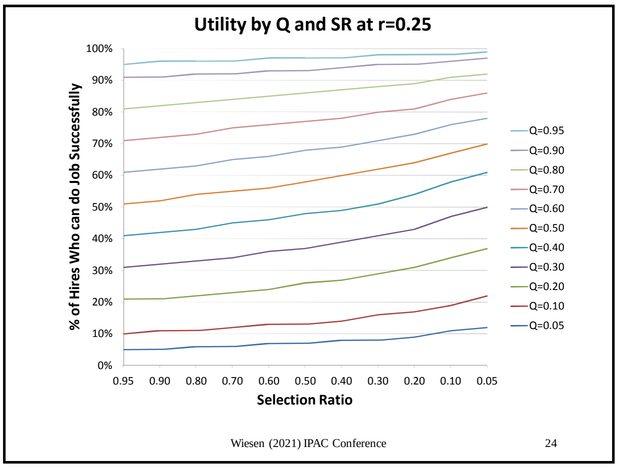
(Based on Taylor & Russell, 1939, page 575)



Expectancy Chart, Q=.9

Group	Chances of hires being successful (r=.25)	Chances of hires being successful (r=.20)
top 20%	95%	94%
top 40%	94%	93%
top 60%	93%	92%
top 80%	92%	91%
All	90%	90%

(Based on Taylor & Russell, 1939, page 575)



Management's View of Tests

- Initial view: Tests work
- Recruit lots of applicants and hire the best
- Tests are a fair way to hire employees
- Will hire really good employees
- Experienced view: Tests do not work well
  - Too many hiring errors

Critique of Management's View

- Tests work but only to a modest extent
- Recruitment should focus on quality
- There will be many mistakes in hiring
  - False positives
  - False negatives
- If we omit KSAPs that have lower *d*, the test is invalid!

Unfairness Overrides Validity

- "If ... excluding some components … has a noticeable impact on selection rates for groups ... the intended interpretation of test scores ... would be **rendered invalid**."

  AERA, APA, NCME (2014, page 21, col 1, par 1, emphasis added)

Evaluating Composite Scores

- Combine tests with lower and higher r
- Utility and *d* for this combination
- Need formulas for:
  - Validity of a composite
  - *d* of a composite
  - Utility of a composite
- Assume the two tests are uncorrelated

Validity of the Sum of 2 Tests

- Correlation of a sum of two weighted measures with a third measure

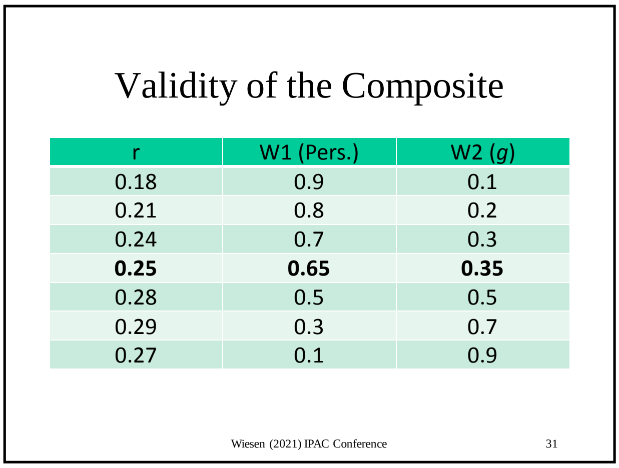
$$r_{c(ws)} = \frac{w_1 r_{c1}\sigma_1 + w_2 r_{c2}\sigma_2}{\sqrt{w^2{}_1\sigma^2{}_1 + w^2{}_2\sigma^2{}_2 + 2r_{12}w_1\sigma_1 w_2\sigma_2}}$$

(Guilford, 1965, page 427, formula 16.25)

Does a Personality Test Dilute *g*?

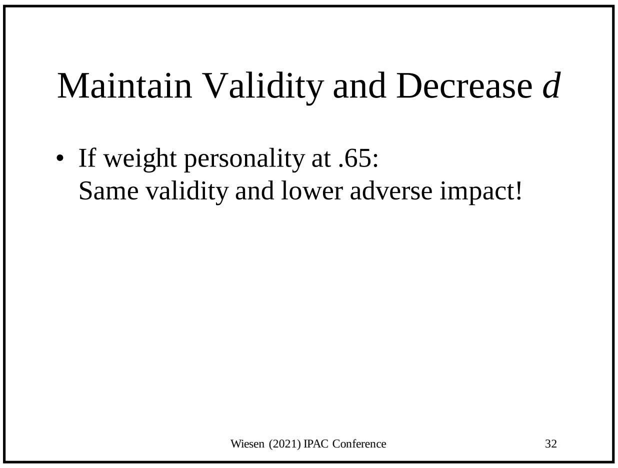
- Will a personality decrease the r due to *g*?
- Assume r = .15 for personality
- Assume r = .25 for *g*
  - r = .24 for police officers
  - I recalculated, to omit unreliability of predictor

Aamodt (2004), Table 3.1, page 36, rho=.27



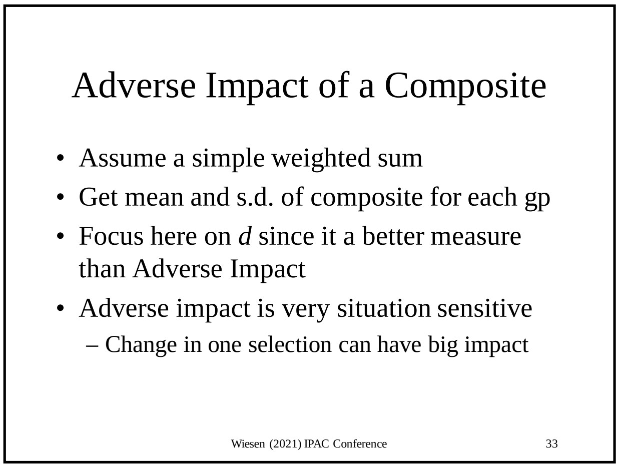
Validity of the Composite

r	W1 (Pers.)	W2 (*g*)
0.18	0.9	0.1
0.21	0.8	0.2
0.24	0.7	0.3
0.25	**0.65**	**0.35**
0.28	0.5	0.5
0.29	0.3	0.7
0.27	0.1	0.9



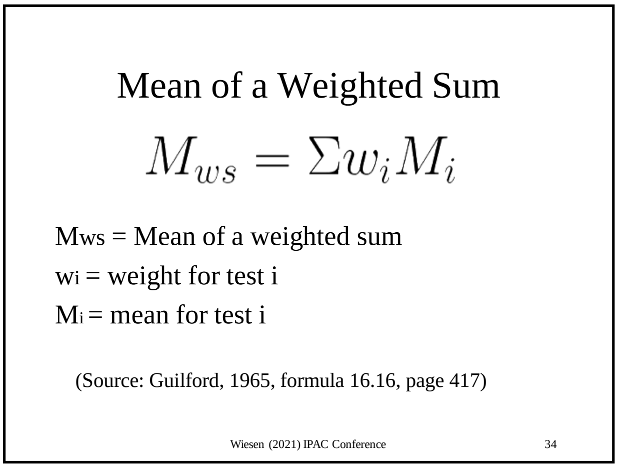
Maintain Validity and Decrease *d*

- If weight personality at .65:
  Same validity and lower adverse impact!



Adverse Impact of a Composite

- Assume a simple weighted sum
- Get mean and s.d. of composite for each gp
- Focus here on *d* since it a better measure than Adverse Impact
- Adverse impact is very situation sensitive
  - Change in one selection can have big impact



Mean of a Weighted Sum

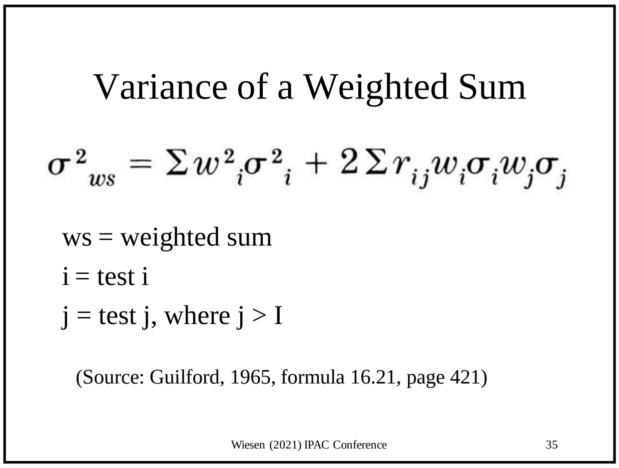
$$M_{ws} = \Sigma w_i M_i$$

M_{WS} = Mean of a weighted sum

w_i = weight for test i

M_i = mean for test i

(Source: Guilford, 1965, formula 16.16, page 417)



Variance of a Weighted Sum

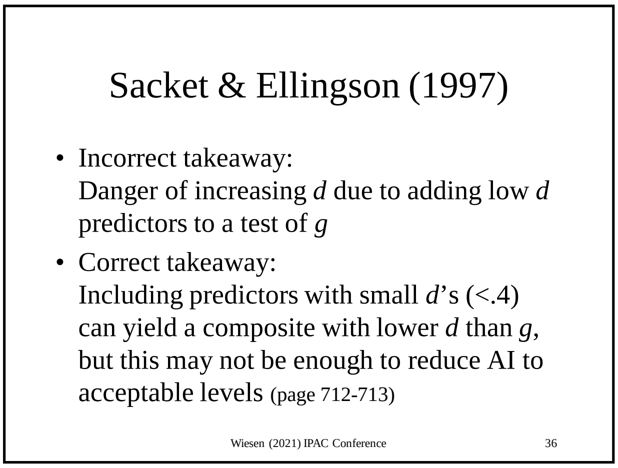
$$\sigma^2_{ws} = \Sigma w^2_i \sigma^2_i + 2\Sigma r_{ij} w_i \sigma_i w_j \sigma_j$$

ws = weighted sum

i = test i

j = test j, where j > I

(Source: Guilford, 1965, formula 16.21, page 421)



Sacket & Ellingson (1997)

- Incorrect takeaway:
  Danger of increasing *d* due to adding low *d* predictors to a test of *g*
- Correct takeaway:
  Including predictors with small *d*'s (<.4) can yield a composite with lower *d* than *g*, but this may not be enough to reduce AI to acceptable levels (page 712-713)

Sacket & Ellingson, Formula 3

$$d = \frac{\sum_{i=1}^{k} w_i d_i}{\sqrt{\sum_{i=1}^{k} w_i^2 + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_i w_j r_{ij}}}$$

(Corrected last term in denominator; typo in journal)

d for Equally Weighted Sum

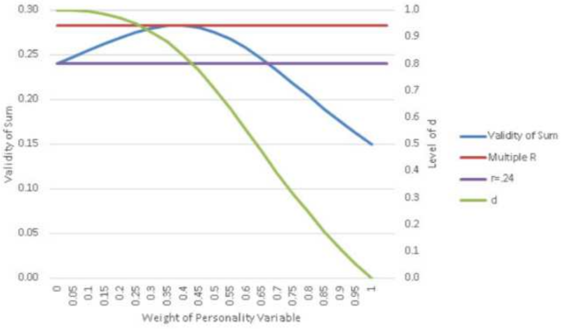
$$d = \frac{d_1 + d_2}{\sqrt{2 + 2r_{12}}}$$

Sacket & Ellingson, 1997, Formula 2

Composites of Two Tests

- An example
- Assume tests with r = .24 and r = .15
  - e.g., *g* and a personality factor, uncorrelated
- Assume *d*s of 1 and zero, respectively
- When form a composite, what happens to:
  - Validity
  - *d*

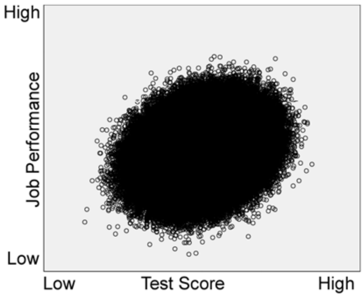
Figure 1. Validity of Sum and *d* by Weight of Personality Variable



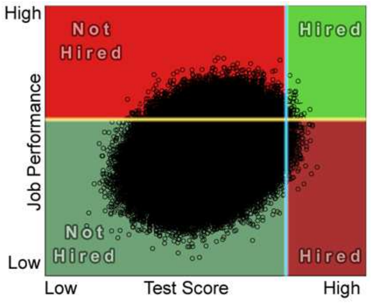
Some Professional Literature

- Sackett, Shewach, Keiser (2017)
  "In contrast to Schmidt and Hunter's … reporting … .51 for ability and .37 for ACs, we found … mean validity of **.22 for ability** and **.44 for ACs**."
- Assessment centers seem to have higher validity than tests of *g*, in general.
  - Why not rank on the test with highest validity?

Predictive Validity of *g*, r=.24



Decisions, Right and Wrong



Unmeasured Abilities

Let’s assume there are untested KSAPs:

- Creative problem solving: 10% deficient
- Oral communication: 10% deficient
- Ability to get along w others: 10% deficient
- Conscientiousness: 10% deficient
- ~34% lack abilities not tested by M/C test

Revaluate False Positive Rate

- Expectancy chart: 61 to 64% true positives
- But 34% of these are deficient on non-*g*
- These abilities probably are independent
- So, reduce the 64% by 34% = 42%
- 42% true positives

Revaluate False Positive Rate

- Conclusion:
  Most POs hires based on *g* are false positives
  - 58% false positives based on a typical test of *g*

What Happens with Higher Q?

- We hire better people
- Less room for improvement over chance
  - Cannot do much better than hiring randomly
  - **Utility is lower**

Expectancy Chart, Q = .9

Group	Chances of hires being successful (r=.25)	Chances of hires being successful (r=.20)
top 20%	95%	94%
top 40%	94%	93%
top 60%	93%	92%
top 80%	92%	91%
All	90%	90%

(Based on Taylor & Russell, 1939, page 577)

Expectancy Chart, Q = .5

Group	Chances of hires being successful (r=.25)	Chances of hires being successful (r=.20)
top 20%	64%	61%
top 40%	60%	58%
top 60%	56%	55%
top 80%	54%	53%
All	50%	50%

(Based on Taylor & Russell, 1939, page 575)

Compare Q= .5 and Q=.9

- Utility of r=.25, Q=.9 is 5% more true pos.
- Utility of r = .2, Q=.5 is 11% more true pos.
- Lower validity can have higher utility
- It depends on Q for the two areas tested
- In PD requiring college, Q for g may be high
- Q for a non-cognitive variable may be low

Takeaways

- Validity sums (validity does not average)
- Adding a low validity test improves validity
- Recruitment can improve utility more than testing
- A low validity test can have high utility
- A high validity test can have low utility
- *g* is not the best predictor of job perf.

Topics Not Covered

- Numeric examples
- Ideas on ways to reduce adverse impact
- Real life applications
- Some of this is on my website: https://appliedpersonnelresearch.com/papers

Q&As

- Feel free to contact me at any time about this topic
  - (617) 244-8859
  - jpw@jpwphd.com

References

- Aamodt, M. G. (2004a). *Research in Law Enforcement Selection*. Boca Raton, FL: Brown Walker Press.
- AERA, APA, NCME. (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

References

- Cascio, W. F. & Aguinis, H. (2011). *Applied Psychology in Human Resource Management*. Boston: Pearson.
- Guilford, J. P. (1965). *Fundamental Statistics in Psychology and Education (4th ed.)* New York: McGraw-Hill.

References

- Guion, R. M. (2011). *Assessment, Measurement, And Prediction For Personnel* Decisions (2nd ed.) Bowling Green State University. New York: Routledge.

References

- Sackett, P. R. & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707-721.

References

- Sackett, P. R., Shewach, O. R. & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102,* 1435–1447.

References

- Taylor, H. C, & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23,* 565-578.

References

- SIOP (2018*). Principles for the Validation and Use of Personnel Selection Procedures, 5th ed*. Bowling Green, OH: Author.