Avoiding Pitfalls: Identifying and Addressing Flawed, Implicit Assumptions in Testing

by
Joel P. Wiesen, Ph.D.
jw@jpwphd.com

2025 IPAC Annual Conference
Atlanta, GA, 7/30/2025

Who Am I?

- Independent consultant
- Worked 15 years for the MA civil service
- A scientist-practitioner
- 48 year member of IPAC
- 45 years as expert for Defense and Plaintiff
- 50 years working in the field of testing

Goals of this Presentation

- Highlight some dubious testing assumptions
- Challenge some conventional wisdom
- Suggest possible improvements
- Suggest applied research topics

Historic Example

- Early 1900's: some top psychologists erred
- Italians and Jews of low intelligence
- Led to restrictive Immigration Act of 1924
- Flaw: Administered IQ tests in English to relatively recent immigrants
- Moral: Kick the tires on accepted wisdom

Contemporary Example

- Old: AC much **less** valid than GMA
- Now: AC much **more** valid than GMA
- Sackett thought the old did not make sense
  - Short M/C test w more validity than longer AC
  - He looked at studies with both MC and AC
    - Studies on only MC had GMA oriented criteria

Source: Sackett, Shewach & Keiser, 2017

Choice of Topics

- Widely applicable
- Appear counter-intuitive
- Likely to be of practical importance

Overview of Each Topic

- Assumption/Presumption
- Discordant Logic or Data
- Discussion
- Research Needs
- Suggestions for Immediate Action

BARS Reliability Implies Validity

- Assumption
  High inter-rater reliability indicates the BARS ratings are accurate/valid
- Discordant data
  High within board rater reliability can coexist with low between board correlation despite thorough rater training

BARS Reliability Implies Validity

- Discordant data
  Two rater boards graded independently
  Job exercise presented via recorded video
  Raters graded recorded candidate responses
  No live grading (post COVID-19)
  Surprisingly low correlations between boards

BARS Reliability Implies Validity

Within and Between Board Correlation					
Year/ Exercise	Reliability Within Board 1	Reliability Within Board 2	Observed Correlation Between Boards	Expected* Correlation Between Boards	Percent (observed/ expected)
2021/1	0.86	0.90	0.48	0.77	62%
2021/2	0.96	0.96	0.71	0.92	77%
2023/1	0.94	0.94	0.64	0.88	72%
2023/2	0.94	0.94	0.70	0.88	79%

* The product of the reliability within Boards

BARS Reliability Implies Validity

- Discussion
  - Low inter-board correlation suggests neither board very valid
  - Sometimes high between board correlations
  - Perhaps 2 boards provide higher validity

BARS Reliability Implies Validity

- Research Needs
  - Do between board rating differences relate to the nature of the exercises, the areas rated, the anchors, or something else?
  - Are the board members able to articulate support for differences in grades?
  - Is there board agreement on invalid content

BARS Reliability Implies Validity

- Suggestions for Immediate Action
  - Use two boards routinely
  - Conduct applied research
  - See my Monday presentation on this topic on my website: http://jpwphd.com/ipac2025

A 100 Item Test is Long Enough

- Assumption
  - A 100-item test provides sufficient reliability
  - Such a test has sufficient content validity
- Discordant data
  - Test outlines with 1 or few items per KSAP
  - Reliability for one KSAP may be very low
  - PHR exam has 200 items and 57 Ks
  - 3.5 items per K on PHR exam; Reliability = .39

A 100 Item Test is Long Enough

- Discordant data
  - 100 item test with alpha = 0.95 implies average inter-item correlation of 0.16
  - Alpha for a KSAP w 4 items = 0.43
  - Alpha for a KSAP with 3 items = 0.36
  - Alpha for a KSAP with 2 items = 0.28

A 100 Item Test is Long Enough

- Spearman-Brown prophecy formula

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}}$$

r_{nn} = reliability for whole test of n items
r_{11} = average intercorrelation of items
n = number of items

A 100 Item Test is Long Enough

- Discussion
  - Dangers of measuring a domain w few items
    - More error in the scores
    - Little indication that the test taker is competent
- Research Needs
  - Calculate reliability for each KSAP

A 100 Item Test is Long Enough

- Suggestions for Immediate Action
  - At least 10 items per KSAP
    - Alpha = 0.66 if inter-item correlation = 0.16
  - Do not measure critical KSAP w 2 MC items
  - Use more than 100 items for M/C test
  - **Need reliable measure of each crucial KSAP**
  - Weight explicitly, not by number of items

One Overall Score is Sufficient

- Assumption/Presumption
  - One overall passing score shows competence
- Discordant Logic or Data
  - Can pass exam with a zero for some KSAPs
  - Low return on studying time for some KSAPs
    - Why study this textbook if only 1 item?

One Overall Score is Sufficient

- Discussion
  - Should these test-takers pass:
  - An EMT who aces anatomy but fails CPR
  - A police sergeant who knows no juvenile law
- Research Needs
  - Use past exams to determine extent of problem

One Overall Score is Sufficient

- Suggestions for Immediate Action
  - Modify job analysis to identify crucial KSAPs
  - Modify the job analysis to identify compensatory and non-compensatory abilities
  - Measure each KSAP reliably
  - Set a passing point for each crucial KSAP

Situational MC Items Mimic Job

- Assumption/Presumption
  - Situational items test K as used on job
- Discordant data
  - Consider this Sergeant item

Situational MC Items Mimic Job

- You respond as a sergeant to a traffic collision involving two vehicles ... One patrol officer is already on scene and directing traffic... As you assess the scene, you notice:

Situational MC Items Mimic Job

- Skid marks leading to one of the vehicles that have not yet been photographed.
- The second driver appears dazed and sits on the curb without being attended to.
- A civilian is recording the scene on a smartphone, standing close to the vehicles.
- The officer looks overwhelmed, shifting between multiple tasks.

Situational MC Items Mimic Job

- Discordant Logic or Data
  - On the job, Sgt must notice or seek out critical information from a rich and often confusing background
  - In M/C item the critical information is provided

Situational MC Items Mimic Job

- Discordant Logic or Data
  - On the job, responses must be generated, not just recognized
  - M/C items measure recognizing, not generating the correct responses

Situational MC Items Mimic Job

- Discussion
  - We use written M/C items because it is easy
  - It is becoming easier to create video items
  - GenAI can help create video items

Situational MC Items Mimic Job

- Research Needs
  - Develop procedure for developing video items with GenAI
  - Need method to evaluate quality of video items
  - Compare test scores on traditional M/C situational items and video items
  - Can AI help grade constructed responses to video items?

Situational MC Items Mimic Job

- Suggestions for Immediate Action
  - Create an IPAC award for students who develop video items with GenAI
  - GenAI workshop at next year's IPAC

Good Items Quote From Sources

- Assumption/Presumption
  - Will win appeals if use source wording
  - Source wording is ideal (e.g., clear, definitive)
- Discordant Logic or Data
  - Such items do not measure application of K
  - Such items do not measure understanding
  - Source wording often unclear out of context

Good Items Quote From Sources

- Discussion
  - Court cases lost based on rote memory items
  - But, how will candidates study for application?
- Research Needs
  - How best to wean candidates from quotes?
  - How to defend key when applying source to a new situation?
    - Will SMEs agree on key?

Good Items Quote From Sources

- Suggestions for Immediate Action
  - Include a certain % of understanding and application items on every promotional exam
  - Issue Rule that SME judgment will be acceptable defense of key
  - Perhaps require agreement of N SMEs
  - Training courses for promotion on application

More Recruiting Helps Diversity

- Assumption/Presumption
  - Additional recruitment of minority applicants means more minorities will be hired
  - Hiring reflects recruitment demographics

More Recruiting Helps Diversity

- Discordant data
  - Lower selection ratios (i.e., smaller percent of candidates hired) will yield more severe adverse impact.
  - Additional recruitment with higher proportion of non-minority candidates, will yield more severe adverse impact.

More Recruiting Helps Diversity

- Discussion
  - Smaller applicant pools will help bottom line
  - Recruit high quality minority candidates
  - Need higher proportion of minority applicants to improve bottom line
  - Effective recruitment is the most practical way to improve overall job performance

More Recruiting Helps Diversity

- Research Needs
  - Which recruitment sources yield the best candidates
- Suggestions for Immediate Action
  - Focus on quality of applicants, not quantity
  - Focus on ethnic mix of applicants
  - Effective recruitment is the easiest way to improve overall job performance

Top Ranks Will Do the Job Well

- Assumption/Presumption
  - Those at the top of the list will be well qualified
  - The test will identify the best candidates even if the candidate pool is of mixed abilities

Top Ranks Will Do the Job Well

- Discordant Logic or Data
  - Assume validity = 0.30
  - Assume 50% of applicants can do the job
  - Assume top 10% on exam are hired
  - Result: 71% of hires can do job
  - Result: 29% of hired cannot do job
  - If hire randomly, 50% of hired can do job

Top Candidates ⇨ Good Hires

- Discordant Logic or Data
  - Same assumptions
    - validity = 0.30
    - top 10% hired
  - Only 22% of hires are in the top 10% on job
  - 3% of hired are in the bottom 10% on job
  - Most of top 10% on job are not hired

Top Ranks Will Do the Job Well

- Discordant Logic or Data
  - Criterion validity often modest , esp. for police (Berry, et al., 2024; Wiesen, 2018)
  - Critical KSAPs ignored (e.g., face memory)
  - Fewer can do job under compensatory grading due to the lack of crucial KSAPs
  - Sometimes only one candidate passes, barely

Top Ranks Will Do the Job Well

- Discordant Logic or Data
  - Mean on police promotional exams in 80's
  - Select items to be really important
  - No training after promotion
  - May promote low score in a small department

Top Ranks Will Do the Job Well

- Discussion
  - Two possible approaches to improve hires
    1. More valid tests
    2. Recruit better candidates
  - Effective recruitment is the most practical way to improve overall job performance
  - Need to improve predictive validity

Top Ranks Will Do the Job Well

- Research Needs
  - Which recruitment sources yield best candidates
- Suggestions for Immediate Action
  - Focus on quality of applicants, not quantity
  - Need to improve predictive validity
  - Promotional training for police

Top Ranks Will Do the Job Well

- Suggestions for Immediate Action
  - Evaluate quality of candidates by recruitment source
  - Conduct criterion related validity studies whenever possible
    - Whenever hire 300 over a several year period

Our Tests Are Fair

- Assumption/Presumption
  - Our tests are fair, perhaps overestimating job performance of minority test takers just a little
  - Based on the Clearly definition of test fairness

  1.Are slope and intercept the same for all groups?

  2. Is regression line the same for all groups?
  - Finding: Prediction is equally accurate for candidates in any group

Our Tests Are Fair

- Discordant data
  - The Cleary definition assumes a fair criterion
  - But men paid more than women
  - But tall paid more than short
  - But attractive paid more than homely
  - Accurately predicting an unfair criterion is not proof of fairness

Our Tests Are Fair

- Discordant data
  - Cleary method has low power for slope
  - We may conclude there is no differential validity when it exists
  - Some research found differential validity (Aguinis et al., 2016).

Our Tests Are Fair

- Discordant data
  - The Thorndike definition is also logical (Thorndike, 1971)
  - Cleary may be asking the wrong question
  - Thorndike asks: Is the selection rate equal among equally qualified candidates no matter what group membership?

Our Tests Are Fair

- Discordant data
  - Among qualified candidates, the selection rate will be smaller for the lower scoring group (AERA et al, 1999, page 79).
  - The B-W difference in mean test score is about twice as large as the difference in job performance. This seems unfair on its face.

Our Tests Are Fair

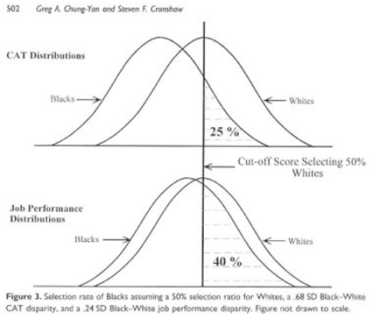

Figure 3. Selection rate of Blacks assuming a 50% selection ratio for Whites, a .68 SD Black–White CAT disparity, and a .24 SD Black–White job performance disparity. Figure not drawn to scale.

Our Tests Are Fair

- Discussion
  - The Cleary definition favors the testing profession and was accepted by the profession
  - Cleary research ignores criterion contamination
  - Thorndike definition is logical
  - Which approach would a court choose?
  - Saying tests are fair in general is simplistic or wrong

Our Tests Are Fair

- Research Needs
  - Renewed focus on definitions of test fairness
  - Ways to implement the Thorndike definition
  - Is it possible to meet both fairness definitions

Our Tests Are Fair

- Suggestions for Immediate Action
  - Use more than one definition of fairness.
  - Calculate power when testing differences in slopes and intercept
  - Focus on utility over validity

Incumbents Are the Best SMEs

- Assumption
  - Incumbents can identify KSAPs needed
- Discordant Logic or Data

  SMEs lack knowledge of KSAPs, such as:
  - Published academic literature on KSAPs
  - Fleishman areas definitions
  - Personality factors, Values

Incumbents Are the Best SMEs

- Discordant data for entry-level
  - Few PO exams test face memory/recognition
  - No recognition of faces in new settings
  - Creativity not measured

Incumbents Are the Best SMEs

- Discordant data for entry-level
  - More competent SMEs rate KSAPs higher
    - Self-serving bias
    - Inexperienced SMEs rate KSAPs lower

  (Cucina, et al. 2012)

Incumbents Are the Best SMEs

- Discordant data for promotion
  - Diverse KSAP lists from different PDs
  - KSAPs in literature do not appear in SME lists
  - Executive function omitted
  - Big Five areas omitted
  - Subsets of the Big Five omitted
  - Low *d* KSAPs omitted
    - Creativity

Incumbents Are the Best SMEs

- Sackett's review of 100 years of research
- Integrity
- Achievement motivation
- Creativity
- Vocational interest

(Sackett, et al., 2017)

Incumbents Are the Best SMEs

- We ignore work in other fields of psych.
- Attention: focus, selective attention
- Simultaneous processing: seeing relationships between stimuli
- Successive processing: use of sequential information

(Agnello, Ryan & Yusko, 2015)

Incumbents Are Best SMEs

- Discussion
  - Unknown origin of use of incumbents for KSAPs
    - Not in landmark Supreme Court decisions
  - Use **incumbents for tasks**
  - Use **supervisors for KSAPs**
    - Goldstein, Zedeck & Schneider (1993, (page 26)
  - Use trainers for KSAPs

Incumbents Are Best SMEs

- Research Needs
  - Compile list of KSAPs from literature
  - Go beyond I/O literature (e.g., cognitive psych)
  - Compare incumbents, supervisors, trainers
  - Which KSAPs can be measured with M/C, oral board, etc.

Incumbents Are Best SMEs

- Suggestions for Immediate Action
  - IPAC develop a KSAP bank with defs. and citations
  - Use trainers as SMEs to develop, rate KSAPs
  - Long tenure supervisors to develop, rate KSAPs
  - Clarify lack of SME agreement in ratings
  - Test developers augment SME input
  - Share KSAPs across jurisdictions

d's Average

- Assumption/Presumption
  - A composite of d=1 and d=0 will have d=0.5

d Defined

- *d* = standardized mean score difference or standardized difference between means

$$d = \frac{mean\ gp\ 1 - mean\ gp\ 2}{pooled\ standard\ deviation}$$

d Defined

Predictor	Black-White d
Structured interviews	0.23
General Mental Ability Tests	0.79
Conscientiousness	-0.07
Sackett et al. (2022)	

d Defined

- Pooled standard deviation

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

d's Average

- Discordant data
  - A composite of d=1 and d=0 will have d=0.71
    (Sackett & Ellingson, 1997, Table 3)
  - Even *d* = 0.2 can cause adverse impact
    (Sackett & Ellingson, 1997, Table 2)

d's Average

- Discussion
  - Mathematics of adverse impact subtle and not always intuitive
  - Need to predict level of *d* numerically
- Research Needs
  - Online *d* calculator to assist in predictions
  - Clarification of the effect of P/F on utility
    (P/F = pass fail)

Utility

- Utility = benefit of testing minus test cost
- Not defined in *Principles* or *Joint Standards*
- Level of job performance central to utility
- Utility a function of
  - Validity
  - Selection ratio (SR)
  - Quality of candidates
    (Cascio & Aguinis, 2011, pg 328)

d's Average

- Suggestions for Immediate Action
  - Calculate *d* for each test and test area
  - Calculate utility for test
  - Consider using a pass/fail approach for measures with large *d*

Global Suggestions

- Establish an advisory board of testing experts for every testing agency
- Fund a testing research position
- IPAC lobby congress for federal funding of local employment test research (such as was available some decades ago)
- IPAC issue compendia of BARS scales, test outlines, etc.

IPAC Member Forum

- Place to discuss discretely
  - Open to members only
- Access it on **ipacweb.org**
  - click on ***three dots*** at right of menu bar
  - select ***Members Only***
  - select ***IPAC Member Forum***

In Closing

- The field is still young
- Many improvements possible
- Many topics need more research

Presentation Citation

Wiesen, J. P. (2025). Avoiding Pitfalls: Identifying and Addressing Flawed, Implicit Assumptions in Testing. *International Personnel Assessment Council*, Atlanta, GA, United States.

Contact Information

- Session URL: **http://jpwphd.com/ipac2025**
- Email: **jw@jpwphd.com**
- Telephone: **(617) 244-8859 (land/no text)**
- **Email and telephone calls welcome!**
- **Q&A's**

Q&A

Learning Objectives

- Describe two major errors in accepted psychometric thought and explain why the erroneous conclusions were accepted and why they are erroneous.
- Describe a major potential content validity weakness that can result from relying on a published reading list to develop test questions for a promotional examination.

Learning Objectives

- Identify the origin of relying on incumbents (AKA SMEs) to identify the knowledges, skills, abilities, and personal characteristics to be included in an employee selection test and the major weakness of such reliance.

References

- Aguinis, H., Culpepper, S.A., & Pierce, C.A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108*, 1045-1059.

References

- Agnello, P., Ryan, R. & Yusko, K. P. (2015). Implications of modern intelligence research for assessing intelligence in the workplace. *Human Resource Management Review*, *25*, 47–55.

References

- American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

References

- Berry, C. M., Lievens, F., Zhang, C., & Sackett, P. R. (2024). Insights from an updated personnel selection meta-analytic matrix: Revisiting general mental ability tests' role in the validity-diversity trade-off. *Journal of Applied Psychology, 109*, 1611-1634.

References

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

References

- Chung-Yan & Cronshaw (2002). A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. *Journal of Occupational and Organizational Psychology*, 75, 489-509.

References

- Cucina, J. M., Martin, N. R., Vasilopoulos, N. L. & Thibodeuax, H. F. (2012). Self-Serving Bias Effects on Job Analysis Ratings. *The Journal of Psychology, 146*, 511–531.

References

- Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 3-34). Jossey-Bass.

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., Lievens, F., Van Iddekinge, C. H., & Kuncel, N. R. (2017). Individual Differences and Their Measurement: A Review of 100 Years of Research. *Journal of Applied Psychology, 102*, 254-273.

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

- Sackett, P. R., Zhang, C., Berry, C. M. & Lievens, F. (2022). Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range. *Journal of Applied Psychology*, *107*, 2040–2068.

References

- Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8*, 63-70.

References

- Wiesen, J. P. (2018). *Tools to increase diversity, utility, and validity in hiring police officers*. Paper presented at the 33rd Annual Society for Industrial and Organizational Psychology (SIOP) Conference.