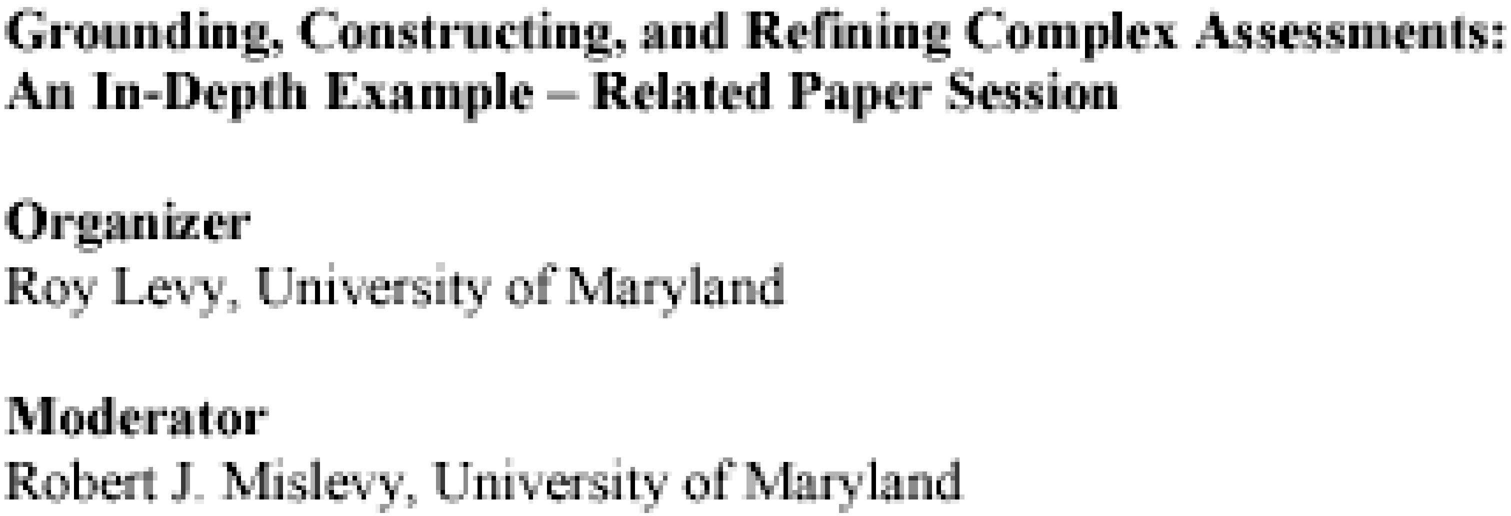
Grounding, Constructing, and Refining Complex Assessments

Discussant's Remarks
2003 NCME meeting Chicago

Bruno D. Zumbo
Professor
University of British Columbia



Grounding, Constructing, and Refining Complex Assessments: An In-Depth Example – Related Paper Session

Organizer
Roy Levy, University of Maryland

Moderator
Robert J. Mislevy, University of Maryland

- I strongly recommend that you get a copy of these papers because many of us will learn a good deal from them.
- The statement that one will learn from these papers is the highest complement one can give a conference paper or any scholarly piece of work.
- The papers are well done and thorough.
- I am not going to make paper-by-paper remarks because, frankly, I have little to add and my queries would be small subtle points in long and involved papers.
- Instead, I am going to try and step back and see if we can take a larger picture of the message(s) in these papers.

- I come to this as a statistician and applied mathematician so, like a drunk making it back to his hotel room from a Chicago bar, I will lean on those lamp-posts to make my way through this material. {Warning: periodically the room may start spinning ….}
- As an applied mathematician, the discipline teaches me, among other things, to use mathematical reasoning to:
  - Make explicit the assumptions of the process being studied
  - Use a mathematical framework and language to formalize a model or embedded models of this process
  - Decompose the "model" into its constituent parts and investigate how I am going to estimate the parameters (or form) of these models

- Throughout this process I ask myself these questions:
  - Has the mathematics helped me understand the process I am modeling?
  - Do the mathematics impose any “unnatural” assumptions or processes on what I am trying to model?
  - Do the results hold up to empirical scrutiny?
- In short, it is about model building and hypothesis checking.
- Of course, good modeling is deeply immersed in the substance of the modeling (or analytic) problem.
- Viewed from this lens, the ECD work I see in this symposium is quite good.

- A few points I would emphasize:
  - Certainly, a question will arise as to whether people reason in a strictly Bayesian fashion? This is an inevitable question you will face when you introduce ECD. Response: We know people reason less formally, but when they reason well this reasoning appears to share key features with I would now call a soft-Bayesian modeling strategy.
  - Perhaps the key in this “Soft-Bayesian modeling” being model checking …. and that needs to be highlighted in the Levy and Mislevy paper.

- – **Leaning on my own line of research, there is an implicit invariance assumption that will be, naturally, imposed with, I believe, any assessment or measurement models. In short, one can ask whether the various models involved are appropriate for all students. This is a big question but worthy of your attention. This applies to the ECD papers and the SNLP papers.**
- – **I encourage you all to get copies of the papers.**

- I like this line of research. It appears to be in the tradition of the very promising line of work done by Embretson, Tatsuoka, Gerhard Fischer, and the exciting line of work going on in computer science and mathematical psychology on knowledge representation and knowledge structures.
- Each new application of the sort of research elicits how and why this strategy works. I encourage you to continue doing this with all sorts of examples.
- I have struggling for some time to capture essence of the quality of this line of research and how it is different than the 100+ years of measurement before it.

Abstract representation of the measurement process

- **I am trying to get a sense how the ECD process and the papers in this symposium construct the assessment / measurement process.**
- **My sense is that there is an essential difference between "test theory" or "measurement / psychometric theory" and ECD. What is that essential difference?**
- **Let me start from an abstraction to first principles. Perhaps it is overly complex but bear with me because it *may* draw out something interesting.**

Probability Spaces, Tests , and Observed Scores

- We begin with a probability space (Ω,A,P), a random variable $X: \Omega \to R$ with finite variance, representing test scores, and a random point $f: \Omega \to \Phi$, where Φ is a set of individuals or experimental objects. The sample space Ω comprises all possible outcomes of a test procedure. The random variable X will be called the **observed score,** as is customary in test theory.

- A random variable defined on a probability space is a familiar object in statistical theory. However, distinctive properties of test scores arise from the interrelationship of the random variable X and the random point f. The function f can be regarded as an assignment of individuals or objects to sample points, which presupposes that one has selected a σ-algebra of subsets of Φ and that inverse images of sets in the collection belong to the σ-algebra A of subsets of Ω.

- Usually, the set Φ is finite or countably infinite, representing a discrete population of individuals, so that the set of all subsets of Φ is a σ-algebra. If Φ is the set of real numbers, the Borel sets are taken as the σ-algebra.

- As in many probability models, a sample space often is implicit, and the primary objects of interest are random variables and probability distributions. These ideas are summarized by the following definition.

Definition 1.

A **test** is a 5-tuple (Ω,A,P,f,X), consisting of a set of outcomes Ω, a σ-algebra A of subsets of Ω, representing observable events, a set function P defined on A, and two point functions f and X defined on Ω, such that (Ω,A,P) is a probability space, f is a random point, and X is a random variable with finite variance.

- At this point in the process the ‘tradition’ at least in some areas of practice (Embretson and other excluded) is to make some passing remark about error of measurement of X …. and pass the batton to the folks who are interested in the “validity of the inferences”.
- So, to build on what Roy and Bob wrote in their paper we (a) build good tasks, (b) pass the items over the wall to the psychometricians and measurement folks, and then (c) toss those scores over the next wall to the folks who want to interpret and do something with the outcomes to worry about validity of the inferences.
- The line research we saw today appears to try and do it differently in some integrative fashion.

- **Definition 2 (ECD notion of Assessment).**

Educational Assessment is a 4-tuple (S,T,E,A), where S denotes the "student model", T the task model, E the evidence model, and A some sort of model that assembles and presents aspects of S and T.

- ***E*** appears to subsume Definition 1 of a test. That is, ***E*** is larger in scope than Defin. 1.
- Likewise, cognitive theorists appear to give priority to the properties of the task model, T, whereas Mislevy and his colleagues appear to give priority to S, the student model. I could be wrong but from my limited reading this is how it appears to me.
- Behrens and his colleagues' papers are focusing on aspects of *E* and *A.* Or at least it appears so.