Curiosities, Pitfalls, and Practical Tips Learned During 45 Years of Work in Testing

Joel P. Wiesen, Ph.D.

Contact: **jw@jpwphd.com, (617) 244-8859**

2022 Annual IPAC Conference, 7/20/2022

Print and Audio Links

- PowerPoints (yet to be posted)
- Audio recording (yet to be posted)
- **http://jpwphd.com/ipac2022**

Questions

- Short questions only during talk
- Will try to address all questions at the end
  - Much material to cover

Outline of This Presentation

- Historical Tidbits
- Statistics Oddities and Pitfalls
- Validity Insights
- Test Development
- The Unexpected
- Assumptions Revisited

Learning Objective 1

- Explain the practical difference between designing a selection system based on test validity or utility.

Learning Objective 2

- Describe a major shortcoming of using z-score equating of structured oral board panels.

Learning Objective 3

- Describe three major weaknesses of using multiple-choice tests for selecting employees.

Historical Tidbits

- We have been wrong before
  - Humility
- Uniform Guidelines
  - Why “Uniform”
  - Definitions of validity
- Job Related and a Business Necessity

We Have Been Wrong Before

- 1910: Asians, Jews are of low intelligence
  - Immigration laws to restrict entry to USA
- 1965: Personality measures are not valid
- 1975: Validity is situation specific
- 1980's: Interviews have no validity
- 1998: *g* is much more valid than Assessment Centers

UGESP

- The **Uniform** Guidelines on Employee Selection Procedures
  - Issued in 1978 pursuant to 1964 CRA
  - Why the word "Uniform"

Why "Uniform" Guidelines

- 1966: EEOC Testing Guidelines
- 1968: OFCCP Testing Order
- 1969: CSC Evaluation of Employees for Promotion and Internal Placement
- 1970: EEOC Guidelines on Employee Selection Procedures
- 1971: OFCCP Revised Testing Order

Why "Uniform" Guidelines

- 1972: CSC Qualifications Standards
- 1972: CSC Applicant Appraisal Procedures
- 1972: CSC Examining Practices
- 1974: OFCCP Amended Testing Order (Documentation)
- 1978: Uniform Guidelines on Employee Selection Procedures

How Title VII Defines Validity

- The definition of validity changed over time
- Three enforcement agencies' views
- CSC (now the US OPM)
  - Office of Personnel Management
- DOL, OFCCP
- EEOC

Title VII Definition of Validity

- EEOC & OFCC: criterion-related validation
  - Objective
  - Wide professional acceptance
    - Content validity debated in the professional literature
- CSC: content validity
  - Developed many exams
  - Many CSC exams based on "rational validation"

Definition of Validity

- 1974 Joint Standards
  "…only rarely is one [criterion, content, construct] alone important in a particular situation."
  "…the content universe includes **all, or nearly all,** important parts of the job."
  - UGESP has similar words for content validity

Definition of Validity

- 2014 Joint Standards:
  "The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test."

Job Related, Business Necessity

- Where did these two terms originate?
- What do these two terms mean?

Job Related, Business Necessity

- Title VII of the 1964 CRA "…demonstrate that the challenged practice is **job related** for the position in question and consistent with **business necessity**"
- EEOC 1970
  - validation must have a "high degree of utility"
  - Perhaps this implements "business necessity"
  - Job related = valid; business necessity = utility

Job Related, Business Necessity

- Another possible view
- Business necessity = valid
- Job related = test resembles job

Statistics, Oddities, and Pitfalls

- z-Score Equating: Oral Boards
- z-Score Equating: Test Components
- Why is rater reliability important?
- Intra and inter-board rater reliability
- Job performance *d* is greater than test *d*
- Adverse impact ratio is a poor measure
- Models predict many will fail on the job

z-Score Equating of Oral Boards

- Need: equate boards for obvious differences in leniency or variance
- Pitfall: superstars no longer shine bright
- Explanation: A really great candidate who is several standard deviations above the others will be scaled back to only 1 or 2 SD
- Solution: Train raters to criterion (Joint Standards, 2014, 4.20, 4.21, 3.8)

z-Score Equating Components

- Can lose the information you may have if any component has anchors related to competence
- Can magnify small differences if takers are similar in ability on one component

Why is Reliability Important?

- Goal is to predict job performance
- Validity is the index we use
- Validity and reliability are related
- Reliability puts a cap on validity
  $r \leq$ SQRT (reliability)

Validity Limited by Reliability

Reliability	SQRT reliability
0.9	0.95
0.8	0.89
0.5	0.71
0.4	0.63
0.2	0.45
0.1	0.32

Intra and Inter-Board Reliability

- Often we have only one board and we calculate inter-rater reliability
  - Estimate reliability based on the one board
  - Estimate the reliability of the board grades
- If we have two boards
  - Estimate reliability based on the two boards
  - Correlate the grades given by the two boards

Intra and Inter Board Reliability

Estimated Board Reliability

	Within Board A	Within Board B	Across Boards A&B
Exam 1	0.97	0.97	0.44
Exam 2	0.97	0.97	0.66
Exam 3	0.88	0.8	0.76
Exam 4	0.84	0.88	0.66

Job Performance *d* < Test *d*

- M-W *d* about .5 for job performance
- M-W *d* about 1.0 for test performance
- Some see this as showing tests are unfair

Job Performance *d* > Test *d*

- Y = r * X
- .5 = .5 * 1
- But the validity of our tests is less than .5
- .4 * 1 = .4
- Why is the difference in job performance larger than predicted?
- Biased measures of job performance?

Job Performance *d* > Test *d*

- But we selected based on a test of *g*
- So the unrestricted *d* must be greater than 1
  - Assume *d* = 2
- Y = r * X
- .5 = .5 (2) - This is clearly wrong
- Why is the difference in job performance larger than predicted?

Adverse Impact a Poor Measure

- Should focus on ***d***, not Adverse Impact
  - *d* = Standardized M-W mean score difference
- Adverse impact (AI) ratio bounces around
  - Influenced by exact number hired, small Ns
- *d* is a more stable measure than the AI ratio
  - Independent of who is selected

Pass-Fail Adverse Impact Unreal

- After a test is given, we can calculate the pass-fail adverse impact (AI)
- But promotion AI impact will be much worse
- AI is more severe with fewer selected
  - Smaller selection ratio
  - Higher cut score

Artificial Way to Avoid AI

- Give a very easy exam
- All tied at the top
- Few test takers complain about high scores or easy exams

Validity Insights

- Validity may not yield high job performance
- Utility - validity relationship
- Public and private sector views of validity
- Many hires fail on the job
- Few high scores on promotional exams
- Employees fail exams for their jobs

Validity - Job Performance

- Test users often assume that high validity and many applicants result in high job performance.
  - **Often this is not so!**
- Utility tells us about job performance
- Validity is only one factor of utility
  - Two other factors drive utility as well

Utility

- SIOP:
  "Projected productivity gains or utility estimates for each employee and the organization due to use of the selection procedure" (SIOP, 2018, 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
  - Best way to improve expected job performance
- Selection ratio (SR)
  - Ratio of openings to applicants
  - Improving SR will worsen adverse impact
- Validity (r)
  - Very difficult to improve validity

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 do not depend on r
- Must pay attention to recruitment
- Cannot recruit more after we see test scores

Recruitment Most Important

- We focus on validity and ignore recruitment
- Validity ceiling is low and it's impact on utility is limited by applicant quality
- Utility should be our focus
- Solution: Get involved in recruitment

Public ≠ Private Sector Validity

- Validity is evaluated for the test's purpose
- Purpose of testing differs by sector
- Private sector goal for testing:
  Improve employee productivity
- Private sector goals for testing:
  Identify test takers who can do the job
  Identify test takers who can do the job best

Many Job Failures

- False positive
  - Hire a person who fails on the job
- Models predict 40-60% of new hires will fail on the job (as police officer)

Expect Many to Fail on the Job

- Due to low r and an applicant group with varying levels of the important KSAPs
- Low r test not good at choosing best
- Solution: Better applicants
  - Better tests, if possible

Low Scores on Promotional Exams

- Often, highest promotional test score is in 80's
- Items chosen to be important, even crucial
  - Miss 10+ crucial items
- Possible explanations
  - No training for new job (esp. promotions)
  - Exams not related to (most) job duties
- Implication: high false positive rate

Employees Fail Xm for Own Jobs

- Employees fail exams for their jobs
- Possible explanations
  - Hired by chance (guessing): false positives
  - Studied and forgot
  - Exams not related to (most) job duties

Test Development

- Definitions of a good item
- Item protests subvert item quality
- Knowledge of law items
- Definition items
- Creativity
- Role of official job specification
- Job analysis results can be unbelievable

Definitions of a Good Item

- No protests
  - Straight from textbook
- Promote good sergeants
  - Extrapolate from textbook
  - Apply knowledge to new situations

Item Protests Subvert Item Quality

- Lay body evaluates item protests
- Easiest way to defend an item is to show it is taken directly from a source document
- Reading lists also subvert item quality

Item Protests Subvert Item Quality

- Items with verbatim quotes from sources measure recognition of wording not application of knowledge
- Such items don't measure application of K
- Better: Use more job simulation questions
  - Rely on SMEs to extrapolate from textbook

Knowledge of Law Items

- Law items often are an exact replication of a case
- No deviation from the court case because no one knows what a court may rule if the facts were somewhat different
- But this omits exactly what a PO or Sgt needs to do to perform the job: apply the law to new situations.

Use of Definition Items

- Definition items are easy to write
- K of definition is only weakly related to application of knowledge
- Avoid definition items, in general
- Use more job simulation questions

MC Tests Do Not Test Creativity

- Creative problem solving important
- M/C tests test recognition of solution
- M/C does not test for thinking of a solution
- Solution: More test modes

No Respect for Job Descriptions

- We emphasize job analysis
- Official job specifications are given deference by courts

Unbelievable Job Analysis Results

- SMEs disagree, sometimes wildly
- Illogical ratings of both tasks and KSAPs
- Requires oral communication (Sergeant):
  - Enters data into and accesses data from computer system
  - Reviews forms are all necessary completed
- Fleishman areas misunderstood by SMEs

Unbelievable Job Analysis Result

- Tasks done daily by a Sergeant:
  - Informs other units of homicide
  - Recommends subordinates for commendation and disciplines them for dereliction of duty
  - Conducts internal investigations
  - Supervise bomb threats

Unbelievable Job Analysis Result

- Can we base our tests on unbelievable job analysis "findings"?

Assumptions Revisited

- Content validity assumptions
- Job performance is stable
- More recruitment can cause Worse AI
  - Fewer minority hires
- Compensatory grading is illogical
- 100 items is long enough
- Are claims of fairness realistic?

Content Validity Assumption

- Content validity ratings may ignore the relationship between validity and reliability
  - SMEs assume we have reliable measures of the KSAPs they rate
  - Lower test reliability yields lower test validity
    $r \leq \text{SQRT (reliability)}$

Content Validity Assumption

- Content validation assumes we can clearly define the job and test content
  - Too often we use brief definitions of test areas
  - It is difficult to specify job content based on tasks, KSAPs, and reading list material
    - Three-way matrix?

Content Validity Assumption

- Content validation is strongest when linkages are made to tasks and KSAPs
- But there may be 100+ of each
- Often we use task and KSAP categories or groupings
  - These higher-level constructs can lose their claim to content validity due to their amorphous nature

Variability in Job Performance

- There is larger within person variance in job performance than between person
- Perhaps we need to re-envision validation research
- Validity correlation assumes that job performance is a constant for a given person

More Recruitment Can Cause Worse Adverse Impact

- Problem: Recruit many and choose the best
- Pitfall: Selection ratio drives adverse impact
- Solution: Recruit better not more applicants; recruit relatively more minority members

More Recruitment ⇒ Fewer Minority Hires

- If recruit more applicants in same proportion of minority/non-minority
  - More severe adverse impact
  - Hire fewer minority test takers

Compensatory Grading Illogical

- Grade is based on # correct
- Tests cover many unrelated topics
- Can hire someone w gaps in KSAPs
- Consider multiple passing points for crucial KSAPs

Our Tests Are Too Short

- Test outline topics with only 1 or 2 items
- Few items ⇒ unreliable measure
- Unreliable measure ⇒ invalidity
- Solution: longer tests

Unrealistic Claims of Fairness

- We claim our tests are fair despite evidence that job criteria are biased
  - Women paid less than men for same work
  - Short people paid less than tall
  - Homely people paid less than handsome
- Perhaps our tests are unbiased predictors of biased criteria, thus not really “fair”

Closing

- Secrecy is harming our field
- Learning objective answers

Secrecy Slows Advancement

- Consultants refine their products
- Best work is not shared
  - BARS
  - M/C items
  - Work sample items
- Field advances slowly without sharing
- IPAC goal is to share work

Learning Objective 1 w/ Answer

- Explain the practical difference between designing a selection system based on test validity or utility.
- When selecting a test based on utility, the test chosen may not have the highest validity

Learning Objective 2 w/ Answer

- Describe a major shortcoming of using z-score equating of structured oral board panels.
- The grades for true superstars will be lowered

Learning Objective 3 w/ Answers

- Describe three major weaknesses of using multiple-choice tests for selecting employees.
- Have avoidable adverse impact
- Have low validity
- Do not test for application of knowledge
- Do not test for creative problem solving

Bonus Slides

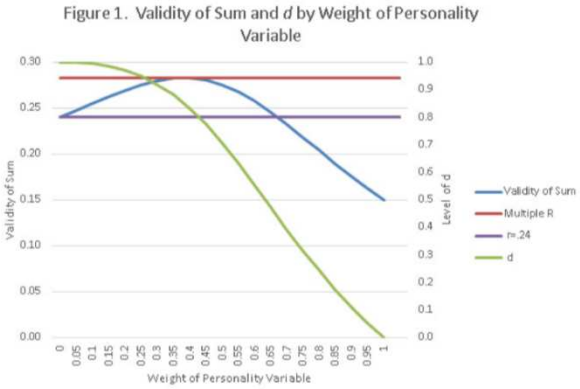
- If time

Does a Personality Test Dilute *g*?

- Will a personality decrease the r due to *g*?
- Assume r = .25 for *g*
- Assume r = .15 for personality
- Assume *d*s of 1 and zero, respectively

Adverse Impact of a Composite

- Assume a simple weighted sum
- Focus here on *d* since it a better measure than Adverse Impact
- When form a composite, what happens to:
  - r
  - *d*



Q&As

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

References

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