Testing for Differential Item Functioning with No Internal Matching Variable and Continuous Item Ratings

Michelle Y. Chen [a,b], Wendy Lam [c], Bruno D. Zumbo [b]

a: Paragon Testing Enterprises b: University of British Columbia c: Independent Consultant

Summary

Differential item functioning (DIF) is a general concern in testing programs as it is closely tied to test validation (Zumbo, 2007). However, typical writing assessments usually pose unique challenges in DIF investigations.

- Building on work by Zumbo (2008), a method to test DIF for a continuously scored writing test with only two prompts on each test form is proposed and demonstrated with real test data.
- This study informs and addresses the limited use of DIF evaluations in writing tests.

Background

DIF investigations

- DIF occurs when test takers from different groups of the same ability level have different chances of achieving the same score levels on a task.
- Many techniques and procedures have been developed to test for DIF (e.g., Rogers & Swaminathan, 1993; Zumbo, 1999).
- Typical DIF methods are designed for binary or polytomous scores and relied on internal matching scores such as total or corrected total scores.

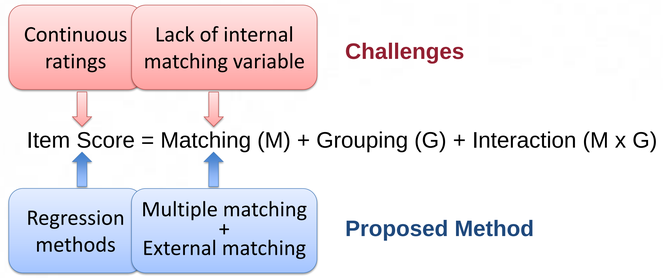
Writing assessments

- Writing ability is usually measured through performance assessments, in which test takers need to compose an essay or other forms of written expression to respond to the writing prompt.
- When test takers produce a writing sample in a test setting, they engage in a complicated process, and their performance can be affected by many internal and external factors other than writing ability.
- Writing assessments often only have two or at most three prompts (and hence writing samples).
- The ratings of a writing sample are usually polytomous and the final score can be a continuous metric in some cases.

Challenges and A Proposed Method

Developing a DIF analysis strategy requires that two major issues be addressed:

(a) define matching variable; and (b) accommodate the continuous responses.



An Example: Gender DIF Investigation

The CELPIP-General Test

- The Canadian English Language Proficiency Index Program - General (CELPIP-General) test intends to measure functional English language proficiency in four domains: reading, listening, speaking, and writing.
- CELPIP-General is a high-stakes test as CELPIP-General scores can be used as mandated evidence of English language proficiency for Canadian citizenship and immigration applications.
- All test takers taking this writing test respond to two different writing tasks. Each task score is a continuous variable which can theoretically be any numerical value between 0 and 12.3.

Samples used in this example

- Eighty-one writing tasks were included in this study. These tasks appeared in 42 writing test forms which were administered in 2014 and 2015.
- Each writing task was answered by at least 120 test takers from each gender group (Total N = 25,656).
- A total of 56 writing raters were involved in rating these writing samples, with each sample rated by two to three raters.
- The correlations among different components of the test (e.g., writing and listening) are fairly high (>0.73). It is possible to use listening and reading scores as matching variables to investigate writing DIF.

Analysis

- For each analyzed task, three regression models were defined for predicting the task scores.

Model 1. $\text{Writing\_task\_score} = b_0 + b_{11} \times (\text{Listening}) + b_{12} \times (\text{Reading})$

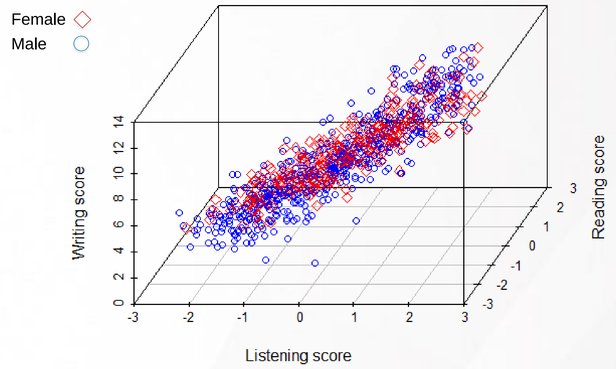
Model 2. $\text{Writing\_task\_score} = b_0 + b_{11} \times (\text{Listening}) + b_{12} \times (\text{Reading}) + b_2 \times (\text{Gender})$

Model 3. $\text{Writing\_task\_score} = b_0 + b_{11} \times (\text{Listening}) + b_{12} \times (\text{Reading}) + b_2 \times (\text{Gender}) + b_{31} \times (\text{Listening by Gender}) + b_{32} \times (\text{Reading by Gender})$

- The nested models were compared by comparing their sums of squares of residuals. A significant improvement from Model 1 to Model 2 or 3 signifies Gender DIF on that item.
- Differences in R^2 between nested models were used to quantify the magnitude of DIF effect.

Results

- Twenty-nine out of 81 tasks (35.8%) were flagged as potential DIF items. The magnitude of the Gender DIF effect on these flagged items was considered small with change of the R^2 less than 0.02.
- The following figure demonstrated a writing task flagged as showing uniform DIF with a change of $R^2 = 0.01$.



Strength of the proposed method

- Linear regression can model task scores directly without shifting to probabilities of specific score categories.
- Linear regression models are flexible. Both uniform and non-uniform DIF effect can be modeled.
- Linear regression models provide effect size measures such as R^2, differences in R^2 between nested models, and regression coefficients which offer useful and intuitive descriptions of DIF effects.

Future directions

- Sensitivity and accuracy of this proposed method still need to be tested.
- Additional studies would be useful for considering how these results compare to those obtained from other testing programs and different DIF detection approaches.
- Another technique that maybe helpful in constructing a matching variable is to make use of available demographic and background information, possibly in combination with scores on the set of performance tasks. One strategy for combining multiple measures into a single composite matching variable is propensity score matching (e.g., Rosenbaum & Rubin, 1985; see Zwick, 1992, for a DIF application).

Contact Information

Michelle Chen mchen@paragontesting.ca