

# **New Ways to Reduce Adverse Impact and Preserve Validity with Additional Discussion of Court Challenges to Testing**

Joel P. Wiesen, Ph.D.

[jwiesen@appliedpersonnelresearch.com](mailto:jwiesen@appliedpersonnelresearch.com)

MAPAC 2010 Conference

Silver Springs, MD

September 21, 2010

# Two Parts of Presentation

- Monte Carlo research on new ways to use test data
- Observations on legal challenges to testing

# Acknowledgements/Thanks

- Program committee
- Coauthors of the Monte Carlo research
  - Herman Aguinis, Ph.D.
  - Tuvshingus Batdelger, Ph.D.
- Support staff for Monte Carlo research
  - Aaron Brown
- Various attorneys

# Overview

- Developed completely new selection methods based on existing tests
- Compared new with traditional selection methods in a large Monte Carlo study

# Test Battery Studied

- General Mental Ability (GMA, m/c test)
- Conscientiousness (CONSC)
- Physical Performance Test (PPT)
- Structured Interview (SI)

# What is New?

- Tests are typical
- Novelty involves ways to use the test data
- Not all tests contribute to the grade of each applicant
- Choose tests based on the strengths or weaknesses of the applicant

# New Selection Methods

- Greatest Strength Method (GSM)
- Two Greatest Strengths Method (GSM2)
- Drop the Lowest Score (DROP)
- Composite without GMA (COMP2)

# Comparison Methods

- GMA alone
- Composite of all tests (COMP)
- Random (RAND)

# Evaluation Areas

- Validity
- Adverse Impact (AI)
- Standardized Mean Group Difference (d)
- Mean Job Performance (MJP)

# Quick Look at Findings

- Preserve much validity
- Reduce AI
- Some caveats

# New Methods Explained

- Determine z-score for each test
- Calculate method grades based on z-scores

## Greatest Strength Method (GSM)

- Determine test with greatest z-score
- Grade = that z-score
- Fail any candidate with a low score on any test
- Rank candidates based on grade

## Two Greatest Strengths Method (GSM2)

- Determine the 2 tests with greatest z-scores
- Grade = composite of those 2 z-scores
- Fail any candidate with a low score on any test
- Rank candidates based on grade

## Drop the Lowest Score Method (DROP)

- Determine test with lowest z-score
- Grade = composite of remaining 3 z-scores
- Fail any candidate with a low score on any test
- Rank candidates based on grade

## Compensatory Omitting GMA (COMP2)

- Omit GMA (i.e., m/c test)
- Grade = composite of other 3 z-scores
- Rank candidates based on grade

# Comparison Methods Explained

- Determine z-score grade for each test
- Calculate method scores based on grades

# GMA Test Alone

- Grade = z-score for GMA (i.e., m/c test)
- Rank candidates based on grade

# COMP

- Grade = composite of all 4 z-scores
- Rank candidates based on grade

# RAND

- Grade ignores all test -scores
- Rank candidates randomly

# Simulation Study Methodology

- Specify intercorrelations
- Generate data with these intercorrelations
- Create gender and EEO groups
- Create mean score differences
- Compute grades using 7 different methods
- Make selections under the various methods
- Evaluate validity, AI, etc.

# Intercorrelation Inputs

	SI	PPT	CONSC	Job Performance
GMA	.31	0	.03	.51
SI		0	0	.48
PPT			0	.35
CONSC				.22

## Create Mean Score Differences

	GMA	ORAL	PPT	CONSC	Job Perf.
Women (Case 1)	0	0	-1.25	0	0
Women (Case 2)	0	0	-1.25	0	-.4375
Ethnic Group	-.72	-.31	0	.07	-.27

# Some Variables Considered

- Selection ratio (SR)
  - Lower SRs typically yield worse AI
- Proportion of ethnic minority applicants (EEO)
- Applicant group size

# Selection Ratio (SR)

- .01
- .05
- .15
- .20
- .30
- .50
- .90

## Proportion of Ethnic Minority Applicants (EEO)

- .05
- .10
- .20
- .30
- .40
- .50

# Applicant Group Size

- 500
- 1,000
- 2,000
- 10,000
- 750,000

# Number of Replications

- Replicate until consider 750,000 cases
  - 1,500 replications for N of 500
  - 750 replications for N of 1,000
  - 375 replications for N of 2,000
  - 75 replications for N of 10,000
  - 1 replication for N of 750,000

# Design Summary

- 7 levels of SR
- 6 levels of EEO
- 7 methods of using data
- 2 genders
  - Case 1: No gender difference in MJP
  - Case 2: Gender difference in MJP
- 2 EEO groups
- 5 levels of sample size

# Results

- Results varied somewhat by level
- Will show results first for
  - Total sample of 750,000
  - SR = .2
  - EEO = .2

# Validity

Method	Validity	AI EEO	AI Gender	MJP Case 1	MJP Case 2
<b>GMA</b>	<b>0.51</b>				
GSM	0.52				
GSM2	0.62				
DROP	0.61				
COMP2	0.57				
COMP	0.69				
RAND	0.00				

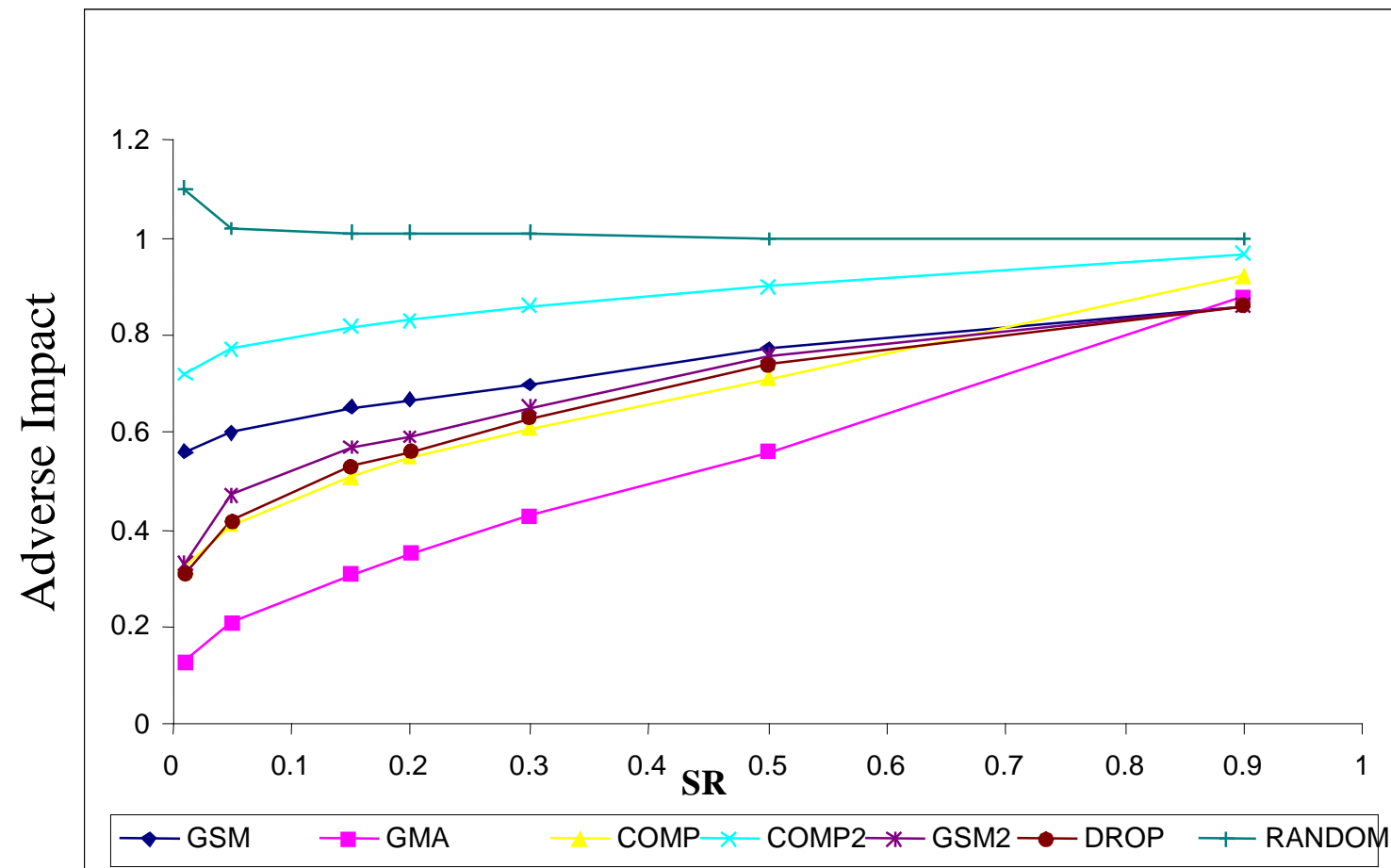
# Adverse Impact (AI)

Method	Validity	AI EEO	AI Gender	MJP Case 1	MJP Case 2
<b>GMA</b>	<b>0.51</b>	<b>0.31</b>	<b>1.00</b>		
GSM	0.52	0.63	0.66		
GSM2	0.62	0.56	0.61		
DROP	0.61	0.53	0.55		
COMP2	0.57	0.83	0.35		
COMP	0.69	0.53	0.44		
RAND	0.00	1.00	1.00		

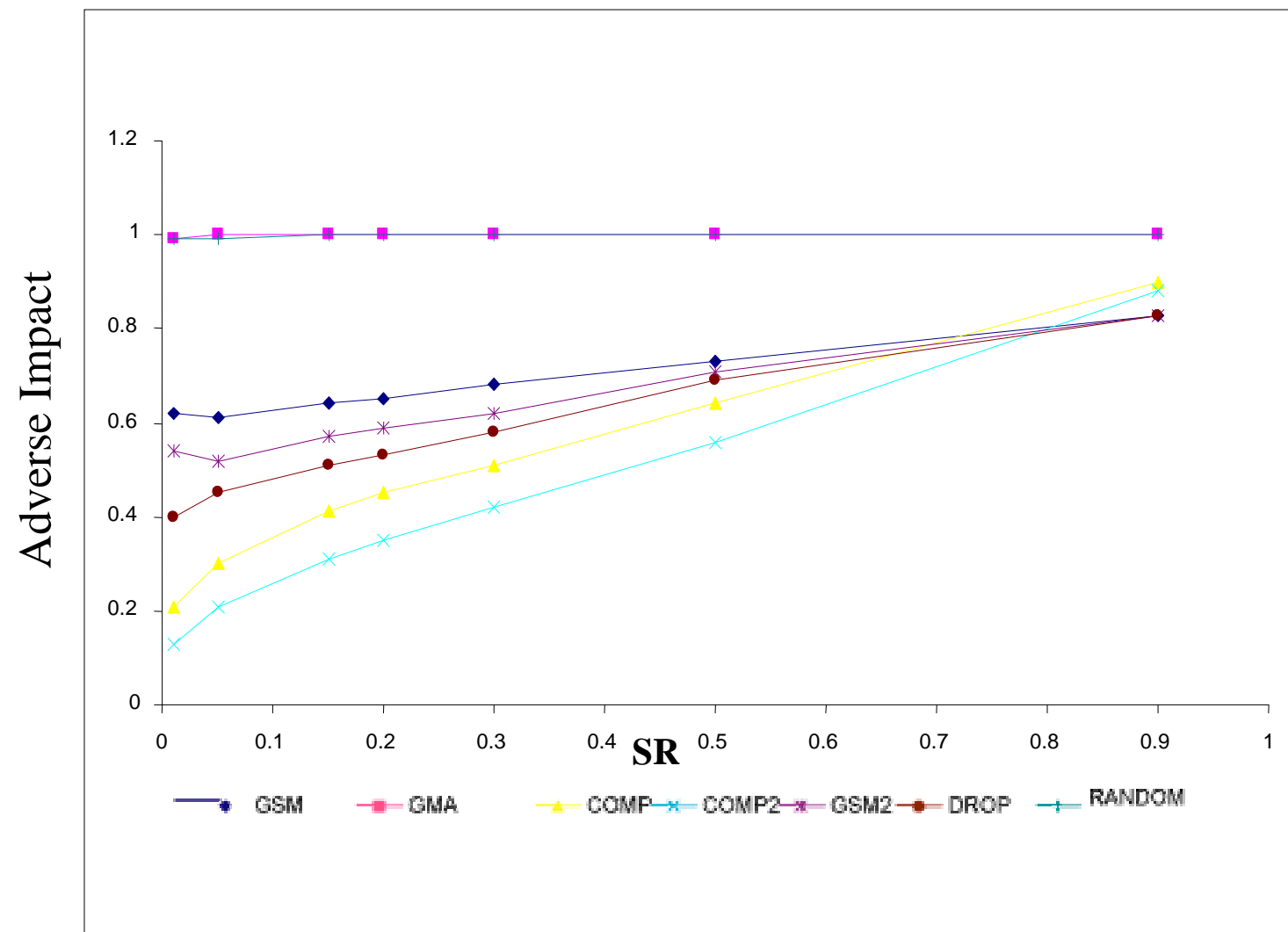
## Mean Job Performance (MJP)

Method	Validity	AI EEO	AI Gender	MJP Case 1	MJP Case 2
<b>GMA</b>	<b>0.51</b>	<b>0.31</b>	<b>1.00</b>	<b>0.67</b>	<b>0.45</b>
GSM	0.52	0.63	0.66	0.74	0.57
GSM2	0.62	0.56	0.61	0.85	0.68
DROP	0.61	0.53	0.55	0.90	0.74
COMP2	0.57	0.83	0.35	0.75	0.63
COMP	0.69	0.53	0.44	0.93	0.79
RAND	0.00	1.00	1.00	-0.06	-0.27

# AI for EEO, by Method and SR (EEO50, N750,000)



# AI for Gender, by Method and SR (EEO50, N750,000)



# Summary of Validity Results

- Validity highest for COMP
- Validity for novel methods also high
- Validity for novel methods greater than GMA m/c test alone

# Summary of AI Results

- Traditional approaches have relatively severe AI for EEO or gender or both
  - GMA has AI for EEO
  - COMP has AI for EEO and gender
- Novel approaches reduce AI for BOTH EEO and gender
  - Especially GSM

# Caveats

- High variability in AI
  - Often AI = zero, especially with
    - Small N
    - Low SR
    - Low EEO
- Mathematical model approach

# Conclusions

- Novel, simple ways to use test scores
  - GSM, GSM2 and DROP have promise
  - Will result in occupational diversity in terms of skills
- Reduce adverse impact
- Maintain much validity
- One approach, not “the solution”

## Note on Occupational Diversity

- More diverse mix of strengths/weaknesses in employees
- Each employee may contribute based on strengths
- May facilitate teamwork

# Final Thoughts on this Research

- This is a simplified summary of the results of a large Monte Carlo study. A full paper is in preparation.
- Call for collaboration in real life applications

A more complete summary of this research will be available at:  
<http://appliedpersonnelresearch.com/pubs.html>

# Court Related Experiences

- Introduction
- Diverse topics
  - Critiques
  - Evaluations of the critiques
  - Your thoughts

# Introduction

- How plaintiffs' lawyers view things
  - Fresh look
  - Focus on gaps/weaknesses
  - Bright

# Passing Point: Take 1

- Situation: Banded exam for Police Sgt and Lieut
- Lieutenant applicants take all Sgt questions and more
- Score Lieutenant applicants on Sgt exam
  - Use these scores to evaluate or set cut score on Sergeant exam

# Passing Point: Take 1

- Attorney: Circular reasoning
- All current Sgts scored high on earlier Sgt exam (they all were appointed)
- Of course current Sgts will score high on the Sgt portion of a Lieut exam
  - Original test could be musical ability or height
  - Only showing test reliability over time

# Passing Point: Take 1

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

## Passing Point: Take 2

- Situation: Angoff ratings used with a M/C exam for entry level job, e.g., firefighter
- Exam uses Fleishman ability areas
  - Inductive Reasoning
  - Deductive Reasoning
  - Memorization
  - Spatial Orientation
  - Visualization

## Passing Point: Take 2

- Attorney: SMEs have no experience on which to base judgements
- Task too complex

## Passing Point: Take 2

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

## Passing Point: Take 3

- Situation: SMEs who were making Angoff ratings were given applicant mean difficulty for each item.

## Passing Point: Take 3

- Attorney: Applicant test performance is unrelated to competence. Applicant data biased the Angoff ratings.

## Passing Point: Take 3

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Passing Point: Take 4

- Situation: Typical “modified” Angoff used to help set pass point

## Passing Point: Take 4

- Attorney: Angoff question is not logical
- “What percent of minimally competent” is less compelling than a clearer mastery oriented approach (“Is this essential to do the job?”)
- Question should be closer to a mastery approach: Is this needed to do job?

# Passing Point: Take 4

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Ranking: Take 1

- Situation: Multiple Angoff ratings collected for a M/C exam for entry level job, e.g., firefighter
- Collected Angoff ratings for:
  - Minimally competent employee
  - Fully acceptable employee
  - Superior employee

# Ranking: Take 1

- Attorney: Demand characteristics led raters to give higher Angoff ratings for superior than for acceptable employees

# Ranking: Take 1

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

## Ranking: Take 2

- Situation: Rank applicants based on a job-knowledge test for Police Sgt.
- Assume the job analysis found these knowledges to be important

## Ranking: Take 2

- Attorney and Experts: Sliver of the job
- Ranking focuses on small differences in test scores while there are large differences in KSAPs not measured, such as:
  - interpersonal skills
  - creative problem solving
  - planning
  - integrity

## Ranking: Take 2

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

## Ranking: Take 3

- Situation: Rank applicants based on a job-knowledge test for Police Sgt.
- Assume the job analysis found these knowledges to be important

## Ranking: Take 3

- Attorney and Experts: Ranking focuses on small differences in test scores.
- Such small differences are unreliable
- Such small differences have basically no utility

## Ranking: Take 3

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Readability: Take 1

- Situation: Some apparently unnecessarily difficult words used in test questions
  - “Aligned” in a spatial ability question
- Non-testing usage
- Airline safety instructions
  - “Placard”
  - “Lavatories”
  - “Obstruction”

# Readability: Take 1

- Attorney: No hard evidence that the word is difficult
- Word frequency dictionaries are from 1+ generations ago.

# Readability: Take 1

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Readability: Take 1

- Developed new word frequency database
  - NY Times
  - NY Daily News
  - Books on grade specific reading lists
- Web site in development
  - **<http://texteval.com> (free)**
  - Signup to be notified when functional

## Readability: Take 2

- Situation: Determine readability of training material and set exam reading level at no higher than that level
- Firefighter training academy manuals

## Readability: Take 2

- Critique: Can use dictionary and get assistance when reading training material.
- No help allowed when taking test

## Readability: Take 2

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Intercorrelation of KSAPs

- Situation: Test measured Fleishman ability areas
- Statistician found higher intercorrelations between items measuring different abilities than the same ability
- Sub-scores correlated

# Intercorrelation of KSAPs

- Attorney: Test is not measuring what it intended to measure

# Intercorrelation of KSAPs

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Validity Generalization

- Situation: Claim all tests valid for all jobs
- Some I/O psychologists very certain this is correct

# Validity Generalization

- Attorney: Uniform Guidelines don't recognize VG
- No employer would ever be liable under UG
  - Alternate I/O view: Minimize inferential leap

# Validity Generalization

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Video Tests

- Situation: Use of written m/c tests with no serious consideration of other testing mediums
- Some research suggests the testing medium affects AI
  - Schmitt & Quinn (2010, Table 16.2)

# Video Tests

- Attorney: No serious consideration of testing medium other than written M/C tests
- Claim that video-based testing is a medium that should be considered for all tests/jobs

# Video Tests

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Other Alternative Methods

- Situational Judgment Tests
- Structure Oral Interviews
- Face recognition
- Memory
- Oral communication with ethnic speakers

# Relying on Law or Agreements

- Situation: Components or weights for exam components determined by labor agreement or state or local law
- Defendant: Cannot ignore legal constraints of state law or labor agreement

# Relying on Law or Agreements

- Attorney: Federal law takes precedence.

# Relying on Law or Agreements

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Poor Research Accepted

- Situation: Job analysis with unbelievable findings.
- oral communication, interpersonal skills, problem identification/analysis, judgment, planning and organizing needed for tasks:
- Cares for and cleans weapons.
- Enters data into and accesses data from computer system.

# Poor Research Accepted

- 100% agreement on all 20,709 job analysis ratings
- 11 SMEs chosen to represent diverse job assignments

# Poor Research Accepted

- Attorney: Judge is not interested in details of research
- Judge will not understand.

# Poor Research Accepted

- Is this a problem?
  - If so, large or small?
- What can we learn from this situation?
- How to deal with this situation?

## 4/5 Rule Statistically Flawed

- Situation: Unpublished Monte Carlo study (Silva, 2010) suggests that, even with  $d = 0$ , AI often is less than .80

## 4/5 Rule Statistically Flawed

- Attorney: All jobs with small N would be exempt from Title VII

# 4/5 Rule Statistically Flawed

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

## *d* for Test Larger Than for Job

- Situation: B-W difference in job performance is smaller than B-W difference in test performance
- $d=.27$  for job performance
- $d=-.3$  for tenure
  - McKay & McDaniel (2005)
- $d = .72$  for job performance (Potosky, et al., 2005)

## *d* for Test Larger Than for Job

- Attorney: This is not fair
- Focus more on tenure as criterion?

## *d* for Test Larger Than for Job

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

## Other Soft Underbelly Areas

- Logic calls for multiple cut points
- Criteria may be biased
- Low job performance –job rating correlation
- Low utility of ranking based on small differences in test score
- Unknown weights for KSAPs in M/C test
  - No standardizing, just add number correct
- Training compensates for ability

## Other Soft Underbelly Areas

- Test scores may not even be ordinal
- Can ask, should we equally weight each item in a m/c exam?
- Items not equally important or valid
- Items intercorrelated, so weight some KSAPs higher than planned

# Multiple Passing Points

- Situation: Test measures multiple KSAPs (e.g., 5 KSAPs)
- The KSAPs are independent and very different from one another
- Should there be a passing point on each area?
- (Outtz, 2010, suggests this)

# Multiple Passing Points

- We do not do this.
- Why?

# Multiple Passing Points

- Is this a fair critique?
- Is this a problem?
  - If so, large or small?
- What can we learn from this critique?
- How to deal with this critique?

# Criteria May Be Biased

- Situation: There are indications that many criteria are biased
- Short people paid less than tall people
- Pretty people paid more than homely people (both genders)
- Women paid less than men

# Criteria May Be Biased

- Implicit tests of bias show that even people who claim not to be biased often show bias
- Studies of Race of Supervisor and Employee
  - Black and white supervisors rated white employees about the same
  - White supervisors rated black employees much lower than black supervisors did.
  - (Stauffer & Buckley, 2005)

# References

- McKay, P.F. & McDaniel, M.A. (2005) A Reexamination of Black–White Mean Differences in Work Performance: More Data, More Moderators. *Journal of Applied Psychology, 91*, 538- 554.

# References

- Outtz, J. L. (2010) Addressing the Flaws In Our Assessment Decisions. In J.C. Scott & D.H. Reynolds, (Eds.) *Handbook of Workplace Assessment*. Washington D.C.: SIOP.

# References

- Schmitt, N. & Quinn, A. (2010) Reductions in Measured Subgroup Mean Differences: What Is Possible? In Outtz, J.L. (Ed.) *Adverse Impact; Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge Taylor & Francis Group.

# References

- Potosky, D, Bobko, P. & Roth, P.L. (2005)  
Forming Composites of Cognitive Ability  
Alternative Measures to Predict Performance and  
Reduce Adverse Impact: Corrected Estimates and  
Realistic Expectations. *International Journal of  
Selection and Assessment*, 13, 304- 315.

# References

- Silva, J.M. (2010) *Report of Jacinto M. Silva, Ph.D. in PEDRO LOPEZ, et al., v. City of Lawrence, Massachusetts, et al., dated May 28, 2010.*
- Stauffer, J.M. & Buckley, M. R. (2005) The Existence and Nature of Racial Bias in Supervisory Ratings. *Journal of Applied Psychology, 90, 586- 591.*