Opportunities to Improve Testing Research and Practice

Joel P. Wiesen, Ph.D.
jwiesen@appliedpersonnelresearch.com
IPAC 2011 Conference
Washington, D.C.
July 18, 2011

Overview

- Consider 10 topics
- Describe each topic
  - Pros
  - Issues
- Suggest possible approaches
  - Opportunities

Topics

- 1. Validity decreases over time
- 2. Criterion bias
- 3. Validity generalization weaknesses
- 4. Readability
- 5. Job analysis for content validity

Topics
(continued)

- 6. Rater reliability in grading: score changes
- 7. Test fairness definition
- 8. Unknown weights within tests
- 9. Norm vs. criterion-referenced testing
- 10. Within person variability

1. Validity Decreases Over Time

- Pros:
  - Not immediately apparent (but stay tuned)

Validity Decreases Over Time

- Issues:
  - Cognitive ability should loom large forever
    - validity decreases over time, even for complex jobs (Farrel & McDaniel, 2001)
  - Validity of psychomotor exceeds that of cognitive ability after 7 years for all jobs (Farrel & McDaniel, 2001)
  - Practice leads to lower validity for CAT (Murphy, 1989)

Validity Decreases Over Time

- Issues:
  - Validity of GPA decreases with time (Roth, BeVier, Switzer, & Schippmann, 1996)
  - But if GPA is measuring intelligence, why does the validity not increase with time (on the theory that the cumulative knowledge curves for smarter and less smart people will diverge more with time)?

Validity Decreases Over Time

- Issues:
  - Thorough training mitigates differences in ability (Schmidt, Hunter & Outerbridge, 1986)
  - Older Air Traffic Controllers show both
    - Decreased essential abilities
    - Unimpaired job performance (Nunes & Kramer, 2009)

Frank Schmidt

- *g* is measured indirectly by verbal and quantitative skills
- People learn these based on the investment of their General Mental Ability (GMA, or *g)*
- ..."adults differ dramatically in where they invested their GMA."
  (SIOP, 2011)

Implications

- Issue: Tests of *g* are fair only if test takers
  - Have equal opportunity to learn, and
  - Have invested equal effort in learning verbal and quantitative skills.

Validity Decreases Over Time

- Issues:
  - 1978 error in scoring/reporting the ASVAB resulted in the military hiring 200,000 in the lowest 10%, all of whom should have failed by virtue of their low scores.
  - The 4 services instituted workplace literacy programs costing $70 million ($350/person!). Performance and promotions were “almost normal.” (DuBay, 2004, page 5)

Validity Decreases Over Time

- Opportunities:
  - Training mitigates differences in ability
  - Screen less and train more to decrease adverse impact

Validity Decreases Over Time

- Opportunity:
- Explore norming tests based on:
  - Socioeconomic variables
  - Educational background
    - e.g., a test score of 100 would be interpreted differently depending on educational background
  - High school quality

Validity Decreases Over Time

- Opportunities:
  - Perhaps screen using one or more of these:
    - Conscientiousness
    - High school rank
    - Multiple cutoffs
    - Greatest strength approach (Wiesen & Aguinis, 2010)
    - Completion of training programs (Wiesen, 2010)

Validity Decreases Over Time

- Opportunities:
  - Perhaps screen using one or more of these:
    - Structured random sample
      - Allow for more in-depth screening (e.g., orals, training)

2. Biased Criteria

- Pros:
  - Impetus to reevaluate validity literature

Biased Criteria

- Issue: Indications of Criterion Bias
  - Short people paid less than tall (Judge & Cable, 2004)
  - Pretty people paid more than homely people (both genders, Marlowe, Schneider & Nelson, 1996)
  - Women paid less than men
  - Implicit bias research

Biased Criteria

- Opportunities:
  - Strive for better criteria
  - Perhaps test bias literature will make more sense

3. Validity Generalization (VG)

- Pros:
  - See big picture (not obscured by chance)
  - Shortcomings of single studies less serious (if not systematic across studies)

Validity Generalization

- Issues:
  - Few validity studies for most job titles
  - Coarse classification of test areas
    - Can test verbal ability many ways
  - Ignores criterion bias in underlying studies
  - Only corrects upward
  - Some VG findings are counter-intuitive

VG: Only Corrects Upward

- Possible downward corrections
  - Validity of job performance vs. training
  - Validity after learning job (vs. while learning)
  - Publication bias (negative studies not submitted)
  - Criterion contamination (supervisors know selection scores)
  - Biased criteria
  - Method contamination

Counter-Intuitive VG Findings

- Some studies report validity for more complex jobs is not higher than for low complexity jobs
  (Berry, Clark & McClure, 2011; Hartigan & Wigdor, 1989).
- Other studies show the contrary
  (Hunter, 1980, as cited by Berry et al., 2011)
- There must be as yet unnoticed factors

Counter-Intuitive VG Findings

- After 7 years, psychomotor skill is MORE valid than cognitive ability
  (Farrell & McDaniel, 2001)
- Project A found a negative weight for reading test for mechanics
  (Wise, McHenry & Campbell, 1990)
  - Would NOT be predicted by VG literature

Validity Generalization

- Opportunities:
  - Do more criterion related validity studies
    - Will allow for finer classification of jobs
  - Describe test areas fully in VG studies
  - Correct both upward and downward when doing VG studies

4. Readability

- IPAC session Tuesday at 10:30 on a new approach to readability
- http://wordsly.com

5. Job Analysis- Content Validity

- "Content Validation is Useful for Many Things, but Validity Isn't One of Them"
  - This is the title of a journal article
- "There is little empirical support for the hypothesis that the match between job content and test content influences validity"
- Intercorrelations of test areas ignored

Murphy (2009)

Job Analysis for Content Validity

- Creativity is not in vogue. Rather problem solving.
  - But creativity may be important and not related to g
- We ignore advances in other branches of psychology
  - Executive function
  - Cognitive psychology

Job Analysis for Content Validity

- Opportunities:
  - More research on the accuracy of job analysis for identifying test content
  - Challenge SMEs; they can be too compliant
  - Consider intercorrelation of test areas
  - Consider advances in other branches of psychology

6. Rater Reliability

- Two aspects
  - Reliability of .9 may be too low
  - Structured rating scales that increase rater reliability may reduce validity

Is Reliability of .9 Enough

- Consider the applicants viewpoint
  - Want consistency in grading
  - Want consistency in **ranking**

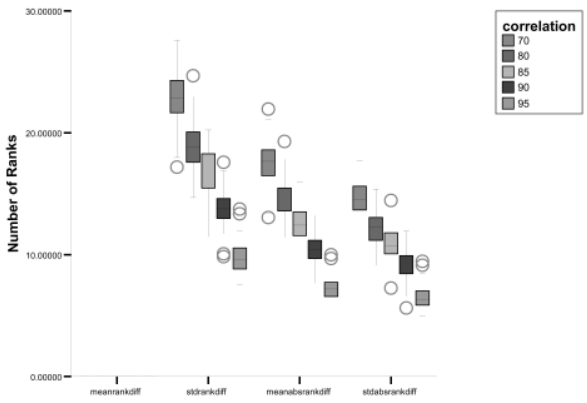
Monte Carlo Study

- Generated groups of 100 applicants
- Rescored their tests
- Calculated changes in scores and rank order
- Repeated 100 times

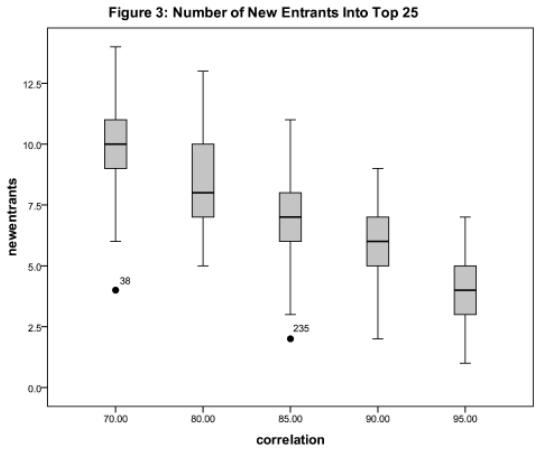
Monte Carlo Study

Figure 2: Difference in Rank Order

Group Means and Standard Deviations



Monte Carlo Study



Monte Carlo Study

- Conclusion
  - Applicants will not be satisfied with the consistency of grading, even with reliability of .9

Structured Rating Scales and Validity

- Identify superior and poor actions
- Count the number of each type of action
- Subtract one from the other
- Problem: Possible lower validity
  - One bad action can outweigh many good ones
  - Equality of weight is assumed
  - Intelligent integration is less reliable

Rater Reliability

- Opportunities
  - Improved validity due to better scoring

7. Test Fairness: Beyond Cleary Definition

- Cleary defines fairness in terms of correlation/regression
- Thorndike definition is more intuitive
  - Select applicants in each group in proportion to job success rates for that group
- The profession accepted Cleary 30 years ago; now some are reconsidering Thorndike

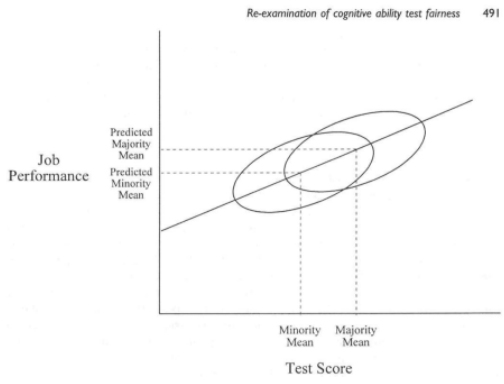
Problem

- **Qualified minorities are rejected at a higher rate than qualified non-minorities**
  - Cleary accepts this as inevitable
  - Thorndike focuses on this as unfairness
- This happens even when a test is fair according to Cleary definition

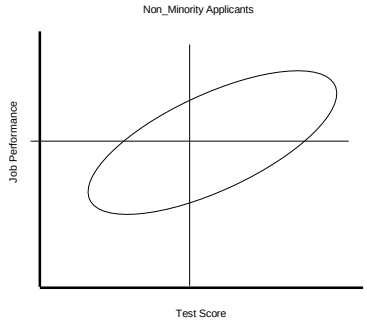
Higher False Rejection for Minorities

- Higher rejection rate for competent minorities
- Consider the next three plots
  - The first plot is from
    Chung-Yan & Cronshaw (2010).

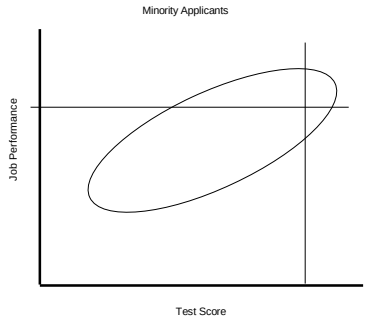
Test vs. Job Performance



False Rejection Rate: Majority



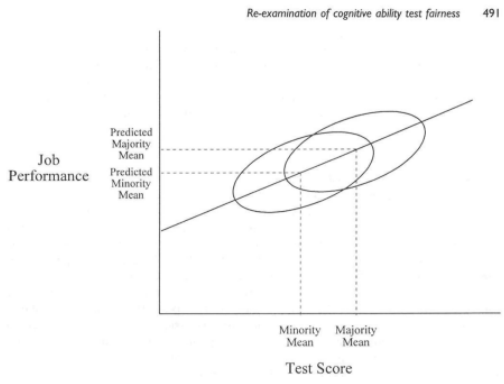
False Rejection Rate: Minority



Credits

- The next two graphics are from Chung-Yan & Cronshaw (2010).

Test vs. Job Performance Difference



Low Power Test Bias Research

- Issue:
  - Differential validity studies show NO test bias
  - Such studies have low power (Aguinis, Culpepper & Pierce, 2010)

Test vs. Job Performance Difference

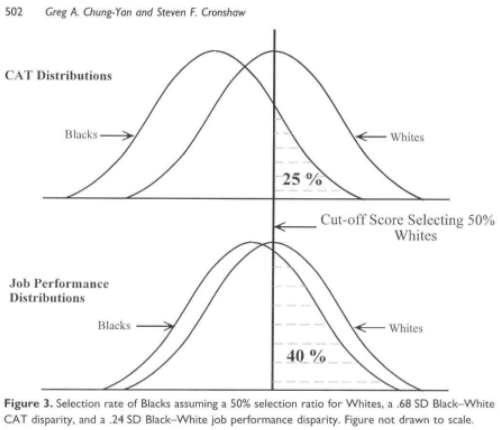


Figure 3. Selection rate of Blacks assuming a 50% selection ratio for Whites, a .68 SD Black–White CAT disparity, and a .24 SD Black–White job performance disparity. Figure not drawn to scale.

Interpretations

- Cleary
  - This is a fair test (equal slopes and intercepts)
  - Predictions of job performance unbiased
- Thorndike
  - This is an unfair test
  - Minority applicants would do better on the job than the test indicates

Some Implications

- Cleary
  - Regression gives best prediction
  - Highest utility for employer
- Thorndike
  - False rejection rates higher for minorities
  - Equally qualified majority and minority applications hired/promoted at unequal rates
  - Disproportionate burden of false rejections

Thorndike Easy to Implement

- Compare group differences on predictor and criterion
  - Difference between group means divided by pooled estimate of standard deviation
- Caution
  - Score variance or skewness can also contribute to unequal false rejection rates

Observations

- "Unqualified declarations that tests do not discriminate against minority groups can be misleading for laypeople."
  Chung-Yan & Cronshaw (2010)
- Only when we suspect unfairness in testing will we work to address the possible unfairness.

Recommendations

- Temper Cleary with Thorndike
- Await final decision from the Courts
  - Legal definition of fairness

8. Unknown Weights Within Tests

- Pros
  - Can make advances in scoring
  - May be able to lower adverse impact

Unknown Weights in Tests

- Issue
  - Test outlines specify number of items/area
    - Grade is the sum of the number of correct
  - The ability with the largest variance gets greater weight, despite the intended weight
  - We use z scores to combine CAT and PPT
    - Why not for math, reading, problem solving, etc.

Unknown Weights in Tests

- Opportunities
  - Passing score for each part of test
  - Could use GSM of Wiesen & Aguinis 2010
  - Could combine by z score for each area
  - Could combine based on raw score
  - Could combine based on score distributions
    - Extreme score anchors. (Livingston and Kim, 2008)
    - Beyond L&K, set anchors at 75th & 25th percentiles

9. Norm vs. Criterion Reference Tests

- Pros:
  - Opportunity to increase validity
  - Opportunity to live up to users' expectations

Norm vs. Criterion-Referenced Tests

- Issue:
  - Omitting items that do not discriminate may omit testing for some essential abilities
    - Our users expect the grades to indicate competence
  - Courts and civil service laws presume that passing points on our norm-referenced tests indicate competence

Norm vs. Criterion-Referenced Tests

- Opportunities
  - Passing score for each part of test
  - Could use GSM of Wiesen & Aguinis 2010

10. Within Person Variability

- Pros
  - Impetus to new testing research

Within Person Variability

- Issue
  - Current criterion-related validity research ignores within person variability
  - There is considerable within-person variability in job performance
  - Reliability of objective criteria is only .55 from week to week, explaining only 30% of variation in performance from week to week (Hunter, Schmidt & Judiesch, 1990, pg. 30 & Table 1)

Within Person Variability

- Opportunity
  - Improve validity studies
  - Higher validity
  - Explain more variance
  - Better job performance
    - Focus on consistency or highest performance depending on the job

Summary

- 1. Validity decreases over time
  - Suggests importance of training/experience
- 2. Criterion bias
  - Our claim of fair tests may be based on biased criteria.

Summary

- 3. VG
  - Illogical findings
  - Ignores deflationary corrections
  - Be careful consumers
  - Conduct more criterion studies
  - Correct down as well as up
  - Employees with different skill sets can do a job.
    - older air traffic controllers

Summary

- 4. Reliability
  - "Acceptable" levels of reliability can yield many and large changes in ranking on re-testing or re-scoring.
- 5. Job analysis for content validity
  - May not work
  - Consider advances in other branches of psychology

Summary

- 6. Rater reliability in grading: score changes
  - Rank order unstable, even with high reliability
- 7. Test fairness definition
  - We (Cleary) accept higher false rejection rate for qualified minorities
- 8. Unknown weights within tests
  - Consider various ways to combine scores

Summary

- 9. Norm vs. criterion-reference testing
  - We use norm referenced tests
  - Courts/statutes envision criterion-referenced
- 10. Within person variability
  - Ignored by current validity research

Other Suggestions

- Public organizations might consider
  - Advisory board of testing experts
  - Full-time research position in HR departments

Q & A's and Comments

- Questions/comments from the attendees

Call me any time to talk about this subject:
(617) 244-8859

Copies of this presentation are (or will be) available at:
http://ipac.org

References

- Aguinis, H., Culpepper, S.A. & Pierce, C.A. Revival of Test Bias Research in Preemployment Testing. *Journal of Applied Psychology, 95,* 648-680.
- Berry, C. M., Clark, M. A., & McClure, T. K. (2011, March 28). Racial/Ethnic Differences in the Criterion-Related Validity of Cognitive Ability Tests: A Qualitative and Quantitative Review. *Journal of Applied Psychology*. Advance online publication. doi: 10.1037/a0023222

References

- Chung-Yan, G.A. & Cronshaw, S.F. (2010). A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. Journal of Occupational and Organizational Psychology75, 489-509.
- DuBay, W.H. (2004). *The Principles of Readability*. Impact Information; Costa Mesa, CA. Retrieved January 25, 2008 from http://www.impact-information.com/impactinfo/readability02.pdf

References

- Farrell, J. N., & McDaniel, M. A. (2001). The stability of validity coefficients over time: Ackerman's (1988) model and the general aptitude battery. *Journal of Applied Psychology, 86*, 60–79.
- Judge, T.A. & Cable, D.M. (2004). The Effect of physical height on workplace success and income. *Journal of Applied Psychology, 89*, 428-441.

References

- Hunter, J.E., Schmidt, F.L. & Judiesch, M.K. (1990). Individual Differences in Output Variability as a Function of Job Complexity. *Journal of Applied Psychology, 75*, 28-42.
- Livingston, S.A. & Kim, S. (2008) *Small-Sample Equating by the Circle-Arc Method*. ETS, Princeton, NJ. (ETS RR-08-39)

References

- Marlowe, C. M., Schneider, S. L., & Nelson, C. E. (1996). Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased? *Journal of Applied Psychology, 81*, 11-21.
- Murphy, K.R. (1989) Is the Relationship Between Cognitive Ability and Job Performance Stable Over Time? *Human Performance, 2,* 183-200.

References

- Murphy, K.R. (2009) Content Validation is Useful for Many Things, but Validity Isn't One of Them, Industrial Organizational Psychology, 2, 453-464.
- Nunes, A. & Kramer, A.F. (2009) Experience-Based Mitigation of Age-Related Performance Declines: Evidence From Air Traffic Control. *Journal of Experimental Psychology: Applied, 15*, 12-24.

References

- Roth, P.L., BeVier, C.A., Bobko, P., Switzer III, F.A. & Tyler, P. (2001). Ethnic Group Differences in Cognitive Ability in Employment and Educational Settings: a Meta-analysis. *Personnel Psychology, 54,* 297-330.
- Schmidt, F.L., Hunter, J.E. & Outerbridge, A.N. (1986). Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance. *Journal of Applied Psychology, 71,* 432-439.

References

- (SIOP, 2011) An Interview With Frank L. Schmidt. *The Industrial-Organizational Psychologist,* April. (Downloaded 5/22/2011 from http://www.siop.org/tip/april11/04schmidt.aspx)
- Wiesen (2010) *A Hypothetical, Novel Employee Selection System to Reduce Adverse Impact and Improve Job Performance for Fire Lieutenant: Musings of a Practitioner.* The Industrial Organizational Psychologist, April.

References

- Wiesen, J.P. & Aguinis, H. (2010, April). New methods for reducing adverse impact and preserving validity. In H. Aguinis (Chair), *Solutions for solving the adverse impact-validity dilemma*. Symposium conducted at the meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.
- Wise, L.L, McHenry, J. & Campbell, J.P. (1990) Identifying Optimal Predictor Composites and Testing for Generalizability Across Jobs and Performance Factors. *Personnel Psychology, 43,* 355-366.