Multiple-Group Calibration in SAS®:
PROC IRT and SAS/IML®
Multiple-Group Calibration in SAS®: PROC IRT and SAS/IML®
Kyungyong Kim1, Seohye Park1, Jinah Choi1, and Hongwook Seo2
1University of Iowa and 2ACT

ABSTRACT

- Item response theory (IRT) has been gaining popularity in the field of educational and psychological measurement.
- IRT assumes that the probability of a correct response to an item is a function of person and item parameters.
- In IRT, many applications require a common ability scale. However, the origin and unit of measurement of the ability scale are undetermined.
- IRT estimation programs typically choose the ability scale so that mean and SD of the person parameters are 0 and 1 for the group at hand.
- When estimating parameters for two different test forms with nonequivalent groups, this common procedure yields parameter estimates that are on two different ability scales.
- Multiple-group calibration, which is supported by the IRT procedure in SAS/STAT® (SAS Institute Inc., 2013), is one approach that is often used to handle this issue.
- To conduct multiple-group calibration, the two test forms must share some common items.
- The purpose of this paper is to compare the performance of PROC IRT, a multiple-group calibration program written using SAS/IML®, and a commercial software flexMIRT® (Cai, 2013) in terms of the recovery of item parameters.

METHODS

- A simulation study was conducted to compare the performance of PROC IRT, SAS/IML®, and flexMIRT® using the two parameter logistic (2PL) model (Birnbaum, 1968):
  \[ P_{ij}(\theta) = \frac{1}{1 + \exp[-1.7a_j(\theta_i - \theta_j)]} \]
  where \( a_j \) and \( b_j \) are the discrimination and difficulty parameters for item \( j \), and \( \theta_i \) is the ability parameter for person \( i \).
- The study factors for the simulation study were as follows:

<table>
<thead>
<tr>
<th>Study Factor</th>
<th>Form Y / Group 1 (Reference)</th>
<th>Form X / Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Length</td>
<td>40 unique items + 20 common items</td>
<td>40 unique items + 20 common items</td>
</tr>
<tr>
<td>Sample Size</td>
<td>3,000</td>
<td>3,000</td>
</tr>
<tr>
<td>Ability Distribution</td>
<td>N(0, 1)</td>
<td>N(0.5, 1)</td>
</tr>
</tbody>
</table>

- The recovery of the item parameters were evaluated using three statistics: bias, standard error (SE), and root mean squared error (RMSE).

RESULTS

Table 1. Program comparison

<table>
<thead>
<tr>
<th>Program</th>
<th>SAS® PROC IRT</th>
<th>SAS/IML®</th>
<th>flexMIRT®</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation Method</td>
<td>Marginal MLE</td>
<td>Marginal MLE</td>
<td>Marginal MLE</td>
</tr>
<tr>
<td>Ability Distribution</td>
<td>(Both Groups)</td>
<td>Normal</td>
<td>Empirical</td>
</tr>
<tr>
<td>Optimization Method</td>
<td>Quasi-Newton</td>
<td>Newton-Raphson</td>
<td>Newton-Raphson</td>
</tr>
</tbody>
</table>

Note: For SAS/IML® and flexMIRT®, the distributions of ability for both groups are estimated empirically without assuming any shapes.

Table 2. Simulation results

<table>
<thead>
<tr>
<th>Par.</th>
<th>Program</th>
<th>Form Y</th>
<th>Form X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SE</td>
<td>RMSE</td>
</tr>
<tr>
<td>( a )</td>
<td>SAS® PROC IRT</td>
<td>.027</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>SAS/IML®</td>
<td>.002</td>
<td>.048</td>
</tr>
<tr>
<td></td>
<td>flexMIRT®</td>
<td>-.012</td>
<td>.047</td>
</tr>
<tr>
<td>( b )</td>
<td>SAS® PROC IRT</td>
<td>-.225</td>
<td>.046</td>
</tr>
<tr>
<td></td>
<td>SAS/IML®</td>
<td>-.005</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>flexMIRT®</td>
<td>.009</td>
<td>.049</td>
</tr>
</tbody>
</table>

- Overall, the values of bias for the item parameter estimates obtained with SAS® PROC IRT tended to be larger than those obtained with SAS/IML® and flexMIRT®.
- This tendency was more noticeable for the \( b \)-parameters than the \( a \)-parameters.
- As a result, SAS® PROC IRT yielded item parameter estimates with the largest values of RMSE.
- Item parameter estimates obtained with SAS/IML® and flexMIRT® were comparable in terms of all three evaluation criteria.
Multiple-Group Calibration in SAS®: PROC IRT and SAS/IML®

Kyungyong Kim¹, Seohée Park¹, Jinah Choi¹, and Hongwook Seo²

¹University of Iowa and ²ACT

---

**RESULTS CONTINUED**

![Conditional Bias, SE, and RMSE Values](image)

Figure 1. Conditional bias, SE, and RMSE values for the α- and β-parameters.  
Note: In all figures, the results for SAS/IML® are depicted with green, the results for SAS® PROC IRT are depicted with orange, and the results for flexMIRT are depicted with red.

- SAS® PROC IRT systematically overestimates the α-parameters and underestimates the β-parameters for all items in Forms X and Y.
- Between the two item parameters, SAS® PROC IRT produced significantly less accurate β-parameter estimates in terms of both the conditional bias and RMSE statistics.
- SAS/IML® and flexMIRT yielded nearly unbiased estimates for both the α- and β-parameters.

---

**CONCLUSIONS / LIMITATIONS**

- The performance of SAS/IML® and flexMIRT are very similar in terms of the recovery of α-parameters.
- SAS® PROC IRT yields inaccurate β-parameter estimates.
  
  ➔ For multiple-group calibration, the main difference between SAS® PROC IRT and the other two programs is the specification of the ability distributions during the estimation process. SAS® PROC IRT assumes abilities for both groups follow a normal distribution, whereas SAS/IML® and flexMIRT® estimate the ability distributions concurrently with the item parameters.
- SAS® PROC IRT requires significantly more computation time than the other two programs.

<table>
<thead>
<tr>
<th>Program</th>
<th>SAS® PROC IRT</th>
<th>SAS/IML®</th>
<th>flexMIRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computation Time</td>
<td>3 minutes</td>
<td>0.75 seconds</td>
<td>0.84 seconds</td>
</tr>
</tbody>
</table>

Note: The SAS/IML® program written for this study only provides the item parameter estimates as the final output.

- In most IRT estimation programs, prior distributions can be assumed for the item parameters to guarantee convergence (i.e., marginalized Bayesian estimation; Mislevy, 1986). However, SAS® PROC IRT only supports the marginal maximum likelihood estimation method (Bock and Aitkin, 1981).
- Because of the small number of conditions examined in the simulation study, this study is limited in its ability to generalize the findings to measurement conditions that are not included in the simulation study.

---

**REFERENCES**


Contact Information: kyungyong-kim@uiowa.edu

---

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.