Detecting DIF between Conventional and Computerized Adaptive Testing: A Monte Carlo Study

Barth B. Riley

Adam C. Carle

Chestnut Health Systems

UNIVERSITY OF Cincinnati
Introduction

- Instruments are being transitioned from paper-and-pencil (P&P) to computerized adaptive modes of administration.
- Problems arise when item parameters used by CAT are estimated from P&P.
- Mode effects can diminish measurement reliability and validity and increase error in trait estimates (Pommerich, 2007).
Problem

- Differential item functioning (DIF) refers to differences in level of item endorsement between two or more groups after controlling for differences in ability.
- Most DIF methods are designed for use within mode but not between mode of administration.
- Differences in level of missing item responses between modes.
Purpose and Rationale

- Develop and evaluate approaches to assessing item-level mode effects.
- Bayesian methods can provide more accurate results compared to conventional approaches.
  - Take into account uncertainty in trait and item parameter estimates.
DIF Procedure

1. Estimate $\theta$ using item response data pooled across administration modes (CAT and P&P).
2. Using $\theta_i$ obtained in Step 1, estimate the posterior distributions of mode-specific item parameters.
3. For each item common across modes, assess the difference in the posterior distributions of the item parameters (i.e., between $\beta_j^{\text{CAT}}$ and $\beta_j^{\text{P&P}}$).
Comparing Posterior Distributions

• Two approaches.
  • Modified robust Z statistic (Huynh & Meyer, 2010).
  • 95% Credible Interval for mean difference between $\beta_j^{\text{CAT}}$ and $\beta_j^{\text{P&P}}$. 
Modified Robust Z

\[ \text{Robust } Z_j = \frac{\text{Med}(\beta_{j}^{\text{CAT}} - \beta_{j}^{\text{P&P}})}{0.74(\text{IQR}[\beta_{j}^{\text{CAT}} - \beta_{j}^{\text{P&P}}])} \]

- \text{Med} = \text{median of the differences in the CAT and P&P item parameters based on their posterior distributions.}
- \text{IQR} = \text{interquartile range of the difference.}
## Priors and Generated Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination</td>
<td>$LN(0.0,0.5)$</td>
<td>1PLM: 1.0 2PLM: $LN(0.0,0.5)$</td>
</tr>
<tr>
<td>Difficulty</td>
<td>$Normal(0.0,2.0)$</td>
<td>Uniform(-3.0,3.0)</td>
</tr>
<tr>
<td>Ability</td>
<td>$Normal(0.0,1.0)$</td>
<td>$Normal(0.0,1.0)$</td>
</tr>
</tbody>
</table>
Monte Carlo Study

- Two sets of P&P item parameters generated using previous criteria fitting one- (1PLM) and two-parameter (2PLM) IRT models.
- Item response data generated using each set of parameters.
- Parameters then estimated using maximum likelihood (Mplus).
Monte Carlo Study cont.

- CAT item response data were generated using the following variables:
  - % of DIF items (10% vs. 30%).
  - Magnitude of DIF $|\beta_j^{\text{CAT}} - \beta_j^{\text{P&P}}| = 0.42$ vs. 0.63.
  - Mean difference in $\theta$ between CAT and P&P samples (0 vs. 1 logit).
  - Direction of DIF was randomized.
  - 10 datasets generated per condition.
- CAT simulations: Firestar 1.33 (Choi, 2009).
- Bayesian Analysis: WinBUGS 1.43 (Spiegelhalter et al., 2007).
- Sample Size:
  - P&P Data: $N = 1,000$.
  - CAT Data: $N = 3,000$. 
Robust Z: False Positive Rate

![Bar chart showing the False Positive Rate for different conditions.](image)
Robust Z: True Positive Rate
CrI: False Positive Rate

![Graph showing false positive rate for different DIF Δθ categories (Med. DIF Δθ = 0, Med. DIF Δθ = 1, Lrg. DIF Δθ = 0, Lrg. DIF Δθ = 1) with 1PL and 2PL models.]
Crl: True Positive Rate
Performance by Item Difficulty

Predicted Mean % vs P&P Item Difficulty

- RZ TP%
- Crl TP%
- RZ FP%
- Crl FP%
True & False Positive Rates by CAT Item Usage

- Robust Z TP%
- Robust Z FP%
- CrI TP%
- CrI FP%
Conclusions

• Both procedures evidenced adequate control of false positive DIF results.
  • Exception: low difficulty items (< -2.5 logits).
  • Not significantly affected by % of DIF items.
  • Was affected by mean trait level difference.
• Crl evidenced slightly higher power to detect DIF, but also higher false positive rate.
Conclusions cont.

- Power to detect DIF varied considerably, and was affected by several factors, including:
  - Item usage.
  - DIF size.
  - IRT model.
  - Mean difference in trait estimates.
  - Item difficulty.
Future Research

• Test robustness of procedures to data that do not conform to prior assumptions.
  • Skewed ability and item parameter distributions.
• Detecting non-uniform DIF.
References


Thank You

For more information, please contact:

Barth Riley
bbriley@chestnut.org