Contrasting Various Explanatory Approaches to Think About and Analyze Test Data to Address Fairness: Causal, Contextual, Ecological

Bruno D. Zumbo

Amery D. Wu

University of British Columbia

Vancouver, Canada

The Fifteenth Annual Maryland Conference:

Test fairness in the new generation of large-scale assessment

Sponsored by the Maryland State Department of Education and Maryland Assessment Research Center; Maryland

Overview

- We acknowledge that test fairness is a large and broad topic. We will focus on measurement invariance.
- We will focus more specifically on how one thinks about and analyzes test data for measurement invariance (in our case, DIF).
- Our purpose is to compare and contrast three recent "*explanatory*" approaches to DIF with an eye to learning:
  - How do the methods conceptualize the problem of explaining DIF?
  - How do these methods observe and analyze covariates/background information for explaining DIF?
  - What do the methods do in the end (purpose)?

Agenda

1. Third generation DIF
2. A brief overview of three recent approaches to explaining DIF
   i. Statistical matching on propensity scores
   ii. Mixture models for ecology of item responding (*LAQ*, 2015)
   iii. The loop-back of ethnographic- psychometric methods (*IJT*, in press)
3. Contrasting the three approaches & conclusions therefrom

SECTION 1

THIRD GENERATION DIF

Third Generation DIF (Zumbo, 2007)

- **First Generation:** motivated by a testing problem (fairness), clarifying terminology (e.g., DIF, impact, bias), first statistical methods
- **Second Generation**: Explosion of statistical methods, detecting/flagging DIF

Third Generation DIF (Zumbo, 2007)

- The third generation of DIF is most clearly characterized as conceiving of DIF as occurring because of some characteristic of the test item and/or testing situation that is not relevant to the underlying ability of interest (and hence the test purpose).

- Refining/extending statistical methods to distinguish item bias from item impact and providing explanations as to why DIF was occurring- revisiting the first generation of DIF

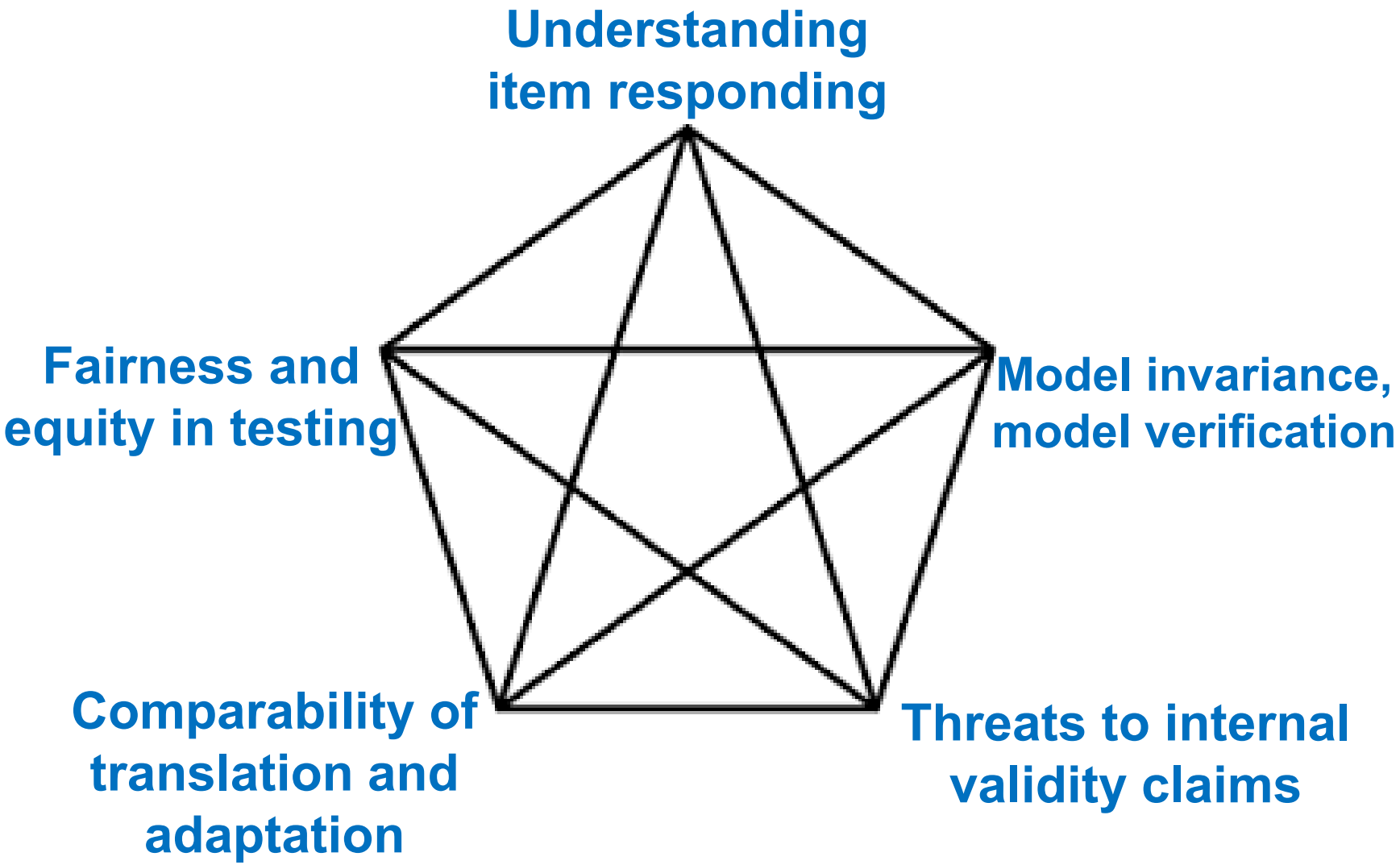
- The third generation of DIF is best represented by its uses, the praxis of DIF. There are five general interconnected uses that embody the third generation praxis of DIF analyses and motivate both the conceptual and methodological developments in third generation DIF.

Third Generation DIF (Zumbo, 2007)

- Not suggesting distinct historical periods and a strictly natural linear stepwise progression towards our current thinking. The phases overlap and the processes go back-and-forth

- Our presentation today will try and highlight this 3rd Generation of DIF

Five Interconnected Uses of 3rd Generation DIF



Uses in Third Generation DIF

1. Fairness and equity in testing.

This purpose of DIF is often driven by policy and legislation. The groups (e.g., visible minorities or language groups) are defined ahead of time before the analyses.

2. Dealing with a possible "threat to internal validity."

In this case, DIF is often investigated so that one can make group comparisons and rule-out measurement artifact as an explanation for the group difference.

3. Investigate the comparability of translated/adapted measures.

This use of DIF is of particular importance in international, comparative, and cross-cultural research. This matter is often referred to as construct comparability.

Uses in Third Generation DIF

4. Trying to understand item response processes.

In this use, DIF becomes a method to help understand the cognitive and/or psychosocial processes of item responding and test performance, and investigates whether these processes are the same for different groups of individuals.

5. Investigating lack of invariance

In this purpose DIF becomes an empirical method for investigating the interconnected ideas of: lack of invariance, model-data fit, and model appropriateness in model-based statistical measurement frameworks like IRT and other latent variable approaches.

Section 2

Brief Overview of Three Recent Approaches to Explanatory DIF

(Three examples of 3rd Generation DIF)

Point of comparison … reminder of conventional usage of DIF

- Used to flag DIF items in operational testing programs, or when writing a test manual or report
- The grouping variable (e.g., race, or gender) is selected based on policy and is known ahead of time
- Uses a statistical method such as MH, LogReg to flag potentially problematic items
- Not quite sure what to do with flagged items (put on ice, delete, re-write?)

A brief Overview Three Recent Papers

1. **Focuses on statistical matching of group membership**

2. **Focuses on an conceptual model for ecology of item responding, and the use of statistical mixture models (*LAQ*, 2015)**

3. **Focuses on the loop-back between ethnographic-psychometric methods (*IJT*, in press)**

Paper #1. Statistical Matching of Group Membership

"***Demonstration of Propensity Score Methods Used in Logistic Regression DIF Analysis***"

Paper #1. Statistical Matching of Group Membership

- The goal is to examine the causal effect (of the grouping variable) on DIF, e.g., whether test translation (English to French) is the cause of DIF.

- The goal is to reduce imbalance in pre-test (pre-existing) covariates between the groups of test takers -- thereby reducing the bias in quantifying the 'treatment effect'.

- Propensity Score Matching: It is a data preprocessing step - matching the two groups on the covariates prior to applying a model (DIF) to estimate a causal effect.

Paper #1. Statistical Matching of Group Membership

- Randomized experimental studies, which have equivalent groups before the experiment, allow researchers to more easily justify a causal claim.
- Most DIF analyses are based on observational groups. Such studies most often do not enjoy the benefits of equivalent groupings before the intervention (e.g., taking a test in different languages).
- The post intervention difference (different probabilities of answering correctly) is not readily attributable to the intervention.

Paper #1. Statistical Matching of Group Membership

Formally, propensity score is the conditional probability of assigning an individual to the treatment condition given a set of observed covariates. The expression for propensity score is:

$$e_i(X_i) = \Pr(Z_i = 1 \mid X_i)$$

where $e_i(X_i)$ denotes propensity score for each individual i; Z_i is an indicator for treatment conditions and $Z_i = 1$ refers to being assigned to the treatment group; X_i is a vector of scores on the observed covariates. The propensity scores are usually estimated by logistic regression:

$$P(Z = 1|X = x) = p(x) = \frac{1}{1+e^{-(\beta_0+\beta(x))}} \qquad (1)$$

Paper #1. Statistical Matching of Group Membership

- Dorans and Holland (1993), Bowen (2011), Lee and Geisinger (2014) have suggested that propensity score matching, a multivariate matching method, might be a good solution instead of matching directly on multiple observed variables.
- Propensity score matching and/or stratification methods can help achieve groups equivalence with respect to participants' pre-test differences.

Paper #1. Statistical Matching of Group Membership

- A big challenge for conventional DIF analyses is that they can only detect DIF, but can not explain why DIF occurs.
- For example, DIF can occur between two groups of test takers writing a test in different languages. However, it is not clear whether DIF is caused by translation or other linguistic, educational, and/or cultural background accompanying the test takers.

Paper #1. Statistical Matching of Group Membership

- Liu, Zumbo, Gustafson, Huang, Kroc, & Wu (2015) extended PSM and stratification methods in the context of the effects of translation.
- Item #13 was an item detected as DIF by all DIF methods, while item #22 produces controversial results from different methods.
- Item #22 initially showed “translation DIF” with conventional methods; which was no longer showing DIF after PSM matching … translation is not the reason for DIF.

Paper #1. Statistical Matching of Group Membership

Table 3. Results of DIF Analyses for Items #13 with Raw Data, Matched Data, and Stratified Data

	Estimate	exp(coef)	s.e.	z value	Pr(>|z|)	
Conventional Logistic Regression (Raw Data)						
language	0.465	1.592	0.083	5.598	< 0.001	***
total	1.186	3.274	0.098	12.040	< 0.001	***
language*total	0.180	1.197	0.107	1.689	0.091	
Conditional Logistic Regression DIF (Pair Optimal Matching)						
language	0.465	1.59	0.105	4.43	< 0.001	***
total	0.923	2.52	0.168	5.49	< 0.001	***
language*total	0.194	1.21	0.154	1.26	0.210	
Conditional Logistic Regression DIF (full Optimal Matching)						
language	0.373	1.45	0.0925	4.03	< 0.001	***
total	1.184	3.27	0.1375	8.61	< 0.001	***
language*total	0.141	1.15	0.1337	1.05	0.290	
Logistic Regression DIF for the Stratification						
language	0.436	1.547	0.092	4.743	< 0.001	***
total	1.151	3.163	0.107	10.789	< 0.001	***
language*total	0.208	1.231	0.115	1.811	0.070	

Note. Significance codes: *** = p-value $\leq$ 0.001; ** = p-value $\leq$ 0.01; * = p-value < 0.05

Paper #1. Statistical Matching of Group Membership

Table 5. Results of DIF Analyses for Items #22 with Raw Data, Matched Data, and Stratified Data

Regular Logistic Regression (Raw Data)						
	Estimate	exp(coef)	s.e.	z value	Pr(>|z|)	
language	**1.210**	**3.354**	**0.568**	**2.131**	**0.033**	*
tot2	0.155	1.167	0.021	7.390	<0.001	***
language*tot2	-0.056	0.945	0.038	-1.479	0.139	
Conditional Logistic Regression DIF (Pair Optimal Matching)						
language	0.375	1.455	0.248	1.515	0.130	
tot2	1.889	6.615	0.407	4.639	0.000	***
language*tot2	-0.276	0.758	0.282	-0.982	0.330	
Conditional Logistic Regression DIF (full Optimall Matching)						
language	0.388	1.475	0.209	1.86	0.063	
tot2	1.921	6.828	0.268	7.17	0.000	***
language*tot2	-0.246	0.782	0.249	-0.99	0.320	
Logistic Regression DIF for the Stratification						
language	**0.397**	**1.487**	**0.189**	**2.099**	**0.036**	*
tot2	2.116	8.301	0.200	10.595	<0.001	***
language*tot2	-0.333	0.717	0.187	-1.778	0.075	

Note. Significance codes: *** = p-value ≤ 0.001; ** = p-value ≤ 0.01; * = p-value < 0.05

Paper #1. Statistical Matching of Group Membership

- **Item 13: the differences in probability of responding correctly is due to translation. [Conclusion: DIF and hence item bias attributable to translation]**

- **Item 22: once there is correct matching by propensity scores, there is no difference in probability of responding correctly. Therefore, differences in probability of responding correctly is no longer statistically significant. The item is not biased due to test language.**

Paper #2. Ecological Mixture Model for Item Responding

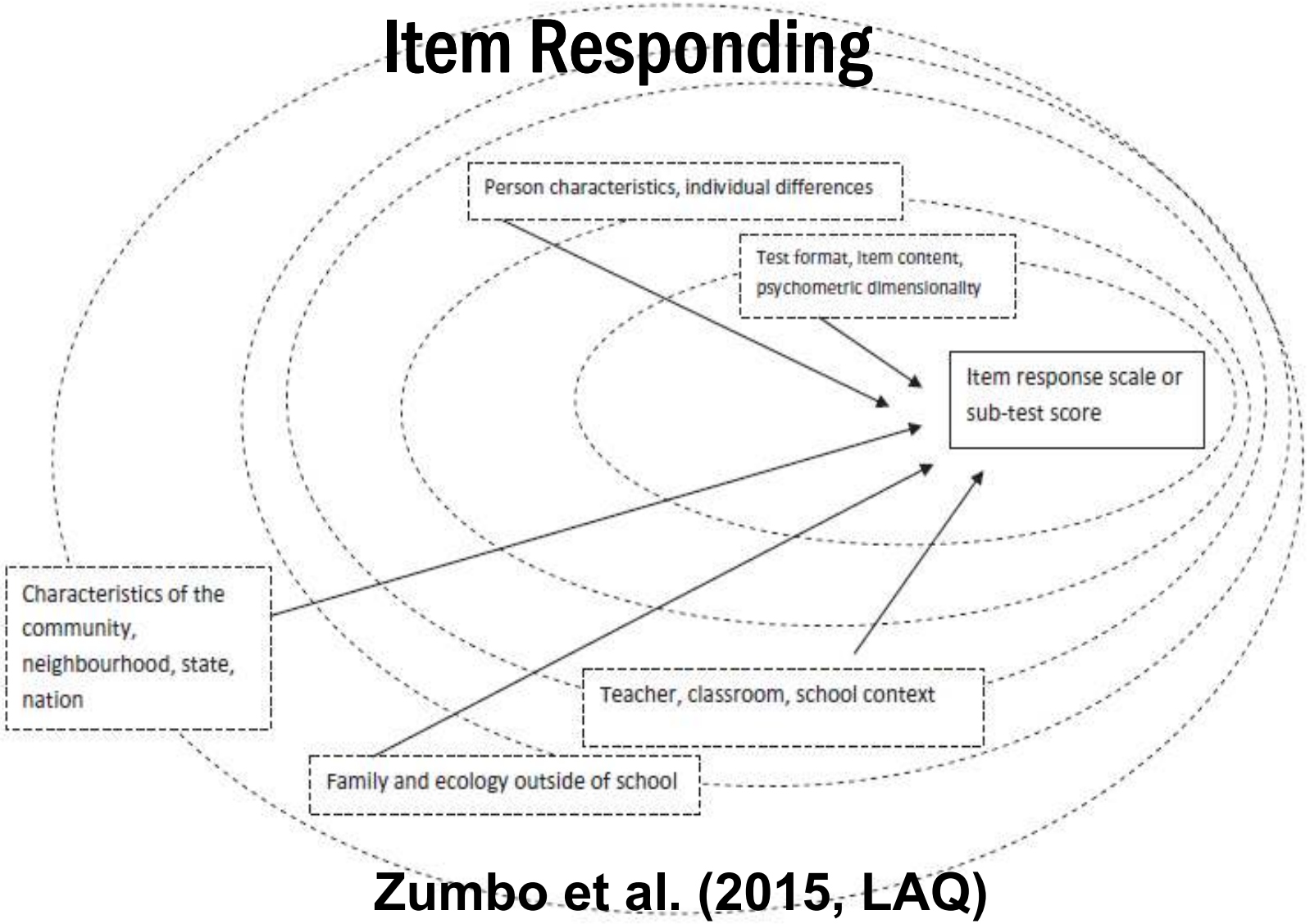
"A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding"

(Zumbo et al., 2015)

Paper #2. Ecological Mixture Model for Item Responding

- An ecology of item responding includes:
  - test format, item content, and psychometric dimensionality;
  - person characteristics and typical individual difference variables such as cognition;
  - teacher, classroom, and school context;
  - the family and ecology outside of the school; and finally
  - characteristics of the community, neighbourhood, state, and nation.

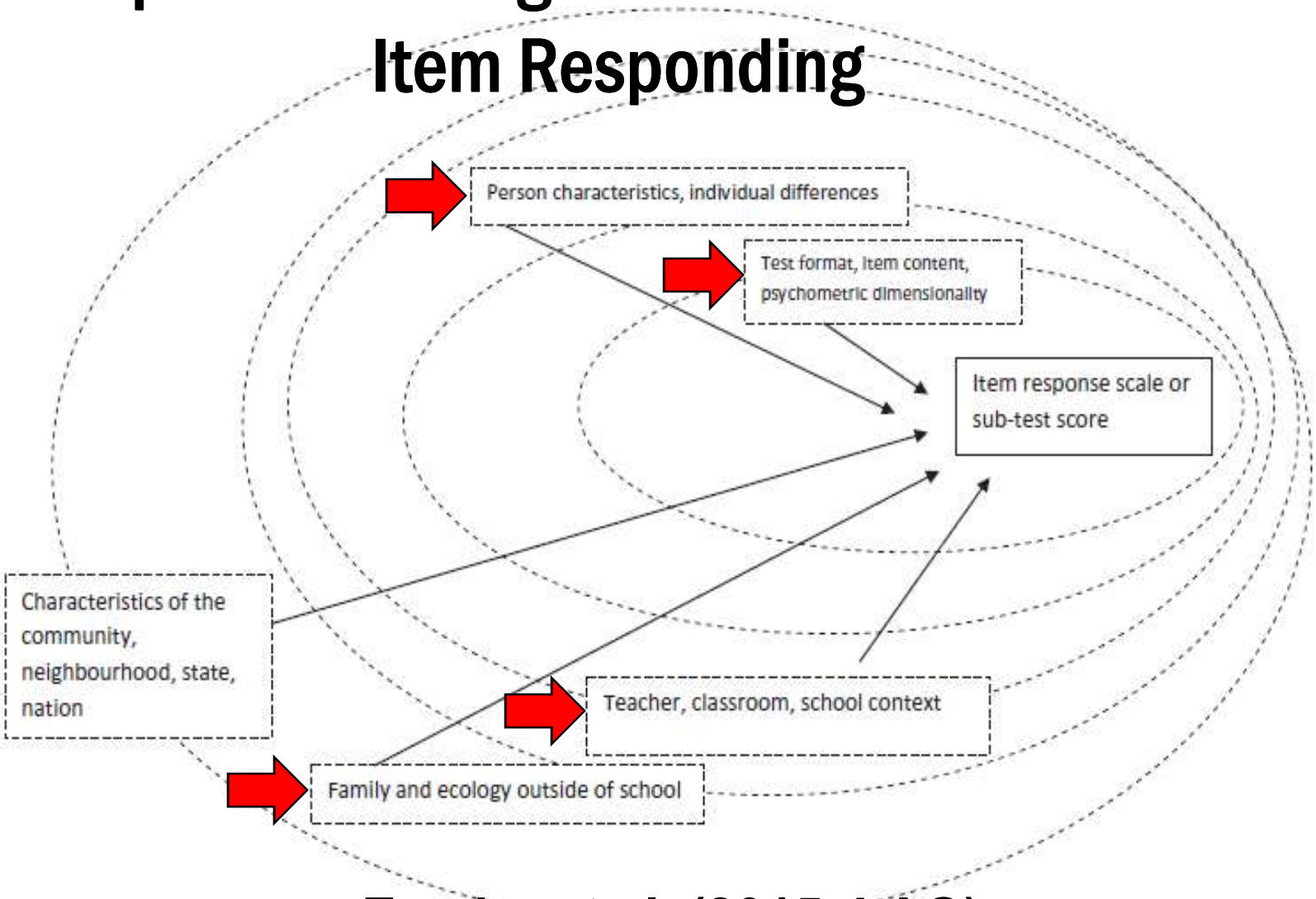
Paper #2. Ecological Mixture Model for Item Responding



Zumbo et al. (2015, LAQ)

FIGURE 1 An ecological model for item responding.

Paper #2. Ecological Mixture Model for Item Responding



Zumbo et al. (2015, LAQ)

FIGURE 1 An ecological model for item responding.

Paper #2. Ecological Mixture Model for Item Responding

- Influenced by ecological systems theory (e.g., Bronfenbrenner,1979)

- Conventional first and second generation DIF practices have focused on the first oval with some modest attempts at the second oval as sources for explanation for DIF.

Paper #2. Ecological Mixture Model for Item Responding

- In this study, a mixture DIF analysis was investigated, i.e., examining whether there are unknown groups of test takers that showing different DIF patterns
- That is, we examined whether DIF was moderated by the latent class.
- Variables from four layers of the ecological system were incorporated to understand the profiles of the latent classes.

Paper #2. Ecological Mixture Model for Item Responding

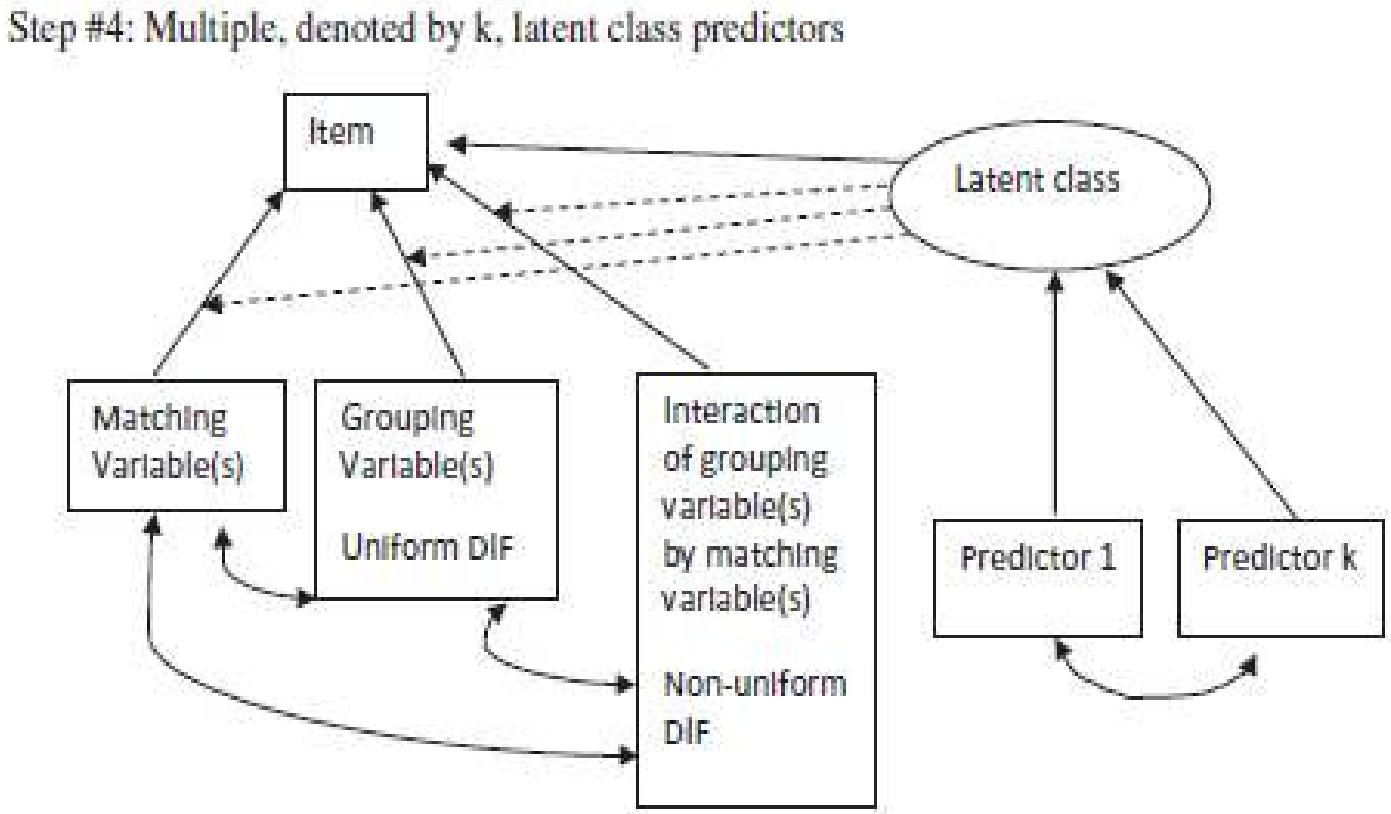


FIGURE 2 The steps in the latent class logistic regression DIF method and the path diagram of the models.

Paper #2. Ecological Mixture Model for Item Responding

- In contrast to the conventional DIF analysis in which we learned that the test language DIF simply favoured students taking the English version.
  - Using latent class logistic regression DIF we learned that test language DIF is moderated by a latent class variables – in our case two latent classes for which the language DIF effect was in opposite directions.

Paper #2. Ecological Mixture Model for Item Responding

- What are the predictors that profile the difference between the latent classes?
- We found a statistically significant predictor from each layer of the ecological model of item responding.

TABLE 1
Outcomes of Step 3, Latent Moderated Logistic Regression DIF With One Predictor at a Time

Levels of Explanatory Variables in the Ecological Model	*Sample size*	*Statistical Significance of Predictor for the Latent Class Variable*
Item Content		
Meta-cognition: student reports of the usefulness of the strategy "summarizing" of a long and rather difficult two page text	1,693	$p = .013$
Meta-cognition: student reports of the usefulness of the strategies such as concentration, quickly read, discuss with others for understanding and memorizing the text	1,695	n.s.
Person characteristics, individual differences		
Gender/Sex	1,750	n.s.
Like read – fiction	1,704	n.s.
Like read – non-fiction books	1,699	$p = .001$
Joy/like reading	1,683	n.s.
Teacher, classroom, school context		
At school – Group work	1,690	n.s.
Time – Language lessons	1,685	n.s.
Time – Other language lessons	1,684	n.s.
Teachers stimulation of reading engagement	1,704	n.s.
Teacher student relations	1,708	n.s.
At school – homework	1,688	$p = .025$
Family and Ecology outside of School		
Index of economic, social and cultural status	1,706	n.s.
Amount of time spent reading for enjoyment	1,701	$p = .019$
Highest parental education in years	1,658	n.s.
Wealth	1,711	n.s.
Highest educational level of parents	1,696	n.s.
Home educational resources	1,711	n.s.
Online reading	1,714	n.s.

Note. The sample sizes differ because there were differing missing values in the predictors; "n.s." denotes a statistically nonsignificant predictor.

Paper #3. Ethnographic-Psychometric Methods

"An Anthropologist Among the Psychometricians: Assessment Events, Ethnography, and Differential Item Functioning in the Mongolian Gobi" (Maddox et al., in press, IJT)

Paper #3. Ethnographic-Psychometric Methods

- The purpose was to explores the potential for ethnographic observations to inform the results of DIF analysis
- In 2010, a standardized, large-scale adult literacy assessment took place in Mongolia as part of the United Nations Educational, Scientific and Cultural Organization Literacy Assessment and Monitoring Programme.
- An ethnographer worked closely with psychometric researchers to investigate the sources and explanations of item responding.
- Ethnographic observations took place in Mongolia over a three-week period (October and November 2010), with the ethnographer and interpreters accompanying testing teams.

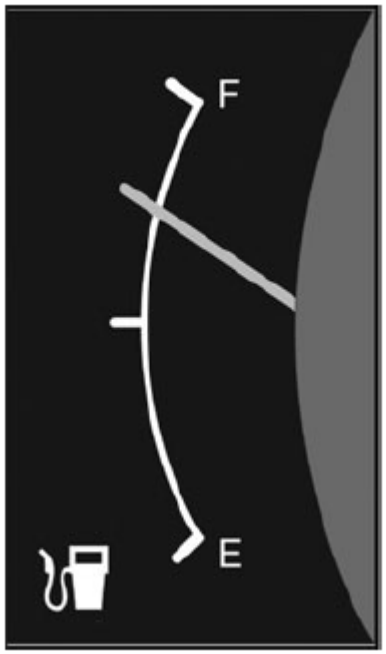
Paper #3. Ethnographic-Psychometric Methods

- Twenty five assessment events were observed in people's homes. Informal interviews and observations also took place to ask about people's experiences of the assessment and their everyday literacy practices.

Paper #3. Ethnographic-Psychometric Methods

- Post-hoc explanation for DIF items of international assessment by subject-matter (literacy) experts who are often geographically or cultural distant from the testing situation often has limited value.
- Embedding ethnography in the operational testing context could help in understanding DIF
- Neither statistical DIF analyses nor the ethnographic approaches are privileged. Both are meant to serve each other.

An Example – Gas/Fuel Gauge



- Designed by Statistics Canada
- The item presents a visual stimulus of a gas/fuel gauge and provides information on the total capacity of the gas/fuel tank.
- The respondents are asked to read the gauge and to calculate the amount as a proportion of the total capacity.

An Example – Gas/Fuel Gauge

- The gas/fuel gauge uses the initials of F and E to indicate Full and Empty. The Mongolians Cyrillic script does not share the letters F and E. This is likely to advantage those who more frequently read gas/fuel gauges.
- The modified MH-DIF test confirmed an ethnographic hypothesis of the likely presence and plausible explanation of gender-based DIF. Mongolian women did less well in this test item when compared to men at the same level of predicted ability.
- The ethnographer was advised that in Mongolia women less frequently drive cars or motorbikes and are less familiar with reading gas/fuel gauges.

Another Example – Timed Parking

- Designed by Statistics Canada. The Timed Parking test item asks the respondents to calculate how long a car has been parked in an automated car park using information on the stimulus including a parking ticket with a time (indicating when a person arrived at the car park 9:45) and the face of a clock (indicating the time when the person departed, 11:20).
- The ethnographic data and DIF analysis provide contrasting perspectives about the performance of the test item in the Mongolian setting.
- The ethnographic data illustrates the subtle but important impact of cultural norms and practices for the way people utilize local knowledge and engage with problem-solving tasks in the test items.

Another Example – Timed Parking

- Mongolia has wide open spaces and absence of paved roads. When people want to park in rural areas they just pull up somewhere (or tie up their horse). The practice of timed parking (automated ticket machines, payment per hour) is therefore unfamiliar to Mongolian respondents.

- In Mongolia the winter temperature drops to –50C. For that reason covered parking is extremely desirable. However, underground parking is generally associated with parking at home (e.g., in the basement of urban housing complexes). As far as we know, there is currently only one underground car park in Mongolia and it does not use an automated ticket machine.

- In Mongolia people tend to only pay for overnight parking.

Another Example – Timed Parking

- The item showed no gender, urban/rural herder/non- herder based DIF.
- The ethnographic study identified an unexpected source of difficulty associated with the time parking test time, but DIF was not detected because the source of difficulty was experienced by most Mongolian respondents.
- If countries is to be compared, it is expected that DIF will be flagged.

Section 3

Contrasting the Approaches with an Eye toward Conclusions

Contrasting & Conclusions

- All 3 of the papers have an 'explanatory' objective, but they differ in what they mean by 'explanation'; different lenses.
  - Propensity score is focused on [manipulation/experiment view of] causal explanations, and the balancing of pre-existing difference, removing the confounders by an elaborate statistical form of matching.
  - Latent class approach is focused on multilayered ecological variables as explanatory … with a Bronfrenbrener lens of ecology … very broad lens of explanation

Contrasting & Conclusions

- **Ethnographic approach:**
  - There is a rhetorical move from how the environment affects the person to a type of interactivism in which the test taker is situated within these (dis)enabling conditions and highlights processes and cultural norms and practices for the way people utilize local knowledge and engage with problem-solving tasks.

How are Covariate Variables Constructed and Collected?

- Propensity score and ecological mixture models: constructed by the researcher; not collected while completing the test ... before or after the test
- Ethnographic approach: the research did not start with a list of variables that were planned to be collected. Relevant information emerged in the observations.

How the 'Covariates' and 'Background' Variables Are Treated?

- Propensity score approach treats them as confounding variables to the causal claim of DIF.
  - The intention is to remove the influence of the covariates.
- Ecological mixture approach uses them as (i) information to profile the latent classes, and (ii) to understand the latent class moderated DIF.
  - The intention is to incorporate the covariate information in order to understand DIF.

How the 'Covariates' and 'Background' Variables Are Treated?

- In the ethnographic approach, the covariate information '*emerged*' from the ethnographic observations; used to inform working hypotheses for DIF tests.
  - focuses on contextualized explanation from a perspective of sociocultural anthropology manifested through the interactions at the test administration (immersed in the 'test culture' as an ethnography)
  - proposes a psychometric-ethnography (or ethnographic-psychometrics) where both are on an even footing; one does not solely serve the other.
  - Background variables are not removed (or controlled, nor treated as predictor/explanatory variables in the statistical sense) but rather emerge as part of the ethnographic observations to explain DIF.

Conclusion

- All of these approaches are instantiations of 3rd Generation DIF, which is meant to focus on a deeper understanding of DIF.
- Together, all three examples, are not simply about quality control but rather about a nuanced understanding of context, culture, and constructed explanations of the of item responding.

Conclusion

- As a community of scholars we tend to equate explanations in the language of 'causal' claims with 'treatment effects' and 'experimental manipulation'.
- One of our points is that there is a need for an expanded view of explanation beyond causal claims via treatment effects.

Some Central Thought Questions

- Some central questions are:
  - Why are we doing the DIF analyses?
  - What do we mean by fairness?
    - Is it a matter of 'equality' of treatment effects?
    - Is it a matter of 'equity', from what perspectives?
    - Is it a matter of explanatory evidence of DIF; in that greater understanding of variation in test performance can lead to more equitable assessment practices?

Thank You For Your Attention!

Bruno D. Zumbo, PhD, Paragon UBC Professor of Psychometrics & Measurement

bruno.zumbo@ubc.ca

Amery D. Wu, PhD, Assistant Professor of Measurement, Evaluation & Research Methodology

amery.wu@ubc.ca