

Balancing Performance and Diversity in Personnel Selection

Paul R. Sackett

Professor of Industrial/Organizational
Psychology

University of Minnesota

The Performance – Diversity Tradeoff

Common issues in employment testing, admissions testing, and licensure and certification testing:

- use of cognitively-loaded tests
- large body of validity evidence
- substantial group mean differences by race/ethnicity
- no systematic predictive bias
- hence a dilemma for those valuing both validity and diversity

d: A Common Index for Group Mean Differences

$d = \text{mean difference} / \text{standard deviation}$

Example

Test 1: 3 point Black/ White mean difference

Test 2: 10 point Black/White mean difference

Test 1: standard deviation = 3

Test 2: standard deviation = 50

Test 1 $d = 1.0$

Test 2 $d = .2$

Common d Values

White-Black d on cognitive tests: 1.0

Male-Female d on Dominance: .5

Male-Female d on muscular endurance: 1.5

Male-Female d on Conscientiousness: -.3

Majority Selection Ratio Relative to Minority Selection Ratio as d Varies

Majority Group Selection Ratio

<u>d</u>	<u>.10</u>	<u>.50</u>	<u>.90</u>
0.0	.10	.50	.90
0.2	.07	.42	.86
0.4	.05	.35	.81
0.6	.03	.27	.75
0.8	.02	.21	.68
1.0	.01	.16	.61

Ratio of Minority to Majority Selection Rates as d Varies

Majority Group Selection Ratio

<u>d</u>	<u>0.1</u>	<u>0.5</u>	<u>0.9</u>
0.0	1.00	1.00	1.00
0.2	0.69	0.84	0.98
0.4	0.46	0.69	0.90
0.6	0.30	0.55	0.84
0.8	0.19	0.42	0.72
1.0	0.11	0.32	0.68

Proposed Strategies for Reducing d : Incorporating minority preference

- These approaches attempt a compromise: some loss of performance is accepted in the interests of minority hiring
- Examples:
 - Bonus points
 - Separate cutoffs
 - Banding
 - Within-group norming

U.S. Legal Climate re Preference

- Historically, preference permitted after a finding of discrimination
- The Civil Rights Act of 1991 bans score adjustment
- Adarand v. Peña restricts preference to identifiable victims of past discrimination by the employer
- California Proposition 209 bans preferences in employment and education
- Recent Michigan cases (Gratz, Grutter v. Bollinger) permit “soft” preferences in educational admission
- Change in the diversity argument: from reparation for past wrongs to the positive value of diversity

Example: U of California - San Diego Medical School Admissions

	White	Black
1996	127/2430 (5.2%)	7/212 (3.3%)
1997	139/2293 (6.1%)	0/196 (0%)

If $d = 1.0$:

Maj. Selection Ratio	Min. Selection Ratio
90%	61%
75%	37%
50%	16%
25%	5%
10%	1.1%
5%	0.4%

So: with 196 Black applicants and 5% maj. selection ratio: we would expect to select $0.4\% \times 196 = .78$

Conclusion regarding strategies with minority preference

- Within-group percentile scoring is the technically optimal compromise between diversity and performance
- Achieves representative selection at smallest performance cost
- But: banned in U.S. for employment decisions by the Civil Rights Act of 1991

Framing the *Alternatives*

- Preference vs. Merit
 - or
- Diversity via Preference vs. Diversity by any means possible
 - Drop valid predictors with adverse impact
 - Use low cutoffs

Seven proposed strategies for reducing d that do not involve minority preference

- Identify and remove biased items
- Use coaching programs
- Provide more generous time limits
- Alter test taking motivation
- Change test format (eliminate paper and pencil)
- Supplement with additional measures
- Expand the criterion

4 strategies with disappointing results

- Identify and remove biased items?
 - virtually no net effects
- Use coaching programs?
 - improves scores for all; differences remain
- Provide more generous time limits?
 - may increase differences!
- Alter test taking motivation?
 - Steele's stereotype threat merits attention

Claude Steele's stereotype threat theory

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: employment testing is such a situation

Stereotype Threat Paradigm

- 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
- Some wrongly interpret this as “subgroup difference is eliminated”
- Key question: only in the lab?

Strategy 5: Changing test format

- Does format matter?
- Sackett (1998) -> no
 - assessment centers for lawyers and teachers do not result in less adverse impact than written tests
- Chan and Schmitt (1997) -> yes
 - far greater adverse impact in written format than in video format

Clinical Legal Skills Assessment Center

- client interview
 - opening statement
 - closing argument
 - cross examination
 - discovery plan
 - settlement proposal
- Also: video-based scenarios requiring judgment

Black-White d and r with Bar Exam

	<u>d</u>	r with Bar
Bar ($r_{xx} = .91$)	.89 (.93)	
AC ($r_{xx} = .67$)	.76 (.93)	.56 (.72)
Video ($r_{xx} = .64$)	.89 (1.11)	.68 (.89)

Chan and Schmitt (1997)

- video-based test of interpersonal skills: watch vignettes, choose course of action
- Obtained $d = .22$
- Transcribed the video: now read the vignettes, choose course of action
- obtained $d = .91$
- Implication: Format indeed matters

Reconciling Sackett with Chan and Schmitt

- If focal construct construct is highly cognitively loaded, format has little impact
- If focal construct is not highly cognitively loaded, format matters greatly. A cognitively loaded measurement method can substantially increase adverse impact

Strategy 6: Supplement Existing Tests with Alternative Tests

- Question: to what degree can subgroup differences be reduced by using a combination of existing and new tests?
- Will rely on the psychometric theory of composites to address this

Quiz

- Test A produces a d of 1.0
- Test B produces a d of 0.0
- A and B uncorrelated; both on same scale

Question: What is the d for the composite A+B?

- a) 1.0
- b) .71
- c) .50
- d) 0.0

Answer : .71

- Lesson : a supplemental predictor will have a smaller effect on d than many think

Q: What about adding more than one new test?

A: d for 2-test composite = .71; for 3-test composite = .58; for 4-test composite = .50

- Lesson: d will be sizable even with multiple new tests

Combining Law School Grades and Bar Exam Scores

- Bar exam $d = 1.0$
- Law school grade $d = .8$;
- r between bar exam and grades = $.5$

Question: what will be the d if a unit-weighted composite of bar exam and grades is used?

- a) 1.0
- b) .9
- c) .8
- d) 1.04

Conclusions about adding supplemental tests

- If equally weighted, reduces d less than some expect
- Diminishing returns for adding multiple additional tests
- New tests would have to be very heavily weighted to produce small d
- If correlated with existing tests, adding new tests can increase d

Strategy 7: Expand the criterion

- Cognitive tests best predict maximum performance
- Noncognitive measures play a larger role in predicting typical performance
- Overall performance = task performance + citizenship – counterproductivity
- So: if important aspects of performance are missing, may overlook valid measures with low d

Conclusions: What does not help balance diversity and performance ?

- Identify and remove biased items
- Use coaching programs
- Provide more generous time limits
- Alter test taking motivation

Conclusions: What can help balance diversity and performance ?

- Change test format
- Supplement with additional measures
- Expand the criterion
- Approaches that involve minority preference are an effective compromise, if legally viable

- There is no ready single solution