Conference on Statistical Methods for Innovative Testing and Learning

Date:

July 7th - July 8th, 2018

Venue:

Room 903, School of Social Work (SSW)
Columbia University, 1255 Amsterdam Avenue
New York, NY, 10027

Organizing Committee:

Yunxiao Chen, Emory University
Guanhua Fang, Columbia University
Qiwei He, Educational Testing Service
Jingchen Liu, Columbia University
Zhiliang Ying, Columbia University

Sponsor:

Department of Statistics, Columbia University

Schedule

Time (July 7th)	Speaker	Title
8:00 - 9:00		Onsite Registration & Breakfast (SSW 903)
9:00 - 9:10		Opening Remarks
9:10 - 9:40	Cees Glas	Item shells, item cloning and rule-based item generation
9:45 - 10:15	Daniel Bolt	A psychometric Model for Discrete-Option Multiple-Choice (DOMC) items
10:20 - 10:40		Coffee Break
10:40 - 11:10	Miyako Ikeda	Programme for International Student Assessment (PISA)
11:15 - 11:45	Kentaro Yamamoto	Multi Stage Adaptive Testing for PIAAC and PISA
11:50 - 2:00		Lunch
2:00 - 2:30	Jimmy de la Torre	A Note on the Equivalence of the Unidimensional Item Response Theory and Cognitive Diagnosis Models
2:35 - 3:05	Gongjun Xu	Identifiability of Restricted Latent Class Models
3:10 - 3:40	Ying Cheng	Detection of Inattentiveness in Questionnaire or Survey Data
3:45 - 4:10		Coffee Break
4:10 - 4:40	Xiaoou Li	Joint Maximum Likelihood Estimation for High-dimensional Exploratory Item Response Analysis
4:45 - 5:15	Chun Wang	Variable-Length Stopping Rules for Multidimensional Computerized Adaptive Testing

Time (July 8th)	Speaker	Title
8:00 - 8: 30		Breakfast
8:30 - 9:00	Michelle LaMar	Modeling Sequences of Actions using Markov Decision Processes
9:05 - 9:35	Jiangang Hao	Addressing the data challenge for innovative testing and learning
9:40 - 10:10	Qiwei He	Using Process Data to Explore Consistent Patterns across Problem-Solving Items in PIAAC
10:15 - 10:40		Coffee Break
10:40 - 11:10	Georgios Fellouris	Efficient mastery-based learning using sequential change-detection
11:15 - 11:45	James Corter	Alternative Strategies and Strategy Choice in Mathematics Problem-Solving

Program

Title: Item shells, item cloning and rule-based item generation

Speaker: Cees Glas, University of Twente

Abstract: Item cloning is the procedure where an number items are derived from blueprint using a number of transformation rules. Generation via an item shell is an analogous procedure, but here the items are generated on the fly and every item is unique. Rule-based item generation takes this a step further by formalizing the generation process more strictly: An item family is defined by a set of so-called radicals which are defined based on cognitive analysis of the item domain. The results from the analysis are then used to devise rules for the generation of new items An example of a radical is whether or not Bayes rule has to be applied to solve a statistics item. Radicals can be used to automate item generation. In addition, the radicals can be assumed to be important determinants of item difficulty. This also motivates the psychometric model used for calibration and test administration in this approach. The IRT model used consists of two levels. At the higher level explanatory variables account for the radicals and the item difficulties. The lower level accounts for the variability in the item parameters of the manifest items accounting for the effects of item cloning or generating items via an item shell. Both frequentist and Bayesian models for calibration and item administration will be discussed and an example pertaining to the generation of statistics items will be presented.

Title: A Psychometric Model for Discrete-Option Multiple-Choice (DOMC) items

Speaker: Daniel Bolt, University of Wisconsin - Madison

Abstract: Discrete option multiple choice (DOMC) items differ from traditional multiple-choice (MC) items in the sequential administration of response options (up to display of the correct option). DOMC can be appealing in computer-based test administrations due to its protection of item security and its potential to reduce testwiseness effects. We propose a psychometric model for DOMC items that attends to the random positioning of key location across different administrations of the same item, a feature that has been shown to affect DOMC item difficulty (Eckerly, Smith & Sowles, 2017). Using two empirical datasets having items administered in both DOMC and MC formats, we consider the variability in key location effects across both items and persons. The proposed model exploits the capacity of the DOMC format to isolate both (a) distinct sources of item difficulty (i.e., related to both the identification of keyed responses and the ruling out of distractor options) and (b) distinct person proficiencies related to the same two components. Practical implications in terms of the randomized process applied to schedule item key location in DOMC test administrations are considered.

Title: Programme for International Student Assessment (PISA)

Speaker: Miyako Ikeda, The Organisation for Economic Co-operation and Development

Abstract: This year PISA conducts its seventh global data collection and assessment. Since the first assessment in 2000, PISA has focused on the extent to which 15-year-old students have acquired key knowledge and skills that are considered essential for full participation in society. PISA has informed national education policies and practices in many countries by establishing benchmarks, providing internationally comparable indicators and identifying performance trends over time. While PISAs fundamental goal has remained unchanged for the last twenty years, the project has evolved to keep its results relevant to changing social needs and policy orientations. This presentation initially describes the key messages emerging from PISAs results and findings and provides some examples on using the PISA results for policy making. It continues by highlighting shifts in survey design and in the focus of analyses that have taken place over time. The presentation closes with a brief outline of possible new directions and other innovative aspects for the future.

Title: Multi Stage Adaptive Testing for PIAAC and PISA

Speaker: Kentaro Yamamoto, Educational Testing Service

Abstract: The talk describes the development of MST(Multi Stage Adaptive Testing) for individual testing in 2002 for PDQ (Individualized Assessment of Prose, Document and Quantitative Literacy) based on the previous population survey (IALS) and more recent evolutionary implementations for the population surveys of PIAAC (Programme for the International Assessment of Adult Competencies) in 2012 and PISA (Programme for International Students Assessment).

Title: A Note on the Equivalence of the Unidimensional Item Response Theory and Cognitive Diagnosis Models

Speaker: Jimmy de la Torre, The University of Hong Kong

Abstract: At present, most existing educational assessments are developed and analyzed using unidimensional item response theory (IRT) models. To obtain information that can be used for diagnostic purposes, the same assessments have also been retrofitted with cognitive diagnosis models (CDMs). However, it remains unclear the extent to which two disparate psychometric frameworks can be simultaneously used to analyze the same assessment. To address this issue, we propose a framework for relating the two classes of psychometric models, as well as boundaries as to when this can be done. Specifically, we impose certain conditions on the higher-order generalized deterministic inputs, noisy and gate (HO-GDINA) model, and reformulate its success probability as a function of the higher-order ability. It can be shown that with appropriate constraints, the HO-GDINA model reduces to the four unidimensional IRT models when only a single attribute is required. When two or more attributes are required, the item response function of the HO-GDINA model can be well approximated by unidimensional IRT models. We investigate a number of factors (e.g., slope and intercept of the higher-order structure, guessing and slip of the item parameters, sample size) to determine their impact on the quality of item parameter approximation, as well as ability estimation. Preliminary results show that a wider range of intercept values leads to a lower bias in the IRT parameter estimates and a slightly higher correlation between the true and estimated abilities. Additionally, more discriminating items and larger sample size lead to a higher correlation between the true and estimated abilities.

Title: Identifiability of Restricted Latent Class Models

Speaker: Gongjun Xu, University of Michigan

Abstract: Latent class models have wide applications in social and biological sciences. In many applications, pre-specified restrictions are often imposed on the parameter space of the latent class models, through a design matrix, to reflect practitioners' diagnostic assumptions about how the observed responses depend on the respondents' latent attributes. Such restricted latent class models, though widely used in cognitive diagnosis assessment, suffer from nonidentifiability due to the models' discrete nature and complex restricted structure. This talk considers the identifiability issues of the restricted latent class models and addresses several open questions in the literature by developing a general framework for the identifiability of the model parameters. The theoretical results are applied to establish for the first time the identifiability of several examples from cognitive diagnosis applications.

Title: Detection of Inattentiveness in Questionnaire or Survey Data

Speaker: Ying Cheng, University of Notre Dame

Abstract: Careless or inattentive responding is frequently observed in questionnaire or survey data, which jeopardizes test validity and more broadly the replicability and generalizability of research findings. It is therefore very important to detect such response behavior. The most frequently encountered type of careless response behavior is back random responding (BRR). Literature suggests that BRR is challenging to detect, with reported power of detection around .5 or lower. Change point analysis (CPA), which is a widely used statistical process control method, can be applied to item response or item response time data or both to detect if aberrant behavior exists in a response pattern. In this talk I will first review existing CPA methods that have been applied in psychometrics to detect intra-individual change, and then introduce two new methods that can be applied to item response or response time data to detect BRR. Simulation results indicated that the proposed new methods are able to detect BRR behavior with much higher power (.80 or above) than existing methods, while keeping the Type-I error rate well under control.

Title: Joint Maximum Likelihood Estimation for High-dimensional Exploratory Item Response Analysis

Speaker: Xiaoou Li, Univeristy of Minnesota

Abstract: Multidimensional item response theory is widely used in education and psychology for measuring multiple latent traits. However, exploratory analysis of large-scale item response data with many items, respondents, and latent traits is still a challenge. We consider a high-dimensional setting that both the number of items and the number of respondents grow to infinity. A constrained joint maximum likelihood estimator is proposed for estimating both item and person parameters, which yields good theoretical properties and computational advantage. Specifically, we derive error bounds for parameter estimation and develop an efficient algorithm that can scale to very large datasets. The proposed method is applied to large scale personality assessment data sets. Simulation studies are conducted to evaluate the proposed method. This is a joint work with Yunxiao Chen and Siliang Zhang.

Title: Variable-Length Stopping Rules for Multidimensional Computerized Adaptive Testing

Speaker: Chun Wang, Univeristy of Minnesota

Abstract: In computerized adaptive testing (CAT), a variable-length stopping rule refers to ending item administration after a pre-specified measurement precision standard has been satisfied. The goal is to provide equal measurement precision for all examinees regardless of their true latent trait level. Several stopping rules have been proposed in unidimensional CAT, such as the minimum information rule or the maximum standard error rule. These rules have also been extended to multivariate CAT and cognitive diagnostic CAT, and they all share the same idea of monitoring measurement error. Recently, Babcock and Weiss (2012) proposed an absolute change in q (CT) rule, which is useful when an item bank is exhaustive of good items for one or more ranges of the trait continuum. Choi, Grady and Dodd (2010) also argued that a CAT should stop when the standard error does not change, implying that the item bank is likely exhausted. Although these stopping rules have been evaluated and compared in different simulation studies, the relationships among the various rules remain unclear, and therefore there lacks a clear guideline regarding when to use which rule. This talk presents analytic results to show the connections among various stopping rules within both unidimensional and multidimensional CAT. In particular, it is argued that the CT rule alone can be unstable and it can end the test prematurely. However, the CT rule can be a useful secondary rule to monitor the point of diminished returns. To further provide empirical evidence, two simulation studies are reported using the multidimensional graded response model (MGRM) with a real item bank from the Activity Measure for Post-Acute Care in the patient-reported outcomes domain.

Title: Modeling Sequences of Actions using Markov Decision Processes

Speaker: Michelle LaMar, Educational Testing Service

Abstract: Frequently process data contain evidence of complex decision making that is central to the focal measurement construct. Rather than distill these data into abstract features, this research aims to model the decision making directly, enabling inferences about the decision maker and the decision making process. Using Markov decision processes (MDP) as generative models, the likelihood of specific sequences of action can be calculated. These models can then be used for both parametric and latent-class inferences. The generalized utility of such a framework will be discussed, along with its limitations, and a few examples.

Title: Addressing the data challenge for innovative testing and learning

Speaker: Jiangang Hao, Educational Testing Service

Abstract: Digitally Based Assessments (DBAs), such as game- and simulation-based assessments, generate a large amount of data that contains evidence of a test takers proficiency along the targeted constructs. The rich response-process data provide unique opportunities for validating and improving the precision of the scores based on the response products as well as furnishing evidence for new constructs. However, extracting the right evidence from the complex process data and modeling them to support the measurement of the constructs of interest can be very challenging. In this talk, I introduce the glassPy data analytics framework developed at ETS for addressing the process-data challenges and show some examples based on our innovative assessment tasks.

Title: Using Process Data to Explore Consistent Patterns across Problem-Solving Items in PIAAC

Speaker: Qiwei He, Educational Testing Service

Abstract: Computer-based testing provides possibility in capturing process data such as action sequences and response times accompanying with response data. Such information is particularly valuable when examining interactive problem-solving tasks that have been increasingly used in large- scale assessments, such as PISA and PIAAC. Although exploring sequence patterns for a specific item is a good starting point, investigation of generalized patterns across multiple tasks bears the promise of identifying factors that are associated with test takers problem-solving behaviors across a variety of contexts and frames.
This presentation illustrates an approach to generalize action patterns across multiple tasks by deriving the degree of similarity between individuals action sequences and predefined optimal sequences with a cluster of problem-solving items in PIAAC. Consistent behavioral patterns were found for a large number of respondents across countries. The results suggest a promising avenue for further exploring the relationship between problem-solving strategies, item difficulty and proficiency estimates in large-scale assessments.

Title: Efficient mastery-based learning using sequential change-detection

Speaker: Georgios Fellouris, University of Illinois Urbana-Champaign

Abstract: A fundamental goal in mastery-based instruction is to minimize the time for skill acquisition and quickly understand when mastery has taken place. We formulate this problem mathematically as a generalization of the Bayesian sequential change-detection problem, where the change (time of mastery) is a latent event that should not only be detected, but also accelerated. Specifically, we assume that observations are collected sequentially as responses to instructional choices made in real time. These choices not only determine the distribution of student responses, but also influence the time of learning. The problem we consider is the minimization of the expected time to learning, while controlling the probability of a false detection. We propose a simple, intuitive, low-complexity solution, which achieves the optimal performance up to a first-order asymptotic approximation under a large class of learning models. The efficiency of the proposed approach will also be illustrated with simulation studies. This is joint work with Yanglei Song (University of Illinois, Urbana-Champaign).

Title: Alternative Strategies and Strategy Choice in Mathematics Problem-Solving

Speaker: James E. Corter, Teachers College, Columbia University

Abstract: The existence and use of multiple strategies for problem solving, particularly in mathematics, poses continuing challenges for diagnostic assessment (recent progress on the issue notwithstanding), raising validity and identifiabiity issues. Thus, the potential for multiple strategy use should be kept in mind when designing diagnostic assessments, otherwise test designers and psychometricians may find themselves "retrofitting" even tests designed as diagnostic assessments. Some examples of multiple strategy use in mathematics problem solving are given, concluding with an exploration of multiple strategy use in Tatsuoka's classic mixed-fraction subtraction data using an extended NIDA model (joint work with Yun Jin Rho and Matthew Johnson).