Statistical and psychometric methods for measurement: G Theory, DIF, & Linking

Andrew Ho, Harvard Graduate School of Education

The World Bank, Psychometrics Mini Course 2

Washington, DC. June 27, 2018
Learning Objectives from Part I

• How do we “develop and validate” a scale?

  – What is validation?
  – What is reliability?
  – What is factor analysis?
  – What is Item Response Theory?
  – How do we do all this in Stata, and interpret the output accurately?
Learning Objectives for Part II

1. Generalizability Theory:
   – *How can we describe and improve the precision of teacher evaluation scores?*

2. Differential Item Functioning:
   – *How can we detect items that are not functioning well across countries?*

3. Linking:
   – *How can we compare scores across different tests?*
Sz-Shyan Wu is not a Cuban baseball star or a dissident musician. But in urging the United States government to grant him a work visa, the New York State Education Department is arguing that Mr. Wu, too, has talents so rare that bureaucracy must be cut and a red carpet rolled out.

Mr. Wu is a psychometrician or, in plain English, an expert on testing. And testing experts are in high demand.

With federal law requiring wider testing of schoolchildren, the nation faces a critical shortage of people like Mr. Wu with the mathematical, scientific, psychological and educational skills to create tests and analyze the results. The problem has sent states, testing companies and big school districts into a heated hiring competition, with test companies offering salaries as high as $200,000 a year or more plus perks.
In recent years, interest has grown in using observational systems as a means to several ends, including program evaluation, and impact evaluations. Although education policymakers have developed numerous observational systems, many developers fail to systematically assess the reliability of their instruments. In this article, the authors explore the question of whether observational systems can be reliable and cost-efficient scoring systems. The article describes a research project that might be developed and improved upon. One example that applies generalizability theory to observational instruments is the MET project, which uses a case study approach to improve observational systems in educational settings.
1. It’s about SCORES.
   - Reliability and precision are properties of SCORES not TESTS.

2. What is reliability?
   - Reliability is the correlation between scores across a replication:
     \[ \rho_{XX'} \]

3. The “Rule of 2”
   - If you don’t have at least two replications (of items, lessons, raters, occasions), you cannot estimate relevant variance.

4. The “G Study”
   - A multilevel model with crossed, random effects gives us “variance components.”

5. The “D Study”
   - Once we know variance components, we can increase score precision strategically, by averaging over replications.
• Canonical measurement data have a crossed-effects design:

\[ y_{ijk} \]

is a response replication \( i \) to item \( j \) by person \( k \).

• We are used to seeing many replications \( i \) (students in schools). In measurement we generally have only 1 replication per person-item combination, so we can drop the subscript.
The Measurement Model

- A response \( i \) to item \( j \) by person \( k \)

\[
y_{ijk} = \mu + \zeta_j + \zeta_k + \epsilon_{ijk};
\]

\[
\zeta_j \sim N(0, \psi_1);
\]

\[
\zeta_k \sim N(0, \psi_2);
\]

\[
\epsilon_{ijk} \sim N(0, \theta).
\]

- Note: Only 1 score per person/item combination

- \( \mu \) — Overall average score
- \( \zeta_j \) — Item location (easiness), \( \psi_1 \) — variance of item effects
- \( \zeta_k \) — Person location (proficiency), \( \psi_2 \) — variance of person effects
- \( \epsilon_{ijk} \) — Person-item interactions and other effects, \( \theta \) — error variance
Two Intraclass Correlations (Reliabilities)

- A response $i$ to item $j$ by person $k$:
  $$y_{ijk} = \mu + \zeta_j + \zeta_k + \epsilon_{ijk};$$
  $$\zeta_j \sim N(0, \psi_1);$$
  $$\zeta_k \sim N(0, \psi_2);$$
  $$\epsilon_{ijk} \sim N(0, \theta).$$

- $\mu$ — Overall average score
- $\zeta_j$ — Item location (easiness). Variance: $\psi_1$
- $\zeta_k$ — Person location (proficiency). Variance: $\psi_2$
- $\epsilon_{ijk}$ — Person-item interactions and other effects. Variance: $\theta$

- Intraclass correlation: $\rho = \frac{\psi_2}{\psi_2 + \theta}$. The correlation between two item responses within persons. The proportion of relative response variation due to persons.

- Intraclass correlation: $\rho_\alpha = \frac{\psi_2}{\psi_2 + \frac{\theta}{n_j}}$. Cronbach’s alpha: The correlation between two average (or sum) scores within persons. The proportion of relative score variance due to persons.
From Brennan (2002), 15 teachers, 3 raters, 5 items.

<table>
<thead>
<tr>
<th>person</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>rater and item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
From Brennan (2002), 15 teachers, 3 raters, 5 items.

See [here](http://www.ats.ucla.edu/stat/stata/faq/doublewide.htm) and the .do file for reshaping data from double/triple-wide formats.

```
local i = 1
foreach var of varlist ritem1-r3item5 {
  rename `var' score`i'
  local i = `i'+1
}
reshape long score, i(person) j(seq)
recode seq (1 6 11=1) (2 7 12=2) (3 8 13=3) (4 9 14=4) (5 10 15=5), gen(item)
recode seq (1/5=1) (6/10=2) (11/15=3), gen(rater)
drop seq
order person item rater score
label variable item ""
label variable rater ""
table person item rater, c(mean score)
save pxixr_long.dta, replace
```
The “$p \times i \times r$” Measurement Model

- A response by person $p$ to item $i$, rated by rater $r$:
  \[
y_{pir} = \mu + \zeta_p + \zeta_i + \zeta_r + \zeta_{pi} + \zeta_{pr} + \zeta_{ir} + \varepsilon_{pir,e};
  \]
  \[
  \zeta_p \sim N(0, \psi_p); \zeta_{pi} \sim N(0, \psi_{pi}); \zeta_{pr} \sim N(0, \psi_{pr})
  \]
  \[
  \zeta_i \sim N(0, \psi_i); \zeta_r \sim N(0, \psi_r); \zeta_{ir} \sim N(0, \psi_{ir})
  \]
  \[
  \varepsilon_{pir,e} \sim N(0, \theta).
  \]
- The Stata “mixed” command will estimate 7 variance components.
```stata
mixed score || _all: R.person || _all: R.item || _all: R.rater || _all: R.pXi
   || _all: R.pXr || _all: R.iXr, variance reml nolog

|          | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|----------|--------|-----------|-------|-------|----------------------|
| _cons    | 2.475556 | 0.344058  | 7.20  | 0.000 | 1.801214 3.149897    |

Random-effects Parameters

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>_all: Identity var(R.person)</td>
<td>0.3777783</td>
<td>0.1591872</td>
<td>0.1654093 0.8628078</td>
</tr>
<tr>
<td>_all: Identity var(R.item)</td>
<td>0.1344455</td>
<td>0.1046619</td>
<td>0.0292355 0.6182755</td>
</tr>
<tr>
<td>_all: Identity var(R.rater)</td>
<td>0.1876201</td>
<td>0.1972506</td>
<td>0.0238993 1.472899</td>
</tr>
<tr>
<td>_all: Identity var(R.pXi)</td>
<td>0.0422223</td>
<td>0.0273449</td>
<td>0.011865 0.1502505</td>
</tr>
<tr>
<td>_all: Identity var(R.pXr)</td>
<td>0.0490476</td>
<td>0.0281733</td>
<td>0.0159104 0.1512009</td>
</tr>
<tr>
<td>_all: Identity var(R.iXr)</td>
<td>0.0139683</td>
<td>0.0159899</td>
<td>0.0014816 0.1316866</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>0.2649205</td>
<td>0.0354015</td>
<td>0.2038774 0.3442406</td>
</tr>
</tbody>
</table>
```

© Andrew Ho
Harvard Graduate School of Education
How do we interpret variance components?

<table>
<thead>
<tr>
<th>Source</th>
<th>var</th>
<th>sd</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.3778</td>
<td>0.6146</td>
<td>35.31%</td>
</tr>
<tr>
<td>i</td>
<td>0.1344</td>
<td>0.3667</td>
<td>12.56%</td>
</tr>
<tr>
<td>r</td>
<td>0.1876</td>
<td>0.4332</td>
<td>17.53%</td>
</tr>
<tr>
<td>pi</td>
<td>0.0422</td>
<td>0.2055</td>
<td>2.04%</td>
</tr>
<tr>
<td>pr</td>
<td>0.0490</td>
<td>0.2215</td>
<td>2.37%</td>
</tr>
<tr>
<td>ir</td>
<td>0.0140</td>
<td>0.1182</td>
<td>0.58%</td>
</tr>
<tr>
<td>pir,e</td>
<td>0.2649</td>
<td>0.5147</td>
<td>10.08%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>Variance in teacher proficiency, persistent across items and raters</td>
</tr>
<tr>
<td>i</td>
<td>Item variance. Some item scores are higher across teachers and raters</td>
</tr>
<tr>
<td>r</td>
<td>Rater variance. Some rater scores are higher across teachers and items</td>
</tr>
<tr>
<td>pi</td>
<td>Some teachers score higher on some items than others, across raters.</td>
</tr>
<tr>
<td>pr</td>
<td>Some raters score some teachers higher than others, across items.</td>
</tr>
<tr>
<td>ir</td>
<td>Some raters score some items higher than others, across teachers.</td>
</tr>
<tr>
<td>pir,e</td>
<td>Some raters score some teachers higher on some items, and random error.</td>
</tr>
</tbody>
</table>
• Identify the variance components that lead to changes in relative position.
• Includes all variance components that intersect with \( p \).
• Why not \( \sigma_i^2 \)? Items that are more or less difficult... for everybody.
• Why not \( \sigma_r^2 \)?
• Why not \( \sigma_{ir}^2 \)?
• Variance components refer to single-unit replications, so we must divide by relevant numbers of items/raters to obtain error for average scores (over items/raters).
• Mnemonic: Divide by the \( n \)'s in subscript in the numerator (besides \( p \)).

\[
\sigma_\delta^2 = \frac{\zeta_{pi}}{n_i} + \frac{\zeta_{pr}}{n_r} + \frac{\zeta_{pir,e}}{n_in_r}
\]

\[
E\rho^2 = \frac{\zeta_p}{\zeta_p + \sigma_\delta^2}
\]
Identify the variance components that lead to changes in absolute position.

Includes all variance components besides $\sigma_p^2$.

Why include $\sigma_i^2$? Items that are more or less difficult for everybody will lead to changes in absolute position.

Why include $\sigma_r^2$?

Why include $\sigma_{ir}^2$?

$$\Phi = \frac{\zeta_p}{\zeta_p + \sigma_\Delta^2}$$
D Study for the Relative Standard Error of Measurement, $\hat{\sigma}_\delta$

$$\sigma^2_\delta = \frac{\zeta_{pi}}{n_i} + \frac{\zeta_{pr}}{n_r} + \frac{\zeta_{pir,e}}{n_in_r}$$

- 1 Raters
- 2 Raters
- 3 Raters
- 4 Raters
- 5 Raters

© Andrew Ho

Harvard Graduate School of Education  18
D Study for the Generalizability Coefficient for Relative Error, $\mathbf{E} \hat{\rho}^2$

$$\mathbf{E} \rho^2 = \frac{\zeta_p}{\zeta_p + \sigma_{\delta}^2}$$

Note Excel Template

© Andrew Ho
Harvard Graduate School of Education
Table 10

<table>
<thead>
<tr>
<th>SCENARIO</th>
<th>Rater Combination</th>
<th>Lesson 1</th>
<th>Lesson 2</th>
<th>Lesson 3</th>
<th>Lesson 4</th>
<th>Lesson 5</th>
<th>Lesson 6</th>
<th>Total Full Observations</th>
<th>Implied Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCENARIO 1</td>
<td>Own Administrator</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer (Full Obs)</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCENARIO 2</td>
<td>Own Administrator</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer (Full Obs)</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCENARIO 3</td>
<td>Own Administrator</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>.69</td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer 1 (Full Obs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer 2 (Full Obs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCENARIO 4</td>
<td>Own Administrator</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer 1 (Full Obs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer 2 (15 Min)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer 3 (15 Min)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Same-Cert Peer 4 (15 Min)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Generalizability Theory:
   – How can we describe and improve the precision of teacher evaluation scores?

2. Differential Item Functioning:
   – How can we detect items that are not functioning well across countries?

3. Linking:
   – How can we compare scores across different tests?
1. Define a “matching criterion” internal or external to the item or test whose differential functioning you wish to assess.
   – Generally a total test score (or $\theta$), an external test score, a test score minus the target item (“rest” score), or a chosen subset of items that serve as a reference.

2. Examine group differences conditional on (at identical values of) the matching criterion.
   – Finding a trustworthy matching criterion is a challenge.
   – It must be free of differential functioning itself, lest it distort the estimation of differential functioning.

3. Flag items with DIF for review by a content committee.
   – Even if you find DIF, it may be construct-relevant!
   – As in Mini Course I, content is king.
If you use a total score (or $\theta$) to match, DIF ends up being relative. DIF across all items will average to (approximately) 0.
Uniform DIF
Conditional on theta, this item is more difficult for the blue group than the red group.

Nonuniform DIF
Conditional on theta, the item is more difficult for the blue group than the red group, particularly for higher scoring examinees.
The Cochran-Mantel-Haenszel Test

- Stratify people by total score (the matching criterion)

- Within each stratum, tabulate 2x2 right-wrong by group and calculate the odds ratio within each matched value.

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>$n_{11k}$</td>
<td>$n_{12k}$</td>
<td>$n_{1.k}$</td>
</tr>
<tr>
<td>Focal</td>
<td>$n_{21k}$</td>
<td>$n_{22k}$</td>
<td>$n_{2.k}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{.1k}$</td>
<td>$n_{.2k}$</td>
<td>$n_k$</td>
</tr>
</tbody>
</table>

- Combine the ratios (weighted average odds ratio over $k$)
  - Some report the common odds ratio (1 under null).
  - Others report the log of this odds ratio (0 under null).

$$OR_{MH} = \frac{\sum_{k=1}^{K-1} n_{11k}n_{22k}/n_k}{\sum_{k=1}^{K-1} n_{12k}n_{21k}/n_k}$$

- This test statistic is chi-square distributed with 1 df
MH Test on a Math/Science test, by gender

- Large $\chi^2$ values (above $1.96^2$) indicate statistically significant differences.
- Odds ratios over 1 favor females.
- Odds of a female getting q1 correct is 60% [$OR - 1$] higher than males conditional on total score.
- If $P_m$ were 50%, $P_f$ would be $\left[\frac{OR}{1+OR}\right] = 62%$

<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>Logits log(OR)</th>
<th>50% to... (OR/(1+OR))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>50%</td>
</tr>
<tr>
<td>1.5</td>
<td>0.4</td>
<td>60%</td>
</tr>
<tr>
<td>2.0</td>
<td>0.7</td>
<td>67%</td>
</tr>
<tr>
<td>2.5</td>
<td>0.9</td>
<td>71%</td>
</tr>
<tr>
<td>3.0</td>
<td>1.1</td>
<td>75%</td>
</tr>
<tr>
<td>3.5</td>
<td>1.3</td>
<td>78%</td>
</tr>
<tr>
<td>4.0</td>
<td>1.4</td>
<td>80%</td>
</tr>
<tr>
<td>4.5</td>
<td>1.5</td>
<td>82%</td>
</tr>
<tr>
<td>5.0</td>
<td>1.6</td>
<td>83%</td>
</tr>
<tr>
<td>5.5</td>
<td>1.7</td>
<td>85%</td>
</tr>
<tr>
<td>6.0</td>
<td>1.8</td>
<td>86%</td>
</tr>
<tr>
<td>6.5</td>
<td>1.9</td>
<td>87%</td>
</tr>
<tr>
<td>7.0</td>
<td>1.9</td>
<td>88%</td>
</tr>
<tr>
<td>7.5</td>
<td>2.0</td>
<td>88%</td>
</tr>
<tr>
<td>8.0</td>
<td>2.1</td>
<td>89%</td>
</tr>
<tr>
<td>8.5</td>
<td>2.1</td>
<td>89%</td>
</tr>
</tbody>
</table>
Nonuniform DIF in a MH context

• The `mhodds` command gives a test of nonuniform DIF for any particular item, as well.

<table>
<thead>
<tr>
<th>Odds Ratio</th>
<th>chi2(1)</th>
<th>P&gt;chi2</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.605273</td>
<td>12.94</td>
<td>0.0003</td>
<td>1.237327 2.082636</td>
</tr>
</tbody>
</table>

Mantel-Haenszel estimate controlling for sumscore

Test of homogeneity of ORs (approx): chi2(7) = 4.69
Pr>chi2 = 0.6973

• We conclude that DIF is uniform in favor of females for Item 1, conditional on the sum score on the test.

• If it were nonuniform, use graphs to describe the direction.
Logistic Regression Approaches (Item 4) - Uniform

Conditional on the leave-one-out score on this test, we estimate that males perform .32 logits better than females on Item 4. Odds of a correct answer are \( \exp(.32) = 1.38 \), 38% better for males.
\[
\log \left( \frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 \text{Score} + \beta_2 \text{Female} + \beta_3 \text{Score} \times \text{Female}
\]

|               | Coef.  | Std. Err. | z      | P>|z| |
|---------------|--------|-----------|--------|------|
| sumscore_nq4  | 0.215  | 0.043     | 4.92   | 0.000|
| 1.female      | -0.994 | 0.317     | -3.13  | 0.002|
| female#c.sumscore_nq4 1 | 0.155  | 0.068     | 2.27   | 0.023|
| _cons         | -1.091 | 0.207     | -5.27  | 0.000|
IRT Approaches to DIF (See Camilli, 2006)

• Fit a model with different $b$ parameters for groups
  – Conduct a $z$-test for differences in $b$ parameter estimates.
  – Likelihood ratio tests constraining group parameters to be equal (vs. not).
  – Also: Calculate areas between item characteristic curves.

• There are also Structural Equation Modeling approaches to DIF. See Stata examples 22 & 23 for continuous (not binary) responses, here.
From DIF to Fairness (See Camilli, 2006)

- Fairness in treatment; fairness as lack of measurement bias, fairness in access to the construct, fairness as validity of score interpretations for intended uses.

- Sensitivity Review - From Bond, Moss, and Carr (1996), “a generic term for a set of procedures for ensuring 1) that stimulus materials used in assessment reflect the diversity in our society and the diversity of contributions to our culture, and 2) that the assessment stimuli are free of wording and/or situations that are sexist, ethnically insensitive, stereotypic, or otherwise offensive to subgroups of the population.”

- AERA/APA/NCME Standard 3.2, “Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests’ being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.”
IV. Enough, for Now

Earlier, I suggested that we already had enough psychometrics for current purposes, and efforts in other directions would be more likely to bear fruit. This does not mean that no one should work in these areas, but only that the primary focus of the field should, in my view, shift in other directions. In addition, I most expressly do not mean that we should not apply these methods to current problems, only that further research into their expansion and continued development should be of lower priority. Some areas that could profit from some neglect are:

11. Differential item functioning (DIF)—There are, fundamentally, two approaches developed for studying DIF; observed score methods, of which the Mantel-Haenszel statistic (Holland & Thayer, 1988) is both the best known and the best performing, and model-based methods, in which likelihood-ratio tests provide the probability of DIF (Thissen, Steinberg, & Wainer, 1988, 1993). These two approaches are more than enough to suit virtually any occasion. Journal editors I have spoken with admitted to the same feeling I had as an editor when a new submission arrived on yet another simulation showing the sensitivity of some DIF method to one variation or another of parameter distributions. Enough already!
1. Generalizability Theory:  
   - How can we describe and improve the precision of teacher evaluation scores?

2. Differential Item Functioning:  
   - How can we detect items that are not functioning well across countries?

3. Linking:  
   - How can we compare scores across different tests?
• How can we compare the average academic performance of districts in different states?

• How can we compare the average academic progress of districts in different states?
Mathematics, grade 4
Difference in average scale scores between all jurisdictions and National public, for All students [TOTAL], 2017

National Assessment of Educational Progress (NAEP)
How are these cross-state “linkages” valid?!

- Hanushek and Woessman (2012) link international assessments to the NAEP scale.
- Bandeira de Mello et al. (2011, 2013, 2015) map state proficiency standards to the NAEP scale.
- Kolen and Brennan (2014) review linking methods.
- In the NRC Report, *Uncommon Measures*, Feuer et al. (1999) recommend against a NAEP as a common national measure for student-level assessment.
- We link district distributions to the NAEP scale to support district-level policy analysis.
- We will treat the issue empirically, using a series of validation checks. How?
When the essential counterfactual question can be answered for at least some of the target units.
What do we need to link test score scales?

**Common persons**

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Common populations**

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>1</td>
<td>✓</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Common items**

<table>
<thead>
<tr>
<th>Population</th>
<th>Sample</th>
<th>X</th>
<th>A</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>1</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Q</td>
<td>2</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Holland and Dorans (2006)

We use common-population linking: NAEP provides state estimates
Notation

Given each state’s NAEP mean and SD, $\mu_{\text{naep}}$ and $\sigma_{\text{naep}}$, interpolating for even years $y$ and grades $g \in \{3, 5, 6, 7\}$.

And each district’s relative mean and SD in its own state test, on a standardized scale, $\mu_{\text{state}}$ and $\sigma_{\text{state}}$:

Estimate each district’s absolute position on the NAEP scale, $\mu_{\text{naep}}$ and $\sigma_{\text{naep}}$. 
Linear Linking

State A standardized distribution (Solid Line):
Mean = 0, SD = 1
District distributions (Dotted Lines)

State A NAEP Distribution (Orange Line):
Mean = 210, SD = 33
District NAEP-Linked Distributions (Dotted Lines)

Linearly rescale district distributions (dotted lines), such that the overall state distribution matches the state distribution on NAEP

© Andrew Ho
Harvard Graduate School of Education
Linear Linking

State B standardized distribution (Solid Line): Mean = 0, SD = 1
State B NAEP Distribution (Purple Line): Mean = 280, SD = 20

District Distributions (Dotted Lines)

Lineraly rescale district distributions (dotted lines), such that the overall state distribution matches the state distribution on NAEP.
How can we correct standardized group means for imperfect reliability, \( \hat{\rho}_{sygb} \)?

Reliability \( \rho = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \).

Observed SDs of 1 overestimate “true” SDs of \( \sqrt{\rho} \), so a district’s mean should be: \( \hat{\mu}_{dygb} / \sqrt{\hat{\rho}_{sygb}} \)

Mean linkage: \( \hat{\mu}_{dygb} = \hat{\mu}_{sygb} + \frac{\hat{\mu}_{dygb}}{\sqrt{\hat{\rho}_{sygb}}} * \hat{\sigma}_{sygb} \)
How can we correct standardized group SDs for imperfect reliability, $\hat{\rho}_{sygb}^{\text{state}}$?

If observed SDs are 1, error variance is:

$$\sigma^2_e = 1 - \rho.$$

Decompose a district’s observed variance:

$$\sigma^2_d = \sigma^2_{Td} + \sigma^2_e$$

So the true district SD on a standardized scale is:

$$\sigma_{Td} = \sqrt{\sigma^2_d - (1 - \rho)}$$

And the true district SD in terms of true state SD units, $\sqrt{\rho}$, is:

$$\sigma_{Td} = \sqrt{\sigma^2_d - (1 - \rho)} / \rho$$
Reliability-adjusted linear linking equations:

\[
\hat{\mu}_{dygb}^{naep} = \hat{\mu}_{sygb}^{naep} + \frac{\hat{\mu}_{dygb}^{state}}{\sqrt{\hat{\rho}_{sygb}^{state}}} \ast \hat{\sigma}_{sygb}^{naep}
\]

\[
\hat{\sigma}_{dygb}^{naep} = \left[ \left( \hat{\sigma}_{dygb}^{state} \right)^2 + \hat{\rho}_{sygb}^{state} - 1 \right]^{1/2} \ast \hat{\sigma}_{sygb}^{naep}
\]
How can we validate the linking?
How can we validate the linking? G4.
How can we validate the linking? G8.
How can we precision-adjust estimates of linked-vs.-true bias, RMSE, and correlations?

\[ \hat{\mu}_{idygb} = \alpha_{0idygb} \left( \hat{\mu}_{idygb}^{naep} \right) + \alpha_{1idygb} \left( \hat{\mu}_{idygb}^{naep} \right) + e_{idygb} \]

\[ \alpha_{0idygb} = \beta_{00} + u_{0idygb} \]

\[ \alpha_{1idygb} = \beta_{10} + u_{1idygb} \]

\[ e_{idygb} \sim N \left( 0, \omega^{2}_{idygb} \right); \ u_{dygb} \sim MVN \left( 0, \tau^{2} \right) \]

\[ \tau^{2} = \begin{bmatrix} \tau_{00}^{2} & \tau_{01}^{2} \\ \tau_{01}^{2} & \tau_{11}^{2} \end{bmatrix} \]

Bias: \[ \hat{B} = \hat{\beta}_{00} - \hat{\beta}_{10} \]

\[ \text{RMSE} = \left[ \hat{B}^{2} + b \hat{\tau}^{2} b' \right]^{1/2} \]

Correlation: \[ \hat{r} = \frac{\hat{\tau}_{01}^{2}}{\hat{\tau}_{00} \hat{\tau}_{11}} \]
Linking error in cross-state district comparisons?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Year</th>
<th>n</th>
<th>RMSE</th>
<th>Bias</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td>2009</td>
<td>17</td>
<td>3.95</td>
<td>2.12</td>
<td>0.96</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td>2011</td>
<td>20</td>
<td>3.69</td>
<td>1.25</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013</td>
<td>20</td>
<td>2.62</td>
<td>0.20</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2009</td>
<td>17</td>
<td>2.92</td>
<td>1.12</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011</td>
<td>20</td>
<td>2.20</td>
<td>0.63</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013</td>
<td>20</td>
<td>3.62</td>
<td>1.67</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2009</td>
<td>17</td>
<td>6.09</td>
<td>4.10</td>
<td>0.93</td>
</tr>
<tr>
<td>Math</td>
<td>4</td>
<td>2011</td>
<td>20</td>
<td>4.97</td>
<td>2.60</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013</td>
<td>20</td>
<td>3.60</td>
<td>1.46</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2009</td>
<td>14</td>
<td>5.21</td>
<td>3.40</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011</td>
<td>17</td>
<td>3.77</td>
<td>2.09</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013</td>
<td>14</td>
<td>4.54</td>
<td>1.47</td>
<td>0.94</td>
</tr>
<tr>
<td>Average</td>
<td>2009</td>
<td>14-17</td>
<td></td>
<td>4.70</td>
<td>2.69</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>2011</td>
<td>17-20</td>
<td></td>
<td>3.79</td>
<td>1.64</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>14-20</td>
<td></td>
<td>3.66</td>
<td>1.20</td>
<td>0.95</td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td>17-20</td>
<td></td>
<td>3.23</td>
<td>1.17</td>
<td>0.96</td>
</tr>
<tr>
<td>Math</td>
<td></td>
<td>14-20</td>
<td></td>
<td>4.77</td>
<td>2.52</td>
<td>0.95</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>14-20</td>
<td></td>
<td>4.07</td>
<td>1.84</td>
<td>0.95</td>
</tr>
</tbody>
</table>
When comparing districts across states, what are five reasons why a high-scoring SEDA linked district may not actually have a high NAEP score, had one been reported?

1. Student sampling.
   - That district samples its higher achieving students, (a) disproportionately over other districts in the state; and, (b) disproportionately on the state test over NAEP

2. Student motivation.

3. Tested Content.

4. Inflation.

5. Cheating.

NOTE: All explanations involve a difference (between districts) in differences (between tests)
Linking U.S. School District Test Score Distributions to a Common Scale

Author/s: Sean F. Reardon, Demetra Kalogrides, Andrew Ho

Year of Publication: 2017

There is no comprehensive database of U.S. district-level test scores that is comparable across states. We describe and evaluate a method for constructing such a database. First, we estimate linear, reliability-adjusted linking transformations from state test score scales to the scale of the National Assessment of Educational Progress (NAEP). We then develop and implement direct and indirect validation checks for linking assumptions. We conclude that the linking method is accurate enough to be used in analyses of national variation in district achievement, but that the small amount of linking error in the methods renders fine-grained distinctions among districts in different states invalid. Finally, we describe several different methods of scaling and pooling the linked scores to support a range of secondary analyses and interpretations.

https://cepa.stanford.edu/sites/default/files/wp16-09-v201706.pdf
Learning Objectives for Part II

1. Generalizability Theory:
   – *How can we describe and improve the precision of teacher evaluation scores?*

2. Differential Item Functioning:
   – *How can we detect items that are not functioning well across countries?*

3. Linking:
   – *How can we compare scores across different tests?*