Paper presented at 2018 ITC

Testing for Differential Item Functioning in Performance Assessments

Michelle Y. Chen[1], Yan Liu[2] & Bruno D. Zumbo[2]

[1] Paragon Testing Enterprises, Inc.

[2] The University of British Columbia

Vancouver, Canada

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

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

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

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: 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

THE UNIVERSITY OF BRITISH COLUMBIA

Paragon TESTING ENTERPRISES

Challenges of DIF Investigation in Performance Assessment

Lack of internal variable for ability approximation

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

- **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

Continuous ratings

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

- **Second challenge**: There is no clear guideline for DIF analysis based on <u>matched data with continuous item scores</u>.
- 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.

Paragon TESTING ENTERPRISES

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.

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

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

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

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

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

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

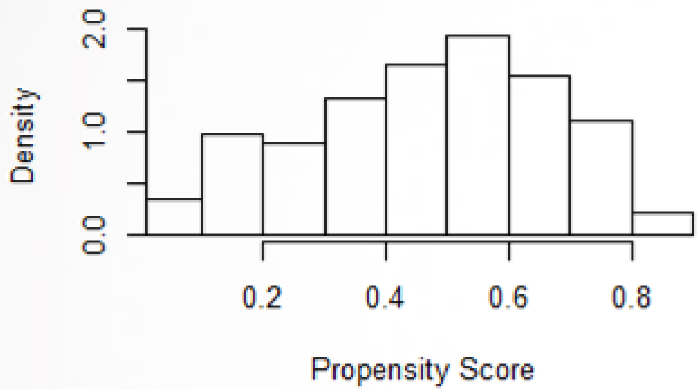
Step-1. Propensity score matching: Selecting covariates

- 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

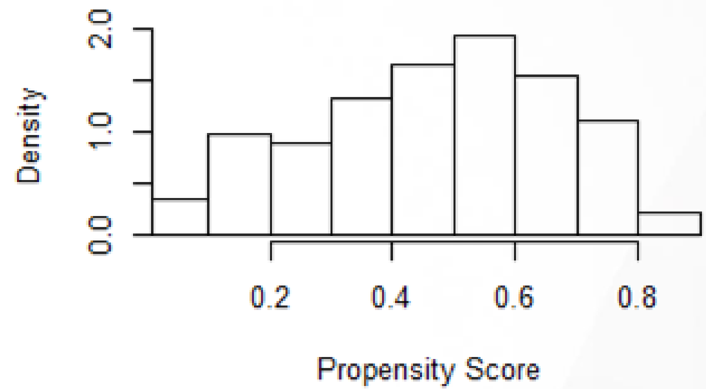
Results: Step-1. Propensity score matching

Optimal Pair Matching

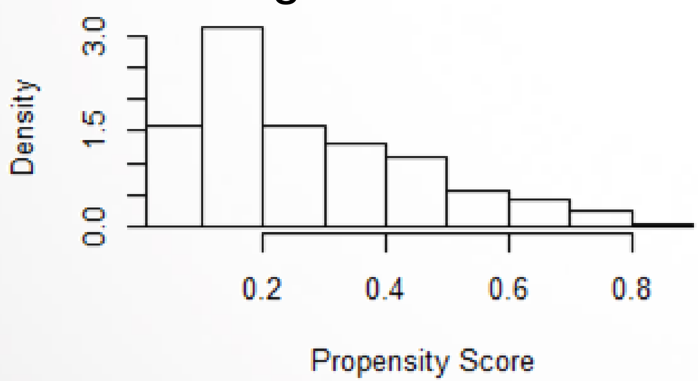
Raw below Undergraduate



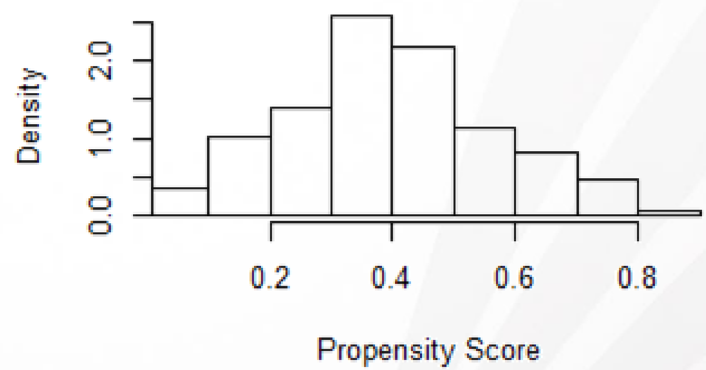
Matched below Undergraduate



Raw Undergraduate or above



Matched Undergraduate or above



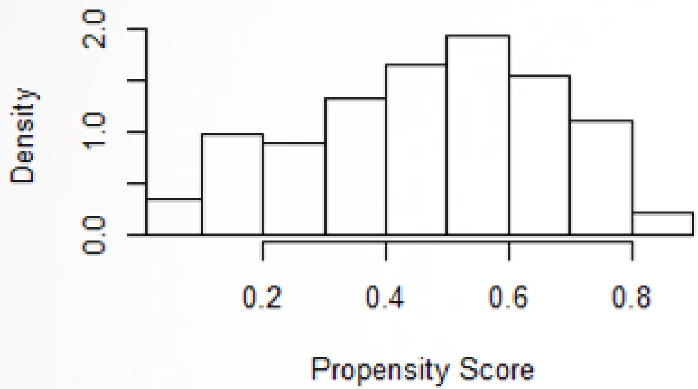
THE UNIVERSITY OF BRITISH COLUMBIA

Paragon TESTING ENTERPRISES

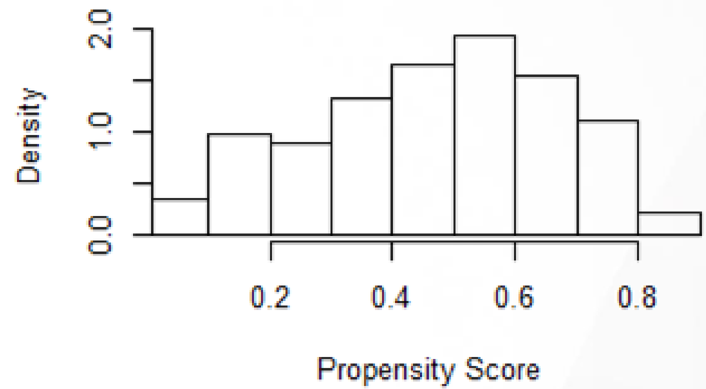
Results: Step-1. Propensity score matching

Optimal Pair Matching

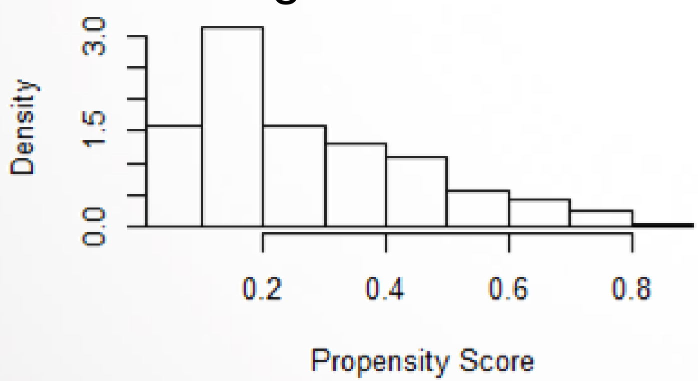
Raw below Undergraduate



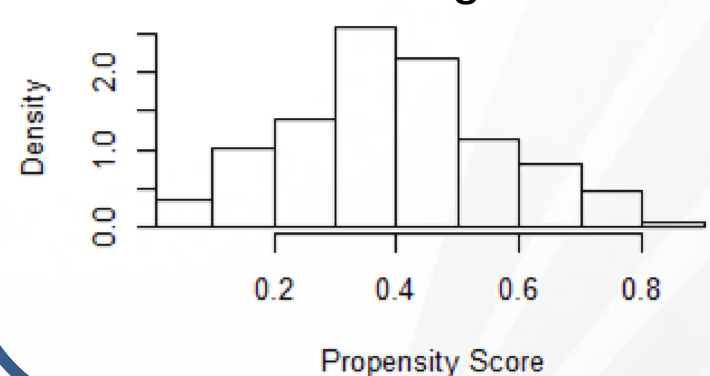
Matched below Undergraduate



Raw Undergraduate or above



Matched Undergraduate or above



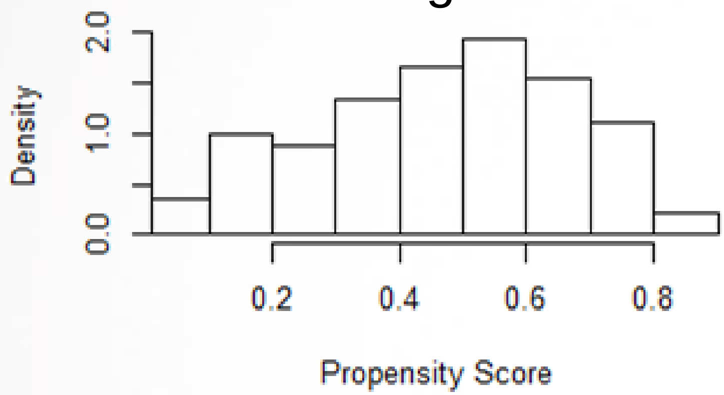
THE UNIVERSITY OF BRITISH COLUMBIA

Paragon

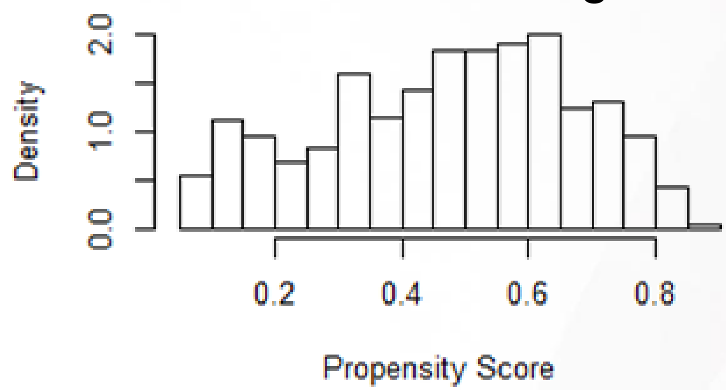
Results: Step-1. Propensity score matching

Optimal full matching: 1 to multiple, multiple to 1

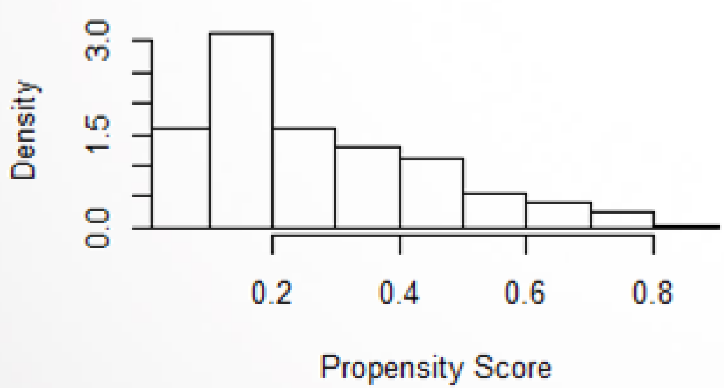
Raw below Undergraduate



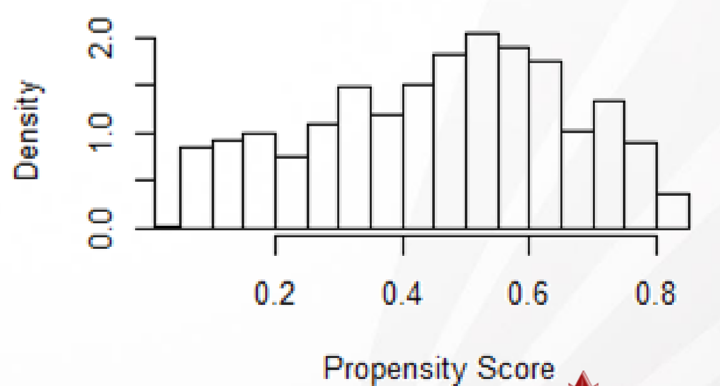
Matched below Undergraduate



Raw Undergraduate or above



Matched Undergraduate or above



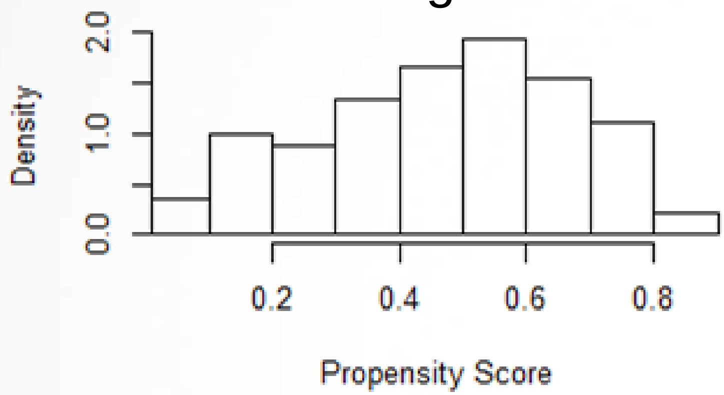
THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

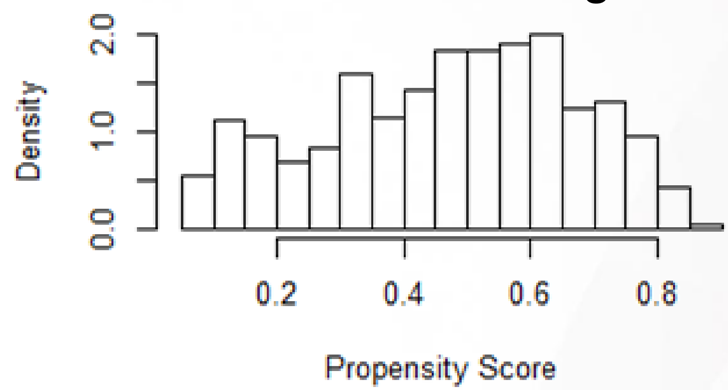
Results: Step-1. Propensity score matching

Optimal full matching: 1 to multiple, multiple to 1

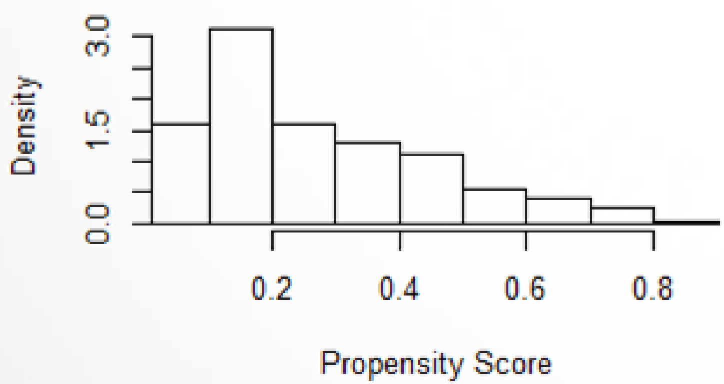
Raw below Undergraduate



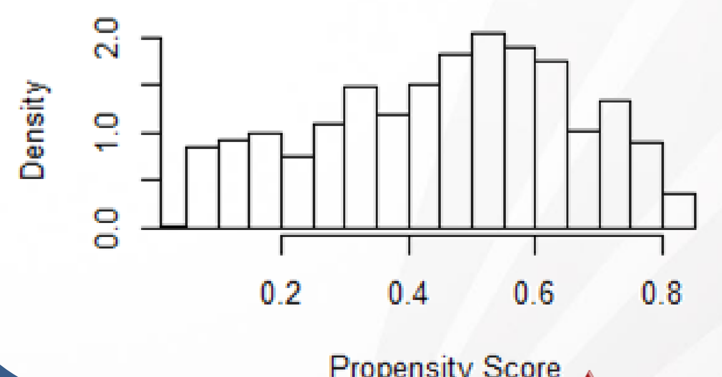
Matched below Undergraduate



Raw Undergraduate or above



Matched Undergraduate or above



THE UNIVERSITY OF BRITISH COLUMBIA

Paragon

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

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

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*** **

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

This graph is prepared for illustration purpose only. Coefficients of fixed effects were used, while random effects were ignored.

THE UNIVERSITY OF BRITISH COLUMBIA

TESTING ENTERPRISES Paragon

2 df likelihood ratio test for DIF detection

Compare two nested models:

- Model 0:

  WritingTask ~ A + u_{0j} + e_{ij}
- Model 2:

  WritingTask ~ A + Edu + interaction(A * Edu) + u_{0j}+ e_{ij}

Note: A: proxy for ability; Edu: education

	df	AIC	BIC	logLik	Deviance	Chi-square (2df)
Model 0	4	4867.1	4888.3	-2429.6	4859.1	
Model 2	6	4841.5	4873.2	-2414.8	4829.5	29.592***

***: 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

- Chen, M. Y., Lam, W., & Zumbo, B. D. (2016). *Testing for differential item functioning with no internal matching variable and continuous item ratings*. Poster presented at the Language Testing Research Colloquium. Palermo, Italy.
- Liu, Y., Zumbo, B. D., Gustafson, P., Huang, Y., Kroc, E., & Wu, A. D. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research & Evaluation, 21*(13). Available online: http://pareonline.net/getvn.asp?v=21&n=13.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement, 27*(4), 361-370.
- Zumbo, B. D. (2008). *Statistical Methods for Investigating Item Bias in Self-Report Measures*, [The University of Florence Lectures on Differential Item Functioning]. Universita degli Studi di Firenze, Florence, Italy.

TESTING ENTERPRISES Paragon