


On Models, Modeling, and Measurement Invariance in Validation Studies: A Stochastic View of Measurement

Invited Plenary Address in the session “Validation of PROs for decision-making” to the 28th Annual Conference of the ***International Society for Quality of Life Research (ISOQOL)***

October 14, 2021  ISOQOL

Bruno D. Zumbo

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



University of British Columbia

The focus of today's presentation

Models

What are they?
I wish to shine a light on
a kind of model-
inferences.

Modeling

How do use modeling and
for what purpose?
What do we do when we
model?

Measurement Invariance

What is it? What is its role at the intersection
of 'models' and 'modeling'?

Foreshadowing the Take-Home Messages:

- I will describe elements of my stochastic view of measurement that invites (*nay, urges*) us to challenge our tacit homogeneity assumptions; assumptions about the uniformity of the measurement of the phenomenon of interest.
 - I describe the interconnected concepts of model, modeling, and measurement invariance.
 - I depict measurement invariance at the intersection of model and modeling.
 - I will describe how **measurement invariance** can be seen not only as central to psychometric practices but as a **key to unpacking validity and validation practices**.
 - Measurement invariance studies are not just about item analysis, quality assurance and fairness, they are **important studies that inform the measurement validity argument**.

Foreshadowing the Take-Home Messages:

- In describing my stochastic view of measurement, I wish to highlight how:
 - 1) to conceptualize measurement invariance from my Draper-Lindley-de Finetti (DLD) framework that invokes a kind of exchangeability,
 - 2) the practice of modeling shores up both the centrality of the 'model' and the concept of model uncertainty, and
 - 3) the space spanned by model uncertainty, sampling uncertainty, and for example uncertainty invoked by repeated testing (and response shift) bound the measurement claims one can make of invariance.
- Reflecting how psychometric models are used, and what they provide for the psychometric modeler, a complementary backup anchor that bounds "what we can say" is described in the form of the DLD framework as a necessary feature of formal models.

Foreshadowing the Take-Home Message:

- Although my view has distinct stochastic features, a useful take-home message is that a **psychometric model is also a kind of predictive machine** for observable qualitative categorical item responses as well as quantities (or categories) that may not be observable in practice, i.e., **latent variables**, but nevertheless are defined within a credible data structure.
 - Although limited in its value and use, the machine analogy is apt because it is made from objective parts of an algorithm intended to interact with elements of an objective external world- e.g., estimators, data structures implied by test design and use, and assumptions. Hence, it encourages the recognition of complementary objective aspects of a model.
 - In short, for our purposes, a **model is more than just platonic mathematical objectives**.

Let's start with 'models'

Models

What are they?

I wish to shine a light on
a kind of model-
inferences

On Models

- As is evident from even a cursory glance at our praxis, **models and the process of modeling** are growing in importance and centrality in the empirical analyses of our measures and as an expanding evidential basis for validation practices to **establish the claims** made from them.
 - I will remain vigilant to signal when I am talking about latent variable models, of the factor analytic and IRT variety.

On Models: Reasons and reasoning behind models in my stochastic view of measurement

- Activities undertaken by a statistical modeler are associated with three common goals of approximation of a measurement process, explanation or understanding of the measurement process, and prediction in support of the inferential leap from item or task responses to claims about the status of a respondent and claims about the instrument itself.
 - As the philosopher of science, Bas van Fraassen (2008) highlighted in his study of the history and philosophy of measurement, the theory of the phenomenon and its measurement cannot be answered independently of each other, and that they co-evolve-- a point that Cronbach and Meehl (1955) highlighted as well.

On Models: Reasons and reasoning behind models in my stochastic view of measurement

- The approach I am describing takes from all three major traditions of statistical modeling:
 - a) the **mathematical scientists'** (probabilist's) approach of associating stochastic models with classes of phenomena,
 - b) the **data analyst's** approach of fitting empirical models to data, and
 - c) the **subjectivist (Bayesian) statistician's** approach of constructing formal relations that represent the uncertainty of the idealized formal setting.

On Models

- Contemporary measurement and validation practices, which are **heavily model-based**, the inferences, in part, arise from and are **supported by the model itself**.
- In short, the statements about the validity of the inferences from the test scores rest, in part, on the measurement model.
 - Discussions of ‘validity’ and ‘validation’ are framed and shaped by the measurement and psychometric models employed, be they classical test theory, item response theory, factor analysis, mixture models, or some hybrid.

On Models (As Zumbo, 2007a, stated:)

- The function of the psychometric model is to **step in when the data are incomplete**.
 - In an important sense, we are going from what we have to what we wish we had.
 - If we had available the complete data or information, then we would know the true score, or theta in IRT models, and no statistics beyond simple summaries would be required.
 - There would be no need for complex models to infer the unobserved score from the observed data and, hence, no need to check the adequacy and appropriateness of such inferences.

On Models

- The measurement model helps us **travel from the item responses** to the test takers' response processes and/or their status on the latent variable of interest.
 - Therefore, not surprisingly, one's test score interpretations may change depending on the psychometric statistical model being used.

Characteristics of 'modeling'

Models

What are they?
I wish to shine a light on
a kind of model-
inferences

Modeling

How do use modeling and
for what purpose?
What do we do when

Transition from Models to Modeling (As Zumbo, 2007a, stated:)

- The occurrence of complete data or full information, as I describe it, is not commonly encountered, if ever, in the practice of measurement.
- Naturally, this leads to the common experiences that are the defining characteristics of what we call modeling:
 - the data you have is **never really the data you want** or **need** for your attributions, recommendations or decisions.

Transition from Models to Modeling (As Zumbo, 2007a, stated:)

- No matter how much data you have, it is never enough because without complete information you will always have some error of measurement or fallible indicator variable.
- We get around data and information limitations by **augmenting our data with assumptions**.
 - In practice, we are, in essence, using the statistical model to creating new data to replace the inadequate data.

$$\text{DATA} = \text{MODEL} + \text{RESIDUAL}$$

↑
A mathematician may ask what this symbol means in this setting?

On Modeling (#1: the role of the model)

In the process of empirical modeling one, in essence, begins with an array of numbers denoting responses to items or tasks for each respondents,

- it could be argued that the psychometric model “provides” the inferences one can make by being the vehicle for going from **what we have** to **what we wish we had** –
- that is, we have item or task responses but, as examples, we wish we had the process of item responding or, the score on the latent variable being measured by the test.

On Modeling (#2: more than just symbols)

- We wish to emphasize IRT modeling practice as psychological theorizing about item responses wherein the item parameters are more than just symbolic letters; that is,
 - item parameters carry psychological information about the interaction between the test taker and the characteristics of the item or task and hence **provide a window into response processes** as a **source of validity evidence**.

On Modeling (#2: more than just symbols)

- The item parameters are essential to this process as they form the kernels of an IRT characterization of item responding.
 - We therefore need to know what these item parameters represent in a psychological sense.
 - You will see that **studies of response shift** (e.g., by Prof. Sprangers and her colleagues, and others) **take this psychological theorizing very seriously** when they **consider the matter of response shift as a type of lack of measurement invariance**.

On Modeling (#3: the role of desire and love, and magical thinking, in model choice)

- At this point a word of caution about modeling practices echoed by Zumbo (2007a, 2017) is important.
 - In the narrative poem Metamorphoses in Greek mythology, Pygmalion is a legendary figure of Cyprus, a king and a sculptor. He made a sculpture that he found so perfect that he fell in love with it, which came to life through his love.
 - Many of us are like the mythical character Pygmalion and fall in love with our models; in good part we fall in love with what we want our model be.
 - We are very much like Pygmalion in that we believe that our particular model of interest becomes real – like him, through our love we make it real.

On Modeling (#4: ritualistic cultural behavior, in model choice)

- The danger in this type of magical thinking, however, is that in our cases these (psychometric and measurement) models are used in to produce scores used in decision making or intervention planning.
 - There is no such thing as a zero-stakes use of an assessment, survey, or instrument.
- This type of behavior inspired by love for our model, and magical thinking, results in psychometric model use as a kind of ritualistic cultural behavior.
 - This **continues unabated** because these model choices and practices appear (at least to some measurement practitioners) to be **objective and exact**, they are easily and readily available in statistical software packages, students are taught to use them, and journal reviewers and editors demand them, and reinforced in journals or associations filled with those who think like us.

On Modeling (#5: Eros and psychometrics)

- Lest I been seen as cold-hearted and lacking (mathematical) desires and impulses:
 - I do not wish to be seen as correcting the grammar in love letters.
 - To be clear, in these remarks I am not thinking of any one measurement practice (e.g., Rasch or RMT) even if that practice corresponds to my description, but rather all model choice and practices.
- We are not the first to fall under Eros' spell.
 - I am of the vintage to remember all too well the time of the *LISRELites* in the desert of social and behavioral research casually making causal claims.

On Modeling- A Central Message

- One of my central messages is that not only are there a variety of models and modeling practices in scientific practice, but that **models are empirical commitments**.
 - Because models (which, in part, include the parameter estimation strategy) are empirical commitments, it is measurement specialists who need to take partial responsibility for the decisions that are being made with the models they provide to others.

On Modeling- A Central Message

- As Zumbo and Rupp (2004) state: *Everyone knows that a useful and essential tool such as an automobile, a chainsaw, or a statistical model can be very dangerous if put into the hands of people who do not have sufficient training and handling experience or lack the willingness to be responsible users. (p. 87)*
- And passing reference to being “hard-nosed” and “theoretical” and claiming the ground of (a naïve form of) “objectivity” does not persuade me to ignore the empirical commitment. Nor should it impress others.
 - A priori claims of the special status of your model do not obviate the matter of models as empirical commitments.

On Modeling- Model Uncertainty In Its Varied Real Forms

- We should start to **regularly consider model uncertainty**.
 - Gustafson and Clarke (2004) and others have approached the varied forms of uncertainty, including model uncertainty, by considered a partitioning of posterior variance to assess prior influence.
 - They decompose components of posterior variance at each level of a Bayesian model.
 - Somewhat in the spirit of an ANOVA decomposition.

On Modeling- Model Uncertainty In Its Varied Real Forms

- Say a (Bayesian) analysis is to be undertaken in the face of **uncertainty about** the correct parameter value within a parametric model, the correct model within a collection or space of models, and the correct space within a collection of spaces.
 - For instance, in the context of estimating an unknown function the different spaces might correspond to different types of basis functions,
 - the different models might correspond to different subsets of basis functions of a given type, and
 - the different parameter values might correspond to different coefficients for the subset of basis functions.

Let's describe 'measurement invariance' at the intersection

Models

What are they?
I wish to shine a light on
a kind of model-
inferences

Modeling

How do use modeling and
for what purpose?
What do we do when

Measurement Invariance

What is it? What is its role at the intersection
of 'models' and 'modeling'?

Measurement Invariance (DLD framework)

- In a series of invited addresses, papers, and book chapters over the last 20 years, Zumbo has developed the Draper-Lindley-De Finetti (DLD) framework

(see, for example, Kroc & Zumbo, 2020; Shear & Zumbo, 2013; Zimmerman & Zumbo, 2001; Zumbo, 2001, 2002, 2007a, 2013, 2016, 2021).

Measurement Invariance (DLD framework)

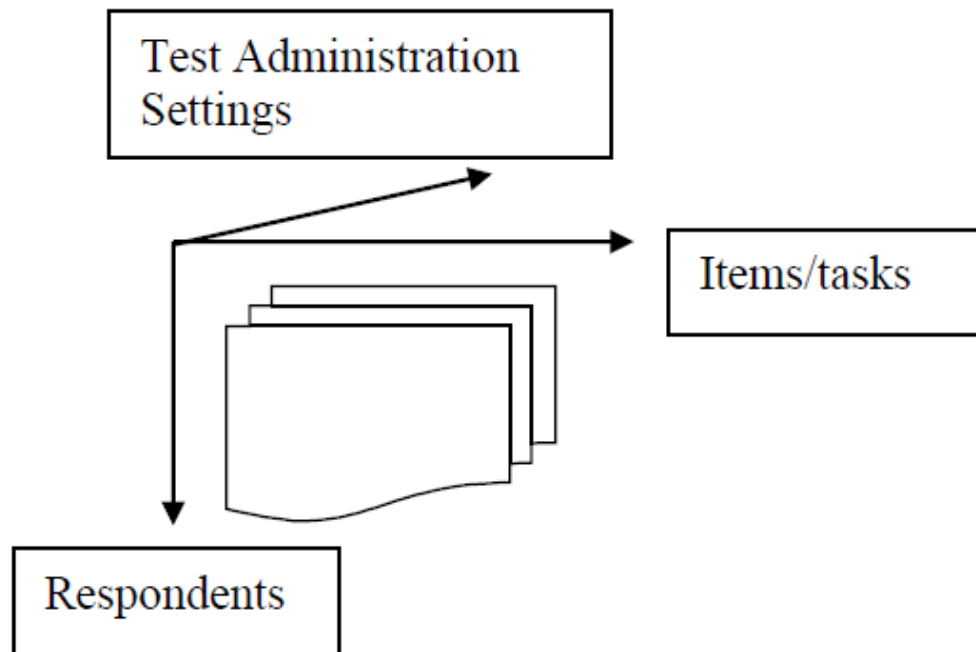
- Building on Draper's (1995), Lindley's (1972), and de Finetti's (1974-1975) Bayesian predictive approach to inference
 - Zumbo's DLD framework highlights the necessity to be explicit about the sorts of inferences one makes, and that one can make from a test design and implementation.

Measurement Invariance (DLD framework)

- As described in Zumbo (2021), at the heart of his DLD framework is de Finetti's notion of 'exchangeability' to describe a certain sense in which, for example, test scores are treated as random variables in a probability specification are thought to be similar.
 - The rigor and logic of the psychometric exchangeability framework (DLD framework) are grounded in test validity wherein measurement invariance (DIF or RS) set bounds on our inferences and claims.

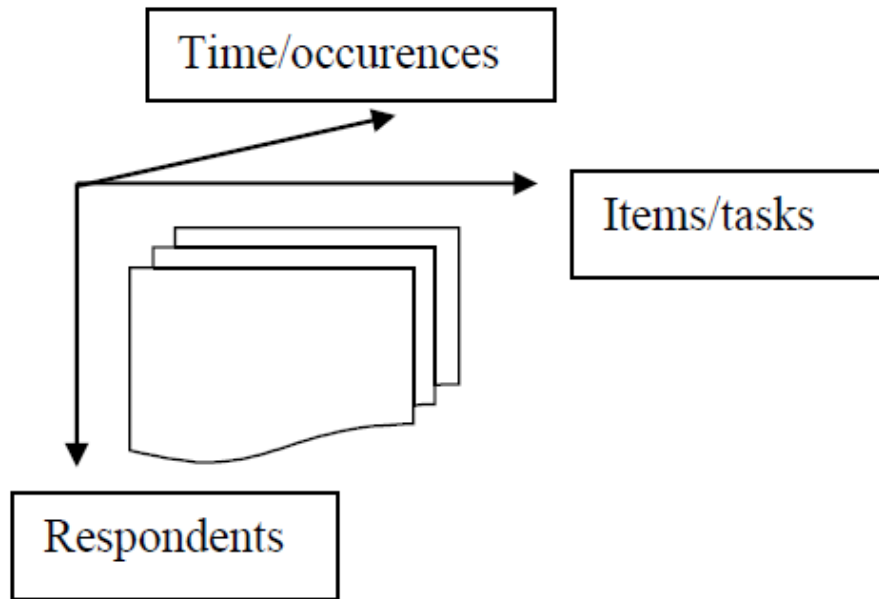
Measurement Invariance (DLD framework)

One can characterize this phenomenon from my Draper-Lindley-de Finetti framework invoking a kind of exchangeability.



Measurement Invariance (DLD framework)

One can characterize this phenomenon from my Draper-Lindley-de Finetti framework invoking a kind of exchangeability.



**LET'S TURN NOW TO THE
PSYCHOMETRIC VIEW OF
MEASUREMENT INVARIANCE**

Measurement Invariance in the Psychometric Literature

- The statistical problem of measurement can be characterized as involving two key tasks:
 - to find a set of indicators (items, scales, tasks, performances, or more generally referred to as measurement opportunities) that we believe that the interaction of probability spaces of respondents and items will imply,
 - to find a methodology for constructing a summary measure or scalar measure of the outcome of the interaction from these indicators.
- Most of the work on psychometric approaches has been characterized in terms of a **latent variable**, in the factor analytic sense, but that is **not necessary**.
 - One could just as easily speak of a scalar measure, more generally.

Measurement Invariance in the Psychometric Literature

- **Absence of measurement invariance**, 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 (DIF) or repeated testing (RS)** and concerns the measurement model relating **observed scores to scalar measure**— more commonly described as underlying latent variables.
- To simplify matters, 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 scalar measure.

Measurement Invariance in the Psychometric Literature

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

$$f(Y|\eta, s) = f(Y|\eta),$$

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,

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

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,

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

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).

**A SMALL DETOUR:
ON THE IMPACT OF DIF (AND
OTHER FORMS OF LACK OF
INVARIANCE) WHEN USING
LATENT VARIABLE OR
OBSERVED SUM SCORES:**

**ARE DIF AND RS RESTRICTED TO USES OF
LATENT VARIABLE SCORES?**

On the impact of DIF (and other forms of lack of invariance) when using observed sum scores

- I have heard it said many times that there is a lot of DIF research and little consequences in real life with everybody using sum scores.
- This claim is **not supported** by the psychometric research literature. See, for example:
 - Li, Zhen; Zumbo, Bruno D. (2009). Impact of Differential Item Functioning on Subsequent Statistical Conclusions Based on Observed Test Score Data *Psicológica*, 30(2), 343-370
 - Hidalgo, M.D. Benítez, I., Padilla, J.L., & Gómez-Benito, J. (2017). How Does Polytomous Item Bias Affect Total-group Survey Score Comparisons? *Sociological Methods & Research* 46(3), 586-603.
 - Rouquette, A., Hardouin, J.B., & Coste J. (2016). Differential Item Functioning (DIF) and subsequent bias in group comparisons using a composite measurement scale: A simulation study. *Journal of Applied Measurement*, 17, 312-334
 - Hidalgo, M.D., Galindo-Garre, F., Gómez-Benito, J. (2015). Differential item functioning and cut-off scores: Implications for test score interpretation. *Anuario de Psicología*, 45(1), 55-69.

On the impact of DIF (and other forms of lack of invariance) when using latent variable scores

- Within a latent variable framework, there are some interesting simulation studies that have tried to document the consequences of lack of invariance (such as DIF or drift or time) on predicted latent variable scores.
- Andre Rupp and I worked out the analytic solutions in a few cases and although I am pleased with the findings, they get knotty pretty quickly even in straightforward cases.
- **Model misspecification** is most important in this setting.
 - Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*, 63- 84.
 - Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49*, 264-276.

**RETURNING TO A PROMISING NUANCED
FRAMEWORK AND ANALYSES ARE
EMERGING**

Measurement Invariance: A kind of Multi-group model central to invariance testing

- In what follows, we review approaches to investigating invariance that can be broadly classified into ones that **treat the grouping variable as known**, and ones that **treat it as latent** and therefore attempt to model it.
- One can show these models with logistic regression, IRT or factor analysis model

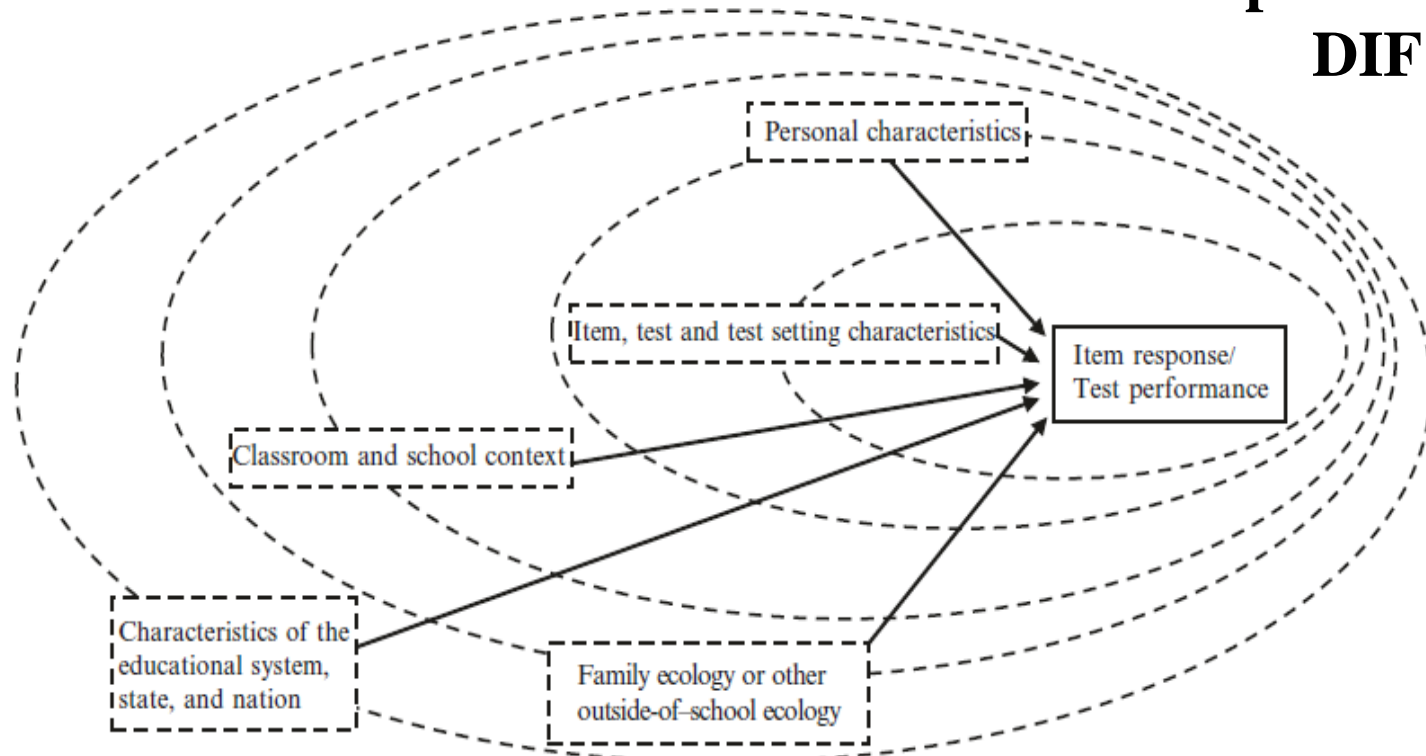
An alternative theory of Lack of MI (DIF or RS)

- Building on his 15- year program of research, Zumbo et al. (2015) introduced the **ecology of item responding** as an **alternative theory of the lack of measurement invariance** (DIF or RS) that informs an explanation-focused view of test validation (Zumbo, 2007a, Zumbo, 2007b), and hence an explanation-focused view of DIF or RS.

Figure 1. An Ecological Model for Item Responding

**Note: Five concentric ovals but could be more, or others.
[Bronfenbrenner]**

**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 or RS.**



Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- Conventional 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, 2007b).
- In Zumbo's (2007b) *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).

Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- In considering these this alternative theory, please keep in mind two routes to arriving at mixture models:
 - Factor Analysis → Mixture Models
 - Kernel Density Estimates → Mixture Models
- The methods described herein are applications and extensions of a family of statistical models, finite mixture models that have emerged out of developments in the statistical sciences over the last half-century (e.g., Clogg & Goodman, 1984; Dayton & Macready, 1988; Goodman, 1974; Lazerfeld & Henry, 1968; Rost, 1988).

Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- Unlike earlier algorithms for clustering respondents or variables, Finite Mixture Models postulate a formal statistical model for the population.
 - The statistical model assumes that the population consists of subpopulations or clusters. In each subpopulation (or cluster), the observed variables have different multivariate probability density functions resulting in a finite mixture density for the population.

Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- To provide you with mathematical intuition from the model, we describe the model without tending to the subtleties or conditions such as model identification. It is well-known that one can write the density function for a C-component (also characterized as a c-class) finite mixture as:

$$f(y|\mathbf{x};\theta_1,\theta_2,\dots,\theta_c;\pi_1,\pi_2,\dots,\pi_c) = \sum_{j=1}^c \pi_j f_j(y|\mathbf{x};\theta_j)$$

where $0 < \pi_j < 1$, and θ_j , in our case denotes the parameters of a (kernel) model that is being mixed with proportion π_j in class j , where j denotes classes 1 to c .

Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- The maximum likelihood estimates are attained by:

$$\max_{\pi, \theta} \ln L = \sum_{i=1}^N \left(\log \left(\sum_{j=1}^c \pi_j f_j(y | \theta_j) \right) \right), \text{ for } i \text{ represents the } N \text{ students.}$$

In practice, the m.l.e. is are often computed using an EM or Bayesian MCMC.

- Assigning respondents to a latent class: Having estimated the assumed mixture density parameters, one can now turn to assign each respondent with a cluster membership based on the maximum value of the posterior probability.

Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- Muthén as well as Asparouhov & Muthen (2008) describe many hybrid latent variable models, most of which have not yet seen widespread application in DIF or RS studies.
 - Muthen's parameterization of the **Grade of Membership Model** (GoM) is one that allows us to **re-consider fixed class membership**. [Individuals in a population **may belong to multiple subpopulations** (latent classes), not just a single latent class.]
 - Following Erosheva's (2002) landmark exposition of the GoM model, Muthen shows how one could view it as a special case of the two-level mixture model where the individual denoted j takes the role of a cluster j and the multivariate vector of all measurements Y_{ij} is treated as a univariate observations clustered in the individual j .

Challenging our Assumptions of Homogeneity and Focusing on Diversity Using Latent Class Models

- These particular forms of the grade of membership models may be **better suited than (fixed) latent class models** in, for example,
 - studies of fairness gender has, in the main, been characterized in the binary as biological sex wherein (binary) biological sex differences on item performance .
 - As 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.

Mixed membership models, such as GoM, challenge the idea that latent class membership is a fixed attribute described by assignment to only one latent class.

Some final thoughts ...

- The notions of **generalizations and inferences are intimately** tied to the **notion of invariance** in measurement models, and hence to **validity**.
- Simply put, measurement invariance allows us model-based generalization and inferences – noting, of course, that model-based inferences are inferences from assumptions in the place of data, as described above.

Some final thoughts ...

- Much has been said in the item response theory literature about invariance, and Rasch specialists make much hay of that model's invariance and specific objectivity properties with, often, implicit and explicit suggestions that one would have greater measurement validity with the Rasch model (Bond & Fox, 2001; Wright, 1997).

Some final thoughts ...

- This suggested supremacy of the Rasch model is an overstatement.
 - In fact, in several applications of the Rasch model one hears the claim that simply fitting the Rasch model gives one measurement item and person parameter invariance, without mention of any bounds to this invariance.
- There are many advantages to Rasch (or 1-parameter IRT) models in test applications but rarely are they the advantages that advocates of the Rasch model present.

END,
Next slide

Thank you

I am grateful for the continued support from:

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



Canada Research Chairs Program |
Programme des chaires de recherche du Canada



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada



Paragon

TESTING ENTERPRISES
— A PROMETRIC COMPANY

Bibliography for 28th Annual Conference of the International Society for Quality of Life Research (ISOQOL) invited plenary addressing” to the: ‘**On Models, Modeling, and Measurement Invariance in Validation Studies: A Stochastic View of Measurement**’, by Bruno D. Zumbo, University of British Columbia

(i) Draper-Lindley-de Finetti (DLD) Framework

Sawatzky, R., Ratner, P.A., Kopec, J.A., & Zumbo, B. D. (2012). Latent Variable Mixture Models: A Promising Approach for the Validation of Patient Reported Outcomes. *Quality of Life Research, 21*, 637-650. (in addition an online Supplementary Technical Appendix, pp. 1-9, can be found at the journal or click here). DOI: <http://dx.doi.org/10.1007/s11136-011-9976-6>.

Zumbo, B. D. (2021). **A Novel Multimethod Approach to Investigate Whether Tests Delivered at a Test Centre are Concordant with those Delivered Remotely Online**: An Investigation of the Concordance of the CAEL [Research Monograph]. DOI: <http://dx.doi.org/10.14288/1.0400581>

Zumbo, B.D. (2013). On Matters of Invariance in Latent Variable Models: Reflections on the Concept, and Its Relations in Classical and Item Response Theory. In Paolo Giudici, Salvatore Ingrassia, and Maurizio Vichi (Eds.), *Statistical Models for Data Analysis*, (pp. 399-408). New York: Springer.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

(ii) Overview of Measurement & Psychometric Models and Invariance / Model Uncertainty

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or Not to Bayes, From Whether to When: Applications of Bayesian Methodology to Modeling. *Structural Equation Modeling, 11*, 424-451.

Gustafson, P., & Clarke, B. (2004). Decomposing posterior variance. *Journal of Statistical Planning and Inference, 119*, 311 – 327. Doi: 10.1016/S0378-3758(02)00491-3

Kroc, E., & Zumbo, B.D. (2020). A Transdisciplinary View of Measurement Error Models and the Variations of $X=T+E$. *Journal of Mathematical Psychology, 98*, 1-9.

Kroc, E., & Zumbo, B.D. (2018). Calibration of measurements. *Journal of Modern Applied Statistical Methods, 17(2)*, 2-28.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement, 66*, 63-84.

Zimmerman, D. W., & Zumbo, B. D. (2001). The Geometry of Probability, Statistics, and Test Theory. *International Journal of Testing, 1*, 283-303.

Zumbo, B.D. (2017). On Models and Modeling in Measurement and Validation Studies. In B. D. Zumbo and A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 363-370). New York, NY: Springer.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible Modeling of Measurement Data For Appropriate Inferences: Important Advances in Reliability and Validity Theory. In David Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

(iii) Ecological Model of Item Responding / Response Processes / Validity Theory

Hublely, A.M., & Zumbo, B.D. (2017). Response Processes in the Context of Validity: Setting the Stage. In B. D. Zumbo and A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 1-12). New York, NY: Springer.

Sawatzky, R., Chan, E.K.H., Zumbo, B.D., Ahmed, S., Bartlett, S., Bingham, C., Gardner, W., Jutai, J., Kuspinar, A., Sajobi, T., & Lix, L.M. (2017). Modern perspectives of measurement validation emphasize justification of inferences based on patient reported outcome scores. *Journal of Clinical Epidemiology*, **89**, 154-159.

Zumbo, B.D., & Padilla, J.L. (2020). The Interplay between Survey Research and Psychometrics, with a Focus on Validity Theory. In P.C. Beatty, D., Collins, L., Kaye, J.L. Padilla, G. Willis, and A. Wilmot, (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 593-612). Hoboken, NJ: Wiley.

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.

(iv) Contemporary Theories of DIF

Liu, Y., Kim, C., Wu, A. D., Gustafson, P., Kroc, E., & Zumbo, B. D. (2019). Investigating the performance of propensity score approaches for differential item functioning analysis. *Journal of Modern Applied Statistical Methods*, **18(1)**, 1-26.

Sajobi, T.T., Brambhatt, R., Lix, L.M., Zumbo, B.D., & Sawatzky, R. (2018). Scoping Review of Response Shift Methods: Current Reporting Practices, and Recommendations. *Quality of Life Research*, **27**, 1133–1146.

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J. A., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research*, **27**, 1745–1755.

Sawatzky, R., Sajobi, T.T., Brahmbhatt, R., Chan, E.K.H., Lix, L.M., & Zumbo, B.D. (2017). Longitudinal Change in Response Processes: A Response Shift Perspective. In B. D. Zumbo and A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 251-276). New York, NY: Springer.

Sawatzky, R., Ratner, P.A., Kopec, J.A., & Zumbo, B. D. (2012). Latent Variable Mixture Models: A Promising Approach for the Validation of Patient Reported Outcomes. *Quality of Life Research*, **21**, 637-650. (in addition an online Supplementary Technical Appendix, pp. 1-9, can be found at the journal or click here). DOI: <http://dx.doi.org/10.1007/s11136-011-9976-6>.

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.

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.

(v) Impact of DIF on IRT Models and Invariance In IRT Models

Sawatzky R., Ratner P.A., Kopec J.A., Wu A.D., & Zumbo, B.D. (2016). The Accuracy of Computerized Adaptive Testing in Heterogeneous Populations: A Mixture Item-Response Theory Analysis. *PLoS ONE*, **11(3)**, 1-16.

Zumbo, B.D. (2013). On Matters of Invariance in Latent Variable Models: Reflections on the Concept, and Its Relations in Classical and Item Response Theory. In Paolo Giudici, Salvatore Ingrassia, and Maurizio Vichi (Eds.), *Statistical Models for Data Analysis*, (pp. 399-408). New York: Springer.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, **66**, 63- 84.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research*, **49**, 264-276.

Rupp, A. A., & Zumbo, B. D. (2004).. A Note on How to Quantify and Report Whether Invariance Holds for IRT Models: When Pearson Correlations Are Not Enough. *Educational and Psychological Measurement*, **64**, 588-599. {Errata, (2004) *Educational and Psychological Measurement*, **64**, 991}

(vi) On the impact of DIF (and other forms of lack of invariance) when using observed sum scores

Li, Zhen; Zumbo, Bruno D. (2009). Impact of Differential Item Functioning on Subsequent Statistical Conclusions Based on Observed Test Score Data *Psicológica*, **30(2)**, 343-370

Hidalgo, M.D. Benítez, I., Padilla, J.L., & Gómez-Benito, J. (2017). How Does Polytomous Item Bias Affect Total-group Survey Score Comparisons? *Sociological Methods & Research* **46(3)**, 586-603.

Rouquette, A., Hardouin, J.B., & Coste J. (2016). Differential Item Functioning (DIF) and subsequent bias in group comparisons using a composite measurement scale: A simulation study. *Journal of Applied Measurement*, **17**, 312-334

Hidalgo, M.D., Galindo-Garre, F., Gómez-Benito, J. (2015). Differential item functioning and cut-off scores: Implications for test score interpretation. *Anuario de Psicología*, **45(1)**, 55-69.