Comparability of Results across Years

Richard Hill
Center for Assessment
CCSSO LSA Conference
Nashville, TN—June 18, 2007
Basic Issue

- Scores go up (or down) from one year to the next
- What might have caused that change?
  - One possibility is a change (up or down) in the effectiveness of the educational system
  - Another possibility is that something else changed
  - What are those “something elses?”
Requirements for Comparability

- Equivalent test
- Equivalent population
- Equivalent conditions
- Equivalent scoring
Equivalent Test

- Same framework
- Same test specifications
- Solid equating
If the Framework Has Changed

- Are the new questions measuring the same inherent skills as the old ones
  - If yes, continue trend line
  - If not, restart trend line
Determining If Changes Are Consequential

<table>
<thead>
<tr>
<th></th>
<th>Unique to Old</th>
<th>Common across Both</th>
<th>Unique to New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique to Old</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common across Both</td>
<td>N+1</td>
<td>N+2</td>
<td>...</td>
</tr>
<tr>
<td>Unique to New</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N+M+1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N+M+2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Determining If Changes Are Consequential

- Compute correlations of each item with each other item for all five cells
- Convert each correlation to a Fisher z
  \[ Z = .5 \times [ \ln(1+r) - \ln(1-r) ] \]
- Compute mean and standard deviation within the five cells
- Compare means across the five cells (dividing by standard deviations to get effect size)
## Determining If Changes Are consequential

<table>
<thead>
<tr>
<th></th>
<th>Unique to Old</th>
<th>Common</th>
<th>Unique to New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique to Old</td>
<td>A</td>
<td>B</td>
<td>X</td>
</tr>
<tr>
<td>Common across Both</td>
<td>X</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>Unique to New</td>
<td>X</td>
<td>X</td>
<td>E</td>
</tr>
</tbody>
</table>
Solid Equating

- All changes in total results are based *entirely* on changes in the equating items
- Were equating items presented identically?
  - Same item—no alterations
  - Same equivalent position
- Were equating items representative?
Solid Equating (cont’d)

- Was equating sample representative?
- Were any equating items deleting during the post-administration analysis?
Equivalence of Population

- Did enrollments change?
  - All students
  - Significant subpopulations

- Was there a change in the percentage of students tested?
  - All students
  - Significant subpopulations

- High school dropouts
Equivalence of Conditions

- **Stakes**
  - “Stakes changes everything”

- **Accommodation policies**
  - Inclusion rules
  - Accommodation rules

- **Time**
  - Testing time
  - Time of testing
    - School schedules
    - Year-round schools

- **Course-taking patterns**
Equivalence of Conditions (Cont’d)

- If concern, look at results district by district
Equivalence of Scoring

- Constructed-response questions
- Method of arriving at a reported score
Random Fluctuation

- Sampling error (small for most states)
- Standard error of equating (likely to be considerably larger)
- If error is larger than change, look for trends over a longer period of time
Thoughts? Suggestions?

- rhill@nciea.org
- Publication on www.nciea.org