Practitioner’s Approach to Identify Item Drift in CAT

Huijuan Meng, Susan Steinkamp, Pearson Paul Jones, Joy Matthews-Lopez, NABP
Introduction

• Item parameter drift (IPD): change in item parameters over time.

• Possible causes: changes in curriculum and training; candidates’ increasing familiarity with frequently exposed items.

• Impact of IPD: affect psychometric quality of IRT applications
  ▪ CAT: item selection; ability estimate
  ▪ Pretest item calibration

• Evaluate IPD: maintain a stable scale and ensure the quality of item calibration
CAT Program and Data

- Data: a fixed-length CAT using 3P model
- Number of candidates: **15,000**
- Test length: **150** operational (scored) items
- Number of items in the data: **1,921**
- Number of items in IPD check: **1,208** items (N>=500)
- Baseline scale: item pool
- Purpose: develop procedures that can be used to efficiently identify items drifting away from the baseline scale in a real CAT data.
IPD Literature (1)

- IPD procedures have often been examined in the fixed-form test data.
- DIF, drift, IRT model misfit: all demonstrate the lack of invariance of item parameters in the data
- IPD identification in CAT research:
  - Lord’s $\chi^2$ statistic
  - CUSUM method
  - Raju’s NCDIF
IPD Literature (2)

- Lord’s $\chi^2$ statistic (2P & 3P): use parameter differences and 2 sets of asymptotic variance-covariance matrices of maximum likelihood estimators for original and new item parameters; fit in the general framework of Wald test.

- CUSUM procedure: a sequential series of Wald tests, in which standardized parameter differences are sequentially added for each time period.

- Issues with Lord’s $\chi^2$ & CUSUM:
  - Unavailability of asymptotic variance-covariance matrix for original item parameter estimates
  - Impact of item sample size on the magnitude of the asymptotic matrix

- Raju’s NCDIF: rely on Monte Carlo technique
  - A large number of replications—time consuming
  - Numerous item parameter sets from the asymptotic variance-covariance matrix for newly calibrated parameters—restriction can’t be guaranteed
G² Statistic

- G² is a likelihood ratio chi-square statistic.

\[ G^2_j = 2 \sum_{h=1}^{n_g} \left[ \frac{r_{hj} / N_{hj}}{P_j(\hat{\theta}_{hj})} + \frac{(N_{hj} - r_{hj}) \log_e \left( \frac{N_{hj}}{N_{hj} - r_{hj}} \right)}{1 - P_j(\hat{\theta}_{hj})} \right] \]

- G² can’t be computed for an item with insufficient cases.
- G² issue: with a large sample size, an item can be flagged with even a trivial model misfit.

Observed proportion correct

Model-based proportion correct

Interval theta mean

Observed proportion incorrect

Model-based proportion incorrect
Item plot (1): Flagged Item (P<0.01)
Item plot (2): Non-flagged Item (P<0.01)
G² Computing and Item Plotting

- BILOG-MG plot
  - Phase 2 output: poor quality
  - IRT Graphics tool: inconvenience

- Visual inspection: subjective and time-consuming

- Quantitative evaluation: more efficient and more objective

- BILOG-MG: no detailed interim computation results for G²

- For each item: compute a G², produce a plot, then use the discrepancies between observed and model-based values to refine statistical test results and to categorize items.
Initial IPD Identifications: $G^2$ Statistic

- $G^2$ comparison: BILOG-MG vs. VUE

<table>
<thead>
<tr>
<th>VUE Flag</th>
<th>BILOG-MG Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>475</td>
</tr>
<tr>
<td>Yes</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>510</td>
</tr>
</tbody>
</table>

Consistency: 1131
$(475+656)/1208=93.6\%$

Inconsistency: 77
$(42+35)/1208=6.4\%$

- alpha=0.01: among 1,208 items, 77 (6%) are classified differently, $G^2$ flagging consistency rate: 94%

- Possible cause: use different interval merging methods
### Further IPD Identifications (1) : Drift Category

- **Indices to check item parameter drift:**
  - **P-DIF:** discrepancy between observed and model-based proportion correct at each theta interval;
  - **Drift:** average of P-DIFs across all intervals;
  - **Absolute drift:** average of absolute P-DIFs across all intervals.

<table>
<thead>
<tr>
<th>Drift Category</th>
<th>VUE G^2 Flag</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>OK</td>
<td>361 (76%)</td>
<td>112 (24%)</td>
</tr>
<tr>
<td>E</td>
<td>10</td>
<td>111</td>
</tr>
<tr>
<td>EE</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>EEE</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>HH</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>HHH</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>V</td>
<td>145</td>
<td>344</td>
</tr>
<tr>
<td>VV</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>VVV</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>517</td>
<td>691</td>
</tr>
</tbody>
</table>

- **473 (39%)** drift OK
- **179 (15%)** getting easier
- **19 (2%)** getting harder
- **537 (44%)** Mixed directions
  - 404 (75%): easier
  - 133 (25%): harder
Further IPD Identifications (2) Standard Indices

- **Standard P-DIF**: $P$-DIF / standard error of model-fit ICC value at each interval

\[ Standard\ Error\ (P(\hat{\theta}_{hj})) = \sqrt{P(\hat{\theta}_{hj}) \cdot (1 - P(\hat{\theta}_{hj})) / N_{hj}} \]

- **Standard Drift Flag (Yes/No)**:
  Yes: standard P-DIF mean $\leq -1.645$ or $\geq +1.645$

- **Absolute Standard Drift Flag (Yes/No)**:
  Yes: absolute standard P-DIF mean $\geq +1.645$

- **Lower Asymptote Flag (Yes/No)**:
  Yes: 2 lower standard P-DIF values $\geq +2$ or $\leq -2$

- **Upper Asymptote Flag (Yes/No)**:
  Yes: 2 upper standard P-DIF values $\geq +2$ or $\leq -2$

- **Medium and large drift Flag (Yes/No)**:
  Yes: drift category is NOT OK, E, H, or V
Each of the 1208 items is placed under one of two categories: Recalibration or Anchor.

Decision rule:

If (VUE $G^2$ flag = Yes or BILOG-MG $G^2$ = Yes or Drift Category ≠ OK) and (any of the drift flags = Yes) then the item is placed in the Recalibration category (509), otherwise it is placed in the Anchor category (699).

<table>
<thead>
<tr>
<th>Category</th>
<th>Anchor</th>
<th>Recalibration</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$G^2$ No</td>
<td>$G^2$ Yes</td>
<td>$G^2$ No</td>
</tr>
<tr>
<td>OK</td>
<td>358</td>
<td>59</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>10</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>EE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EEE</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>HH</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HHH</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>136</td>
<td>104</td>
<td>9</td>
</tr>
<tr>
<td>VV</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>VVV</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>504 (72%)</td>
<td>195 (28%)</td>
<td>13 (3%)</td>
</tr>
</tbody>
</table>

504 + 195 = 699
13 + 496 = 509
Item plot (1): Anchor Item (Flagged by G^2)

Item 623: VB332300 | Drift: OK (D=-0.01 | ABSD=0.04) | PVALUE: 0.27
N=1917  IRT-a=0.867  IRT-b=1.243  IRT-c=0.059
VUE-CHI=26.5  VUE-DF=9  VUE-P=0.002  VUE-Flag=Yes
BLG-CHI=26.3  BLG-DF=9  BLG-P=0.002  BLG-Flag=Yes

Graph showing the probability against theta with ICC, observed proportion correct, and upper and lower 95% tolerance intervals.
Item plot (2): Recalibration Item (Not flagged by $G^2$)

Item 1044: VB330110 | Drift: V (D=-0.05 | ABSD=0.06) | PVALUE: 0.86
N=1042  IRT-a=0.565  IRT-b=-1.543  IRT-c=0.225
VUE-CHI=18.3  VUE-DF=9  VUE-P=0.032  VUE-Flag=No
BLG-CHI=18.8  BLG-DF=9  BLG-P=0.027  BLG-Flag=No

Probability

Theta

ICC

*** Observed Proportion Correct

Upper 95% Tolerance Interval

Lower 95% Tolerance Interval

PEARSON
Item plot (3): Recalibration Item ICC

Item 472: VB243779 | N=572 | Drift: VV | Decision: Recalibration
OA=0.648  OB=2.129  OC=0.291 NA=0.567  NB=0.813  NC=0.243
VUE-CHI=46.4  VUE-DF=9  VUE-P=0.000  VUE-Flag=Yes
BLG-CHI=47.9  BLG-DF=9  BLG-P=0.000  BLG-Flag=Yes

Probability

Theta

-4  -3  -2  -1  0  1  2  3  4

Old ICC
New ICC
Upper 95% Tolerance Interval
Lower 95% Tolerance Interval

** ** Observed Proportion Correct
Summary

- Using both $G^2$ statistic and criteria derived from the discrepancies between observed and model-based proportion correct, we check parameter drift for 1,208 operational items.

- Plots for those items have been produced and scanned; in general, the real data support our final classification of items and recalibration outcomes.

- Although the results can be confounded by item model misfit in original data calibration, it is still considered as a practical way of identifying drift items in a real CAT data.

- A simulation study should be conducted to further examine the accuracy of this approach.

- Finally, we will not completely replace parameters for all flagged items with newly calibrated values; instead, we have procedures to determine whether using recalibration results for an item directly or updating an item parameters by reconciling original and new values.
Questions? Comments?

Thank You!

huijuan.meng@pearson.com