On Computational Psychometrics as a Validity Framework for Process Data

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

Zumbo, B.D. (2020, August). *On Computational Psychometrics as a Validity Framework for Process Data*. Presented at the Annual Meeting of the National Council for Measurement in Education (NCME), San Francisco, CA. [Online Virtual Conference].

Opening Remarks

- I wish to congratulate the session organizer, chair, and the presenters for the thought-provoking nature of their work.

Organizer: Alina von Davier, Duolingo

Chair: Ada Woo, TreeCrest Assessment Consulting

Presenters:

- Yuchi Huang, ACT
- Alina von Davier & Burr Settles, Duolingo
- John Whitmer, Chi2 Labs

- **I believe that work of this calibre that challenges our thinking and many of the orthodoxies in the field deserves serious consideration and reflections and not an on-the-fly "peer review" of typical discussions.**

As we saw today …

- Alina A. von Davier & Burr Settles - Duolingo
  - Emerging trends in assessment: AI powered capabilities. We saw that the power is in the integration of AI tools.
  - Alina von Davier’s notion of *Computational Psychometrics* is unique in the field because it is
    - **psychometrics** centered,
    - **theory based**: modern conceptualization of the construct
    - ***Data driven*** *Algorithm empowered : adaptive, computationally efficient*

As we saw today …

- John Whitmer, Chi2 Labs
  - *Learning Analytics: Making the Transition from Prediction to Action*
  - John's presentation and the related papers I read have been **helpful to me in understanding "learning analytics"**; a term that has been confusing to me.
  - The highlight for me, was the consistency of findings despite high variability in LMS usage.
    - The rhetorical turn to "action" in this field is vital
    - My own preference are explanatory aspirations.

As we saw today …

- Yuchi Huang, ACT
  - We were introduced to Sphinx, a human-AI hybrid system for scalable production of reading comprehension passages in English from writers' samples/prompts to be used in in a variety of learning and assessment.
    - Sphinx is a ground-breaking natural language generation system designed to create reading passages in a computationally efficient manner and can be used in a plethora of learning and assessment contexts.

Where Do We Go from Here?

I will turn to Reflections #0 to #4 to summarize the most salient observations that came to mind for me on the theme of this symposium.

Note: This is not a survey of all the research literature. The reference list at the end of this document will, therefore, for the most part reflect my program of research.

Reflections on the Theme of this Session

- **#0 Test taker response data is needed**
- ***test-free test scores*** (as recently seen, for example, in the UK A-levels scandal) do not fit into computational psychometrics and, more generally, are inappropriate in testing practices.
    - How does one define measurement, misclassification, or predictive error with no test taker data in the model?
        - Group based models used for individual prediction is a problem.
    - *Even a "kludged-up" prediction error, for a test taker, will be large with no test taker data. Duh!*
    - *Predicting test outcomes from available data and information (and not test taker responses) is off the table.*
  - – *In all the papers in this symposium include a type of measurement model and hence measurement error or misclassification as a core feature of their frameworks.*

Reflections on the Theme of this Session

- **Reflection #0 [corollary: as I stated in 2007; we never have the data we want …**
- Chris Anderson, the former editor of *Wired* magazine, famously wrote (2008) that "… companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. *Indeed, they don't have to settle for models at all.*"
  - He went on to say, "We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."
- **My position is that this model-free approach has shown itself to be wrong-headed and bad advice for testing and assessment practices.**

Reflections on the Theme of this Session

- **#1 Orthodox psychometric theory and modeling has long been charged with neglecting reality**
  - *The Times They Are A-Changin*
    - Bob Dylan recorded this song in October 1963 and it quickly became a call to action, *The Times They Are A-Changin'*.
    - It summed up the anti-establishment feelings of the of the time.

Reflections on the Theme of this Session

#1 Orthodox psychometric theory and modeling has long been charged with neglecting reality

- **There are two reasons for keeping this complaint at the forefront of our minds.**
  - First, the criticism is frequently made, frequently acknowledged and just as frequently ignored.
  - Second, and more pertinently, the last 20 years have seen many developments in data-driven algorithms, virtual learning, and complex assessment environments.

Reflections on the Theme of this Session

- **#2 In modeling test and assessment data the addition of an apparently weaker a priori structure to the actual data often produces an apparently stronger a posteriori structure**.
  - We construct different models for different purposes, with different formalisms and equations to describe them (Zumbo, 2017a).
    - Which is the right model, which the ‘true’ set of equations? The question is a mistake.

#2 …the addition of an apparently weaker a priori structure to the actual data often produces an apparently stronger a posteriori structure.

- …the great appeal of measurement models is that in practice 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.
  - The function of the psychometric model in measurement and validity research 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 through validation.

(quotation from: *On Models and Modeling in Measurement and Validation Studies*, Zumbo, 2017a)

#2 …the addition of an apparently weaker a priori structure to the actual data often produces an apparently stronger a posteriori structure.

- …the great appeal of measurement models is that in practice
  - We get around data and information limitations by augmenting our data with assumptions. In practice, we are, in essence, using the statistical model to create new data to replace the inadequate data.
  - For example, the most common data augmentation assumption in psychometrics is that the dependencies (e.g., correlations) among items are accounted for by an unobserved continuum of variation – of prominence in item response theory and factor analysis models

(quotation from: *On Models and Modeling in Measurement and Validation Studies*, Zumbo, 2017a)

Summary to this point …

- I am highly skeptical of both
  - test taker data-free results, and
  - model-free results
- Models of various sorts play important roles in computational psychometrics as well as the the algorithm based, machine-learning, and data driven (data science) approaches.

Reflections on the Theme of this Session

#3 We need to know what we mean by "validity", if not how will we know if we have achieved it. *[I believe of value to the presenters]*

- Over the past 60 years (1960 – 2020) concepts of validity have grown increasingly expansive, and methods of validation have become increasingly complex and multi-faceted.
  - See Shear & Zumbo (2014) and Zumbo & Padilla (2020) for brief historical overviews that aim at improving validation practices.

Reflections on the Theme of this Session

#3 We need to know what we mean by "validity", if not how will we know if we have achieved it. ***[This is important for anyone doing computational psychometrics]***

- In 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 on the measurement model.

(*On Models and Modeling in Measurement and Validation Studies*, Zumbo, 2017a)

Reflections on the Theme of this Session

#3 … what we mean by “validity”, if not how will we know if we have achieved it.

- … if one wants to advance the theorizing and practice of measurement, one needs to articulate what they mean by “validity” to go hand-in-hand with the process of validation (Zumbo 2007). As has been noted several times in the validity theory literature (e.g. Messick 1989; Shear and Zumbo 2014; Zumbo 1998, 2007, 2009), when explicit definitions of validity are not provided, the discipline has tended to conflate validity theory and validation methods. It is therefore important to distinguish them to avoid overly focusing on methods and techniques for data analysis in the absence of a conceptual foundation. (Zumbo & Padilla, 2020)

Reflections on the Theme of this Session

#3 … what we mean by “validity”, if not how will we know if we have achieved it.

- Where were started as a discipline rarely helps …
  - *Cronbach and Meehl’s (1955) description of construct validity is not easily distinguished as either a definition of validity or a process of validation. Cronbach and Meehl clearly articulated, for example, how one might go about gathering evidence during the process of validation. But they also emphasized that, “Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator” (Cronbach & Meehl, 1955, p. 282).*
  - *Despite this call for a holistic framework of scientific inquiry, validity remained a fragmented concept, and the type of validity one demonstrated was most often a product of the method used to document validity (Hubley & Zumbo, 1996).*

(quoted form Shear & Zumbo, 2014)

Reflections on the Theme of this Session

#3 … what we mean by "validity", if not how will we know if we have achieved it.

- Zumbo & Chan (2014) showed that rarely do validation studies state or describe a framework or theory of validity to guide their studies. And, citing Cronbach and Meehl (1955) does not provide sufficient guidance.
- As Shear and Zumbo (2014) state:
  - The absence of guiding theories of validity is more troubling than the absence of any one particular concept of validity. In the absence of a clear guiding theory of validity, it is difficult to evaluate whether a particular program of validity research has accomplished its aims. This absence complicates comparisons from findings across different validity studies because they may not be trying to accomplish the same goal. **[today I add that this also impedes replication and accumulation of evidence]**
  - It also undermines the statement in the *Standards* that validity is "the most fundamental consideration in developing and evaluating tests" (AERA et al., 1999, p. 9) because it may not be clear what exactly a concern for validity entails.
  - The argument-based approach to validation provides one framework for structuring validation research, ***but still seems to require a theory of validity*** that can serve as a guiding aim.
    - Further developments on both of these fronts seem more important than advocating that a particular concept of validity be adopted.

Reflections on the Theme of this Session

#3 … what we mean by "validity", if not how will we know if we have achieved it.

- At this point it is more important to a have a guiding validity theory than which one you fancy.
  - It is also important not to confuse a definition of validity with the techniques and methods used to obtain such evidence.
    - My own leanings are a combination of an ecological model of item and test responding and the ***explanation-focused view of validity*** bridges the inferential gap from the test data to response processes and provides inferential strength to the conclusions based on the empirical data modeling (see, e.g., Stone & Zumbo, 2016; Zumbo, 2007, 2009, 2017b).

Reflections on the Theme of this Session

#3 … what we mean by "validity", if not how will we know if we have achieved it.

- I emphasize that the aim of validation practices is:
  - identifying the determinants (or explanatory theory) of task / item / test score variation … the explanation is the basis of any strong validity claims

    My colleagues and I take an ecological systems approach that give us inferential strength.

Reflections on the Theme of this Session

#3 … what we mean by “validity”, if not how will we know if we have achieved it.

- *In a series of essays, Zumbo describes his view of “validity” as the explanation of the variation in survey or questionnaire response data, and “validation” as the process of developing and testing the explanation (Stone and Zumbo 2016; Zumbo 2007, 2009, 2015, 2017b; Zumbo and Hubley 2016, 2017; Zumbo et al. 2015, 2017).*

Reflections on the Theme of this Session

#3 … what we mean by “validity”, if not how will we know if we have achieved it.

- *In Zumbo’s (2009) explanation-focused view of validity, the aim is explanation, within a pragmatic philosophical tradition and not conventional causal views of validity. In the tradition of philosophy of science, causation is only one possible view of explanation (Zumbo, 2007, 2009). This view of test validation is reflective of Messick’s (1989, 1995) sense of substantive validity, which focuses on evidence about the process of responding (i.e., how and why people respond), as central to validation*. (quotation from Zumbo, 2017b)

Reflections on the Theme of this Session

- It should be noted that by ‘inferential strength’ I mean the amount of support that the evidence or reasons provide the conclusion about **response processes** (and hence validity);
  - and is therefore considered a matter of degree such that the more support (the more evidence or reasons) there is for a conclusion, the stronger the argument for the conclusion.

(*On Models and Modeling in Measurement and Validation Studies*, Zumbo, 2017a)

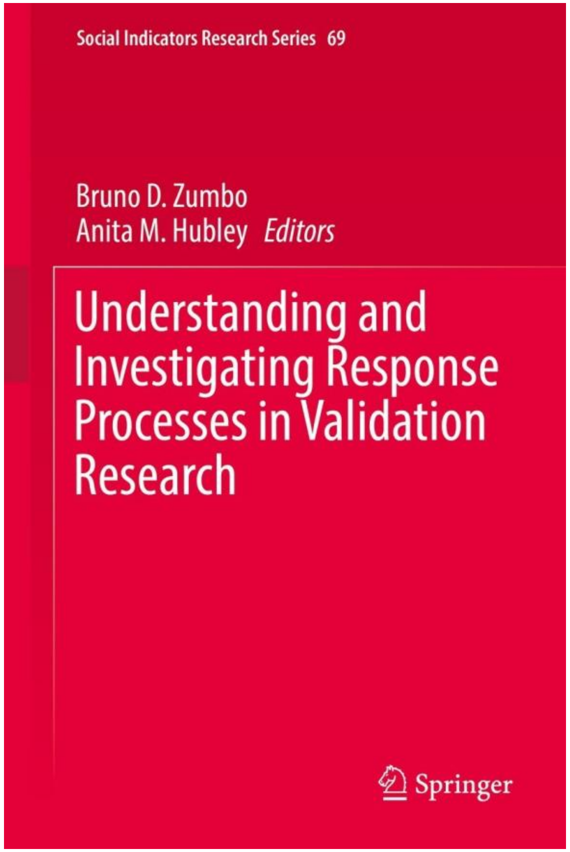
Focus on response processes … central in our view of validity [from Maddox & Zumbo, 2016]

- **Not An Entirely New Idea**:
- (Cronbach, 1949, in his text on testing) "One of the most valuable ways to understand any test is to administer it individually, requiring the subject to work the problem aloud […] the tester learns just what mental processes are used in solving the exercises, and what mental and personality factors cause errors." (p. 54) Recently reminded of this quotation by Paul Newton.
- Messick (1995, in *American Psychologist*) described construct validity as comprising "the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables" (p. 743).
  - He noted that, historically, most attention has been placed on evidence involving essentially internal structure, convergent and discriminant coefficients, and test-criterion relationships, but that evidence of expected performance over time, across settings or groups, and as a result of experimental manipulation would be more illuminating. (documented widely in Zumbo & Chan, 2014)

Response Processes

‘.. one may think broadly of response processes as the mechanisms that underlie what people do, think, or feel, when interacting with, and responding to, the item or task and are responsible for generating observed test score variation’. (Zumbo & Hubley, 2017, p. 2).

In a recent book by Zumbo & Hubley, (2017) published by *Springer Press.*



Social Indicators Research Series 69

Bruno D. Zumbo
Anita M. Hubley *Editors*

Understanding and Investigating Response Processes in Validation Research

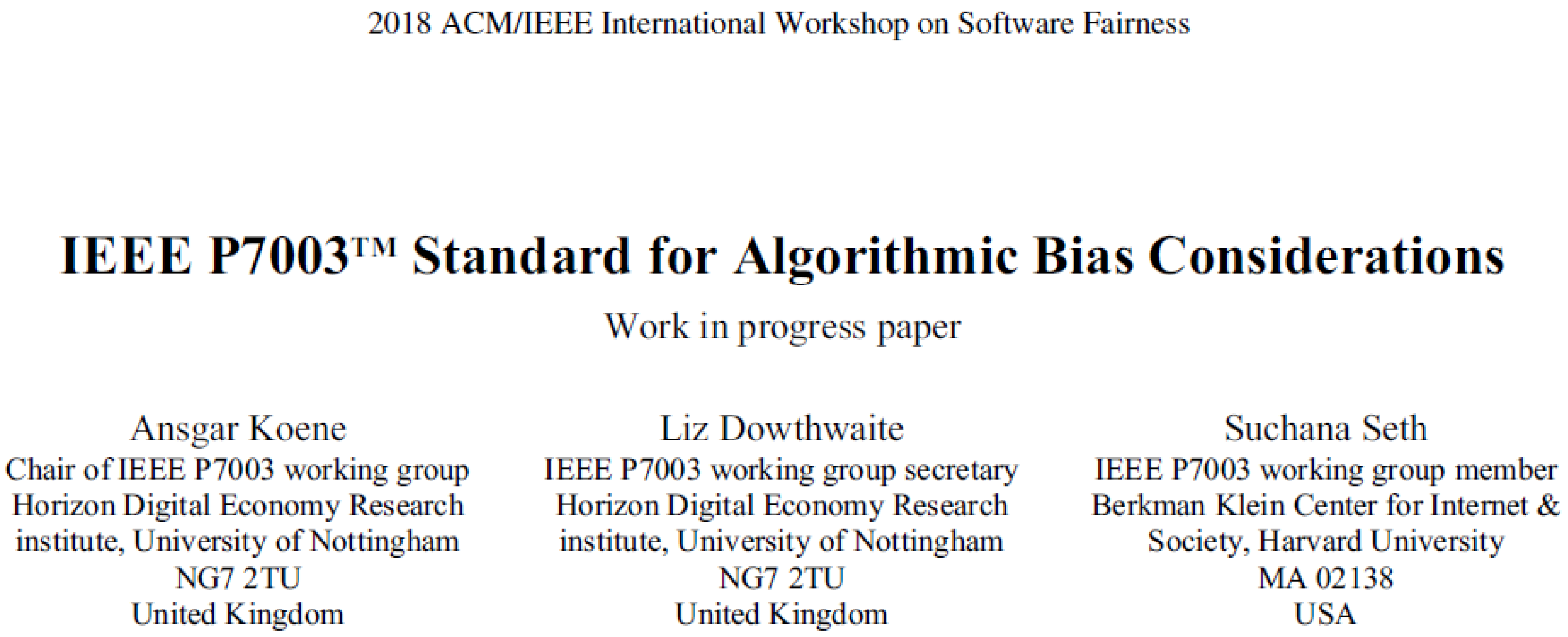
Springer

Reflections on the Theme of this Session

#4 Consequences of Assessment are important and, I have argued, central to validity (e.g., Zumbo, 2007, 2015; Zumbo & Hubley, 2016)

- Artificial intelligence, machine learning and algorithmic bias is a topical issue in social media and popular press due to key events.
- Rather than simply guarding against these harms **passively**, these algorithmic and machine-learning systems should be used **proactively** to advance equity in assessment and testing.
- But their industry advances that need to be considered.

Some good starts have been made in the software industry



2018 ACM/IEEE International Workshop on Software Fairness

IEEE P7003™ Standard for Algorithmic Bias Considerations

Work in progress paper

Ansgar Koene
Chair of IEEE P7003 working group
Horizon Digital Economy Research
institute, University of Nottingham
NG7 2TU
United Kingdom

Liz Dowthwaite
IEEE P7003 working group secretary
Horizon Digital Economy Research
institute, University of Nottingham
NG7 2TU
United Kingdom

Suchana Seth
IEEE P7003 working group member
Berkman Klein Center for Internet &
Society, Harvard University
MA 02138
USA

Closing Summary

- I am highly skeptical of both
  - test taker data-free results, and
  - model-free results
- Models of various sorts play important roles in computational psychometrics as well as the the algorithm based, machine-learning, and data driven (data science) approaches.
- A guiding framework (theory) of validity is needed to inform validation practices.
  - Explanation-focused validity theory is well-suited for computational psychometrics and related data-based algorithm-driven testing.
- Consequences of assessment, and particularly assessment that is algorithm driven, data-based, and computational psychometrics needs to be considered *proactively*.

References

(please contact me if you would like a reprint of any of my papers or chapters in my edited books)

- Chen, M.Y., & Zumbo, B.D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B. D. Zumbo and A.M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 53-68). New York, NY: Springer.
- Hubley, 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.
- Shear, B.R., & Zumbo, B.D. (2014). What Counts as Evidence: A Review of Validity Studies in Educational and Psychological Measurement. In Bruno D. Zumbo, and Eric K.H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 91-111). New York: Springer.
- Stone, J., & Zumbo, B.D. (2016). Validity as a Pragmatist Project: A Global Concern with Local Application. In Vahid Aryadoust, and Janna Fox (Eds.), *Trends in Language Assessment Research and Practice* (pp. 555-573). Newcastle: Cambridge Scholars Publishing. http://brunozumbo.com/wp-content/uploads/2020/08/Validity-as-a-Pragmatist-Project-Stone_Zumbo-2016.pdf
- Zimmerman, D. W., & Zumbo, B. D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing, 1*, 283–303.   for a reprint: http://brunozumbo.com/wp-content/uploads/2017/12/Zimmerman_Zumbo_2001.pdf
- 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. http://faculty.educ.ubc.ca/zumbo/papers/Zumbo_Validity_Chapter_Reprint.pdf
- Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte, NC: IAP – Information Age Publishing, Inc..

References

(please contact me if you would like a reprint of any of my papers or chapters in my edited books)

- Zumbo, B.D. (2014). What Role Does, and Should, the Test Standards Play Outside of the United States of America? *Educational Measurement: Issues and Practice, 33*, 31-33.
- Zumbo, B.D. (2015, November). *Consequences, side effects and the ecology of testing: Keys to considering assessment 'in vivo'*. Keynote address, annual meeting of the Association for Educational Assessment – Europe (AEA-Europe), Glasgow, Scotland. http://brunozumbo.com/aea-europe2015
- Zumbo, B.D. (2017a). 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. (2017b). Trending away from routine procedures, towards an ecologically informed 'in vivo' view of validation practices. Measurement: Interdisciplinary Research and Perspectives 15 (3–4): 137–139.
- Zumbo, B.D., & Chan, E.K.H, (Eds.) (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*. New York: Springer. Click here for information about this edited book.
- Zumbo, B.D. and 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.
- Zumbo, B. D., & Hubley, A.M. (Eds.). (2017). *Understanding and Investigating Response Processes in Validation Research*. New York, NY: Springer. For information about the book at Springer Press' website please click http://www.springer.com/us/book/9783319561288.
- 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.

Thank you

- **Please contact me for a copy of these slides.**
- **bruno.zumbo@ubc.ca**

For a full list of publications, please see http://faculty.educ.ubc.ca/zumbo/cv.htm