Balancing Covariates to Disentangle/Detect DIF, Item Bias, and Item Impact

Amery D. Wu[1]

Yan Liu[2]

Bruno D. Zumbo[1]

Presented at the 2016 Modern Modeling Methods Conference

Department of ECPS, The University of British Columbia

[2]Harvard Medical School

Background

Confusions in Terminology

- DIF vs. Bias. For some, DIF and item bias are synonymous. For others, DIF is *not* necessarily equated with bias.

- Bias/DIF vs. Impact. For some, item Impact is the group difference irrespective of the presence of DIF and bias. For others, it is believed that item impact can not be studied if DIF or bias is present.

- There is no *clearly shared understanding* of the three terms: DIF, item bias, and item impact.

Our Views

- The three terms denote distinct concepts although they are closely interconnected.

- The three concepts can be disentangled if they are all addressed as group comparison at the item-level.

- Methods of investigating these three phenomena are mostly comparative studies based on observational data, i.e., group membership can not be randomly assigned.

Unresolved

- Conceptually, the same term is used to mean different ideas. Different terms are used to mean the same idea.
- Statistically, there are not yet fairly straightforward methods for item bias and item impact, despite a variety of methods for DIF (see Ackerman, 1992; Shealy & Stout,1993).

Motivation

- If the connections and distinctions among these terms can be <u>ironed out conceptually</u>, a set of <u>integrated statistical procedures</u> can be identified to empirically disentangle and detect the three phenomena.

Definition Refined

- **(Group) Item Bias. An item is biased against a group if the differences in the item score are <u>caused</u> by factors that could <u>invalidate the comparison</u> with the other group(s).**
  For example, an item measuring math ability is biased against the other language-speaking groups if their lower item score is caused by test translation.

- **(Group) DIF is the <u>statistical differences</u> in endorsing or answering an item between groups who possess <u>an equal amount of the attribute</u> that a given item measures.**

- **(Group) Item impact is the group difference(s) in the item scores <u>caused</u> by the measured attribute, <u>if and only if the item is a valid measure of the attribute</u>.**
- For example, an item impact is expected between English-first speakers and English learners <u>if an given item is a valid measure of language proficiency</u>.

Goals

***Based on our definitions* of DIF**, *bias, and impact…….*

- The purpose of this presentation is to propose a methodology for disentangling/detecting DIF, item bias, and item impact.
- The presentation focuses on a proof of concept via two parts:
  (1) an explanation of the logic and rationale underlying the proposed methodology, and
  (2) a demonstration with real data example.
- The technical details are presented in another session at this conference and written in a manuscript. They are available upon request.

A Proof of Concept for the Proposed Methodology

Logic for Detecting Group Item Bias

- DIF signals the possibility of bias, but can not verify the existence of bias.

- The technique of DIF <u>can *not*</u> tell whether "differences in the item score are <u>caused by factors that could invalidate the group comparison</u>" as we define item bias.

- These factors are <u>confounders</u> for group comparison. They are confounders because they are <u>unwanted pre-existing group differences</u> that have an effect on the variation of the item scores.

Logic for Detecting Group Item Bias

(Continued)

- To show an item is biased, we need to show that the item functions differently after controlling for the confounders that could invalidate the group comparison.
- These confounders are referred to as “covariates” under the convention of Neyman-Rubin’s or Rubin’s causal model.
- To detect item bias, we first need to balance the covariate distributions between the groups and then detect DIF.
- IF DIF still exists between the two groups of individuals with balanced covariates, we can conclude, with strong credibility, that the item is biased against a group.

Logic for Detecting Group Item Impact

- Item impact can not be studied if item bias is present.

- This is because impact, as we defined it, is the group difference(s) in the item score caused by the measured attribute, if and only if the item is a valid measure of the attribute.

- Presence of item bias refutes the premise that the given item is a valid measure of the attribute. To detect item bias, we need to show items are not biased.

Logic for Detecting Group Item Impact

(continued)

- Based on our definition of item impact, we also need to show that the group difference is indeed caused by the grouping variable after showing the item is not biased.

- Likewise, we need to control for the factors that can confound the causal claim.

- To detect item impact, we first need to balance the covariates distributions between the groups and then test group difference.

- If item score difference still exists between the two groups of individuals with balanced covariates, we can conclude, with strong credibility, that the groups have an impact on the item.

Summarizing Proposed Procedure

Step-0: Testing conventional DIF (DIF without balancing covariates)
- Yes → Conduct expert review for bias.
- No → Go to Step-1 and then Step-3 for detecting item impact.

(Step-0 can be can be skipped.)

⇩

Step-1: Balancing covariate distributions
- Yes → Go to Step-2 for detecting bias or impact statistically.
- No → Bias or impact can not be detected statistically.

⇩

Step-2: Testing causal DIF between covariate-balanced groups (testing bias)
- Yes → Item is considered biased against a group statistically.
- No → Go to Step-3 for detecting impact.

⇩

Step-3: Testing item score difference between covariate-balanced groups (testing impact)
- Yes → The groups are considered having an impact on an item statistically.
- No → The groups are considered *not* having an impact on an item statistically.

Illustrative Study

Items (Y)

We tested the proposed procedures on the 25 dichotomously scored items from the Grade-8 Mathematic booklet one of TIMSS 2007.

Grouping Variable (G)

The sample consists of a total of 822 students from Canada.
The students took one of the two versions of the test:

French = 1 (focal group, N_1= 281)
English = 0 (reference group, N_0= 541)

Attribute Measure (T)

The observed rest total score was treated as the proxy for students' math ability (attribute to be measured).

Covariates (X_j)

Nine background variables from TIMSS 2007 were used as covariates ((j =9). Their distributions were to be balanced between the two test language groups:

- number of books at home (nbook)
- use of calculator (calculator)
- parents' education (parentEdu)
- availability of computer (computer)
- time on mathematics homework (timehw)
- positive affect to mathematics (affect)
- valuing mathematics (valuing)
- self-confidence in math (slfconf)
- perception about school safety (safty)

Analysis

Step-0. Conventional DIF (without balancing covariates)

Binary logistic regression

$$Logit\ (P(Y)|T,G) = b_0 + b_1T + b_2G + b_3T * G$$

Step-1. Balancing covariates – Propensity scores matching

a. Propensity scores (e)

Propensity scores are multivariate estimates of balance scores, such that

$$G \perp X \ | \ e$$

Propensity scores are estimated by logistic regression

$$e = Logit\ (P(G)|X_j) = b_0 + \sum_{1}^{j} b_j X_j$$

Analysis

Step-1. Balancing covariates (*continued)*

a. Propensity scores

b. Matching

- Individuals in the focal group are matched with individuals from the control group who have close propensity scores.
- We matched the groups on the estimated propensity score using full-optimal matching by R *MatchIt* package

c. Checking covariate balance

i. graphs of propensity score distributions

ii. percent bias reduction $\frac{Bias_{pre} - Bias_{post}}{Bias_{pre}}$ where

Bias = $|M_1(X_j)-M_0(X_j)|$. Pre and post refer to the status matching

Note. Bias here means the difference in the covariates, rather than the group item bias that being investigated.

Analysis

<u>Step-2: Testing causal DIF (bias) between covariates-balanced groups</u>

Binary logistic regression testing uniform and non-uniform DIF

$$Logit\ (P(Y)|T, G) = b_0 + b_1 T + b_2 G + b_3 T * G$$

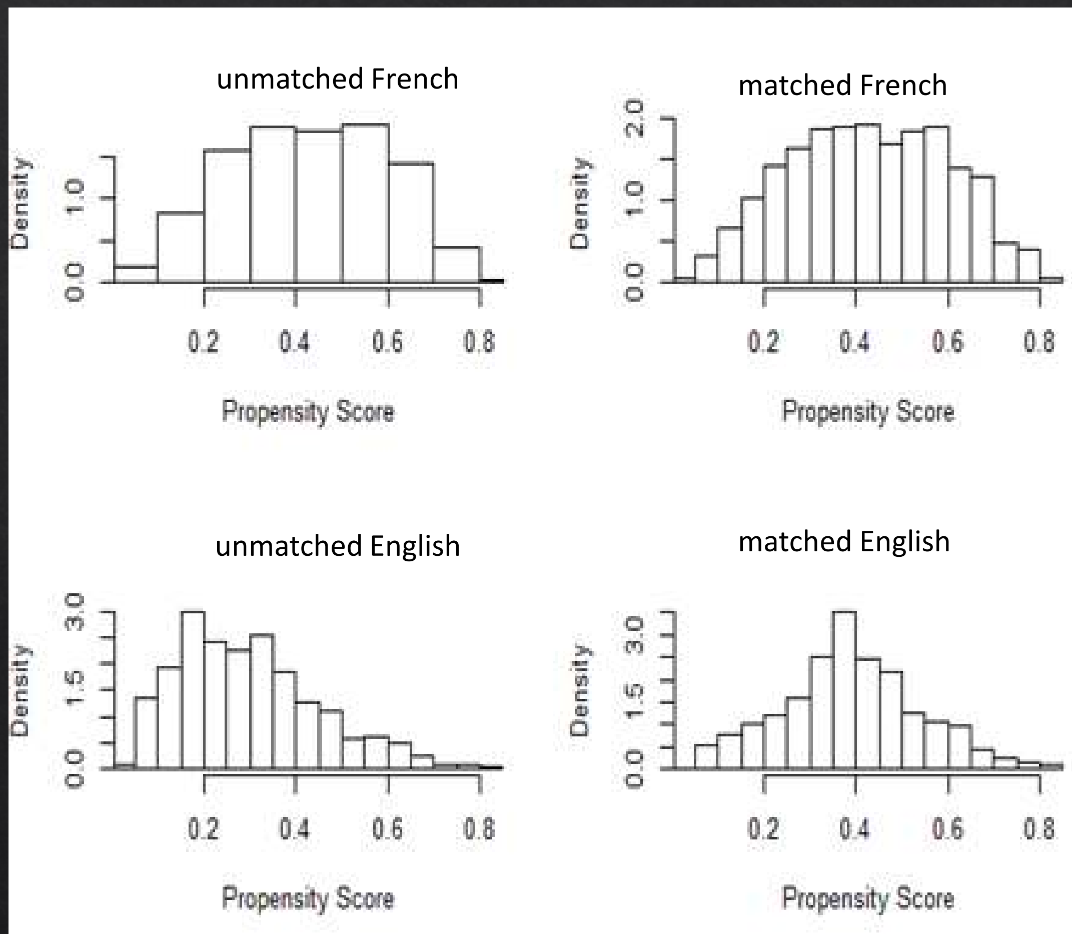
<u>Step-3: Testing item difference (impact) between covariate-balanced groups</u>

Group difference was tested for impact using logistic regression

$$Logit\ (P(Y)|G) = b_0 + b_1 G$$

Results- Covariates Balancing

Propensity Scores Distributions



Results - Covariates Balancing

% Bias Reduction

	M	Before Matching		After Matching		% Bias Reduction
	Focal	Reference	Difference	Reference	Difference	
Covariate	0.430	0.296	0.134	0.388	0.043	68.2
nbook	1.722	2.381	-0.658	1.883	-0.161	75.6
calculator	2.466	2.141	0.326	2.395	0.071	78.1
parent edu	3.238	3.218	0.020	3.226	0.012	40.2
computer	3.626	3.669	-0.043	3.670	-0.043	-0.9
timehw	0.989	1.198	-0.209	1.080	-0.091	56.6
affect	1.231	1.100	0.132	1.178	0.053	59.5
valuing	1.765	1.784	-0.019	1.766	-0.001	97.5
slfconf	1.392	1.392	0.000	1.378	0.014	-
safty	1.463	1.390	0.073	1.418	0.044	39.1

Results - Bias or Impact

TIMSS #	Item	Step-0, Testing Conventional DIF	Step-1. Balancing covariate distributions	Step-2. Testing Causal DIF (Bias)	Step-3. Testing group difference (Impact)	Conclusion
1	1	N		~	N	No Impact
2	2	N		~	Y	Having Impact
3	3	N		~	N	No Impact
4	4	Y		N	N	No Impact
5	5	N		~	Y	Having Impact
6	6	N		~	N	No Impact
7	7	Y		N	Y	Having Impact
8	8	N		~	N	No Impact
9	9	N		~	N	No Impact
13	10	Y		Y	/	Biased
14	11	N		~	N	No Impact
15	12	N		~	N	No Impact
16	13	N		~	N	No Impact
17	14	Y		Y	/	Biased
18	15	N		~	N	No Impact
19	16	Y		Y	/	Biased
20	17	N		~	N	No Impact
21	18	N		~	N	No Impact
22	19	Y		Y	/	Biased
23	20	N		~	N	No Impact
24	21	N		~	N	No Impact
25	22	N		~	N	No Impact
26	23	N		~	N	No Impact
27	24	N		~	N	No Impact
28	25	Y		N	N	No Impact
Test positive		7 out of 25 items		4 out of 7 conventional DIF items	3 out of 21 unbiaed items	

Results

Summary & Interpretation

Conclusion	No. of Items	%	Direction of Bias/Impact
Biased	4	16%	All four items were uniformly biased. Two items were biased against the French-test group; the other two items were biased against the English-test group.
Having Impact	3	12%	For all three items having group impact, the group taking the French-version performed better.
No Impact	18	72%	Eighteen items were considered not having group impact. The two groups performed equally well on these items.
Total	25	100%	

Discussion

- The proposed procedures are suggested in conceptual terms and do not prescribe specific statistical techniques for testing DIF, balancing covariates, or testing group difference.
- The key to the success of covariates balancing is the selection of the covariates It important to detect whether conclusions of causal DIF and impact) are sensitive unobserved covariates (hidden biases).
- The methodology does not replace judgement-based methods. In fact, it takes judgement to determine what are confounders in making group causal claims.
- What counts as a confounder is a consideration of what is irrelevant to and will invalidate the group comparison. Information in construct-relevant covariates that are deemed valid causes for group difference should not be controlled for. Be careful NOT to throw out the baby with bath water.

Questions and Comments

Thank you for your Participation.