Assessing Differential Item Functioning (DIF) for a Large Number of Groups

Michelle Y. Chen[a], Taylor Asbury[a], Lok H. Chau[a], William Tang[a], Bruno D. Zumbo[b]

a: Paragon Testing Enterprises b: University of British Columbia

Summary

- This study introduces a new method for investigating differential item functioning (DIF) based on generalized linear mixed models (GLMM).
- The proposed DIF method treats the grouping variable as a random effect, allowing one to simultaneously model the DIF effect across a large number of groups.
- Using an example of DIF for a listening item attributable to test takers' first language (L1) backgrounds, we demonstrated our method with 46 L1 groups.

Background and Motivational Question

- Differential item functioning (DIF) analyses are used to ensure the fairness of tests based on internal test structure (AERA et al., 2014).
- Commonly, DIF analyses compare two groups—a reference and a focal group.
  - When more than two groups are involved in the comparison, multiple pairwise DIF tests for each focal group are routinely performed one item at a time.
  - This pairwise strategy may be reasonable in the often-discussed case of 4 or 5 groups leading to a maximum of 6 or 10 pairwise DIF tests, respectively.
- However, for large-scale assessments of diverse populations, we need to investigate DIF across a large number of groups to support fair and valid score interpretations. For example:
  - The Canadian English Language Proficiency Index Program (CELPIP) – General Test is designed to measure the functional language proficiency required for successful communication in general Canadian social, educational, and workplace contexts.
  - The CELPIP Test scores are used for Canadian immigration and citizenship, and professional designation.
  - In our operational setting, we investigate DIF attributable to test-takers' self-reported first language (L1) where a large number of possible L1 groups exist.
  - Across items and forms, test takers reported more than 100 first language groups.
  - In our demonstration, 46 self-declared L1 groups were represented.

Data Source

We demonstrate the DIF method on a multiple-choice listening item.

- 14,611 test taker responses
- 46 self-declared L1 groups with group sizes ranging from 30 to 3003 (mean=317.63 and SD= 514.33).

A Generalized Linear Mixed Modeling (GLMM) Approach for DIF Investigation

Uniform DIF: A random-intercept GLMM model.

$$Logit(p) = (\mu + U_{0l}) + \beta\,(proficiency) + \epsilon, \quad \text{where } U_{0l} \sim N(0, \tau)$$

Overall DIF effect (uniform & non-uniform DIF).

$$Logit(p) = (\mu + U_{0l}) + (\beta + U_{1l})\,(proficiency) + \epsilon,$$

$$\text{where } \begin{bmatrix} U_{0l} \\ U_{1l} \end{bmatrix} \sim N(0, \Omega), \quad \text{and} \quad \Omega \sim \begin{bmatrix} \sigma^2_{u_0} & \\ \sigma^2_{u_{01}} & \sigma^2_{u_1} \end{bmatrix}$$

Results

Table 1. DIF Effects across L1 Groups

Random Effects	Uniform DIF Model		Overall DIF Effect Model	
	Variance	SD	Variance	SD
L1 background (Intercept)	0.270	0.519	0.353	0.594
L1 background (Slope)	--	--	0.001	0.024
(Unadjusted) ICC for random effects	0.055		0.086	

Note: L1 = self-reported first language; SD = standard deviation; ICC = intraclass correlation coefficient

Discussion

- Taken conventional pairwise comparison approach, we would have to run 1035 pairwise DIF tests to investigate DIF across the 46 self-declared L1 groups.
- There are clear statistical advantages if one recasts the solution to comparisons of a large number of groups from a framework of statistical modelling to quantify DIF among all groups simultaneously.
- The GLMM is an extension of Swaminathan and Rogers' (1990) approach to DIF between two groups. It models the DIF effect across a large number of groups simultaneously by treating the group-specific effect as random.