Testing for Differential Item Functioning in Performance Assessments

Michelle Y. Chen\textsuperscript{1}, Yan Liu\textsuperscript{2} & Bruno D. Zumbo\textsuperscript{2}

\textsuperscript{1} Paragon Testing Enterprises, Inc.
\textsuperscript{2} The University of British Columbia
Vancouver, Canada
Background: Performance assessment (I)

- Performance assessments often require test takers to create answers or products that demonstrate their knowledge and skills (Rudner & Boston, 1994).
Background: Performance assessment (II)

Features of Performance Assessment

- Authentic
- Intended to assess higher level (cognitive) skills
- Likely use open-ended tasks
- Each task may require a relatively long time to complete
  - Number of tasks is small
- Performance is often evaluated by multiple raters using scoring rubrics
  - Test scores may be on a continuous scale
Background: Performance assessment (II)

Features of Performance Assessment

• Authentic

• Intended to assess higher level (cognitive) skills

• Likely use open-ended tasks

• Each task may require a relatively long time to complete
  ➢ Number of tasks is small

• Performance is often evaluated by multiple raters using scoring rubrics
  ➢ Test scores may be on a continuous scale
Background: Performance assessment (II)

Features of Performance Assessment
• Authentic
• Intended to assess higher level (cognitive) skills
• Likely use open-ended tasks
• Each task may require a relatively long time to complete
  ➢ Number of tasks is small
• Performance is often evaluated by multiple raters using scoring rubrics
  ➢ Test scores may be on a continuous scale
Features of Performance Assessment

- Authentic
- Intended to assess higher level (cognitive) skills
- Likely use open-ended tasks
- Each task may require a relatively long time to complete
  - Number of tasks is small
- Performance is often evaluated by multiple raters using scoring rubrics
  - Test scores may be on a continuous scale
Background: Performance assessment (II)

Features of Performance Assessment

- Authentic
- Intended to assess higher level (cognitive) skills
- Likely use open-ended tasks
- Each task may require a relatively long time to complete
  - Number of tasks is small
- Performance is often evaluated by multiple raters using scoring rubrics
  - Test scores may be on a continuous scale
Background: Differential Item Functioning (DIF) I

• Motivated by fairness issues in testing
• Can also be used in
  ➢ Quality assurance; e.g., drift analysis
  ➢ Establishing measurement invariance to allow group comparison
  ➢ Investigating comparability of different versions of a measure; e.g., translation effect
Background: Differential Item Functioning (DIF) II

Many DIF methods have focused on dichotomously scored items or polytomously scored items with few possible scoring categories.
Background: Differential Item Functioning (DIF) III

Logistic regression method:

Item Score = Total Score (A) + Grouping (G) + Interaction (A x G)

Proxy for Ability  Uniform DIF  Non-Uniform DIF
Challenges of DIF Investigation in Performance Assessment

First challenge: There is no well defined ability approximation variable because performance assessments are typically short with 1 or 2 tasks.

Researchers have used external variables to approximate ability scores (e.g., the total score of other related subjects).

It is also possible to match the two groups: e.g., covariance adjustment, exact matching, and propensity score matching.
Challenges of DIF Investigation in Performance Assessment

- **Second challenge**: There is no clear guideline for DIF analysis based on matched data with continuous item scores.
- For DIF analysis with covariance adjustment, linear regression can be applied.
- However, for exact matching or propensity score matching, we have not found any published studies providing statistical solutions or guidance.

Item Score = Ability (A) + Grouping (G) + Interaction (A x G)
Research Purpose (Propensity Score DIF for Performance Assessment)

Extends on the current literature of DIF investigation in performance assessments—(multiple) matching of other, correlated, sub-scales or tests.

Describes a propensity score DIF method that handles continuous scores in cases that lack well-defined ability approximation variables.
Demonstration

- Investigates DIF due to different levels of education in a writing task.
- Our example uses 1450 test takers’ data from a high-stakes English writing test which consisting of two tasks.
Participants

1450 Adult test takers
21% were females
A wide variety of language backgrounds
Education level:
• 487 below undergraduate level (coded 1);
• 963 undergraduate level or above (coded 0).
Measure:

• A measure of functional English language proficiency in: reading, listening, speaking, and writing.

Focus is on: Writing Test with two tasks

• Task 1 Email & Task 2 Response to a survey question

• Each task score is a continuous variable which can theoretically be any numerical value between 0 and 12.

• In the past we have used linear regression DIF with reading and listening scores as multiple covariates matching (Chen, Lam, & Zumbo, 2016).
Analysis

A 2-step modeling approach

Step-1. Propensity score matching
  • Selecting covariates
  • Estimating propensity score and matching

Step-2. DIF analysis with mixed effects regression models
Step-1. Propensity score matching: Selecting covariates

- **Employment**: student, construction & factory, store & restaurant, office, or unemployed

- **Daily use of English**:
  - Speaking: grocery shopping, talk to friends/coworker/family, meeting, chat online
  - Listening: watching TV and video
  - Reading: read books/reports/news, online social media
  - Writing: write email/assignment/reports/business correspondence

- **Language background**: first language, year of learning English, year living in English speaking countries

- **Test taking experience**: repeater
Step-1. Propensity score matching: Selecting covariates

- **Employment:** student, construction & factory, store & restaurant, office, or unemployed

- **Daily use of English:**
  - Speaking: grocery shopping, talk to friends/coworker/family, meeting, chat online
  - Listening: watching TV and video
  - Reading: read books/reports/news, online social media
  - Writing: write email/assignment/reports/business correspondence

- **Language background:** first language, year of learning English, year living in English speaking countries

- **Test taking experience:** repeater
Step-1. Propensity score matching: Selecting covariates

- Employment: student, construction & factory, store & restaurant, office, or unemployed

- Daily use of English:
  - Speaking: grocery shopping, talk to friends/coworker/family, meeting, chat online
  - Listening: watching TV and video
  - Reading: read books/reports/news, online social media
  - Writing: write email/assignment/reports/business correspondence

- Language background: first language, year of learning English, year living in English speaking countries

- Test taking experience: repeater
Step-1. Propensity score matching: Selecting covariates

- Employment: student, construction & factory, store & restaurant, office, or unemployed

- Daily use of English:
  - Speaking: grocery shopping, talk to friends/coworker/family, meeting, chat online
  - Listening: watching TV and video
  - Reading: read books/reports/news, online social media
  - Writing: write email/assignment/reports/business correspondence

- Language background: first language, year of learning English, year living in English speaking countries

- Test taking experience: repeater
Step-1. Propensity score matching: Selecting covariates

- Employment: student, construction & factory, store & restaurant, office, or unemployed

- Daily use of English:
  - Speaking: grocery shopping, talk to friends/coworker/family, meeting, chat online
  - Listening: watching TV and video
  - Reading: read books/reports/news, online social media
  - Writing: write email/assignment/reports/business correspondence

- Language background: first language, year of learning English, year living in English speaking countries

- Test taking experience: repeater
Results: Step-1. Propensity score matching

**Optimal Pair Matching**

<table>
<thead>
<tr>
<th>Raw below Undergraduate</th>
<th>Matched below Undergraduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Undergraduate or above</td>
<td>Matched Undergraduate or above</td>
</tr>
</tbody>
</table>

![Histograms of Propensity Scores](Image)
Results: Step-1. Propensity score matching

**Optimal Pair Matching**

<table>
<thead>
<tr>
<th>Raw below Undergraduate</th>
<th>Matched below Undergraduate</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="chart1.png" alt="" /></td>
<td><img src="chart2.png" alt="" /></td>
</tr>
<tr>
<td>Raw Undergraduate or above</td>
<td>Matched Undergraduate or above</td>
</tr>
<tr>
<td><img src="chart3.png" alt="" /></td>
<td><img src="chart4.png" alt="" /></td>
</tr>
</tbody>
</table>
Results: Step-1. Propensity score matching

**Optimal full matching: 1 to multiple, multiple to 1**

Raw below Undergraduate

Matched below Undergraduate

Raw Undergraduate or above

Matched Undergraduate or above
Results: Step-1. Propensity score matching

**Optimal full matching: 1 to multiple, multiple to 1**

- Raw below Undergraduate
- Raw Undergraduate or above
- Matched below Undergraduate
- Matched Undergraduate or above
Results: Step-2. DIF analysis using linear mixed effects regression model

Based on matched dataset (Optimal full matching)

Regression model for DIF investigation:

Item Score = Total Score (A) + Grouping (G) + Interaction (A x G)

Proxy for Ability  Uniform DIF  Non-Uniform DIF
Results: Step-2. DIF analysis using linear mixed effects regression model

Based on matched dataset (Optimal full matching)

ICC=0.26

Fixed effects:

|                | Estimate | S.E.  | df  | t value | Pr(>|t|)   |
|----------------|----------|-------|-----|---------|------------|
| Intercept      | 6.686    | 0.070 | 458 | 95.267  | < .001***  |
| A              | 0.074    | 0.003 | 1263| 25.943  | < .001***  |
| Education      | -0.303   | 0.071 | 1250| -4.276  | < .001***  |
| A * Edu        | -0.018   | 0.005 | 1381| -3.805  | < .001***  |
This graph is prepared for illustration purpose only. Coefficients of fixed effects were used, while random effects were ignored.
2 df likelihood ratio test for DIF detection

Compare two nested models:

• Model 0:
  \[ \text{WritingTask} \sim A + u_{0j} + e_{ij} \]

• Model 2:
  \[ \text{WritingTask} \sim A + \text{Edu} + \text{interaction}(A \times \text{Edu}) + u_{0j} + e_{ij} \]

Note: A: proxy for ability; Edu: education

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Deviance</th>
<th>Chi-square (2df)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 0</td>
<td>4</td>
<td>4867.1</td>
<td>4888.3</td>
<td>-2429.6</td>
<td>4859.1</td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>6</td>
<td>4841.5</td>
<td>4873.2</td>
<td>-2414.8</td>
<td>4829.5</td>
<td>29.592***</td>
</tr>
</tbody>
</table>

***: \(p<.001\)
Summary

Building on previous work (e.g., Chen et al., 2016; Liu et al., 2016; Swaminathan & Rogers, 1990; 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.
Discussion: Regression Methods for DIF Investigation

• Directly modeling continuous data; without shifting to probabilities of specific score categories.

• Cluster effect of matched data has been accounted for in mixed effects model.

• Regression-type models are flexible. Both uniform and non-uniform DIF effect can be modeled.

• Propensity score matching allows a large number of covariates to be included to approximate randomized experimental design; Avoid problems with many covariates in the final DIF analysis.
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.
Thank You

Michelle Chen
mchen@paragontesting.ca
Selected Reference


