

Validation, Meta-analyses, and the Scientific Status of Selection

Neal Schmitt

Invited Address

International Testing Council

July 2010

Outline

- Discuss the nature of validity evidence
- Review the bases for validity claims in employee selection research
- Discuss limitations of data base on which these claims are made
- Propose a large scale multi-national collection and sharing of data on individual difference-performance relationships

Validity and Validation

- Validity is the degree to which inferences we draw from test data about future job performance are accurate.
- Standards and Principles both emphasize construct validity
 - Content
 - Criterion-related
 - Construct
 - Consequential

Construct Validity Includes

- Concerns about content
- Concerns about construct
 - Relationships with measures of theoretically similar and dissimilar constructs
 - Freedom from the usual “biasing” factors: e.g., social desirability, faking
 - Process variables

Construct Validity involves several inferences (Binning & Barrett, 1989)

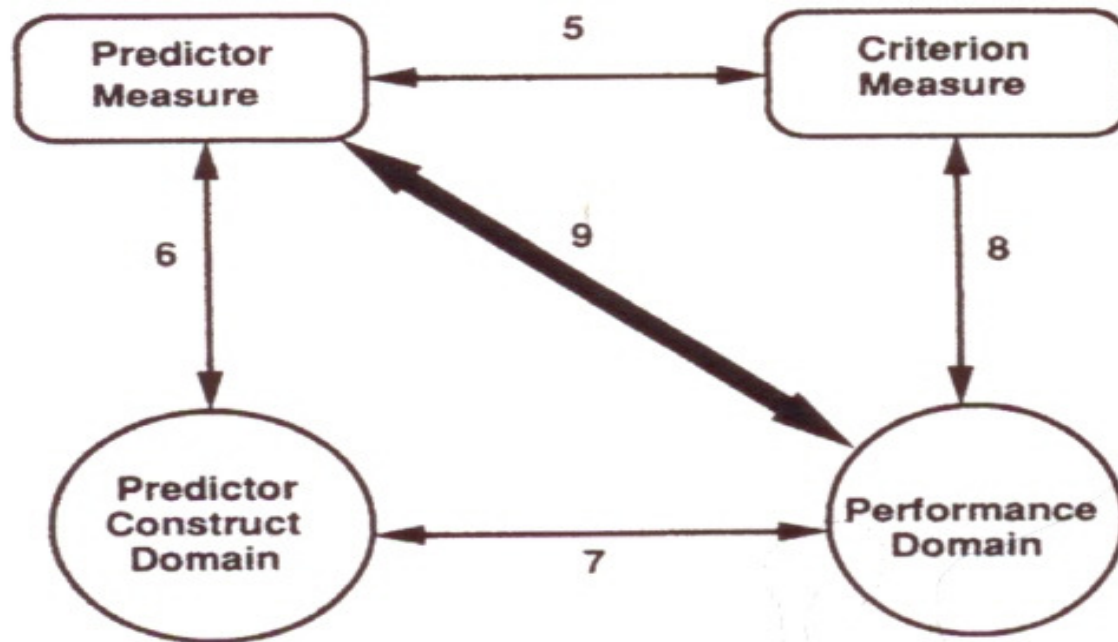


Figure 2. A common conception of the inferences for personnel selection.

Binning and Barrett Linkages

- Predictor measure is an accurate depiction of the intended predictor construct
- Criterion measure is an accurate depiction of the intended criterion construct
- Predictor measure is related to the criterion measure—uncorrected validity coefficient in the usual criterion-related study
- Predictor measure is related to the performance construct – validity coefficient corrected for unreliability in the performance measure though this may be an imperfect estimate for a number of other reasons
- Predictor construct is related to the criterion construct; the primary scientific question

Examples of construct validation

- Mumford et al (PPsych, 1996): Biodata
 - Job-oriented variables (biodata items related to actual job tasks)
 - Worker-oriented variables (biodata items related to constructs thought to underlie job performance)
 - Biodata scales were developed to predict performance in laboratory tasks
 - Biodata scales were developed to predict performance of foreign service officers

Physical Ability

- Arvey et al. (JAP, 1992) and Arnold et al. (JAP, 1980)
 - Job analyses (use of force reports, SME descriptions, survey)
 - Literature review led to hypothesized importance of endurance and strength
 - Collection of data on physical tasks (some face-valid and some not—handgrip and situps)
 - Collection of supervisory ratings on many aspects of physical performance
 - Tested and confirmed a model that hypothesized the same set of endurance and strength factors plus a rating factor – good fitting model

Data base on validity of existing measures is based primarily on criterion-related validity

- Job analysis
- Specification of the performance domain
- Hypotheses about the relevant predictor constructs (KASOs)
- Selection or construction of predictor variables
- Collection of data on predictor and criterion (concurrent and predictive validity)
- Analysis of relationships between predictor and criterion measures
- Conclusions and implementation

Meta-analysis: Summary of Criterion-related research and basis for Validity Generalization Claims

- Prior to the use of VG work, there was a general belief that validity of tests was unique in each situation in which a test was used: Situational Specificity
- Schmidt & Hunter (1977) showed that much of the variability in observed validity coefficients could be explained by artifacts of the study in which validity was estimated
- S & H computed the sample-size weighted averages of existing validity estimates for cognitive ability-performance relationship
- Bare-bones analysis corrected for differences in sample size only and found much of the variability in observed validity coefficients was explained

Validity Generalization

- In addition, computations of credibility intervals around the averages of validity coefficients revealed that for some KASO-performance relationships, one could expect to find non-zero relationships most of the time. That is, if we are examining a situation in which the performance construct is similar to that in the validity data base, we should be able to use the predictor (or a measure of the same construct) with confidence that the validity will generalize.

Results of Meta-analyses: Cognitive Ability

- Hunter and Schmidt efforts were largely restricted to reanalysis of GATB data collected by the US Department of Labor
- Hunter (1983)
- Hunter & Schmidt (1977) used data from Ghiselli (1966)
- Hunter & Hunter (1984)
- Schmidt & Hunter (1998)

Other Sources of Information about Cognitive Ability

- Schmitt et al. (1984) used all published validity studies between 1964 and 1982
- Salgado and colleagues (Salgado et al., 2003; Bertua, Anderson, & Salgado, 2005) provided European data

Results

- Average observed validity coefficients are almost all in the .20s. Most estimates of ρ are in the .40s and .50s.
- The lower bound of the 90% credibility interval is always substantial, from .10 to the .50s meaning it is unlikely that one would find a nonsignificant validity coefficient when using cognitive ability tests to select employees.
- Very similar results are reported for specific cognitive abilities such as memory, numerical ability, perceptual ability, psychomotor ability, spatial ability

Personality Sources

Performance Outcomes

- Barrick and Mount (1991): Review of published and unpublished sources from the 1950s through the 1980s
- Tett, Jackson, & Rothstein (1991): Reanalysis of existing data exploring the role of hypothesis testing and directionality as moderators of validity
- Hertz and Donovan (2000): Published and unpublished research on Big Five only
- Barrick, Mount, & Judge (2001) summarized 15 prior meta-analytic studies
- Salgado (1997) European community

Results: Performance

- Observed validity: .04 to .22 across the Big Five
- Corrected validity: .04 to .33 across the Big Five (highest validity in Tett et al(1991))
- 90% credibility interval did not include zero for Conscientiousness and occasionally Emotional Stability

Results: Other Criteria

- Most frequent alternate criterion is OCB
- Much smaller sample sizes (usually less than 2000)
- Observed validity ranges from .04 to .24
- Corrected validity ranges from .04 to .27
- Other personality traits including customer service orientation, core self-esteem, self efficacy, integrity tests: validities ranged from .14 to .50.

Summary

- We have an extensive data base that supports the use of tests of major constructs, certainly cognitive ability and conscientiousness across most occupations
- Validities are such that significant utility can result from the use of measures of these constructs
- The validity of these constructs is generalizable

Problems are with the nature of the primary data base: Consider Landy comments (2007)

- Without empirical studies, VG dies of its own weight OR meta-analysis can make a silk purse out of a sow's ear, but you need the ears to begin with—so it is wrong to abandon criterion-related validity studies
- We are burdened with the empirical studies of the 50's, 60's and earlier that viewed the world through g and O (overall performance lenses)

First Recorded US Army Efficiency Report

- Lt. Col. Alex Denniston A good natured man
- Major Crolines A good man, but no officer
- Capts. Martel, Crane, Wood All good officers
- Capt. Shotwell A knave despised by all
- Capt. Reynolds Imprudent and of most violent passions
- Capt. Porter Stranger, little known
- Lt. Kerr Merely good, nothing promising
- Lts. Perrin, Scott, Ryan, Elworth Low vulgar men, Irish and from the meanest walks of life
- Ensign Mehan Very dregs of the earth, unfit for anything under heaven.

What's wrong with the empirical studies that underlie our meta-analyses

- Small sample sizes or nonexistent sample sizes (Schmidt & Hunter, 1977)
- Lack of information on the two major artifacts necessitating use of hypothetical distributions: Range restriction and criterion unreliability
- Mostly concurrent criterion-related studies with various design flaws: Sussman & Robertson (JAP, 1989)

Flaws continued

- Lack of sample descriptions limiting the use of the data base in studying the role of demographic characteristics
- Lack of information on organizational characteristics: Concern about organizational characteristics does not imply a return to situational specificity notions
- Conceptualization of performance: Use of single item overall performance index
- Age of the data base: At least in the cognitive area, we have few studies less than 30-40 years old and they range from 30-60 or more years

Are the studies “old”: Cognitive ability example

- <1930 18
 - 1931-1950 42
 - 1951-1960 39
 - 1961-1970 42
 - 1971-1980 30
 - 1981-1990 34
 - >1991 32
-
- Bertua et al
 - Salgado et al
 - Vinchur et al (clerical performance)
 - Schmitt et al. (1964-1982)
 - Schmidt & Hunter (1977) < 1966
 - Hunter (GATB) < 1980 most likely

Why is age a problem?

- Changes in work have implications
- Teamwork – we must measure team performance on the criterion end and attend to interpersonal KSAOs
- Virtual work— we must attend to the context in which work takes place in measuring KASOs
- Contingent and temporary work—is there need for a different system of selection than that used for regular workers
- Cognitive requirements of jobs have likely both increased and decreased
- Technical and nontechnical work
- Constantly changing work
- Global work

Single most important validation study has been Project A

- Goal was to generate criterion variables, predictor measures, analytic methods, and validation data that would result in an enhanced selection and classification system for the military services in the U.S.
- 276 entry level occupations
- In 1990, 800,000 person force
- Screened 300,000 to 400,000 annually to select 120,000 to 140,000
- ASVAB consisting of 10 subtests and a number of composites was, and is, the primary selection tool

Performance Domain (performance is behavior,
not the result of behavior): Binning & Barrett -
Linkage 8

- Task analyses and critical incident analyses
- Generating the critical performance dimensions
- Constructing measures of each dimension
 - Rating scales
 - Job knowledge tests
 - Hands-on job samples
 - Archival records

Data Analyses

- Content analyses of performance measures
- Principal components analyses
- Development of a target model of performance for all jobs
- Confirmatory factor analyses
 - Core technical proficiency
 - General soldiering proficiency
 - Effort and leadership
 - Personal discipline
 - Physical fitness and military bearing

Predictor Domain (Binning & Barrett, Linkage 6)

- Extensive literature review (non-cognitive, cognitive, and psychomotor areas)
- Expert judges in the field were consulted as to predictor-criterion relationships for a variety of criterion content categories
- These judgments were factor analyzed to come up with a preliminary list of predictors thought to be relevant to various performance domains
- Practical time constraints were considered in arriving at a list of paper-and-pencil cognitive ability tests, computer-administered psychomotor tests, and paper-and-pencil noncognitive measures

Validity Estimates

- Both concurrent and predictive samples
- Observed validity (Binning & Barrett, Linkage 5)
- Corrected validity for restriction of range and criterion unreliability (Binning & Barrett, Linkage 9) using actual empirical values

Project A was initiated 25 years ago: What's new?

- Interest in multi-level issues
- Changes in work have escalated
- Interest in the persons evaluated:
Reactions
- Sustainability is an increasingly important issue
- Technology change

COSTELLO CALLS TO BUY A COMPUTER FROM ABBOTT

ABBOTT: Super Duper computer store. Can I help you?

COSTELLO: Thanks. I'm setting up an office in my den and I'm thinking about buying a computer.

ABBOTT: Mac?

COSTELLO: No, the name's Lou.

ABBOTT: Your computer?

COSTELLO: I don't own a computer. I want to buy one.

ABBOTT : Mac?

COSTELLO: I told you, my name's Lou.

ABBOTT: What about Windows?

- COSTELLO: Why? Will it get stuffy in here?
- ABBOTT: Do you want a computer with Windows?
- COSTELLO: I don't know. What will I see when I look at the windows?
- ABBOTT: Wallpaper.
- COSTELLO: Never mind the windows. I need a computer and software.
- ABBOTT: Software for Windows?
- COSTELLO: No. On the computer! I need something I can use to write proposals, track expenses and run my business. What do you have?

Proposal for Major New Exploration of KASO-Performance Relationships

- Large Sample:
 - Multiple organizations
 - Multiple jobs
 - Multiple countries
 - Collection of demographic data: Gender, ethnic status, first language, age, disability status, etc.

Careful conceptualization and measurement of Performance Constructs

- Task Performance
- OCBs
- Adaptive performance with focus on willingness and ability to learn
- Job or organizationally specific outcomes (e.g., vigilance or safety)
- Retention
- Counterproductivity

Predictor Measures

- Careful specification of the predictor domain and hypothesized linkages to performance constructs
- Construct studies and related literature surveys of predictor measures
- Assessment of “biasing” factors and mode of stimulus presentation issues

Design

- Predictive
- Data streaming
- Followup of those not selected
- Data should be collected and archived to allow for assessment and correction for artifacts and relationship changes over time

Collection of context or macro level variables

Organizational characteristics: Job, industry type, customer service requirements, size, age, hierarchy, private vs. public, etc.

Type of work schedule and workplace:
Virtual work, temporary work, etc.

Economy

Country and culture

Allowance for ancillary studies

- Mode of data collection: Web-based versus computer-supervised versus paper and pencil
- How do we manage faking?
- New predictor or outcome measures
- Time and ability-outcome relationships

Examination of Multi-level issues

- Team performance: Is the whole different than the sum of its parts? How do individual difference – team performance measures relate? How does team composition affect performance? Does culture or political system have any influence on these relationships
- Do relationships hold across individuals, teams, units, organizations?
- Are there cross-level effects?
- E.g., what is the effect of culture on selection practices? If people within a culture are more similar on some trait than those from different cultures, validity findings will not generalize.

Placement and Classification Issues

- We know from large scale military studies and simulations that a classification model can improve the utility of selection procedures rather dramatically even when validity differences in composites are relatively tiny. Key here are the intercorrelations between predictors as well as the validity. Utility increases come from decreases in selection ratios for individual jobs.
- We have few applications in civilian arena

Reactions to selection procedures

- Fairness
- Relevance
- Feedback
- Interpersonal treatment
- Consistency
- Reconsideration opportunity
- We know there are cultural differences in the use of and reaction to selection procedures, but have no studies of the impact on validity

Sustainability (Kehoe et al., In press)

- Continued benefit must be recognized
- Organizational fit (is the selection system consistent with the organizational culture)
- Tradeoffs: Quality versus speed and administrative simplicity
- Cost
- Luck associated with who is in charge
- Perceived fairness
- Perceived defensibility

Provision for data-sharing

- Must address individual and organizational confidentiality issues
- Division of labor and maintenance of data base
- Has been done for decades using survey research: “No name” group

Summary and Conclusions

- We have an extensive data base that supports our inferences about the use of many of our selection procedures
- That data base is “old” and deficient in a variety of ways
- Modern theory and research can improve the data base in important practical and scientific ways
- There is a need for a macro cooperative study that is global, cross-cultural, multi-organizational, and continuous

Thank you very much!