Situational Judgment Tests & Disparate Impact: The Big Picture

Linda S. Gottfredson
University of Delaware
SIOP, Los Angeles
April 16, 2005
Fact-Set 1: Structure & Relation of Predictor & Criterion Domains

<table>
<thead>
<tr>
<th>Predictor domain</th>
<th>Criterion domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>Technical</td>
</tr>
<tr>
<td>Non-cognitive</td>
<td>Citizenship</td>
</tr>
</tbody>
</table>
Fact-Set 1: Structure & Relation of Predictor & Criterion Domains

<table>
<thead>
<tr>
<th>Predictor domain</th>
<th>Criterion domain</th>
<th>Simple to complex jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$</td>
<td>Technical</td>
<td></td>
</tr>
<tr>
<td>Non-$g$</td>
<td>Citizenship</td>
<td></td>
</tr>
</tbody>
</table>
**Fact-Set 2: Race & Sex Differences in g and Personality**

<table>
<thead>
<tr>
<th>Predictor domain</th>
<th>$d$: W-B</th>
<th>W-H</th>
<th>W-A</th>
<th>Male-Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$</td>
<td>1.1</td>
<td>.7</td>
<td>-.2</td>
<td>0</td>
</tr>
<tr>
<td>Non-$g$</td>
<td>~0</td>
<td>~0</td>
<td>?</td>
<td>~0</td>
</tr>
</tbody>
</table>

- No evidence of change over place or time
- $g$ (and $d$) not a function of content or format, but cognitive load

- + emotional stabil.
- - agreeable
- - conscientious
<table>
<thead>
<tr>
<th>Predictive validity:</th>
<th>Selection tests &amp; criterion measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$</td>
<td></td>
</tr>
<tr>
<td>Non-$g$</td>
<td></td>
</tr>
</tbody>
</table>

Disp impact—race:
Predictions From 2 Fact-Patterns

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Selection tests &amp; criterion measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive validity:</td>
<td></td>
</tr>
<tr>
<td>$g$</td>
<td>Tech</td>
</tr>
<tr>
<td>Non-$g$</td>
<td>Citizen</td>
</tr>
<tr>
<td>Predictive validity:</td>
<td></td>
</tr>
<tr>
<td>Disp impact—sex:</td>
<td></td>
</tr>
</tbody>
</table>

Disp impact—sex: [Diagram showing arrows between constructs and measures]
MA Results for SJT Predictors
(Nguyen, McDaniel, & Whetzel paper)

Predictive validity:

- $g$
- Non-$g$

Format: Video $\leftrightarrow$ Written
Response: Behavior $\leftrightarrow$ Knowledge

- Consistent with theory on $g$, $g$ load, & $g$ diffs
- But what constructs are formats capturing?
- What constructs do we want?

Disparate impact-race:

- Unwelcome questions for practice
- Can only interfere with picking on $d$
Tweaking Tests Won’t Help Much

Rules of thumb
1. Eliminating $d$ requires eliminating $g$
2. Eliminating $g$ reduces validity (would you want your doctor picked only on personality?)
3. Don’t-ask-don’t-tell governs discussion
4. Law, politics, & employer insist on $\sim 0 \ d$

So, new enthusiasm for changing the criteria
Race-driven, but an important question
MA Results for Performance Criteria (McKay & McDaniel paper)

Predictive validity ($g$ load):

- $g$
- Non-$g$

Contextual
- Task
- Overall rating
  - Work sample
  - Job knowledge

Again
- Consistent with theory on $g$, $g$ load, & $g$ difs
- But what *constructs* are measures capturing?
- What constructs do we want?

Disparate impact:

Will choice of criteria be race-driven?
Two MA Studies: Bottom Line

Conclusions
• Cognitive load is the major source of disparate impact (by race) in both predictors and criteria

Recommendations
• Avoiding the big picture?
  – Pick SJTs with lower $g$ load (but prudently)
• Avoiding the here-and-now?
  – Raise cognitive ability of lower-scoring groups
• But, less so than others. Show the big picture
SJT for College Admission

(Imus, Schmitt, Kim, Friede, & Oswald paper)

• Two similarities
  – Same basic $g$-$d$ tradeoffs in selection
  – That’s why “non-cognitive” predictors are being sought

• One difference
  – Women over-represented in college (60-40)

• College Board efforts—one of two teams highly competent
Academic SJT: Research Design and Results

- Why seek unidimensionality?
- What constructs captured?
- Is GPA the correct criterion?
- Does it select for female personality?
- What incremental validity?
- Maybe really a sex effect?

- Selection tests
  - SJT-36
  - SJT-15
  - GPA

- Performance measures

- Disparate Impact: W-B

- Disparate SJT-36
  - Disparate Impact: W-B
  - Disparate Impact: W-B
  - Disparate Impact: W-B

- GPA

- r=.17

- r=.20

- -.08

- -.03
Bottom Line

• **Conclusions**—Good news for SJTs in admissions
  – Biased items, but balanced so make no difference
  – Some validity
  – No disparate impact (by race)

• **Recommendations**—IRT can be useful with SJTs
  – Avoiding the big picture?
  – Proceeding as if didn’t have the 2 fact-sets?
  – Can expect same disappointments/tradeoffs as in personnel selection for race
  – Primary effect of SJTs may be to further reduce male representation
  – Problem is not a technical one
  – Its roots in $g$ will not be entertained first in this field (health is more promising)
• gottfred@udel.edu

• http://www.udel.edu/educ/gottfredson/