Equity and Fairness at the Nexus of Data Science, Psychometrics, Digital Innovation, and Social Justice

Invited Address to the Quantitative Methods Section,
Canadian Psychological Association convention,
June 17, 2021

Bruno D. Zumbo

Professor & Distinguished University Scholar
Tier 1, Canada Research Chair in Psychometrics and Measurement
Paragon UBC Professor of Psychometrics and Measurement

University of British Columbia

What follows is the text of the 2021 featured invited address to the Quantitative Methods Section of the Canadian Psychological Association (CPA).

Citation:

Zumbo, B.D. (2021, June 17). *Equity and Fairness at the Nexus of Data Science, Psychometrics, Digital Innovation, and Social Justice* [Featured Invited Address]. Quantitative Methods Section of the Canadian Psychological Association (CPA) [virtual conference]. https://brunozumbo.com/?page_id=31

Acknowledgement of funding: I am grateful for the ongoing support from:

- ***Social Sciences and Humanities Research Council of Canada*** *(SSHRC)* for their grant support
- ***Canada Research Chairs Program (CRCP)****,* for the Canada Research Chair in Psychometrics & Measurement – Tier 1
- ***UBC Distinguished University Scholars*** program
- ***UBC*** and ***Paragon Testing Enterprises, a Prometric Company***, for co-funding the Paragon UBC Professor of Psychometrics and Measurement and the Research Initiative.

Opening Remarks

- I am thrilled and honoured to be the 2021 Quantitative Methods Section of CPA, Featured Speaker
  - This is a kind of return visit as I gave the address to this esteemed group of scholars at CPA in 2014.
- Today I will focus on the main ideas defining my program of research appointment 2020-2027 as

Tier 1, Canada Research Chair in Psychometrics and Measurement

with its central theme of Equity and Fairness at the Nexus of Data Science, Digital Innovation, and Social Justice.

  - Please keep an eye out for a series of **key messages [in red font]** highlighted throughout the presentation.

Setting the Stage

- Influenced by varied historical events, cultures, and technological developments, we are facing a whole new world of digital innovation,

  as well as a moral and ethical social justice imperative of the consequences of measurement, surveys, and testing.

Zumbo, B.D., & Hubley, A.M. (2016). Bringing Consequences and Side Effects of Testing and Assessment to the Foreground. Assessment in Education: Principles, Policy & Practice, 23, 299–303.

Setting the Stage

- Today, tests and measures continue to be widely used for research, decision-making, ranking, and policy purposes in the social, behavioural, and health sciences using large-scale testing, regularly administered tests of a population over time, assessment of individuals, as well as social, health, and economic surveys.

Setting the Stage

1. The concept, method, and process of validation are central to social, psychological, and health science research, for without validation, any inferences made from a measure are meaningless.

2. Throughout this presentation, the terms measure, instrument, test, assessment, survey, and scale will be used interchangeably and in their broadest senses to mean any coding or summarization of observed phenomenon.

3. Furthermore, *lest we fall into traditional camps and comfortable silos*, **validity applies equally to tests or measures used in**

- language assessment,
- educational measurement,
- certification and licensure testing,
- social indicators,
- psychological instruments
- health measurement,
- measures of health status,
- patient-reported outcome measures (PROMS),
- patient-reported experience measures (PREMS),

Measurement Invariance in the Psychometric Literature

- Absence of measurement invariance, which is also known as differential item functioning, has been studied extensively both in the context of confirmatory factor analysis and item response theory.
- You will have seen measurement invariance defined with respect to a grouping or selection variable, S, such as gender, and concerns the measurement model relating observed scores to underlying latent variables.
- The measurement model has been treated as the same for all groups in the sense that the probability of observing a given item score is equal for members of different groups who have the same score on the underlying latent variable.

Measurement Invariance in the Psychometric Literature

It is widely seen in the research literature that, more formally, measurement invariance has been defined as

$$f\left(Y\middle|\eta,s\right)=f\left(Y\middle|\eta\right),$$

where observed variables are denoted as Y, latent variables as η, and the grouping variable as S.

Measurement Invariance in the Psychometric Literature

A situation where measurement invariance is absent, that is,

$$f\left(Y\middle|\eta,s\right)\neq f\left(Y\middle|\eta\right),$$

an observed variable *Y* is non-invariant with respect to a grouping variable *S* if the observed score depends not only on the latent variables η but also on *S*, or variable(s) related to *S*.

Measurement Invariance in the Psychometric Literature

A situation where measurement invariance is absent, that is,

$$f\left(Y|\eta,s\right) \neq f\left(Y|\eta\right),$$

an observed variable *Y* is non-invariant with respect to a grouping variable *S* if the observed score depends not only on the latent variables η but also on *S*, or variable(s) related to *S*.

Following the seminal work of Mellenbergh (1989) and Meredith (1993), there are three different types of effects of *S* or variable(s) related to *S*, that may or may not occur simultaneously:

- Constant for all possible scores on η, which results in a group difference in the intercept of the regression of Y on η.
- The effect can increase or decrease as a function of η, resulting in a group difference with respect to the steepness of the regression.
- The regression curves (or non-linear regression) on η are equal across groups, but the residuals of the regression differs.

Measurement Invariance in the Psychometric Literature

- In the decades since this early work by Mellenbergh and Meredith, there has been an enormous amount of research that has articulated very clever and useful analytical methods (e.g., MG-CFA, Bayesian Alignment methods, IRT based approaches, MH, SIBTEST, GLIM models).
- DIF is a statistical phenomenon that can occur in any item of a test.
  - In addition, **differential test functioning (DTF)** occurs when test takers of the same ability do not receive the same overall test score – in some cases greater emphasis is placed on test-level analyses that allow items favoring one group to cancel the DIF of items favoring another group.

Measurement Invariance in the Psychometric Literature

- The seminal work of Mellenbergh (1989) and Meredith (1993) did what it was supposed to do very well.
  - It was intended to formalize the definition of measurement and give us a lens from which to think about invariance and develop methods.
- What has emerged in the last 15 years is a promising nuanced framework and class analytic methods to carry forward in the Mellenbergh-Meredith tradition

PROMISING NUANCED FRAMEWORK AND ANALYSES ARE EMERGING

- Theoretical framework and a class of analytic methods

a) to build on Zumbo's (2007) description of Third Generation DIF by introducing the ecology of item responding, and

b) to also introduce a family of new psychometric DIF methodologies that are particularly well suited for this ecological Third Generation view of DIF.

Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223-233.

Third Generation DIF

In 2007 in *Language Assessment Quarterly* Zumbo described three generations of DIF and introduced Third Generation DIF.

- The matter of wanting to know why DIF occurs is an early sign of the third generation of DIF.
- Third Generation DIF is best characterized by a subtle, but extremely important change in how we think of DIF.
  - In the third generation DIF is conceived 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.

Third Generation DIF

- By adding and highlighting "**testing situation**" as a possible reason for DIF, one greatly expands DIF praxis and theorizing to matters beyond the test structure itself,
  - hence moving beyond the multidimensional model of DIF that emerged from the second generation.
  - In short, *Third Generation DIF* is part of building an ecological model of item responding and assessment.
    - The ecology of item responding, as Zumbo and Gelin (2005) note, allows the researcher to focus on sociological, structural community and contextual variables, as well as psychological and cognitive factors, as explanatory sources of item responding and hence of DIF.
    - Woitschach, Zumbo, and Fernández-Alonso (2019) extend this ecological view of measurement focusing on multilevel model explanation of differential item functioning.

Figure 1. An Ecological Model for Item Responding

Note: Five concentric ovals but could be more, or others.

1st & 2nd Generation DIF practices have focused on the first oval with some modest attempts at the second oval as sources for explanation for DIF.

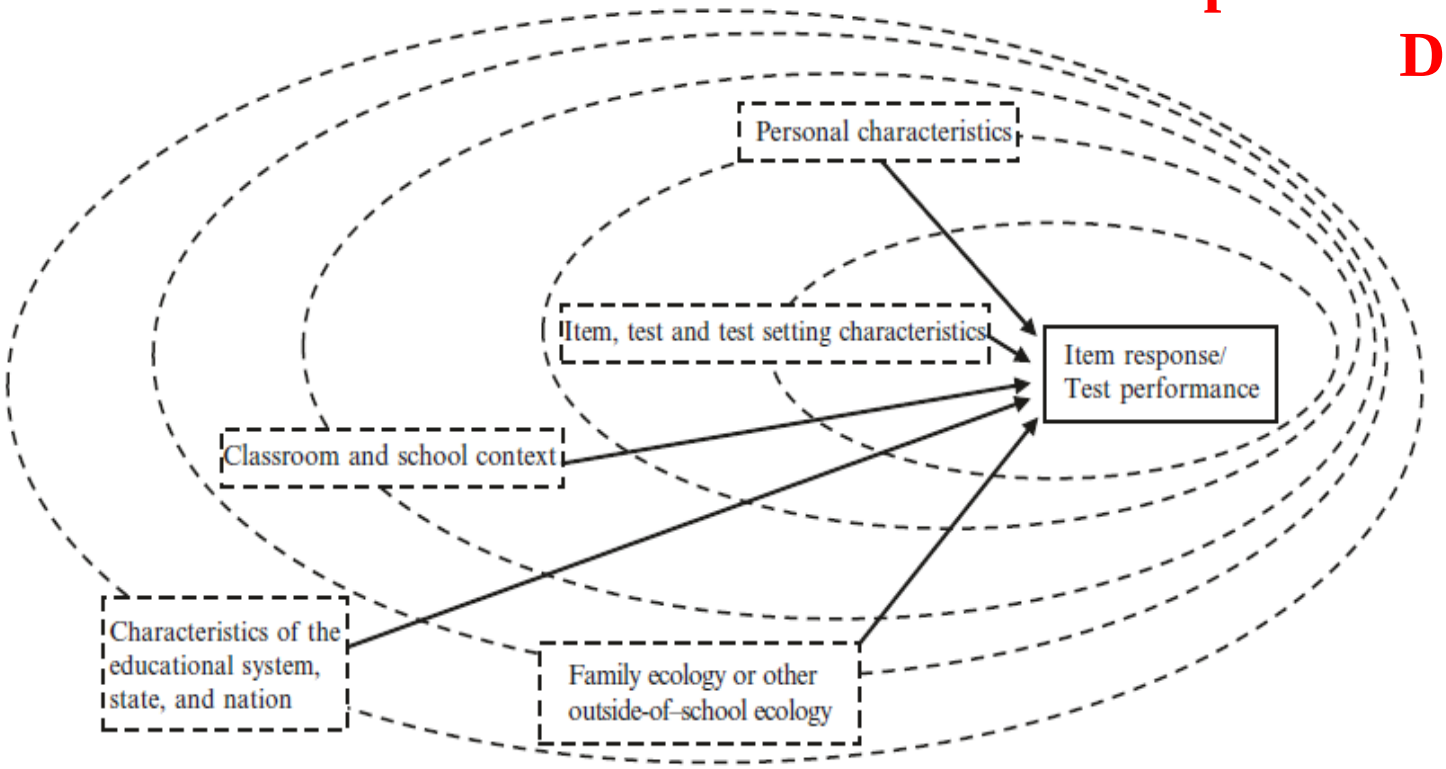

Figure 1. An Ecological Model for Item Responding

- A few points are worthy of note as we transition to the psychometric methods.
  - First, clearly, our model is influenced by ecological systems theory (e.g., Bronfenbrenner, 1979).
  - Linking DIF to the broader issue of measurement validity, the ecological model further articulates what is meant by 'context' in Zumbo's (2009) view of validity as contextualized and pragmatic explanation – the multilayered ecology is the context.
  - Lastly, the ecological model serves as a foundation for the a family statistical and psychometric methodology of DIF analysis.

Figure 1. An Ecological Model for Item Responding

- This theoretical framework that shares salient features of Bronfenbrenner's bioecological model of development (e.g., Brofenbrenner, 1993, Brofenbrenner & Morris, 2006).
  - This model recognizes that item responding is shaped by characteristics of the respondents themselves, the environments in which they are embedded at multiple levels, and the processes that they engage in with these multiple environments.
- We explicitly and systematically allow the model to guide and inform our thinking about factors influencing item responding and scale scores.
  - This is a **substantial departure from prior models** of item responding.

Latent Class Logistic Regression DIF

- Conventional DIF methods focus on manifest grouping variables, such as gender, language of the assessment, and are primarily meant to be used for detecting or flagging potentially problematic items (Zumbo, 2007).
- In *Third Generation DIF* methodology, the use of latent variable mixture models, particularly
  - mixture item response theory (IRT) and mixture Rasch methods, have proven to be useful tools for detecting latent groups and testing postulated explanatory models for potential causes of DIF (e.g., Cohen & Bolt, 2005; De Ayala, Kim, Stapleton, & Dayton, 2002; von Davier, & Yamamoto 2007).

Latent Class Logistic Regression DIF

- We introduced a family of latent class (mixture) logistic regression models that, unlike the previous IRT-based approaches, are extensions of widely used logistic regression DIF methods to allow for latent classes.
  - Traditional models used in DIF logistic regression analysis contain parameters that describe only relationships between observed variables, however, latent class models differ from these by including one or more discrete latent variables. This family of logistic models can deal with binary and ordinal item response, or their combination, in an assessment.
- Two routes to arriving at mixture models:
  - Factor Analysis → Mixture Models
  - Kernel Density Estimates → Mixture Models

Latent Class Models, also travel under the name Mixture Models - Model specification (Grün & Leisch, 2008)

A finite mixture density of Generalized Linear Models with K components is given by

$$h(y|x,\Theta) = \sum_{k=1}^{K} \pi_k f_k(y|x,\theta_k)$$

where Θ denotes the vector of all parameters for the mixture density h().

- The dependent variable is y and the independent variables are x, and f_k is the component specific density function which is assumed to be univariate and from the exponential family of distributions.
- The component specific regression and dispersion parameters

$\theta_k = (\beta_k', \phi_k)$ where β_k are the regression coefficients and ϕ_k

The many ways of being human

- The principle, as I see it, that there are many ways to be human is at the core of how I live and theorize and how I express my Canadian experience and identity.
- Over the last 30 years my experience has been that the field of psychometrics, in particular, has tended to go into a moral panic over gender identity, gender expression, and aspects of cultural expression.
  - At the core of my theorizing and the methods I develop and/or advocate aim to challenge that view and aim to honour the many ways of being human and capturing the human experience.

An Example: Ecological Model of Item and Test Responding

In studies of fairness gender has, in the main, been characterized in the binary as biological sex wherein (binary) biological sex differences on item performance that are eventually explained by item characteristics such as item format and item content.

In what I described in 2007 as my Third Generation DIF "gender" more properly should be considered a social construction, and gender differences on item performance are explained by contextual or situational variables (ecological variables, if you wish), such as institutionalized gender roles, classroom size, socioeconomic status, teaching practices, and parental styles.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4(2)*, 223–233.

Ecological Model of Item and Test Responding

We believe that these richer ecological variables have been largely ignored in relation to explanations for (and causes of) DIF because of the focus on test format, content, cognitive processes, and test dimensionality that is pervasive in the second generation of DIF.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Astivia, O.L.O. & Ark, T.K. (2015). A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding. *Language Assessment Quarterly, 12*, 136-151.

What my approach to fairness and equity implies …

- Traditional views follow a "social address" model of criterion prediction and group differences.
  - This spills over in to test validation; group differences.
- In using the common "social address" approach to group comparisons, classification into groups might be confused with fixed biological or ethnic classification.

  As John Stuart Mill (1848) wrote:

  > Of all the vulgar modes of escaping the consideration of the effect of social and moral influences on the mind, the most vulgar is attributing the diversities of conduct and character to inherent natural differences. (p. 319).

What my approach to fairness and equity implies …

- In a series of chapters and papers from 1998 to 2017, I have made the case that the aim is: identifying the determinants (or explanatory theory) of task / item / test score variation … the explanation is the basis of any strong validity claims.
- I take an ecological systems approach
- Most research on response processes focuses on cognitive factors.
  - We have taken a broader view of response processes proposed by Zumbo & Hubley (2017) and embrace the notion of assessment 'in vivo' to shine a spotlight on test-takers' behaviour, stance, gesture, motivation, and affect besides cognition.
  - My approach highlights response processes, explanation of task / item / test variation, and is 'in vivo' (Zumbo & Hubley, 2017)

Reflections on the Theme of this Session

- Today's presentation hints to the question of "context" and "culture" as sources for abductive explanatory sources.
- In my opinion, "context" and "culture" in assessment research allow us to explore how the various ways of being human interact with measurement and assessment.

Some final thoughts …

- This line of research has me interfacing with anthropologists [Bryan Maddox, Lidia Jendzjowsky]:
- There is no one definition or conceptualization of 'culture' in anthropology.
- Here are some themes:
  - *Culture is the lens through which life derives meaning for individuals.*
  - *It centrally shapes human development and human expression.*

Some final thoughts …

- We should avoid (and challenge) homogenous cultural interpretations.
  - "Although culture is most obviously identifiable through variations in race, ethnicity, and national origins, it is increasingly recognized that diversity in attitudes and behavior within one racial or ethnic group arises due to age and class differences."
  - I believe the field of surveys, testing, and assessment will benefit by considering the utility of understanding age (generational) class differences as a matter of culture.

Some final thoughts …

- Including culture as a central element in assessment research will advance the field significantly because of the way it illuminates connections between macro- and microlevels of human experience and highlight how the various ways of being human interact and express themselves.

Thank you

TESTING ENTERPRISES
Paragon
CANADA'S LEADER IN ENGLISH LANGUAGE TESTING