The Second Workshop on Statistical Methods in Cognitive Assessments

Department of Statistics
Columbia University

May 16-18, 2013

SCHEDULE

Thursday, May 16			Room 903 School of Social Work Building
Afternoon Session			Lihua Yao, Defense Manpower Data Center
2:00	-	3:35	Short Course
3:40	-	3:55	Coffee Break
4:00	-	5:35	Short Course

Friday, May 17			Room 501 Northwest Corner Building
Morning Session			
9:10	-	9:55	William Stout, University of Illinois at Urbana-Champaign
10:00	-	10:15	Coffee Break
10:20	-	11:05	Frank Rijmen, Educational Testing Service
11:10	-	11:55	Shelby Haberman, Educational Testing Service
			Lunch
Afternoon Session			
2:00	-	2:45	Brian Junker, Carnegie Mellon University
2:50	-	3:35	Curtis Tatsuoka, University Hospitals
3:40	-	3:55	Coffee Break
4:00	-	4:45	Sophia Rabe-Hesketh, University of California, Berkeley
			Dinner

Saturday, May 18			Room 501 Northwest Corner Building
Morning Session			
9:10	-	9:55	Chun Wang, University of Minnesota
10:00	-	10:15	Coffee Break
10:20	-	11:05	Bor-Chen Kuo, National Taichung University of Education
11:10	-	11:55	Louis Roussos, Measured Progress

Abstracts

Thursday, May-16, Afternoon Session

[2:00 - 5:35]

Lihua Yao

Multidimensional Item Response Theory: Applications and BMIRT, LinkMIRT, and SimuMIRT Software

BMIRT (Yao, 2003) is a computer program that uses the Markov Chain Monte Carlo (MCMC) method to estimate item and ability parameters in the multidimensional multi-group IRT framework; exploratory and confirmatory approaches are supported. LinkMIRT (Yao, 2004) is a linking software that links two sets of item parameters onto the same scale in the MIRT frame work. SimuMIRT is software that simulates data for various MIRT models.Data requirements and formats, and sample data and input files will be provided to participants prior to the workshop; participants are recommended to go to www.BMIRT.com to download software into the laptop computers that they are required to bring to the workshop. Java run time environment (JRE), which can be downloaded on the internet, is required to be installed into the computer, and the computer path need to be set up following the directions in "1.5.1 Java Run Time Environment" in the document BMIRT2013.pdf.

Friday, May-17, Morning Session

[9:10 - 9:55]

William Stout

A Family of Generalized Diagnostic Classification Models for Multiple-Choice Option-Based Scoring (GDCM-MC), with Applications to RUM & DINA Modeling

A new family of restricted latent class diagnostic classification models (DCMs) for multiple choice item based assessments is presented, denoted by GDCM-MC. Its purpose is to allow researchers and practitioners to more fully capture the useful and fine-grained diagnostic information in a multiple choice item based diagnostic classification assessment, especially when there is useful diagnostic information in which incorrect options students choose. The primary targeted area of application is formative assessment, although summative assessment applications can also occur. In this regard, formative assessments are beginning to be designed where specific misconceptions (which students and teachers would like to correct) make choosing certain distractor options probable. We consider two important specific instances of the broad GDCM-MC family. These are an extended version of the Reparameterized Unified Model (sometimes called the Fusion Model), denoted ERUM-MC,

and an extended version of the DINA model, denoted EDINA-MC. The GDCM-MC family addresses four DCM modeling challenges: (1) modeling option-based scoring, (2) modeling guessing, (3) modeling student misconceptions, and (4) having the capacity to choose the right parametric complexity that balances modeling bias with estimation variability. Simulation studies demonstrate the capability of our MCMC-based procedure to effectively estimate the ERUM-MC model and classify students. Real data applications and further simulations are in progress.

[10:20 - 11:05]

Frank Rijmen

A graphical model framework for higher-order item response theory models

Second-order item response theory models have been used for assessments consisting of several domains, such as content areas, in order to derive both overall and domain abilities from the same model (de la Torre and Song, 2009). We propose a third-order item response theory model for assessments that include subdomains nested in domains, such as topic areas nested within content areas. Using a graphical model framework, it is shown how second and third-order models do not suffer from the curse of multidimensionality. Exploiting the conditional independence relations implied by the higher-order structure, maximum likelihood estimation can be carried out efficiently. We apply unidimensional, second-order and third-order item response models to the 2007 Trends in International Mathematics and Science Study (TIMSS) to accommodate cognitive domains and content domains, and topic areas nested in content domains. Our findings suggest that deviations from unidimensionality are more pronounced at the content domain level than at the cognitive domain level, and that deviations from unidimensionality at the content domain level all but disappear after taking into account topic areas.

[11:10 - 11:55]

Shelby Haberman

A General Program for Item-response Analysis that Employs the Stabilized Newton-Raphson Algorithm

A general program for item-response analysis is described which uses the stabilized Newton-Raphson algorithm. This program is written to be compliant with Fortran 2003 standards and is sufficiently general to handle independent variables, multidimensional ability parameters, and matrix sampling. The latent variables may be either polytomous or multivariate normal. Items may be dichotomous or polytomous.

LUNCH

Friday, May-17, Afternoon Session

[2:00 - 2:45]

Brian Junker

Educational surveys, plausible values, and Goldilocks

Statistical analyses of large scale educational surveys such as NAEP, TIMSS, etc., incorporate multiple imputations, or "plausible values", to make inferences related to cognitive status of surveyed populations. Plausible values (Mislevy, 1991) are random draws from the posterior distribution of a latent cognitive proficiency variable, given a set of demographic and other variables in a "conditioning model". What variables should the survey agency include in the conditioning model? When the latent proficiency variable is the dependent variable in an analysis, the conditioning model is well understood, and standard methodology exists for building the conditioning model. When the latent proficiency variable is an independent variable, however, it is not clear what should be in the conditioning model. In this somewhat informal talk, I will review these ideas and propose an answer for the independent variable case, which suggests that there is not a generic "one size fits all" set of plausible values that survey institutions can publish for all secondary analysts.

[2:50 - 3:35]

Curtis Tatsuoka

Cognitive modeling of neuropsychological functioning, with an application to Alzheimers disease

Cognitive modeling approaches can be useful in medical research, such as for understanding how neuropsychological functioning is affected by neurological disorders. Some issues that arise in these applications are the limited number of measures that are collected, confounding of cognitive profiles in classification, and the complexity of the response data. Bayesian nonparametric density estimation approaches involving Dirichlet process priors will be discussed in the context of latent partially ordered cognitive models. An application to identifying cognitive markers for Alzheimers disease also will be reviewed.

[4:00 - 4:45]

Sophia Rabe-Hesketh

Autoregressive IRT growth model for longitudinal item analysis

A first-order autoregressive item response theory (IRT) growth model is proposed for longitudinal binary item analysis where responses to the same items are conditionally dependent across time given the latent trait. The proposed model is equivalent to a local dependence IRT model that includes interaction parameters for responses at adjacent time points. The initial conditions problem is handled using the method suggested by Heckmann (1981) as implemented by Aitkin and Alfo (2003). The implication of this treatment is investigated with respect to measurement invariance. Asymptotic and finite sample power of the test for the autoregressive parameter are investigated. When the data are generated from a first-order autoregressive IRT growth model, the consequences of ignoring local dependence and the initial conditions problem are also examined. An empirical study is provided using longitudinal data on Korean students' self esteem.

DINNER

Saturday, May-18, Morning Session

[9:10 - 9:55]

Chun Wang

The Promise and Challenge of Computerized Adaptive Testing with Diagnostic Features

Over the past thirty years, obtaining diagnostic information from examinees item responses has become an increasingly important feature of educational and psychological testing. The objective can be achieved by sequentially selecting multidimensional items to fit the class of latent traits being assessed, and therefore Multidimensional Computerized Adaptive Testing (MCAT) or cognitive diagnostic CAT (CD-CAT) are probable solutions to such task. In this talk, I will introduce four promising item selection methods: *D-optimality method based on Fisher information*, *KL information index*, *Shannon entropy method*, and *mutual information method * and show the applications of these methods with a number of psychometric models, including the multidimensional compensatory item response model, the non-compensatory model, the DINA model, and the higher-order IRT model. The potential challenges of large scale implementations of CAT, such as test security and item bank usage, will be discussed and possible solutions will be given.

[10:20 - 11:05]

Bor-Chen Kuo

A Computerized Cognitive Diagnostic Test with Mathematical Multiple Choice and Constructed Response Items

In this talk, a web-based cognitive diagnostic test with mathematical multiple choice and constructed response items will be presented. Constructed response items are developed to evaluate complex concepts or skills (ex. problem solving), which in turn facilitate learners higher-level cognitive abilities. However, human scoring of constructed response items is not only time-consuming by also economically inefficient. To ameliorate current scoring procedure, , an automated scoring mechanism for mathematical constructed response items diagnosing the status of concept/skill and error/bug during problem solving process will be introduced. A test with mixed-type items (multiple choice and constructed response items) is more suitable for practical conditions. To analyze the responses of multiple choice items and the output of automated scoring mechanism of constructed response items, DINA and DINO are modified and applied to model examinees concepts and bugs respectively. The results of simulated and real data experiments showed that the modified DINA and DINO models are applied to mixed-type tests well. The mixed-type tests with automated scoring process of constructed response items show better attribute and pattern correct classification rates than multiple choice items.

[11:10 - 11:55]

Louis Roussos

Multidimensionality: Where is thy sting?

This focus of this talk is the development and evaluation of a new effect size measure for the amount of multidimensionality in a set of test data. The motivation behind the statistic, the description of the statistic, and the development of its population parameter will be presented. Specifically, the statistic is a nonparametric estimate of the standard error of measurement for a raw score on a test that takes into account violations of local independence.A small simulation study is presented. The results indicate that evenlarge amounts of multidimensionality seem to have little effect on the standard error of measurement.
