

**2018 Revision of the SIOP**  
***Principles for the***  
***Validation and Use of***  
***Personnel Selection***  
***Procedures***

Paul Sackett



UNIVERSITY OF MINNESOTA

# Purpose

- Updating, rather than a re-write of the Principles and the references
  - Impetus is the 2014 revision of the *Standards for Educational and Psychological Testing*
- Adding material areas where there have been new developments, such as unproctored testing

# Contrasting the *Standards* and *Principles*

- *Standards:*
  - Several hundred enumerated Standards
  - Prescriptive: applicable Standards are expected to be met prior to operational use of the test
    - Though acknowledgment that there may be specific features of a given setting that may preclude meeting some Standards
- *Principles:*
  - Expository text: no specific enumerated Principles
  - Aspirational, rather than prescriptive
    - Guide to best practices
    - Many references added in this iteration to aid the test developer/user (*Standards* by policy includes no references)

# Committee

- Paul Sackett  
(co-chair)
- Nancy Tippins  
(co-chair)
- Winfred Arthur
- Tanya Delaney
- Eric Dunleavy
- Ted Hayes
- Fred Oswald
- Dan Putka
- Ann Marie  
Ryan
- Neal Schmitt
- Leaetta Hough

# Timeline

- April 2015 – first meeting of Committee
- January 2017 – SIOP Committee review of Principles draft
- April 2017 – membership review
- July 2017 – Send to APA
- August 2018 – Approval by APA Council
- December 2018 – Published as a special supplemental issue of SIOP's journal *Industrial and Organizational Psychology: Perspectives on Science and Practice*
  - Also published by APA: different typesetting and thus different page numbering. Will be confusing!



# Work Analysis

- Job analysis is now called work analysis throughout the document
  - Addresses situations in which formal jobs don't exist because of rapid changes in the external environment, the nature of work, or the processes for accomplishing work
  - Includes competency modelling
  - Work analysis of future jobs is often needed because of new jobs, global expansion, changing technology, etc. and presents special challenges



# **New predictors are covered**

- References to newer methods (e.g., big data, games, computer simulations) as employee selection procedures and emphasis that these should meet these Principles

# Unproctored and Remotely Proctored Testing

- UIT and RPIT are addressed
  - Recommends user adopt mechanisms and procedures to
    - Diminish the chance the assessment content is compromised
    - Reduce the opportunity for cheating on the assessment
    - Facilitate positive identification of the individual completing the test
  - Admonishes the user to understand the pros and cons of this kind of proctoring and stay abreast of the emerging best practices

# Retains Same Perspective on Validity

- Validity is evaluated relative to the inferences to be made
- Five sources of evidence are outlined (as in *Standards*):
  - evidence based on test content
    - Parallels “content validity” in older nomenclature
  - evidence based on relations to other variables
    - Includes criterion-related validity and convergent and discriminant validity evidence
  - evidence based on internal structure
    - Involves evidence regarding dimensionality
  - evidence based on response processes
    - Involves evidence regarding the underlying processes used by candidates to respond to test stimuli
  - evidence based on consequences of testing



# The “Consequences” Issue

## Examples:

- a high school abandoning much of its curriculum to focus exclusively on skills covered in a state-mandated exam
- theft increases when an integrity test is adopted, as supervisors relax monitoring of employees
- the finding of mean differences between identifiable groups
- applicants view a personality measure as an invasion of privacy and form a negative opinion of the employer



# Perspective Taken in *Standards and Principles*

- negative consequences do not in and of themselves detract from the validity of intended test interpretations
- negative consequences prompt examination of causes
- only if linked to a deficiency in the test do consequences affect conclusions about validity
- consequences can affect decisions about test use, even if consequences are not a validity issue

# Confusion about types of validity evidence

- Common misperception: expected to provide all of these types of evidence
- Key point: need to provide evidence to support the inferences you wish to draw from the test, or the claims you wish to make for the test



## **Example of differing claims, requiring different evidence:**

- “Measuring 7 distinct facets of personality in ways unlike any other product on the market, the Sackett Test identifies the top performers in any sales setting, increases diversity, lowers turnover, and leads to enhanced organizational effectiveness”



## **Example of differing claims, requiring different evidence:**

- “Measuring 7 distinct facets of personality in ways unlike any other product on the market, the Sackett Test identifies the top performers in any sales setting, increases diversity, lowers turnover, and leads to enhanced organizational effectiveness”
- “The Sackett test correlates with sales volume, a key performance measure in our organization”

# Use of the term “construct”

- Older view of “construct” as referring to an attribute that is defined in theoretical terms and is not directly observable
  - Uniform Guidelines view certain lines of validity evidence as not applicable to “constructs,” leading to much attention as to whether a given test does or does not measure a construct
- *Standards* view “construct” differently: “the concept or characteristic that a test is designed to measure”
  - “rarely, if ever, is there a single possible meaning that can be attached to a pattern of test responses. Thus it is always incumbent on test developers and users to specify the construct interpretation that will be made” (p. 11)



# Use of the term “construct”

- Example: we see variance in performance on a set of math items.
  - Possibility 1: this reflects differences in math ability
  - Possibility 2: this reflect differences in exposure to the math principles tested, and thus reflects current knowledge
  - Possibility 3: this reflects differences in effort
  - Possibility 4: this reflect differences in the degree to which test performance is affected by stereotyped threat
- So all validity evidence is in some way “construct validity” evidence, in that it provides information about the meaning of test scores.
  - And so the term “construct validity” is not used: it’s all just “validity”



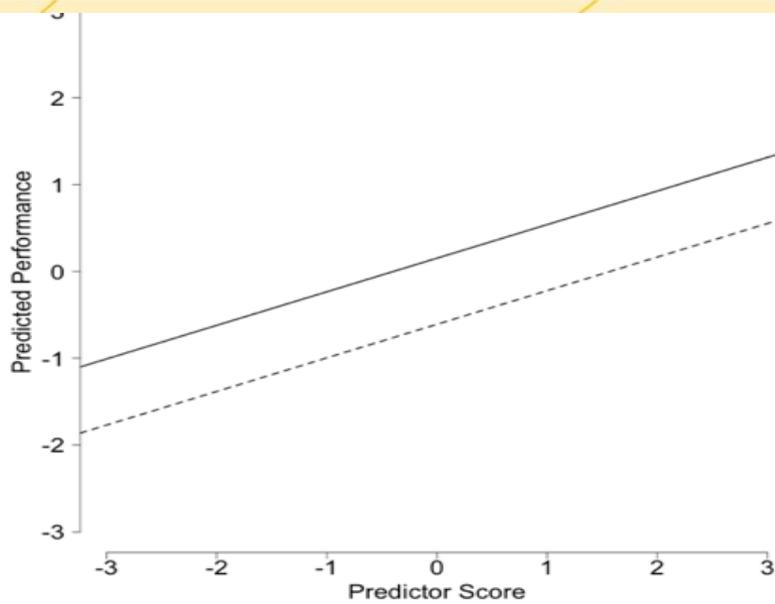
# Adjusting/Correcting Validity Coefficients

- Section on adjusting validity coefficients is expanded to include testing for corrected validities
  - The usual test of statistical significance and standard error or confidence intervals for unadjusted coefficients do not apply to adjusted coefficients
  - Indicates procedures for testing the significance of adjusted coefficients
  - References emerging literature on establishing standard errors and confidence intervals for adjusted scores

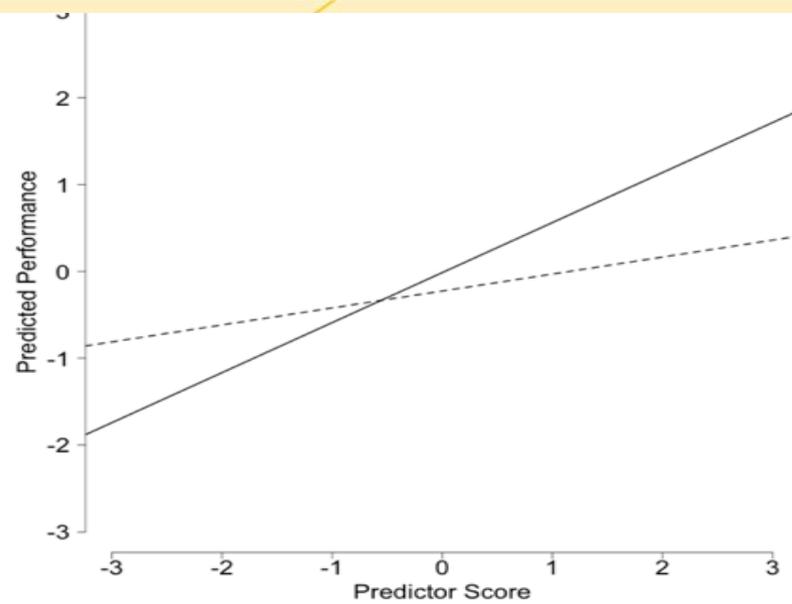
# Significant Changes re Fairness and Bias

- Retains multiple perspectives on fairness:
  - Equitable treatment
  - Lack of bias
  - Rejects equal outcomes
  - New language: “equal access”
- Focuses on “predictive bias”, rather than “measurement bias”
  - Next slide illustrates predictive bias.
  - Solid line - majority group
  - Dashed line – minority group

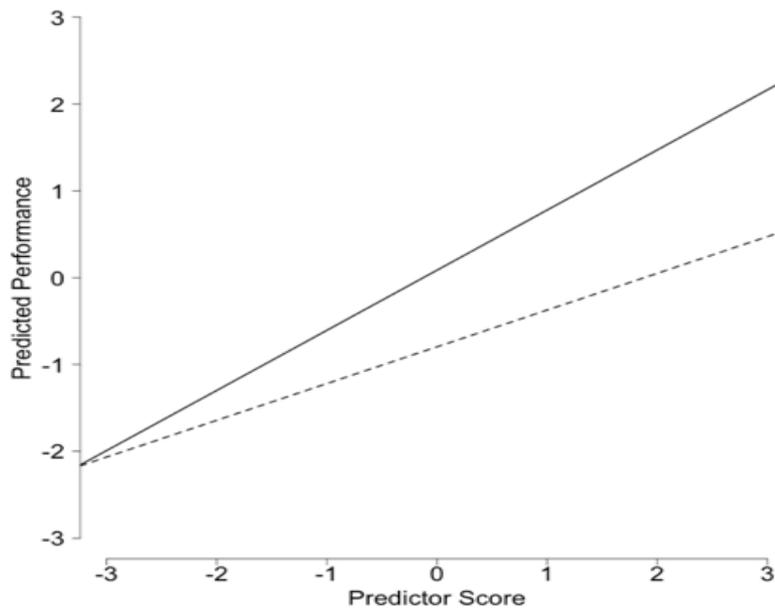




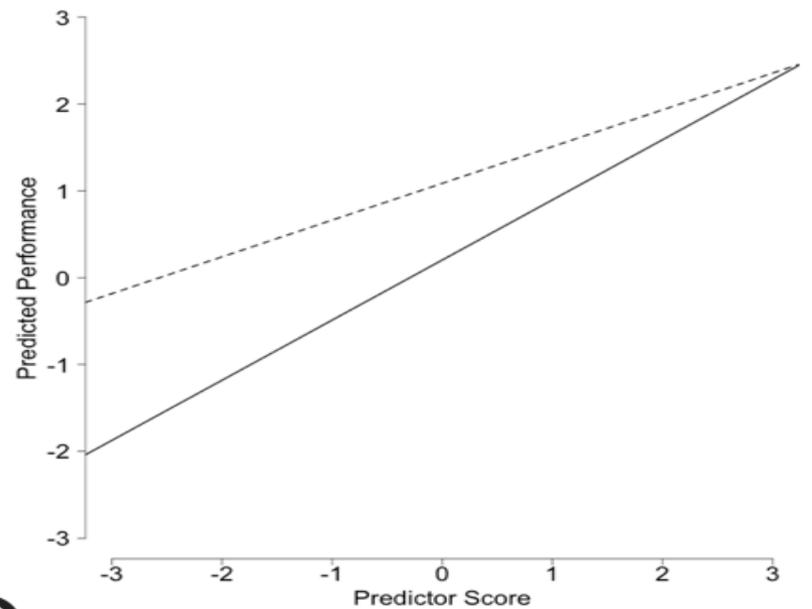
(A) Intercept bias: Minority overprediction



(B) Slope bias: Mixed patterns of over- and under-prediction



(C) Slope bias: Consistent minority overprediction



(D) Slope bias: Consistent minority underprediction

# Predictive Bias

- Differentiates “are regression lines identical?” and “is there bias against a group of interest?”
- Technical discussion of bias now includes discussion of need for unbiased estimate of the intercept difference and operational validity parameters instead of observed parameters
  - This requires correction for range restriction
- Predictive bias should focus on test as used (e.g., if a composite is used, examine bias in the composite)
  - Failure to do this has caused confusion in the literature

# **Aguinis, Culpepper and Pierce (2016): A New Challenge to Predictive Bias Findings**

- Their central point: conclusion in the literature that predictive bias against racial/ethnic minority groups is rare is based on aggregating across samples. This ignores variability from sample to sample.
- Using SAT data from hundreds of schools, they focus on variability from sample to sample.
- They report huge variation (overprediction in some schools, underprediction in others). Claim hundreds of thousands of students have been wronged.

# How Aguinis et al. got it wrong

- They estimate one model with HSGPA, 3 SAT subtests, group membership, and 4 interactions
- Differential prediction describes what would happen if a given predictor was used as the sole basis for selection
- So in interpreting, say, SAT-Critical Reading coefficients in the model, one is asking “is there differential prediction if college admission was based solely on the residual score on SAT-Critical Reading, net of SAT-Writing, SAT-Math, HSGPA, and the interactions?”

# **Sackett, Laczo, and Lippe (2003) offer a clear prescription, now endorsed by the *Principles***

- If multiple predictors are combined in a selection system, differential prediction should be applied to the composite.
- Failure to do so can bias differential prediction analyses when predictors are correlated: to which predictor is the shared variance assigned?
- If this is done with the Aguinis data (e.g., select on a composite of SAT subtests and HSGPA), minority overprediction is consistently found (i.e., underprediction occurs no more frequently than by chance).
  - (findings from my student Jeff Dahlke's 2019 dissertation)

## ***Principles highlight development of an effect size measure for predictive bias***

- There has been a move toward pairing significance tests with a measures of effect size
- Nye and Sackett (2017) and Dahlke and Sackett (2018) developed a class of effect size measures for moderated regression (of which predictive bias is a specific example)
- Expressed in familiar “d” metric
- A measure called “ $d_{\text{mod, signed}}$ ” expresses the average degree of predictive bias against a given group across the range of test scores in standard deviation units



## Computational formula is an integral

$$d_{Mod\_Signed} = \frac{1}{SD_1} \int (\hat{Y}_1 - \hat{Y}_2 | X) f_2(X) dX$$



# A breakthrough: a simple computational formula

- My student Jeff Dahlke's 2019 dissertation derives a really simple computational formula that very closely approximates the calculus-based one:
- $d_{\text{mod\_signed}} = d_y - (r_{xy \text{ maj}} * d_x)$
- Permits estimating  $d_{\text{mod}}$  from published reports without access to original data



# Missing Data

- Missing data discussion expanded to include Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI)
  - These are preferable to traditional leastwise or pairwise deletion
  - The two give comparable findings
  - Still pushback about MI: perceived as “making up data”
  - But only MI permits subsetting of data and creation of interaction terms
- Calls for examining nature of missingness
  - Missing completely at random
  - Missing at random
  - Missing not at random
- This may be new to many. Great summary paper by Newman (2014) is essential reading.

# Outliers

- Calls for checking for univariate and multivariate outliers
  - Sensitivity tests for effect of including or excluding outlier
  - Cautions against removing outliers absent evidence that a data point is erroneous

# Survey of J. Applied Psychology authors (Orr, Sackett, and Dubois, 1991)

## *Attitudes Towards Data Removal*

Percent	Data removal options
29%	All data points always should be included in an analysis regardless of where they lie relative to other data points.
67%	Data points should be removed if they are extreme outliers and there is an identifiable reason that leads you to consider them invalid.
4%	Data should be removed from an analysis if they lie in an extreme area relative to the rest of the data. There does not need to be identifiable reason to believe that they are invalid; extremity is reason enough.



# Conclusion about *Principles*

- Overall, an updating
- Fundamental ideas remain unchanged
- New content related to new ways of administering tests
- New developments on technical issues
  - Assessing predictive bias
  - Dealing with outliers
  - Dealing with missing data



# **“Bonus material” (time permitting)**

- Challenges in getting APA Council approval of Principles

# How Credible Are Claims That Stereotype Threat Results in Biased Test Scores?

- The concept:
  - in a situation in which a stereotype of a group to which one belongs becomes salient, concerns about being judged according to that stereotype arise and inhibit performance
- Hypothesis: high stakes testing is such a situation
  - Stereotypes about women's performance in math
  - Stereotypes about racial/ethnic group performance on cognitive tests

# Original Stereotype Threat Paradigm (Steele and Aronson, 1995)

- Induce threat by manipulation
  - threat: “This is a test of intelligence”
  - non-threat: “This is a problem-solving task developed in our lab”
- Administer test
- Compare “threat” and “non-threat” groups
- Find better minority group performance in non-threat condition



## **Amicus brief in Fisher v. U. of Texas**

“Stereotype threat has been one of the most extensively studied topics in social psychology over the past two decades. In hundreds of studies, scientists have confirmed the existence of stereotype threat and have measured its magnitude, both in laboratory experiments and in the real world.”



# A New Meta-Analysis of Threat Effects

- Shewach, Sackett, and Quint (2019) found 212 true experiments using adult participants
- Conducted a comprehensive meta-analysis, finding an average effect of  $d = -.33$
- However, we identified 3 features present in some experimental studies that would not be present in operational settings
  - Conducted our focal analysis on studies with features plausible in operational settings



# Feature 1: Test scoring

- While most use “ number right” scoring, 10% use “proportion correct among item attempted”
- Would never use this in an operational setting – easily coached
- Mean d w number right scoring:  $-.31$
- Mean d with proportion correct scoring:  $-.51$

## Feature 2: Control group vs. Threat Removal Group

- 60% compare a threat group with a control group
- 40% compare threat group with threat removal group
  - Tell test-takers that there are no group differences on the test
- Would never do this in an operational setting – unethical to lie to test takers
- Mean  $d$  w control group:  $-.28$
- Mean  $d$  with removal group  $-.40$



# Feature 3: Subtle vs. Blatant Threat Manipulation

- 65% use a blatant threat manipulation: tell test takers that mean differences exist on the test they are about to take
- 35% use a subtle threat manipulation
  - Induce threat by “priming” race or gender (e.g., asking them to self-report race/gender just before taking the test)
- Would never use a blatant manipulation in an operational setting
- Mean  $d$  w blatant manipulation:  $-.42$
- Mean  $d$  with subtle manipulation:  $-.16$



# **Our focal analysis: eliminate studies with these problematic features**

- 43 studies with features plausible in operational testing settings
- Mean  $d = -.14$



# Role of Motivation

- The vast majority of ST studies utilize college students participating for course credit
- Performance on these tests does not affect the test-taker beyond the experiment
- Could a plausible interpretation of stereotype threat be motivational?
  - Experimenter tells me my group performs poorly → why should I exert high-levels of effort on this test?
  - Nine studies offer financial incentives to do well.
  - Mean  $d = -.01$

# Role of Motivation: Operational settings

- Furthermore, there are four large  $N$  studies which come from operational testing scenarios
- College placement exams
- Experimental studies – randomly assigned students to experimental conditions which checked gender/race before versus after the placement exam
- Mean  $d = -.01$

## **But: there is also evidence of publication bias**

- Bigger effects in small sample studies
  - Mean  $d$  in 10 largest studies:  $-.07$
- “Trim and fill” analysis: expect symmetric distribution of effect sizes
  - Corrected for asymmetry, mean  $d = -.09$



## Bottom line: weak evidence of a threat effect

- Overall mean effect of  $-.33$
- Mean effect in studies with features plausible in operational settings, corrected for publication bias:  $-.07$  to  $-.09$ 
  - This is not significantly different from zero
- Nine studies with financial incentives
  - Mean  $d = -.01$
- Four true experiments in operational settings
  - Mean  $d = -.01$
- My conclusion: advocates have not made the case that threat systematically affects test scores in operational settings



# References

- Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology, 108*(7), 1045.
- Aguinis, H., Gottfredson, R., and Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods, 16*, 270-301.
- Dahlke, J.A., and Sackett, P. R. (2018). Refinements to the  $d_{Mod}$  class of categorical-moderation effect sizes. *Organizational Research Methods, 21*, 226-234
- Newman, D.A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*, 372-411.
- Nye, C. D., and Sackett, P. R. (2017). New effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods, 20*, 639-664.
- Orr, J. M., Sackett, P. R., and DuBois, C. L. Z. (1991) Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*, 473-486.
- Sackett, P. R., Laczko, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: the omitted variables problem. *Journal of Applied Psychology, 88*(6), 1046.
- Shewach, O.R., Sackett, P. R., and Quint, S. (2019). Stereotype threat effects in settings with features likely vs. unlikely in operational testing settings: A meta-analysis. *Journal of Applied Psychology, 104*, 1514-1534.

