Quantitative Considerations in Balancing Validity, Utility, Fairness, and Adverse Impact

Joel P. Wiesen, Ph.D.
jwiesen@appliedpersonnelresearch.com

IPAC 2017 Conference

Birmingham, AL

July 19, 2016

(updated 8/20/2017)

Overview of Presentation

Define and discuss:

- Validity
- Utility
- Fairness
- Adverse impact

Describe tools to evaluate individual tests and combinations of tests in terms of the above.

Program Abstract

- Make better decisions about combining tests
- Intuition often faulty concerning:
  - 1. Validity
  - 2. Utility
  - 3. Selection ratio
  - 4. Adverse impact
  - 5. Applications; interrelationships of the above

Context for Program Abstract

- Societal problem: Few black police officers
- Cause: Adverse impact
- We are up to our necks in this dilemma
- Goal for today: Cures for this dilemma
  - Cures involve understanding the statistics of employee selection

Hold on to Your Seats

- Findings are Mind Blowing
- Low r test with more utility than higher r
- High weight to low r test yields good validity
- More recruitment yields more adverse impact

Background Information

- Terms and definitions
- Concepts

Validity ≠ Utility

Should we focus on validity or utility?

1. Validity

"The degree to which accumulated evidence and theory support specific interpretations of scores from a selection procedure entailed by the proposed uses of that selection procedure" (SIOP, 2017, glossary, page 72).

Validity Coefficient

"A coefficient of correlation that shows the strength of the relation between predictor and criterion." (AERA, APA, MCME, 1985, glossary, page 94).

2. Utility

"Projected productivity gains or utility estimates for each employee and the organization due to use of the selection procedure" (SIOP, 2017, page 46).

Utility formulas use the validity coefficient.

Utility

Evaluate overall benefit, including:

Cost of recruiting

Cost of testing

Cost of training

Implications for the organization's workforce diversity

(Cascio & Aguinis, 2011, pg 331)

What Drives Utility?

- Quality of applicants
  - Proportion of applicants who can do the job
- Number of applicants and openings
  - Selection ratio
- Validity

(Cascio & Aguinis, 2011, pg 328)

Quality of Applicants

- 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

Quality of Applicants

Moral for testing specialists:

- Pay attention to recruitment!
- Especially in the public sector
  - Cannot recruit more after we see exam scores

Quality of Applicants

- Use **Q** for quality of applicant group
  - Notation:
    Let Q = proportion of applicants who can do job

3. Selection Ratio (SR)

- Number of applicants and openings

  SR = # openings / # applicants
- Lower SR results in better hires
  - Screen out most applicants
  - Hire from the right tail of the normal curve
  - Hire from the extreme part of the right hand tail
- Lower SR results in more severe AI

Validity

- At any SR, higher validity results in:
  - Higher proportion of true positives
  - Lower proportion of false positives

Numeric Examples of Utility

- Don’t focus on details in the charts.
- Will present figures soon.

Examples of Utility, Q=.7

- Assume SR=.1, r=.25, Q=.7
- Proportion hired who can do job = .84
- Assume SR=.1, r=.20, Q=.7
- Proportion hired who can do job = .81

  (Taylor & Russell, 1939, page 576)
- Lose 3% if validity drops from .25 to .20

Examples of Utility, Q=.2

- Assume SR=.1, r=.25, Q=.2
- Proportion hired who can do job = .34
- Assume SR=.1, r=.20, Q=.2
- Proportion hired who can do job = .31

  (Taylor & Russell, 1939, page 574)
- Lose 3% if validity drops from .25 to .20

Textbook Expectancy Chart

Group	Chances of hires being successful (r=.7)
top 20%	90%
top 40%	80%
top 60%	70%
top 80%	60%
All	50%

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

Expectancy Chart, Q=.5

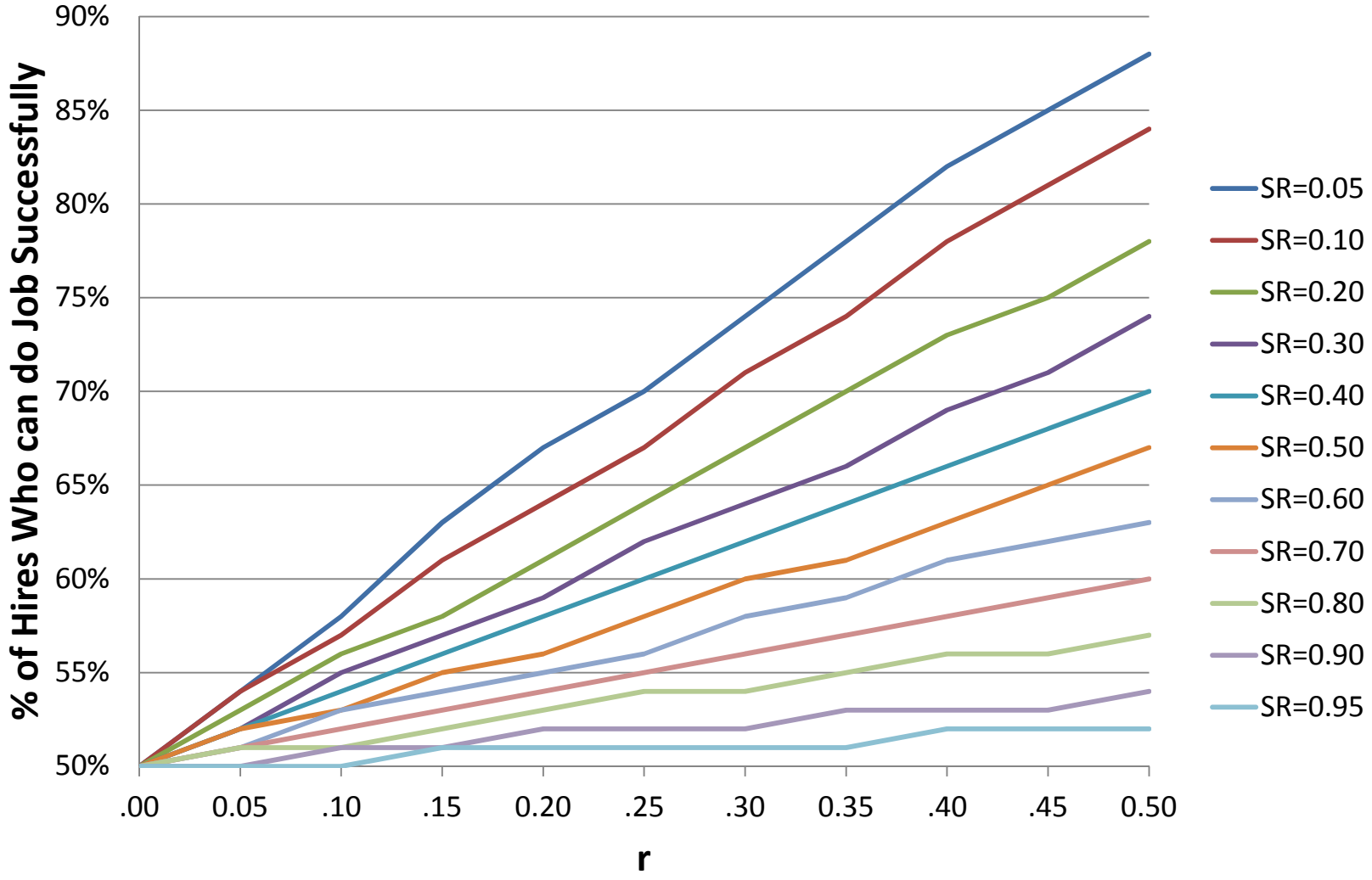
Group	Chances of hires being successful (r=.25)	Chances of hires being successful (r=.20)
top 20%	64%	61%
top 40%	60%	58%
top 60%	56%	55%
top 80%	54%	53%
All	50%	50%

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

Expectancy Chart, Q=.5

- Interpretation:
  utility driven by SR more than r
  - Within typical ranges of SR and r
- Utility reasonably large (11% or more)

Percentage of Hires Expected to Perform the Job Successfully, by Selection Ratio and Validity, for Q=.50

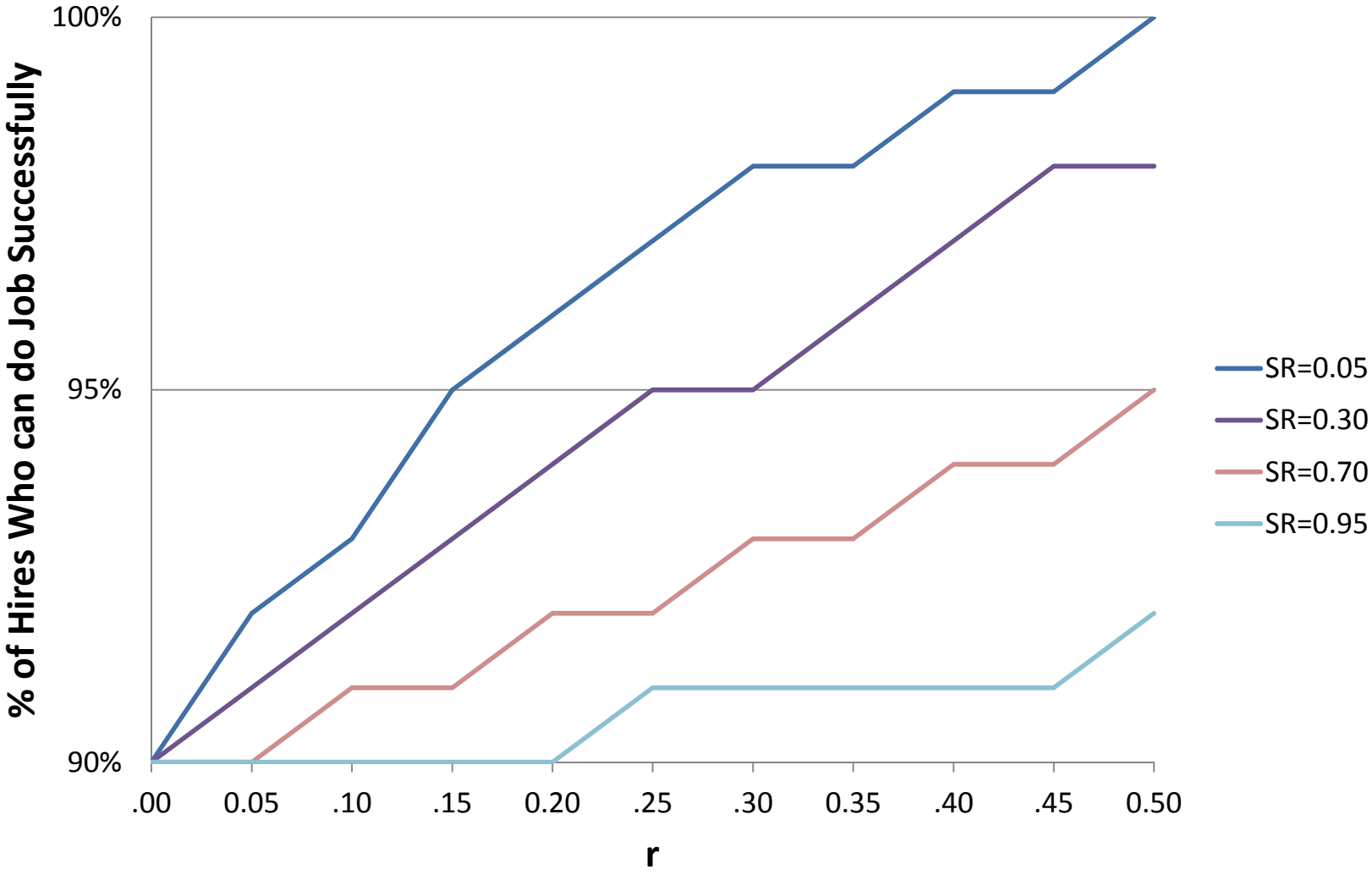


Expectancy Chart, Q=.9

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)

Percentage of Hires Expected to Perform the Job Successfully, by Selection Ratio and Validity, for Q=.90



Expectancy Chart, Q=.9

- Interpretation:
  utility driven by SR more than r
  - Within typical ranges of SR and r
- Utility small, never more than 10%

Focus on Validity or Utility?

- High validity does not guarantee high utility
- Yet utility is rarely discussed
- Utility is more the practical bottom line
- 1970 EEOC Guidelines called for high utility (Guion, 2011, page 128)
  - Superseded by the Uniform Guidelines on Employee Selection Procedures (1978)

Expectancy Chart: Honest View

- Facts
  - Hard to improve validity
  - Hard to change selection ratio
  - Hard to change quality of applicants
- Conclusion
  - Expectancy charts not relevant for civil service?
  - Is useful for multiple hurdle systems

Focus on Validity Self-Serving?

- We emphasize validity over utility because that is what our profession does?
  - We did not take college courses in recruitment
- Our past (optimistic?) claims concerning utility were rejected out of hand by management as implausible.

4. Adverse Impact (AI)

- Goal of management is “no surprises”
  - Predict AI before exam administration

Adverse Impact Definitions

- 80% rule for pass rates
- 80% rule for hiring rates
- Significantly different hiring rates
- Delays to hire date
- Mean score differences (effect size)
- Differences in placement on list

Adverse Impact Definitions

- Effect size definition is stable
  - Standardized mean score differences
- Some other definitions are a moving target
  - Especially adverse impact ratio
    - High variance
    - Changes with each additional hire
- AI ≠ fairness

Definitions of Fairness

- Cleary (industry standard)
  - Regression model
  - No under or over-prediction for individuals
- Thorndike (not widely accepted)
  - Select from each group proportional to those who would be successful on the job
    - Constant ratio approach
  - Focus on fairness for groups

End of Background

- Turn to interesting statistics of employee selection

Overview of Remaining Topics

- A. Fairness of low selection ratios
- B. Compare utility for tests of *g*, personality
- C. Predict adverse impact and validity
  - Combining g and personality tests
- D. Differential validity
- E. Some implications of the above
  - Cures for the Police Officer dilemma

A. One Fairness Issue: SR

- Lower SRs result in:
  - More severe adverse impact
  - Higher job performance (higher utility)
- Should we strive for low SR?

Are Lower SRs Unfair?

- Lower SRs can be seen as unfair
  - More false negatives overall
  - Even more false negatives for minorities
  - "a given selection score …will often result in proportionately more **false negative** decisions in groups with lower mean test scores" (AERA, APA, NCME, 1999, page 79, emphasis added).

Effect of Extending Recruitment

- Scenario: Police officer exam announced and “not enough” minority applicants
- Decide to extend application period
- Unseen implication: lower SR, higher AI
- Facilitates hiring minority POs only if proportionally more additional minority applicants

Police Officer

- Validity for *g* = .24 (meta-analysis)
  - Supervisor evaluations
  - I recalculated, to omit unreliability of predictor
  - (Aamodt, 2004, Table 3.1, page 36, rho=.27)
- Many police departments require a B.A.
  - But far from a majority

B. Utility:
Police Officer Selection

- Utility of g and Personality Tests, and other tests with lower *d* than *g*.

When Do Tests Work Best?

- High validity
- Small selection ratio
- Few applicants can do job

Utility of *g* for PO Selection

- Low validity (r=.24)
- Small selection ratio
- **High proportion of applicants can do job**

Comparison (*g* for PO)

Tests Work Best	Our Situation
High validity	Low validity
Small selection ratio	Small selection ratio
Few applicants can do job	Most applicants can do job

How to Improve Utility

- Let's apply what we have seen today to hiring police officers

Q. What Can We Change?

(A) Validity (r)

(B) Selection ratio (SR)

(C) Quality of applicant group (Q)

(D) None of the above

Key: D

Challenge the Key!

Hidden Assumptions

- Q = .9 or .95 assumes a focus on *g*
- Considering personality, Q drops sharply.
- Q drives U

Utility of Personality for PO

- Low validity
- Small selection ratio
- **Relatively few applicants can do job**

Utility of g vs. Personality (Q=.2)

Group	Chances of hires being successful (g, r=.25, Q=.95)	Chances of hires being successful (personality, r=.15, Q=.2)
top 5%	99%	30%
top 20%	98%	26%
top 40%	97%	24%
top 60%	97%	23%
top 80%	96%	21%
All	95%	20%

(Based on Taylor & Russell, 1939)

Utility of *g* vs. Personality (Q=.5)

Group	Chances of hires being successful (g, r=.25, Q=.95)	Chances of hires being successful (personality, r=.15, Q=.5)
top 5%	99%	63%
top 20%	98%	58%
top 40%	97%	56%
top 60%	97%	54%
top 80%	96%	52%
All	95%	50%

(Based on Taylor & Russell, 1939)

Utility of g vs. Personality (Q=.7)

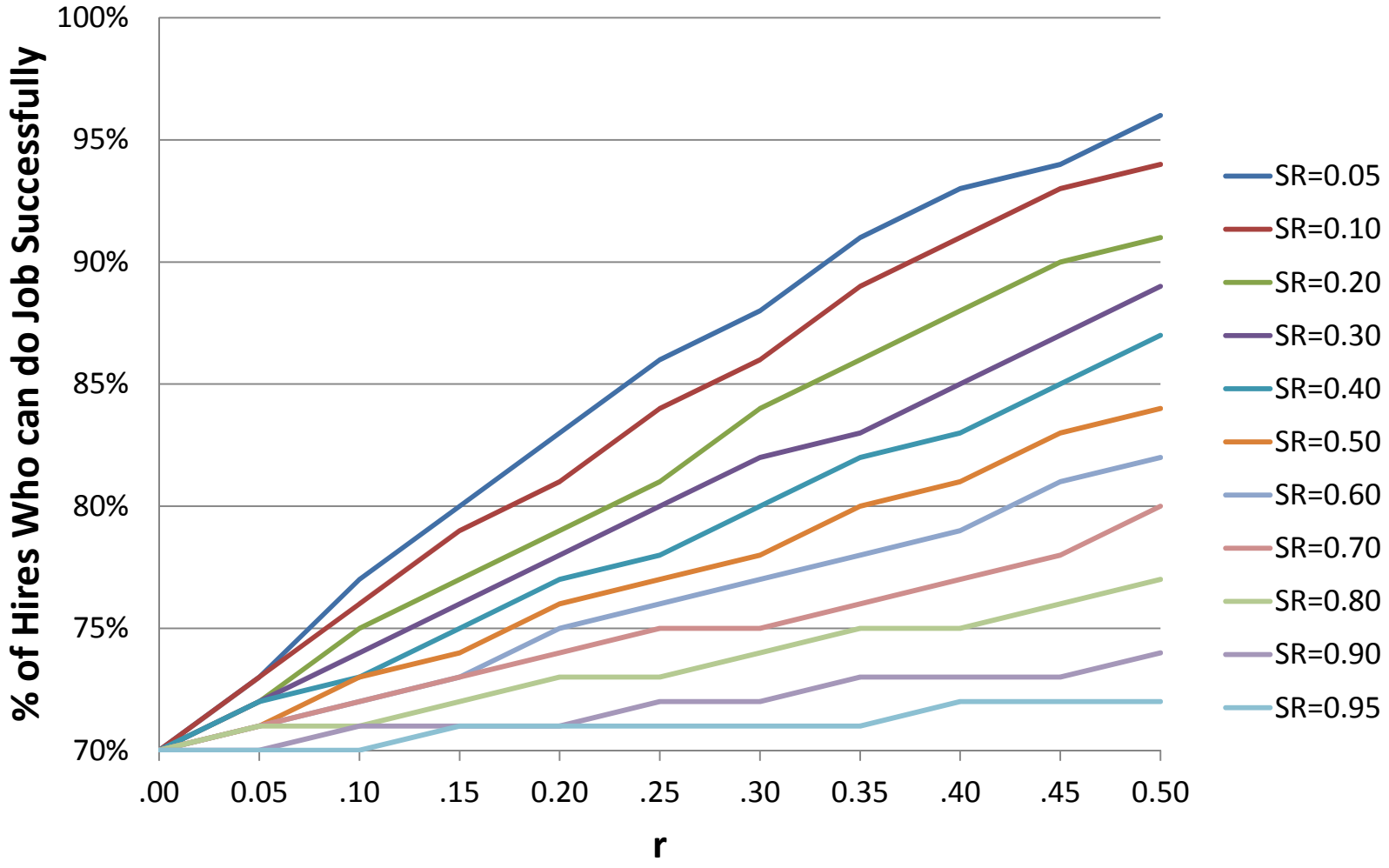
Group	Chances of hires being successful (g, r=.25, Q=.95)	Chances of hires being successful (personality, r=.15, Q=.7)
top 5%	99%	80%
top 20%	98%	77%
top 40%	97%	75%
top 60%	97%	73%
top 80%	96%	72%
All	95%	70%

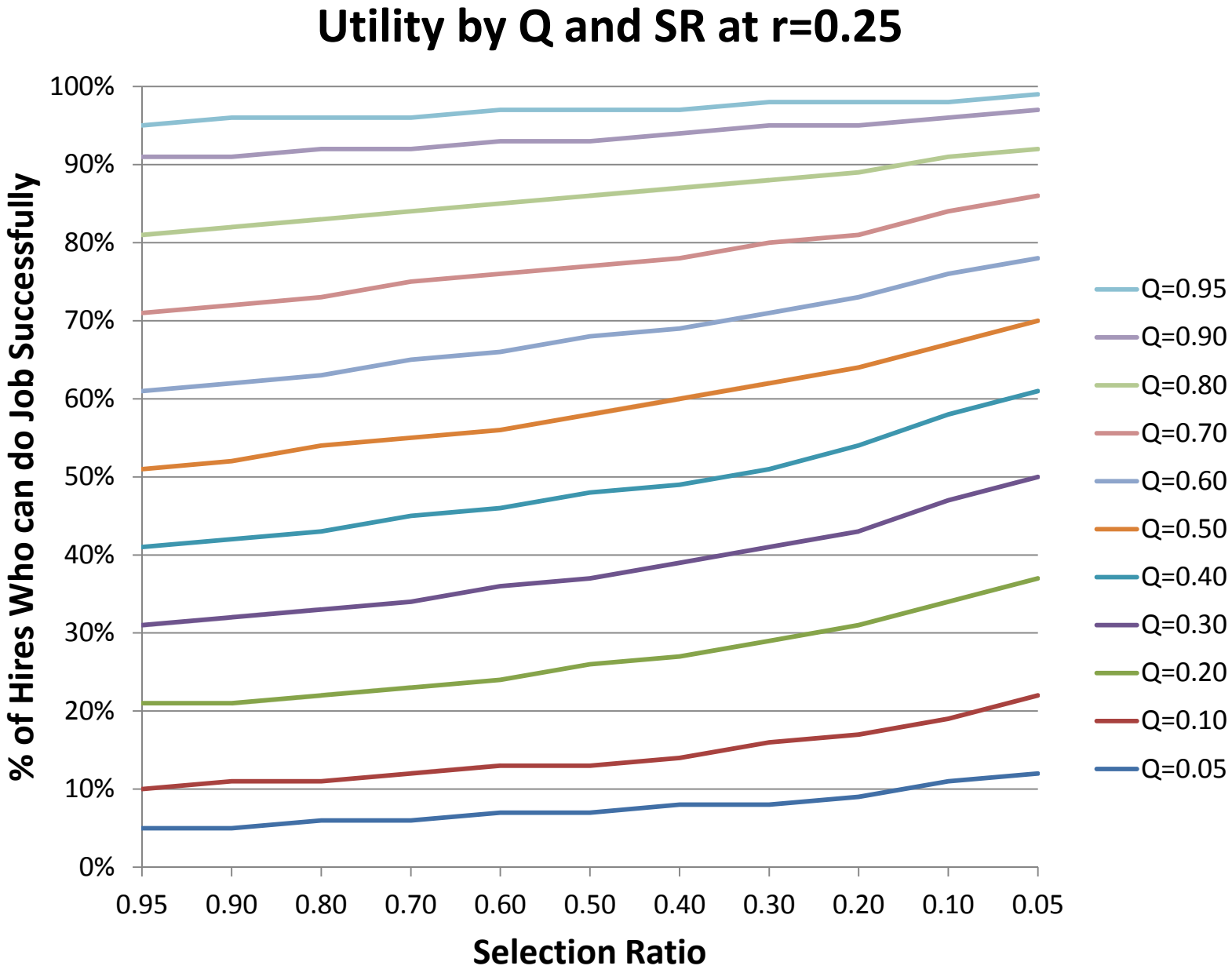
(Based on Taylor & Russell, 1939)

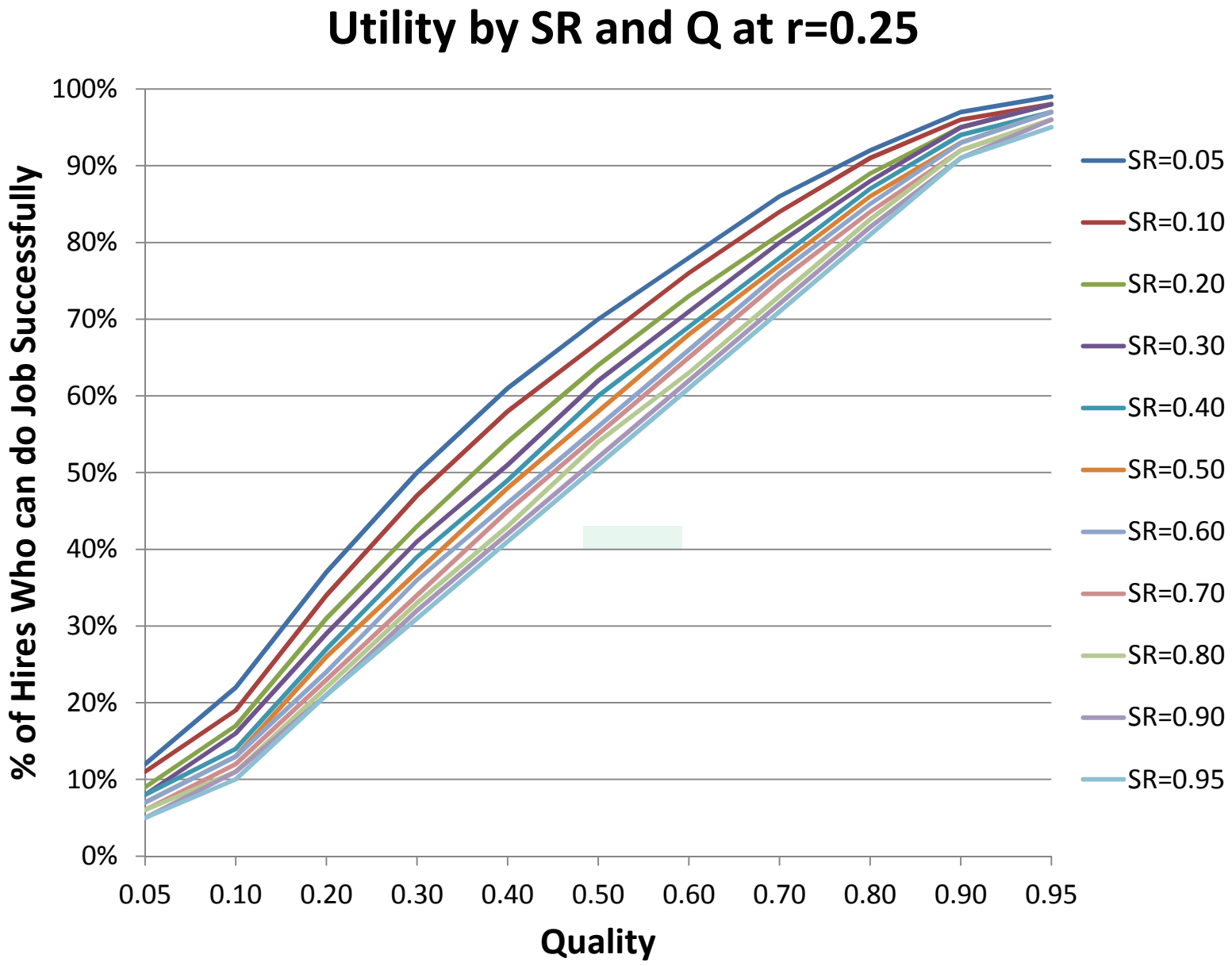
Summary of this Utility Analysis

- Test of g: 4% increase in utility (U)
- Test of personality: 10-13% increase in U
- Despite higher validity of g!

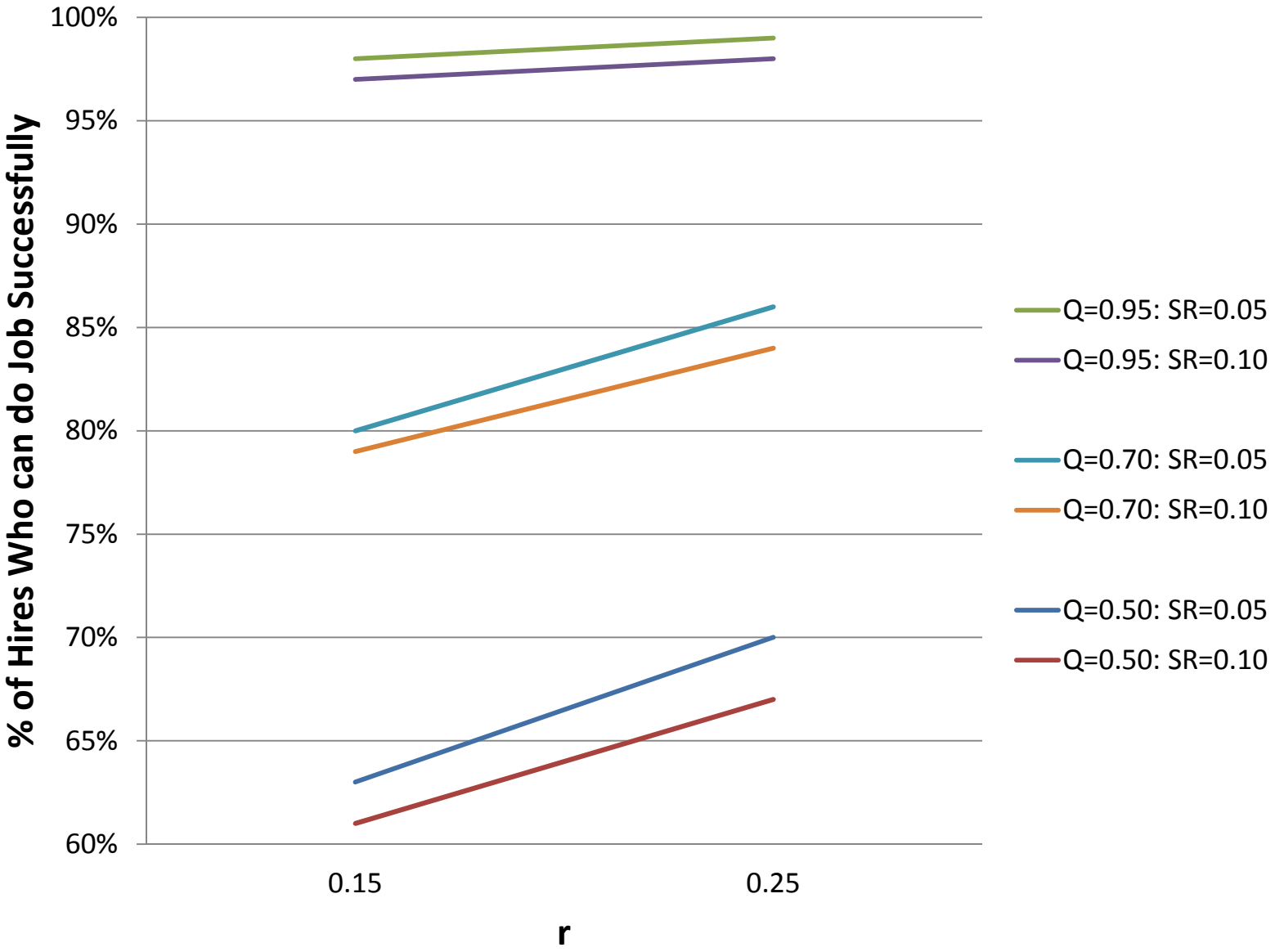
Percentage of Hires Expected to Perform the Job Successfully, by Selection Ratio and Validity, for Q=.70




Utility by Q and SR at r=0.25
% of Hires Who can do Job Successfully
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0.95
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.05
Selection Ratio
Q=0.95
Q=0.90
Q=0.80
Q=0.70
Q=0.60
Q=0.50
Q=0.40
Q=0.30
Q=0.20
Q=0.10
Q=0.05


Utility by SR and Q at r=0.25
% of Hires Who can do Job Successfully
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0.05
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
0.95
Quality
SR=0.05
SR=0.10
SR=0.20
SR=0.30
SR=0.40
SR=0.50
SR=0.60
SR=0.70
SR=0.80
SR=0.90
SR=0.95

Utility for Plausible Values of r, SR, and Q



Some Conclusions Based on T-R

- Tests provide modest utility
- Q drives utility more than r or S.R.
- Low r test can have high utility
- Personality can have higher utility than *g*

(T-R = Taylor Russell)

Implications for Police Dilemma

- Tests with lower r can have higher utility
  - Under certain circumstances
- The circumstances seem to exist for Police

Other Ways to Evaluate Use of Personality Tests

- Convergent findings of varying analyses are always comforting.
- Let's turn to another way to evaluate utility

Expected Mean Job Performance

- Naylor-Shine model for Expected Mean Job Performance
- Posits a linear relationship between validity and utility for all SRs
- Taylor-Russell utility model includes Q

(Source: Cascio & Aguinis, 2011, pg 333)

Naylor-Shine Model

$$\bar{Z}_{y_i} = r_{xy}\frac{\lambda_i}{\phi_i}$$

$\bar{Z}_{y_i}$ is the mean criterion score

r_{xy} is the validity coefficient

λ_i is the ordinate or height of the normal distribution at the predictor cutoff, $\bar{Z}_{x_i}$

ϕ_i is the SR

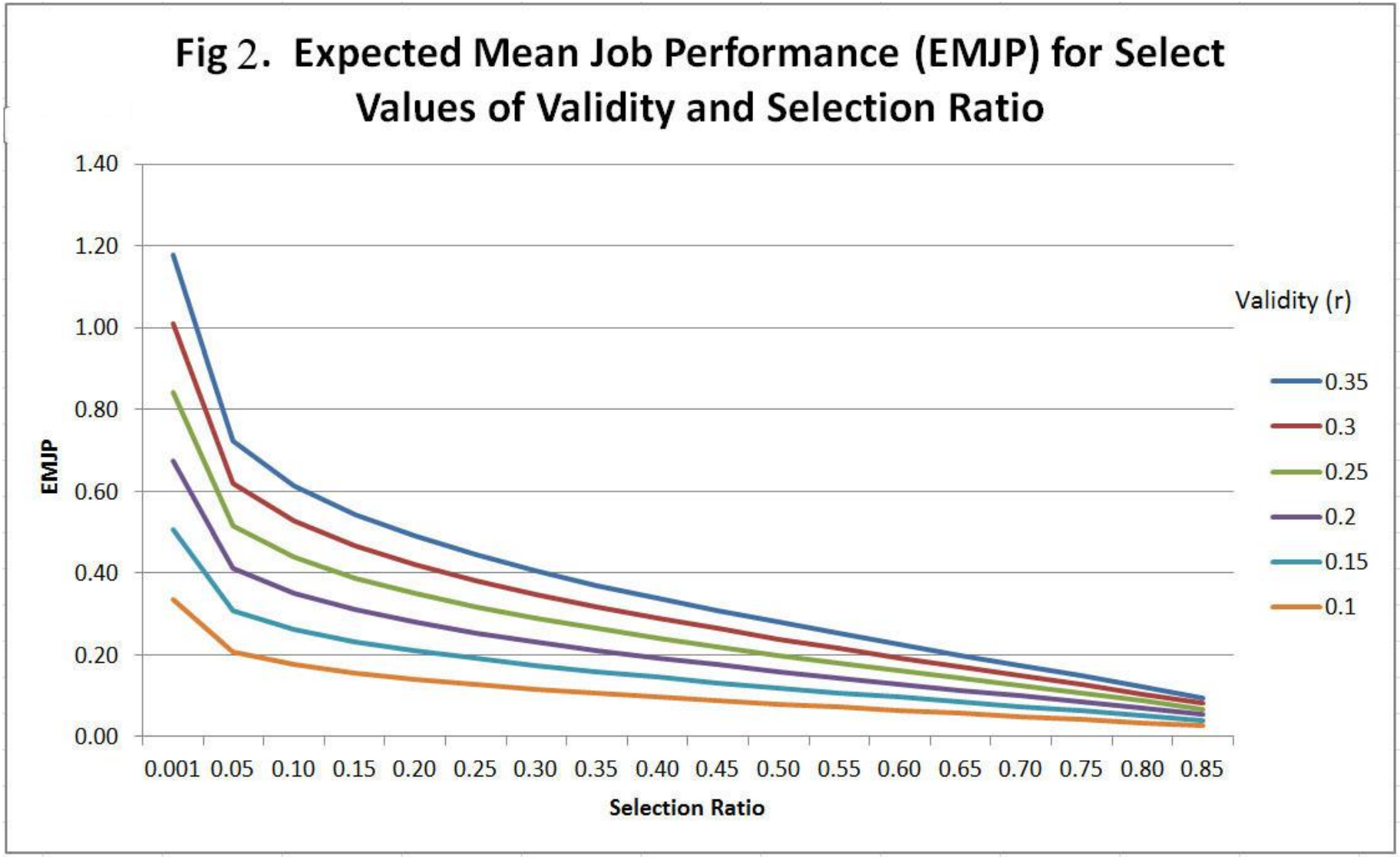


Fig 2. Expected Mean Job Performance (EMJP) for Select Values of Validity and Selection Ratio
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
EMJP
0.001 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85
Selection Ratio
Validity (r)
0.35
0.3
0.25
0.2
0.15
0.1

Some Conclusions Based on N-S

- Modest differences in utility of tests with low and higher validity
- SR drives utility, seemingly more than r
- At high validity and low SR, U is < 1 sd

Which Model is Correct?

- Taylor & Russell with Q

versus

- Naylor-Shine without Q

Which Model is Correct?

- There is a relationship between r and Q
- If there is little variance in the criterion, the observed validity will be low
- Taylor-Russell seems to assume that r in their formula is for the population (rho)
- Naylor-Shine seems to assume r is for the sample

Effect of Weights on r and AI

- First look at relevant formulas
- Then apply formulas
  - How much to weight personality vs *g*

Validity of the Sum of 2 Tests

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

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

(Guilford, 1965, page 427, formula 16.25)

C. Formulas to Calculate *d*'s

- Formula for mean of a weighted sum
- Formula for variance of a weighted sum

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, forumula 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) will 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)

Sacket & Ellingson, Formula 2

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

Estimating Adverse Impact

- Can use Excel to calculate area of a normal curve above a certain score
- The probability that a score is > z:

  =1- NORMSDIST(z-score)
- The z score for a given p value:

  =NORMSINV(p value)

Estimating Adverse Impact

- Create two distributions in Excel to calculate area of a normal curve above a certain score
- Calculate probability that a score is > z
  - Subtract 1.0 from mean of minority distribution
- Form ratio of the two probabilities

Adverse Impact for $d = 1.0$

z score White	p value White	1-p White	z score Minority	p value Minority	1-p Minority	Adverse Impact
-2	0.02	0.98	-1	0.16	0.84	0.86
-1	0.16	0.84	0	0.50	0.50	0.59
0	0.50	0.50	1	0.84	0.16	0.32
1	0.84	0.16	2	0.98	0.02	0.14
2	0.98	0.02	3	1.00	0.00	0.06

Putting It All Together

- Look at validity and AI of combination of *g* and personality tests
- Predictions of:
  - Validity of combination vs *g*
  - Validity compared to multiple R
  - Adverse impact for g, multiple R, other weights

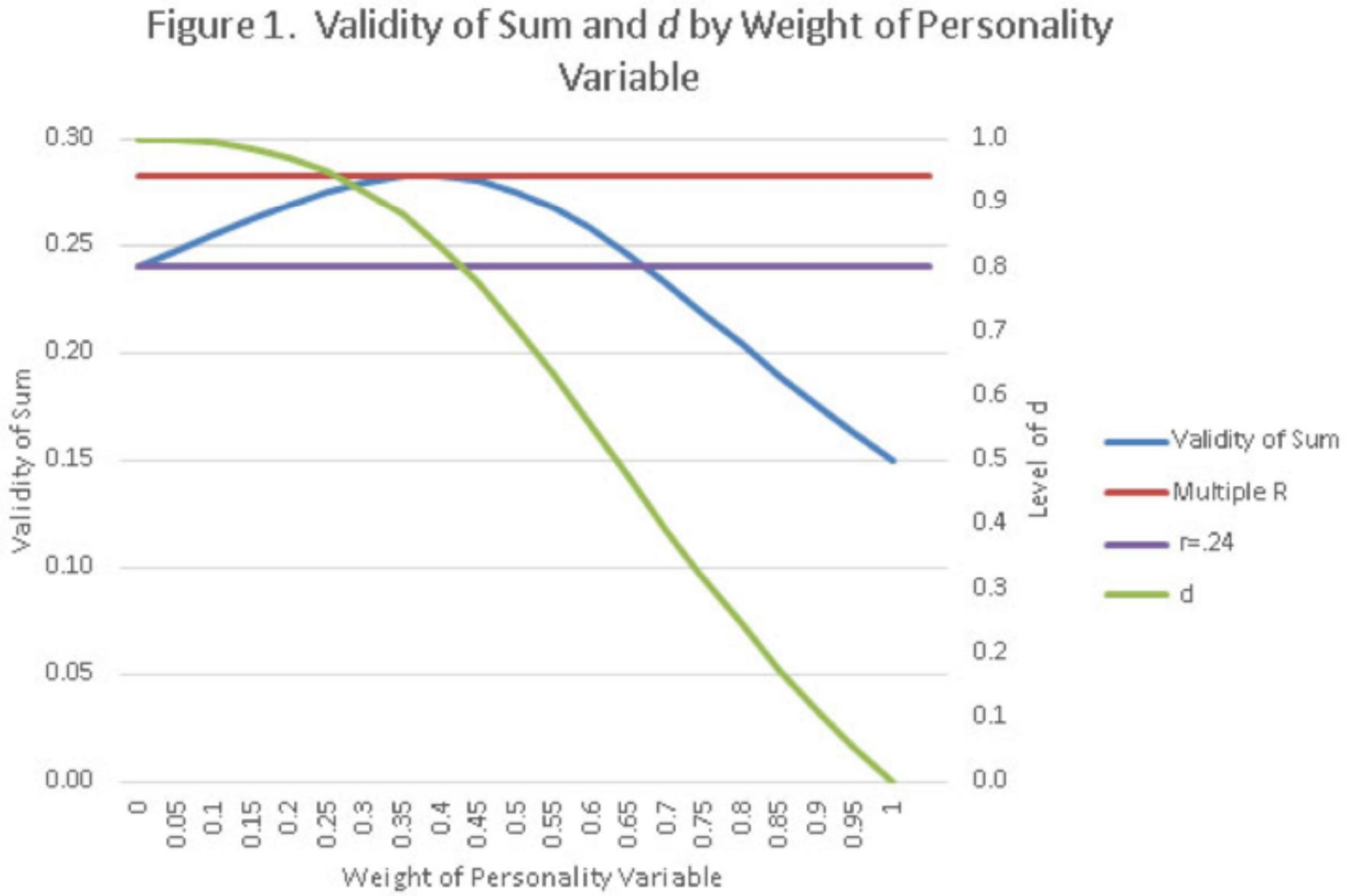


Figure 1. Validity of Sum and d by Weight of Personality Variable
Validity of Sum
Level of d
Weight of Personality Variable
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Validity of Sum
Multiple R
r=.24
d

Tradeoffs

- Look at validity/adverse impact
  - Tradeoff may be non-existent or small
- Look beyond validity/adverse impact
  - May be no tradeoff in utility
  - Utility arguably more important than validity

D. Differential Validity

- Long thought that differential validity does not exist.
- Now literature indicates it does exist

Differential Validity Exists

- SAT used for college admission

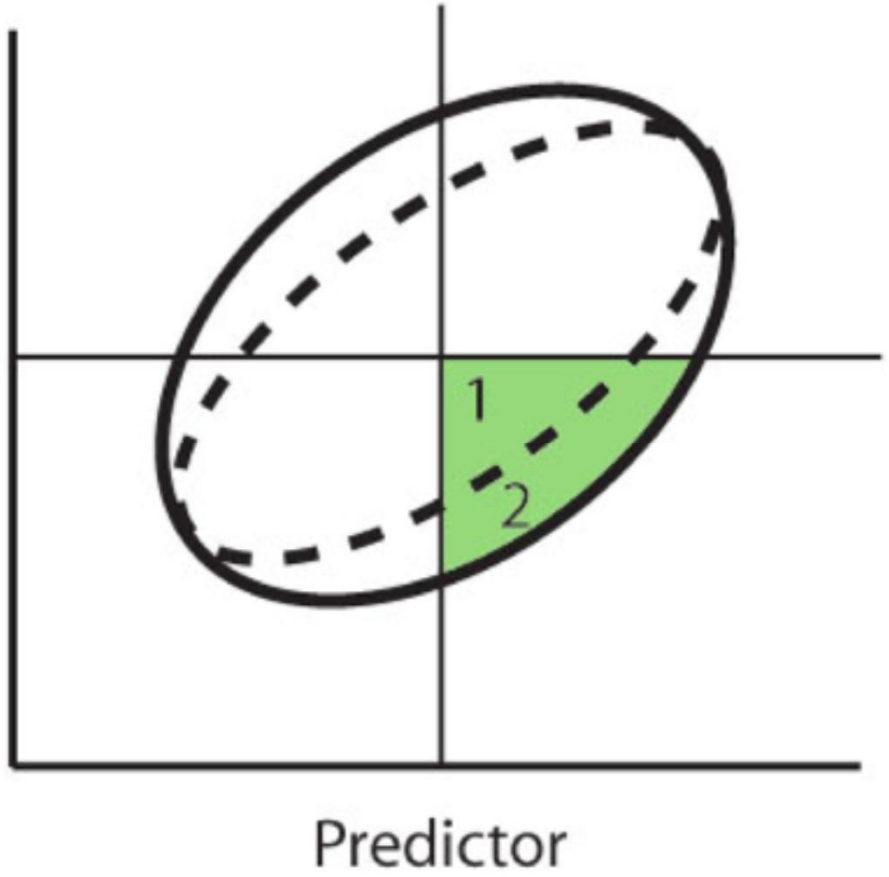
  (Mattern, Patterson, Shaw, Kobrin & Barbuti, 2008, Table 2)
- Cognitive ability tests used for employee selection, for some employment types

  (Berry, Clark and McClure, 2011, Table 1)

Danger of Differential Validity

- Lower mean job performance for group with lower validity, despite same hiring standard for all applicants
  - Reason, more false positives with lower r
  - Also more false negatives, but none hired

False Positives for Two Levels of Validity



E. Cures for the PO Dilemma?

- Pay attention to recruitment
  - High quality candidates, esp. minorities
- Use g on pass/fail basis, esp. when there is a minimum qual of a bachelor's degree
    - Lose little utility since all have high g
- Rank based on personality
    - Personality has high utility due to low Q

Cures for the PO Dilemma?

- Present Chief with predictions of r, U, AI of all options
  - Let Chief make decisions about tradeoffs between AI and U, if needed.

Cures for the PO Dilemma?

- More tools/approaches to hire minority police officers in a psychometrically responsible manner are available.

  (Sources: Wiesen, 2016; Wiesen 2017a, 2017b)

- Click on the top two links here:

http://www.appliedpersonnelresearch.com/papers/

Closing Remarks

- Focus on utility has promise for increasing job performance and improving adverse impact
- Can predict level of adverse impact
  - Should provide utility and AI information as part of proposals or selection system options.
  - If you write RFPs, make sure you ask for this!

Closing Remarks

- Sets on ear much of our thinking
- Lead us down different paths
- Does this raise new legal issues?
- Is focusing on *g* alone now an act of intentional discrimination?
- Should the search for alternatives be guided and evaluated by utility rather than validity?

Your Questions/Comments

- Questions/comments from the attendees

- Copies of this presentation are available at http://ipacweb.org and from the author at jpw@aprpsych.com

References

Aamodt, M. G. (2004) *Research in Law Enforcement Selection*. Boca Raton, FL: Brown Walker Press.

AERA, APA, NCME (1999) *Standards for Educational and Psychological Testing (2nd ed*.) Washington, DC: American Psychological Association.

References (continued)

Berry, C. M., Clark, M. A. & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*, 881–906.

Casio, W. F. & Aguinis, H. (2011). *Applied Psychology in Human Resource Management*. Boston: Pearson.

References (continued)

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 45*(166), 38290-38315. Washington, DC: Equal Employment Opportunity Commission.

Guilford, J. P. (1965). *Fundamental Statistics in Psychology and Education* (4th ed.) New York: McGraw-Hill.

Guion, R. M. (2011). *Assessment, Measurement, and Prediction for Personnel Decisions* (2nd ed). New York: Routledge.

References (continued)

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L. & Barbuti, S. M. (2008). Differential validity and prediction of the SAT. *College Board Research Report No. 2008-4*. New York: The College Board. Downloaded 1/3/2017 from https://collegereadiness.collegeboard.org/pdf/redesigned-sat-pilot-predictive-validity-study-first-look.pdf

Sacket, 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 (continued)

SIOP (2017). *The Principles for the Validation and Use of Personnel Selection Procedures [Draft of 5th ed.]* Downloaded March 22, 2017 from http://siop.org/principlesreview

Wiesen, J. P. (2016, November). Tools to Increase Diversity and Validity in Hiring Police Officers. *The Personnel Testing Council of Metropolitan Washington Newsletter, XII (3)*, 4-11.

Wiesen, J. P. (2017a, March). Tools to Increase Diversity and Validity in Hiring Police Officers - Part II. *The Personnel Testing Council of Metropolitan Washington Newsletter, XII(4)*, 6-15.

References (continued)

Wiesen, J. P. (2017b, July). Tools to Increase Diversity and Validity in Hiring Police Officers - Part III. *The Personnel Testing Council of Metropolitan Washington Newsletter, XIII (1)*, 6-17.