Using Item Response Theory to Assess Measurement Equivalence of 360° Performance Ratings Across Organizational Levels

S. Bartholomew Craig
North Carolina State University

Robert B. Kaiser
Kaplan DeVries Inc.


How to Get the Paper

► Download from www.sbcraig.com
► Request a copy via e-mail from bart_craig@ncsu.edu
► Leave your contact information after the presentations

Background

► 360° performance ratings have become the norm in organizations
  - Primary value lies in opportunity to compare ratings from different sources
  - Comparison implicitly assumes measurement equivalence

Crossing Organizational Levels

► Previous research on equivalence of rating sources does not address organizational levels
► Theoretical bases for possible inequivalence across levels
  - Differing skill requirements (Katz, 1955; Mann, 1965)
  - Application of structure (Katz & Kahn, 1978)
  - Interactive Complexity Theory (Streufert & Swezey, 1986)
  - Stratified Systems Theory (Jacobs & Jaques, 1987)
► Need for new research that disentangles rating source from organizational level

Method

► Archival database of Benchmarks® ratings
  - N = 85,342 raters / 8434 ratees
  - Only peers with unambiguous level used (N = 3763 raters / 974 ratees)
  - Response scale collapsed to four points
► EFA used to empirically derive factors
► CFA used to test between-factors equivalence
► DFIT used to test within-factor equivalence
**Results—EFA**

- EFA suggested 3 factors (83% of variance)
  - Decisive Action Orientation
    - "is action oriented"
  - Relationship Orientation
    - "has a warm personality that puts people at ease"
  - Working Through Others
    - "focuses more on managing other people to accomplish a task than on personally finishing everything the workgroup does"
- Highest loading 4 items from each factor used in subsequent analyses

**Results—CFA**

<table>
<thead>
<tr>
<th></th>
<th>CFI</th>
<th>NNI</th>
<th>RMSEA</th>
<th>$\chi^2$ / df</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Peers</td>
<td>.95</td>
<td>.93</td>
<td>.075</td>
<td>149</td>
</tr>
<tr>
<td>Lowest Level</td>
<td>.94</td>
<td>.93</td>
<td>.075</td>
<td>4.46</td>
</tr>
<tr>
<td>Middle Level</td>
<td>.95</td>
<td>.94</td>
<td>.072</td>
<td>12.4</td>
</tr>
<tr>
<td>Highest Level</td>
<td>.94</td>
<td>.92</td>
<td>.079</td>
<td>3.15</td>
</tr>
</tbody>
</table>

**Results—DFIT**

- All possible pairwise comparisons among levels were tested (3)
- 36 item-level tests conducted (NCDIF)
  - None was significant
- 9 scale-level tests conducted (DTF)
  - None was significant

**Interpretation**

- Conclusions
  - These 3 scales function equivalently at each of these three organizational levels
  - These 3 scales can be meaningfully compared
- Limitations / Future Research
  - Constructs not explicitly chosen to reflect dimensions thought to vary across levels
  - Small number of constructs
  - Small number of items per construct