Special Focus Series:

Statistical Methods in Diagnostic Assessments

Columbia University
April 27-28, 2012

SCHEDULE

Friday, April 27

Morning Session

9:30	-	10:00	Susan Embretson, Georgia Institute of Technology
10:05	-	10:35	David Thissen, The University of North Carolina at Chapel Hill
11:00	-	11:30	Klaas Sijtsma, Tilburg University
11:35	-	12:05	Matthias Von Davier, ETS
			Lunch

Afternoon Session

2:00	-	2:30	Jimmy de la Torre, Rutgers University
2:35	-	3:05	Tao Xin, Beijing Normal University
3:30	-	4:00	Jeff Douglas, University of Illinois at Urbana-Champaign
4:05	-	4:35	Andre Rupp, University of Maryland
4:40	-	5:20	Sandip Sinharay, ETS

Evening

6:30	-	10:00	Dinner
			Howard Wainer, NBME & Wharton School of UPenn

Saturday, April 28

Morning Session

9:30	-	10:00	Wilm J. van der Lindern, CTB/McGraw-Hill
10:05	-	10:35	Bob Henson, University of Michigan
11:00	-	11:35	Hua-Hua Change, University of Illinois at Urbana-Champaign
11:35	-	12:05	Jingchen Liu, Columbia University

Abstracts

Jimmy de la Torre

The G-DINA Model Framework Approach to Fit Evaluation

This presentation covers several aspects of model-data fit evaluation in cognitive diagnosis modeling from a general framework perspective. The first section briefly introduces the G-DINA model as a general cognitive diagnosis model (CDM), and a framework for conducting additional analyses. The second section discusses various aspects of model-data fit that needs to be evaluated in fitting CDMs. Model-data fit evaluation in cognitive diagnosis modeling can be classified as either absolute or relative, and test-level or item-level, and includes examining the appropriateness of specific CDMs and of attribute specifications as indicated in the Q-matrix. The third section presents some findings regarding test-level absolute and relative fit evaluations, and item-level CDM comparison and Q-matrix validation. The final section discusses the implications of these recent developments on the practicability of cognitive diagnosis modeling.

Jeff Douglas

Granularity of Latent Variable Models for Cognitive Diagnosis

Cognitive diagnosis models can be thought of as alternatives to unidimensional and multidimensional item response models when diagnostic information is desired. We revisit the standard assumption of local independence and the interpretation that results for the latent variables of interest. The implications of modeling a vector of binary latent variables, versus a vector of continuous latent variables are discussed, together with corresponding assumptions of a conjunctive, disjunctive, or compensatory response process. A model that combines continuous with binary latent variables is introduced, along with an algorithm for estimation. Simulation results under a variety of correctly specified and misspecified models are given, and a comparison of classifications under different approaches is studied with the Tatsuoka fraction subtraction data. A summary discussion questions how an appropriate model can be identified, as well as when a model may not be needed at all. Joint work with Chun Wang, Grace Hong, and Hua Hua Chang.

André A. Rupp

On the Alignment of Tools from Modern Psychometrics, Educational Data Mining, and Multivariate Statistics for Evidence-based Reasoning in Digital Learning Environments

Digital learning environments create an opportunity to gather a much larger and, potentially, richer set of information about the way students reason when they engage with tasks of different complexities in a particular domain such as network engineering or urban planning. The resulting process and product data that such systems provide hold great promise for the delivery of more actionable, personalize, and timely feedback to support learning. However, in order to cull so-called diagnostic measurement information at different levels of grain-size from these types of data, a coherent approach to evidence-based reasoning is needed. From a design-based research perspective, that work requires the utilization of a coherent framework for evidence-based reasoning such as evidence-centered design. From a statistical perspective, that work requires the integration of multivariate statistics, approaches from modern psychometrics, and tools from the educational data mining community. It furthermore requires the alignment of qualitative and quantitative pieces of evidence within a coherent program of validation to demonstrate that diagnostic score reports and associated interventions are defensible, meaningful, and actionable. This work cannot take place without a constant exchange between subject-matter experts, curriculum developers, teachers, and measurement specialists. Using his experiences from work within such contexts, Dr. Rupp will illustrate the empirical and conceptual challenges to achieve diagnostic measurement objectives, especially in light of real-life practical decision-making constraints.

Klaas Sijtsma

Psychological Measurement Between Physics and Statistics

This contribution discusses the physical perspective on psychological measurement represented by additive conjoint measurement and the statistical perspective represented by item response theory, and argues that both fail to adequately address the real measurement problem in psychology: This is the absence of well-developed theories on psychological attributes. I argue that the two perspectives leave psychology out of the equation and by doing that come up with proposals for psychological measurement that are fruitless. Only the rigorous development of attribute theories can lead to meaningful measurement. I provide several examples of the measurement of well-developed attributes and suggest future directions for psychological measurement.

Sandip Sinharay

A Critical Evaluation of Subscore Reporting: Temptations, Pitfalls, and Some Recommendations

There is an increasing interest in subscores because of their simplicity and their potential diagnostic value. This presentation provides a critical evaluation of reporting of subscores in the context of educational tests. The presentation starts by raising questions about the quality of the currently reported subscores. A review is provided of a recent classical-test-theory-based method that examines whether subscores provide any added value over the total score (Haberman, 2008). A discussion is provided of the extension of the method to subscores based on multidimensional item response theory models. Analysis of data from several operational data sets shows that subscores have to be based on a sufficient number of items and have to be sufficiently distinct from each other to be worth reporting and that several operationally reported subscores actually do not have any added value over the total score. The presentation concludes with some recommendations on reporting of subscores.

David Thissen

From Subtest Scores to Subscores: Multidimensional Item Response Theory and Diagnostic Assessment

Across a variety of assessment contexts, the questions How many scores should be reported? and How should those scores be computed? arise in many different guises. Using long-standing statistical test theory for summed scores, answers to those questions turn around the reliability of the scores, and the degree of added precision that can be obtained with each potential summed score. Alternatively, in the past two decades various statistical models have been developed to provide probabilistic assignment of test respondents to latent classes; in applications these classes have represented the presence or absence of specific skills hypothesized to underlie performance on academic achievement test items. A third approach using multidimensional item response theory (MIRT) models provides a latent-variable alternative to summed-score theory, and a continuous-measurement alternative to latent-class diagnostic models. In the past decade, computational advances have made application of MIRT models a practical endeavor. This presentation positions MIRT within the context of diagnostic assessment, and provides illustrations of its practicality and usefulness; it also raises yet-unanswered questions about the dissemination of multidimensional scores.

Wim J. van der Linden

Is Multidimensional IRT Ready for Use in Diagnostic Assessment?

Although the reporting of full vectors of test scores rather than single scores is quite attractive from a diagnostic point of view, I wonder if we already know enough about multidimensional IRT to warrant successful application. For instance, it is now clear that for multidimensional models to have success probabilities monotonically increasing in the abilities, the abilities have to be compensatory (van der Linden, 2012). But it is hard to view educational achievements are the result of compensatory abilities. Another issue, bound to frustrate scale maintenance in routine application of multidimensional IRT, is lack of identifiability of the ability parameters. More conceptual issues also play a role; for instance, it is common in the educational testing community to confound the lack of constancy of ability during testing with violation of the unidimensionality assumption. I will reflect on several of these issues and propose a hierarchical, multiple unidimensional approach asat least a temporarymore feasible alternative to diagnostic assessment.

Matthias von Davier

Stochastic Approximation for Estimation of Diagnostic Models

The talk will present explorations an estimation method for high dimensional models with discrete latent variables. The data used are from the PIRLS 2006 grade four reading assessment. The PIRLS framework specifies a number of content domains and a number of cognitive processes targeted by the assessment. More specifically, each item in the assessment is classified with respect to exactly one domain and one process. The presentation provides a comparison of an experimental estimation approach, stochastic approximation for discrete multidimensional latent structure models, to customary estimation methods such as the EM algorithm. The basis for comparisons are model that are characterized by high dimensionality, for example multi-trait-multi-method IRT model with domains and subscales as structuring entities, as well as bi-factor and higher-order diagnostic models with a general dimension and subscales and/or cognitive domains as subdimensions.