ADVERSE IMPACT:
A Persistent Dilemma

David Chan
Catherine Clause
Rick DeShon
Danielle Jennings
Amy Mills
Elaine Pulakos
William Rogers
Jeff Ryer
Joshua Sacco
David Schmidt
Lori Sheppard
Matt Smith
David Whitney

Valid Selection Procedures have high adverse impact: Significantly different hiring rates for different subgroups.

Civil Rights Act of 1991 prohibited score adjustments.

How do organizations use valid selection procedures in a way that will produce a workforce that is optimally capable and representative of the diverse groups in our society?

Schmitt, Clause & Pulakos (1996)

Subgroup effect sizes for various abilities: African-American--White comparisons

Ability	N of effects	Range of effect sizes[a]	Weighted effect sizes[a]	Total N[a]
Manual dexterity	3	-0.03 to -0.30	-0.14	1128
Spatial ability	7	-0.09 to -1.20	-0.66	2868
Verbal ability	8	-0.14 to -1.15	-0.55	2024
Math ability	11	-0.23 to -1.16	-0.64	3765
General/Cognitive	16	-0.31 to -1.46	-0.83	7590
Job Sample/Job knowledge	37	+0.16 to -1.01	-0.38	15738
Clerical speed/accuracy	2	-0.05, -0.26	-0.15	341
Mechanical comprehension	1	-0.40	-0.40	430
Interview (motiv./exp.)	6	+0.12 to -0.39	-0.15	1531
Personality	6	+0.22 to -0.46	-0.09	801
Accomplishment rec.	1	-0.33	-0.33	250

What Standardized Group Differences Mean for Minority Selection (Sackett & Wilke, 1994)

Minority Group Selection Ratio When Cutoff for Majority Group is Set at 10%, 50%, and 90%

	Majority Group Selection Ratio		
Standardized Group Difference (d)	10%	50%	90%
0	.100	.500	.900
0.1	.084	.460	.881
0.2	.070	.421	.860
0.3	.057	.382	.836
0.4	.046	.345	.811
0.5	.038	.309	.782
0.6	.030	.274	.752
0.7	.024	.242	.719
0.8	.019	.212	.684
0.9	.015	.184	.648
1.0	.013	.159	.610
1.1	.009	.136	.571
1.2	.007	.115	.532
1.3	.005	.097	.492
1.4	.004	.081	.452
1.5	.003	.070	.413

Schmidt & Hunter Psychological Bulletin (1998)

Predictive Validity for Overall Job Performance of General Mental Ability (GMA) Scores Combined with a Second Predictor Using (Standardized) Multiple Regression

Personnel Measures	Validity (r)	Multiple R
GMA tests[a]	.51	
Work sample tests[b]	.54	.63
Integrity tests[c]	.41	.65
Conscientiousness tests[d]	.31	.60
Employment interviews (structured)[e]	.51	.63
Employment interviews (unstructured)[f]	.38	.55
Job knowledge tests[g]	.48	.58
Job tryout procedure[h]	.44	.58
Peer ratings[i]	.49	.58
T & E behavioral consistency method[j]	.45	.58
Reference checks[k]	.26	.57
Job experience (years)[l]	.18	.54
Biographical data measures[m]	.35	.52
Assessment centers[n]	.37	.53
T & E point method[o]	.11	.52
Years of education[p]	.10	.52
Interests[q]	.10	.52
Graphology[r]	.02	.51
Age[s]	-.01	.51

POTENTIAL SOLUTIONS

1. Include measures of job-related constructs with low or no adverse impact along with, or instead of, traditional paper-and-pencil measures of cognitive ability.

2. Investigate tests for items that are culturally laden and remove those items or options.

3. Use computer or video technology to present test stimuli and collect responses.

4. Use portfolios or accomplishment records to document job-related accomplishments or achievements.

5. Coaching or orientation efforts.

6. Change the way in which test scores are used.

MEASURE OTHER CONSTRUCTS

Traditionally, organizations have often used structured ability tests, usually highly cognitive in nature, as a major component of their selection procedure.

7. Ease of administration and scoring

8. Relatively easy to develop

9. Valid

but ...

large subgroup differences

If other abilities (i.e., teamwork or interpersonal skills) are important, then why not include measures of those abilities, particularly since subgroup differences are small or nonexistent on measures of these constructs?

10. Lack of valid measures of these constructs
11. Expense associated with development and administration
12. They may not change adverse impact as much as one might estimate

Pulakos & Schmitt (1996)

Selection procedures designed to select investigative officers at a large federal agency. All were college graduates, many with advanced degrees. Study was a concurrent criterion-related validation study.

Paper-and-pencil measure of verbal ability (analogies, vocabulary, reading comprehension)

Performance measures of verbal ability

13. Written stimulus material, written response
14. Audio visual stimulus material, written response

Biodata
Situational Judgment
Interview

Results

	Subgroup Difference		
	AA - White	HA - White	Validity
Verbal Ability	1.03	.78	.19
Health Fraud	.91	.52	.22
Munitions	.45	.37	.15
Biodata	-.05	.05	.22
Situational Judgment	.41	.02	.20
Interview	.12	.22	.35
BIO, SJ, INT	.23	.16	.41
BIO, SJ, INT, VA	.63	.48	.43

Adverse impact of combination of variables of differing impact will be a function of:

15. Level of impact of components of a battery
16. Reliability of individual components
17. Intercorrelation of components
18. Selection ratio
19. Others?

Legally,

20. It would be hard to challenge the combination used in this instance, but ....
21. Does adding .02 to test battery justify added level of impact associated with Verbal Ability test?

Schmitt, Rogers, Chan, Sheppard, & Jennings (1997)

Results of analyses of representative empirical estimates of validity and adverse impact of a battery that includes :

22. Cognitive ability
23. Structured interview
24. Biodata
25. Personality (Conscientiousness)

Proportion of Majority and Minority Groups Selected and Adverse Impact (AI) Ratios for Five Selection Ratios Using Table 1 Meta-Analytic Estimates of Intercorrelations and Validities

	Composite of Four Predictors		
Selection Ratio	**Majority**	**Minority**	**AI Ratio**
.90	.92	.77	.84
.70	.74	.50	.66
.50	.55	.30	.53
.30	.35	.14	.41
.10	.13	.03	.28
	Interview, Biodata, and Personality		
	Majority	**Minority**	**AI Ratio**
.90	.90	.87	.96
.70	.71	.65	.91
.50	.51	.44	.86
.30	.31	.25	.80
.10	.11	.08	.72
	Cognitive Ability Only		
	Majority	**Minority**	**AI Ratio**
.90	.93	.69	.74
.70	.77	.39	.51
.50	.58	.21	.37
.30	.37	.09	.25

.10 .14 .02 .14

Schmitt, Rogers, Chan, Sheppard, & Jennings (1997)

Investigated the degree to which predictors with varying characteristics used in combination produce adverse impact and validity in a simulation.

Varied:

26. Number of alternate predictors (1, 2, or 3)
27. Predictor intercorrelation (.00, .25, .50)
28. Levels of subgroup differences on alternate predictors (d = .00, .25, .50)
29. Validity of alternate predictors (.10, .20, .30)

Constants:

30. Cognitive ability - d = 1.00
31. Proportion of lower scoring group in sample = .20
32. Subgroup difference on criterion d = .45
33. Validity of cognitive ability = .29

Results

34. Alternate predictors do reduce impact, but will not remove it. d remains high over a broad range of study factors

35. With alternate predictors that exhibit no subgroup difference, high validity, high intercorrelation, we have lowest d

36. Relative validity of alternate predictors is important in a multiple regression combination

37. See also Sackett & Roth (1996) and Sackett & Ellingson (1997)

INVESTIGATE DIFFERENTIAL ITEM FUNCTIONING

38. Item content or context
39. Item format or structure

Scheuneman (1987) is representative of the findings of this research

10 of 16 hypotheses regarding subgroup by item format interactions were confirmed but interpretation was difficult

Items exhibiting dif equal that expected by chance is the most frequent finding

When dif items are removed, we also remove validity (Roznowski, 1987)

Whitney & Schmitt (1997)

Response options to biodata items were written to reflect the culture of African-American and White subgroups based on the cultural typology of Kluckhorn & Strodtbeck (1961) and the empirical work of Carter (1990)

40. Cultural values were related to response option selection

41. Cultural values were not related to subgroup differences on these measures.

Use alternate methods other than highly verbal paper-and-pencil measures to assess abilities including cognitive ability.

Most of this small body of research confounds format and content (or constructs?) measured?

Why?

Example: What would you do if an angry parent confronted you about the grade their son/daughter received on an examination?

42. Multiple-choice format with alternatives
43. Could require a written essay response
44. Interview question with a required oral response
45. Could require that candidate role play the teacher with a Aparent@ actor. Candidate=s behavior is rated
46. Could present a video enactment of the confrontation along with video enactments of alternative courses of action. Candidate must choose an alternative

Differences in format along several dimensions.

47. Realism
48. Scoring difficulty
49. Written vs. oral response
50. Visual vs. written stimuli
51. Breadth of content that can be sampled

Video versions of tests are usually developed to reduce or eliminate reading requirements of tests

Possibilities include

52. Video stimuli, written response
53. Video stimuli, oral response
54. Written stimuli, oral response
55. Written stimuli, written response

Chan & Schmitt (1997) took a video test developed by HRStrategies to assess three interpersonal skills dimensions and changed it to a written multiple-choice measure using the video scripts.

Video version d = .22
Written version d = .91

The content of the two tests was the same, but the written version introduced a reading comprehension factor.

Mills & Schmitt (1999)

Compared performance of African-American and White applicants for insurance claims personnel on two batteries of tests.

56. Paper-and-pencil

57. Computerized simulation with frequent telephone interruptions

Both measured verbal and cognitive ability constructs.

Multiple groups analyses of the factor structure and mean differences indicated the following:

58. The simulation and paper-and-pencil factor were correlated .73 (No differences across groups).
59. There were large subgroup differences in variance across two test factors (AAs scores were more than twice as variable).
60. Paper-and-pencil tests were more highly correlated for AA group than White group.
61. d for paper-and-pencil tests was .77; d for simulation was .36.

In addition, validity for a small subsample (N = 51) was .28 for paper-and-pencil tests, .31 for simulation.

QUESTION: Why do face valid and content valid measures produce less adverse impact?

ONE HYPOTHESES: Different test formats may have differing motivational effects.

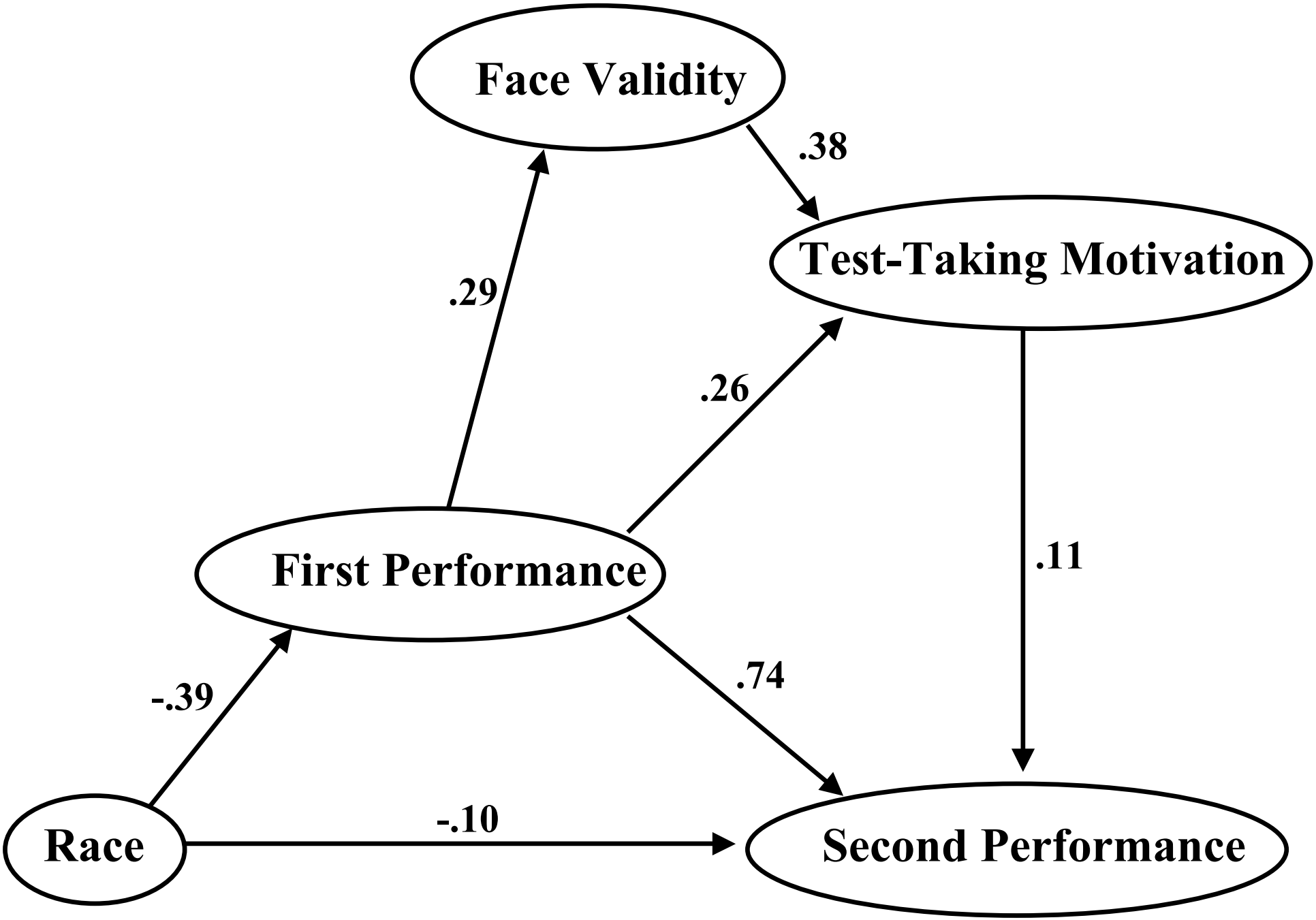
Previous researchers (e.g., Schmitt et al., 1997; Rynes & Connerly, 1993; Smither et al., 1993) have found that

examinees prefer realistic exams or job sample tests over more abstract general tests of ability or personality.

Chan, Schmitt, DeShon, Clause, & Delbridge (1997) reported

62. Relatively large and statistically significant differences in the perceived fairness and self-reported test taking motivation of African-American and White students on a cognitive ability test.

63. These motivational differences had an impact on performance on a second similar test even after performance on the first test was statistically controlled (see Figure).


Face Validity
.38
Test-Taking Motivation
.29
.26
.11
First Performance
.74
-.39
Race
-.10
Second Performance

Documentation of Previous Accomplishments

- Portfolio/authentic assessment
  - Reliability
  - Validity
  - Subgroup differences remain and are sometimes even greater (Linn, Baker, & Dunbar, 1991; Bond, 1995)

Accomplishment records (Hough, 1984)

- More attention to psychometric adequacy and standardization
- Documentation of examinee involvement in accomplishments
- Documentation of accomplishments done at time of application

Interrater reliability in .70s and .80s (Schmidt et al., 1979; Hough, 1984)

Validity = .25 (Hough, 1984)

Subgroup difference = .33 (Hough, 1984)

Completion rates? Also see Schmit & Ryan--greater withdrawal among African-American group than others.

COACHING or Orientation Programs

Kaplan claims a one standard deviation change in SAT scores as result of their preparation course.

Three meta-analyses of educational literature report effect size changes between .10 and .25 (Messick & Jungeblut, 1981; Der Simonian & Laird, 1983; Bangert-Downs, Kulik, & Kulik, 1983). Messick reported larger effect sizes in those studies including African-Americans.

Ryan, Ployhart, Greguras, & Schmit (1998)

- No effect of test preparation on test performance for a Civil Service firefighter exam.

Ryer, Schmidt, & Schmitt (1999)

- Effect of 1**2** hour preparation course on scores on entry-level selection procedures for manufacturing jobs were about .10 overall but there were confusing cross-location results.

ALTERNATE USE OF TEST SCORES

Banding (refers to use of test scores rather than a change in the tests)

Candidates within a band (established by reference to the top scorer) are considered equal and candidates within the band are chosen using criteria other than test scores.

Fixed bands

Sliding bands

Secondary criteria

LEGAL/PRACTICAL CONSIDERATIONS

Complexity of banding may be difficult to explain

An increase in minority hiring may or may not occur

Perception that this approach is a new version of score adjustment

Order of use of test data & secondary criteria

Need for expansion of our notions of organizational effectiveness

SUMMARY & CONCLUSIONS

64. Careful consideration of full range of performance goals and organizational interests

65. Construct & use measures that reflect the full range of required abilities

66. Pay attention to face validity

67. Continue research on alternative testing methods, technologies & constructs

68. Develop job-relevant, psychometrically adequate measures of past achievements

69. Prepare examinees

70. Recognize the existence of subgroup differences on certain ability dimensions & develop programs to remediate these differences