An Iterative Procedure to Detect Item Parameter Drift in Equating Items

Jing Jiang, Boston College
Louis Roussos, Measured Progress
Lei Yu, Measured Progress
Outline

- Background
- Objective
- Methodology
- Real Data Analysis
- Simulation Study
- Summary
Pre-Equating

- The raw-to-scaled score table is created before operational test administration.
- Commonly used when
  - it is important to obtain scores shortly after the test;
  - testing lasts for a long period that it is not practical to wait for all the data to be collected.
- Item parameters may drift from bank values since pre-equating is conducted without knowing how items performed on operational tests (Bejar & Wingersky, 1982; Eignor, 1985; Kolen & Harris, 1990; Tong, Wu & Xu, 2008; Gao, He & Ruan, 2012)
Post-Equating

- The raw-to-scaled score table is created after operational test and after all the data have been collected.
- Item parameters are re-estimated using bank item parameters, new item parameters, and a scale transformation procedure.
  - The transformation should be based on a set of equating items that are invariant in terms of item parameters across tests
Background (cont.)

- **Equating Items**
  - Basis for accurate equating
  - Approaches to detect item parameter drift (IPD) in equating items
    - **Bivariate plot** (e.g., bb-plot, aa-plot)
    - **Delta plot**
    - Robust Z statistic
    - Displacement statistic
    - Difference between ICCs
Objective

- This study aims to propose a new method to detect item parameter drift in equating items, and select stable equating items to ensure the accuracy of post-equating results.
Methodology

Step 1 • Item Calibration
Step 2 • Scale Transformation
Step 3 • IPD Detection
Step 4 • Iterative Procedure
Step 5 • Final Check of Equating Items
Methodology (cont.)

- **Step 1: Item Calibration**
  - Marginal maximum likelihood estimation
  - Three-parameter logistic IRT model

\[ P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + \exp\left(-a_i(\theta_j - b_i)\right)} \]
Step 2: Scale Transformation

Stocking-Lord (SL) transformation

\[
SL_{diff}(\theta_i) = \left[ \sum_{j: V} p_{ij}(\theta_{ji}, \hat{a}_{jj}, \hat{b}_{jj}, \hat{c}_{jj}) - \sum_{j: V} p_{ij} \left( \theta_{ji}, \frac{\hat{a}_{jj}}{A}, A\hat{b}_{jj} + B, \hat{c}_{jj} \right) \right]^2
\]

The estimation proceeds by finding the combination of slope A and intercept B that minimizes

\[
SL_{crit} = \sum_i SL_{diff}(\theta_i)
\]
Methodology (cont.)

- **Step 2: Scale Transformation**
  - Since we don’t have pre-knowledge about the equating items, when generating the SL transformation, all items but the studied item (i.e., all other items) will be served as equating items.
  - For each item, their post-operational test item parameter estimates will be adjusted and put on the pre-equated test scale by using

\[
\begin{align*}
    a_{i,\text{post on pre}} &= \frac{a_{i,\text{post}}}{A_i} \\
    b_{i,\text{post on pre}} &= A_i b_{i,\text{post}} + B_i
\end{align*}
\]
Methodology (cont.)

- **Step 3: IPD Detection**
  - Weighted expected score differences between reference (i.e., pre-equated test) and focal (i.e., post-operational test) groups

\[
\beta = \int \left( P(\theta, R) - P(\theta, F) \right) f_F(\theta) d\theta
\]

| $|\beta| < 0.05$ | negligible |
|-----------------|------------|
| $0.05 \leq |\beta| < 0.1$ | moderate   |
| $|\beta| > 0.1$ | large      |
Methodology (cont.)

- **Step 4: Iterative Procedure**
  - Step 2 (Scale Transformation) and Step 3 (IPD Detection) will be repeated.
  - The iterative procedure will **stop** until
    - no more items are detected as drifting items;
    - or there are less than 20 percent of items.
Methodology (cont.)

- **Step 5: Final Check of Equating Items**
  - The final equating set will be used to generate the Stocking-Lord transformation.
  - All item parameter estimates of post-operational test will be transformed to the pre-equated test scale by applying the SL slope and intercept.
Real Data Analysis

- Objectives
  - Compare the final equating set obtained by the beta analysis with the one that was actually operationally used;
  - Find IPD patterns to help inform the simulation study.
Real Data Analysis (cont.)

- Data: A statewide test in 2014-2015

### Table 1. Raw Score Associated with Cut Points

<table>
<thead>
<tr>
<th></th>
<th>History</th>
<th></th>
<th>Science</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predictive</td>
<td>Actual</td>
<td>Predictive</td>
<td>Actual</td>
</tr>
<tr>
<td>U-LK</td>
<td>23</td>
<td>23</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>Lk-P/S</td>
<td>28</td>
<td>28</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>P/S-A</td>
<td>39</td>
<td>40</td>
<td>33</td>
<td>38</td>
</tr>
<tr>
<td>Max Score</td>
<td>60</td>
<td>60</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>
### Real Data Analysis (cont.)

- **Comparison of Equating Sets**

#### Table 2. Number of Non-Drifting(Drifting) Items in the Final Equating Set and Corresponding $\beta$ Values

<table>
<thead>
<tr>
<th>History</th>
<th>Number of Items</th>
<th>$\beta$</th>
<th></th>
<th>Science</th>
<th>Number of Items</th>
<th>$\beta$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical Selection</strong></td>
<td>17 (1)</td>
<td>-0.048~0.049</td>
<td>(-0.085)</td>
<td>16 (1)</td>
<td>-0.048~0.047</td>
<td>(-0.068)</td>
<td></td>
</tr>
<tr>
<td><strong>Iterative Procedure</strong></td>
<td>22 (0)</td>
<td>-0.020~0.023</td>
<td>(N/A)</td>
<td>8(0)</td>
<td>-0.037~0.493</td>
<td>(N/A)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16 (9)*</td>
<td>0.041~0.095</td>
<td>(0.102~0.189)</td>
</tr>
</tbody>
</table>

* using the criterion of $|\beta| < 0.10
Real Data Analysis (cont.)

- Item Parameter Drift Patterns: 105 items

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Number (Percentage) of Items</th>
<th>Number of Drifting Items</th>
<th>a-drift</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.2</td>
<td>82 (78%)</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>0.2 – 0.4</td>
<td>19 (18%)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>&gt; 0.4</td>
<td>4 (4%)</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Number (Percentage) of Items</th>
<th>Number of Drifting Items</th>
<th>b-drift</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.4</td>
<td>62 (59%)</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>0.4 – 0.8</td>
<td>16 (15%)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0.8 – 1.2</td>
<td>6 (6%)</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1.2 – 1.6</td>
<td>13 (12%)</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>&gt;1.6</td>
<td>8 (8%)</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Item Parameter Drift Patterns
Simulation Study

- **Objective**
  - Examine the power and usability of the iterative procedure in detecting item parameter drift and selecting stable equating items.
Simulation Study (cont.)

- **Manipulated Conditions**
  1) Test Length: 50 items
  2) Sample Size: 2000
  3) Direction: unidirectional
  4) Types of IPD: b-drift only
  5) Percentage of drifting items:
     - 20%
     - 40%
  6) Distribution of IPD magnitudes:
     - U(0, 0.4)
     - U(0, 0.8)
     - U(0, 1.2)
     - U(0, 1.6)
     - U(0, 2.0)
Simulation Study (cont.)

Table 4. Simulation Conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>Percentage of Drifting Items</th>
<th>Distribution of IPD Magnitudes</th>
<th>Magnitudes of IPD (-) (from smallest to largest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>20% (10 items)</td>
<td>U(0, 0.4)</td>
<td>0.023 0.071 0.075 0.111 0.125 0.264 0.271 0.303 0.362 0.394</td>
</tr>
<tr>
<td>02</td>
<td></td>
<td>U(0, 0.8)</td>
<td>0.028 0.129 0.192 0.309 0.361 0.465 0.471 0.535 0.634 0.780</td>
</tr>
<tr>
<td>03</td>
<td></td>
<td>U(0, 1.2)</td>
<td>0.099 0.132 0.235 0.349 0.576 0.661 0.770 0.887 1.020 1.091</td>
</tr>
<tr>
<td>04</td>
<td></td>
<td>U(0, 1.6)</td>
<td>0.093 0.307 0.353 0.495 0.725 0.809 1.134 1.210 1.384 1.476</td>
</tr>
<tr>
<td>05</td>
<td></td>
<td>U(0, 2.0)</td>
<td>0.115 0.462 0.616 0.820 0.933 1.032 1.394 1.599 1.607 1.899</td>
</tr>
<tr>
<td>06</td>
<td></td>
<td>U(0, 0.4)</td>
<td>0.033 0.070 0.074 0.105 0.120 0.134 0.135 0.165 0.171 0.194</td>
</tr>
<tr>
<td>07</td>
<td></td>
<td>U(0, 0.8)</td>
<td>0.082 0.141 0.152 0.184 0.213 0.217 0.231 0.238 0.283 0.292</td>
</tr>
<tr>
<td>08</td>
<td></td>
<td>U(0, 1.2)</td>
<td>0.063 0.095 0.115 0.242 0.267 0.275 0.303 0.306 0.531 0.544</td>
</tr>
<tr>
<td>09</td>
<td></td>
<td>U(0, 1.6)</td>
<td>0.041 0.065 0.126 0.173 0.358 0.406 0.472 0.683 0.797 0.886</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>U(0, 2.0)</td>
<td>0.006 0.138 0.207 0.476 0.568 0.593 0.638 0.694 0.797 0.830</td>
</tr>
</tbody>
</table>

Note: IPD stands for Internal Process Delay.
Simulation Study (cont.)

(1) the percentage of times that each drifting item was detected
(2) The proportion of drifting items that were used as equating items in the final equating set

- Condition 01-05: 20%
- Condition 06-10: 40%

<table>
<thead>
<tr>
<th>Condition</th>
<th>Distribution of IPD</th>
<th>Expected</th>
<th>Actual</th>
<th>Condition</th>
<th>Distribution of IPD</th>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>U(0, 0.4)</td>
<td>1.0</td>
<td>0.778</td>
<td>06</td>
<td>U(0, 0.4)</td>
<td>1.0</td>
<td>0.840</td>
</tr>
<tr>
<td>02</td>
<td>U(0, 0.8)</td>
<td>0.5</td>
<td>0.406</td>
<td>07</td>
<td>U(0, 0.8)</td>
<td>0.5</td>
<td>0.496</td>
</tr>
<tr>
<td>03</td>
<td>U(0, 1.2)</td>
<td>0.333</td>
<td>0.301</td>
<td>08</td>
<td>U(0, 1.2)</td>
<td>0.333</td>
<td>0.363</td>
</tr>
<tr>
<td>04</td>
<td>U(0, 1.6)</td>
<td>0.25</td>
<td>0.231</td>
<td>09</td>
<td>U(0, 1.6)</td>
<td>0.25</td>
<td>0.244</td>
</tr>
<tr>
<td>05</td>
<td>U(0, 2.0)</td>
<td>0.2</td>
<td>0.089</td>
<td>10</td>
<td>U(0, 2.0)</td>
<td>0.2</td>
<td>0.199</td>
</tr>
</tbody>
</table>
(2) The proportion of drifting items that were used as equating items in the final equating set.
(3) Hit rate: the proportion of drifting items that are correctly detected as drifting items

<table>
<thead>
<tr>
<th>Condition</th>
<th>Distribution of IPD</th>
<th>Expected</th>
<th>Actual</th>
<th>Condition</th>
<th>Distribution of IPD</th>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>U(0, 0.4)</td>
<td>0</td>
<td>0.221</td>
<td>06</td>
<td>U(0, 0.4)</td>
<td>0</td>
<td>0.153</td>
</tr>
<tr>
<td>02</td>
<td>U(0, 0.8)</td>
<td>0.5</td>
<td>0.591</td>
<td>07</td>
<td>U(0, 0.8)</td>
<td>0.5</td>
<td>0.512</td>
</tr>
<tr>
<td>03</td>
<td>U(0, 1.2)</td>
<td>0.666</td>
<td>0.696</td>
<td>08</td>
<td>U(0, 1.2)</td>
<td>0.666</td>
<td>0.614</td>
</tr>
<tr>
<td>04</td>
<td>U(0, 1.6)</td>
<td>0.75</td>
<td>0.767</td>
<td>09</td>
<td>U(0, 1.6)</td>
<td>0.75</td>
<td>0.679</td>
</tr>
<tr>
<td>05</td>
<td>U(0, 2.0)</td>
<td>0.8</td>
<td>0.899</td>
<td>10</td>
<td>U(0, 2.0)</td>
<td>0.8</td>
<td>0.752</td>
</tr>
</tbody>
</table>
(3) Hit rate: the proportion of drifting items that are correctly detected as drifting items
Simulation Study (cont.)

(4) False alarm rate: the proportion of non-drifting items that are incorrectly detected as drifting items.
Discussion

- Traditional approaches to detect IPD (e.g., bb-plot, delta plot, robust z statistic) usually compare parameter estimates before linking the pre-equated and post-operational test forms.
- This paper introduces a new iterative procedure to detect IPD after linking two test forms, which provides a more accurate classification of drifting and non-drifting items.
Discussion (cont.)

- This method is effective
  - when 20% or 40% of items functioned differently across test administrations;
  - when there exists large b-drift with drift magnitude up to 2;
  - when there is few pre-knowledge about the equating items.
- It is recommended to conduct this iterative procedure after using traditional methods (e.g., aa-plot, bb-plot, delta plot).
- However, it is more difficult to detect drifting items when 60% of items or more experience item parameter drift, since the scale would be highly contaminated in this case, and it is hard to locate stable items.
Thank you!

jiangjc@bc.edu
Real Data Analysis

- **Comparison of Equating Sets**
  - **Empirical Selection**
    - Criteria for selecting initial equating set
      - Position shift
      - Recent item bank parameters
      - Operational tests v.s. field tests
      - Stability of parameters (e.g., not too large of a shift between 12/13 and 13/14)
      - At least 20% of the total test should be selected
  - Methods for selecting final equating set
    - Check extreme values (e.g., aa-plot, bb-plot, delta plot)
    - Proportionalized TCCs and TIFs

- **Step 5** was conducted to check whether there exist drifting items in the final equating sets.
Real Data Analysis

- Comparison of Equating Sets
  - Iterative Procedure
    - Step 1: item calibration—PARSCALE
    - Step 2: scale transformation—STUIRT
    - Step 3: IPD detection
    - Step 4: iterative procedure
    - Step 5: final check of equating items
Real Data Analysis

- Item Parameter Drift Patterns

Table 3. Number and Percentage of Drifting Items in History and Science Tests

<table>
<thead>
<tr>
<th></th>
<th>History</th>
<th>Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>14/60</td>
<td>24/45</td>
</tr>
<tr>
<td>Percentage</td>
<td>23%</td>
<td>53%</td>
</tr>
<tr>
<td>Total</td>
<td>38/105</td>
<td></td>
</tr>
</tbody>
</table>
Simulation Study

- **Analysis Procedure**
  - **Step 1:** item calibration—mirt R package
    - The simulated item responses for both pre-equated test and post operational test were generated following **3PL model**;
    - Ability parameters were drawn from the standard normal distribution \( N(0,1) \).
  - **Step 2:** scale transformation—equateIRT R package
  - **Step 3:** IPD detection
  - **Step 4:** iterative procedure
  - **Step 5:** final check of equating items