Fifth Conference on the Statistical Methods in Psychometrics

Date:
Nov 17 - Nov 18, 2017

Venue:
C03 (Nov 17), 903 (Nov 18)
School of Social Work (SSW), Columbia University
1255 Amsterdam Ave, New York, NY 10027

Organizing Committee:
Yunxiao Chen, Emory University
Xiaoou Li, University of Minnesota
Brian Ling, Columbia University
Jingchen Liu, Columbia University
Zhiliang Ying, Columbia University

Sponsor:
Department of Statistics, Columbia University

Schedule

Time (17-Nov)	Speaker	Title
8:00 - 8:50		Onsite Registration & Breakfast (SSW C03)
8:50 - 9:00		Opening Remarks
9:00 - 9:35	Irini Moustaki	Pairwise likelihood estimation for confirmatory factor analysis models with ordinal variables and data that are missing at random
9:35 - 10:10	Edward Ip	Statistical methods for quantifying, testing, and modeling local dependency
10:10 - 10:30		Coffee Break
10:30 - 11:05	Hong Jiao	A Joint Multigroup Testlet Model for Responses and Response Times Accounting for Differential Item and Time Functioning
11:05 - 11:40	Shiyu Wang	Using response time to assess learning progress: A hidden Markov Model for Response and Response Time
11:40 - 1:30		Lunch
1:30 - 2:05	Han van der Maas	Relations between psychometrics for complex systems research
2:05 - 2:40	Daniel Bolt	Parameter Invariance and Skill Attribute Continuity in Diagnostic Classification Models: Bifactor MIRT as an Appealing Alternative
2:40 - 3:20		Coffee Break
3:20 - 3:55	Matthias von Davier	Discontinuation Rules in Ability Testing: New Results on Ignorability, Local Dependency, and Bias
3:55 - 4:30	Jeff Douglas	Directions for Learning Research in Cognitive Diagnosis

Time (18-Nov)	Speaker	Title
8:00 - 9:00		Breakfast (SSW 10th floor)
9:00 - 9:35	Gunter Maris	Looking at Deep Neural Networks through IRT glasses
9:35 - 10:10	Jean-Paul Fox	Real Time Performance Monitoring in Serious Gaming
10:10 - 10:30		Coffee Break
10:30 - 11:05	Steven Culpepper	Estimating the Cognitive Diagnosis Q Matrix with Expert Knowledge
11:05 - 11:40	Jimmy de la Torre	Do **I** Complete **Q**?

Program

Title: Pairwise likelihood estimation for confirmatory factor analysis models with ordinal variables and data that are missing at random

Speaker: Myrsini Katsikatsou and Irini Moustaki*, The London School of Economics and Political Science

Abstract: Methods for the treatment of item non-response in attitudinal scales and in large-scale assessments under the pairwise likelihood (PL) estimation framework are proposed. In confirmatory factor analysis (CFA) with categorical observed variables and data being missing at random, multiple imputation followed by the three-stage weighted least squares is recommended to obtain unbiased estimates. The approach requires a practical imputation model along with the model for the observed data. Alternatively, we propose two, easy to implement, strategies for incorporating missing values into the CFA under the PL framework, the complete-pairs (CP) and the available-cases (AC) pairwise likelihood approaches. Both require only a model for the observed data and standard errors are easy to compute. Doubly-robust versions of the PL estimation are also studied, but they are computationally more demanding and do not show a better performance than the CP and the AC in our simulation study. The proposed methods, made available in the R package lavaan, are employed to analyze the UK data on numeracy and literacy collected as part of the OECD Survey of Adult Skills.

Title: Statistical methods for quantifying, testing, and modeling local dependency

Speaker: Edward Ip, Wake Forest School of Medicine

Abstract: The use of latent variables for modeling psychological constructs and abilities has a long and successful history. The operationalization of latent variable models almost always involves evoking the local independence assumption, which states that condition on the latent variable, the multiple responses from an individual to a given stimulus are conditional independent. Local dependency (LD) refers to the scenario for which such an assumption is violated. As test designs become more complex (e.g., testlet, task-based assessment) and the scope of applications of latent variable models rapidly expands (e.g., non cognitive assessment) increasing attention has been given to the assessment and modeling of LD. In this presentation, I will describe statistical tools for measuring, testing, and modeling LD. Motivated by applications in psychology, education, and health sciences, two perspectives will be discussed LD as a nuisance factor and LD as something of substantive interest. This presentation is not meant to be a comprehensive overview of the literature on LD; it will include statistical methods that the presenter has used, together with brief reviews of their origin and history and psychometric applications. These statistical tools include the loglinear model, the random effects model, generalized estimating equation, and hybrid parameterization.

Title: A Joint Multigroup Testlet Model for Responses and Response Times Accounting for Differential Item and Time Functioning

Speaker: Hong Jiao*, University of Maryland, College Park
Peida Zhan, Beijing Normal University
Manqian Liao, University of Maryland, College Park

Abstract: A joint modeling approach of response and response times is often used to understand the trade-off between accuracy and speed. As it is observed in real data analyses, the relationship between accuracy and speed is not invariant across diverse groups of test-takers. Some studies reported positive relation while other studies reported negative correlation between accuracy and speed. Some other studies even observed a curvilinear pattern between the two variables. Previous studies on joint modeling of responses and response times often assume invariance of model parameters across sub-groups of the examinee population. Differential item functioning and differential speed functioning were not considered. As testlets are frequently used unit in many assessment program, local item dependence due to testlets should be accounted for in the joint modeling of response and response times. Several studies have explored using PISA data for the joint modeling of responses and response times, none of the studies has modeled testlet effects and potential differential item and time functioning in PISA data.

This study proposes a multigroup testlet modeling approach jointly for responses and response times which simultaneously take into account of differential item functioning and differential time functioning. The proposed joint models will be first fitted with the PISA response and response time data. Several alternative joint models will be compared to better understand the nature of PISA response and response time data. A follow-up simulation study will mimic the response and response time structures as obtained in the analysis of PISA data to explore model parameter recovery and the impact of ignoring testlet effects and differential item and time functioning on the joint modeling of responses and response times.

Title: Using response time to assess learning progress: A hidden Markov Model for Response and Response Time

Speaker: Shiyu Wang, University of Georgia

Abstract: Analyzing students growth has always been an important topic in educational research. Most recently, the cognitive diagnostic models have been used to track skill acquisition in a longitudinal fashion, with the purpose to provide an estimate of students' learning trajectories in terms of the change of fine-grained skills over time. Response time (RT), that is the amount of time the test taker spends considering and answering each item, has been extensively studied and used in a testing environment as the useful source of information to reflect individual response behavior and item characteristics. In this study, we consider using response time in a learning environment to model students' learning progress. This type of information, in addition to their responses to the test questions, can be valuable source of information to measure students learning trajectory. We propose a methodology framework to model the change of response time together with the longitudinal DCM, which can be provided to characterize individualized students growth pattern and at the same time evaluate the designed learning system. The proposed models are evaluated through a computer-based learning system that is designed to improve students spatial skills.

This is the joint work with Jeff Douglas, Susu Zhang and Steve Culpepper, from University of Illinois at Urbana-Champaign

---

Title: Relations between psychometrics for complex systems research

Speaker: Han van der Maas, University of Amsterdam

Abstract: Humans, like eco-systems, the weather and the stock market, are non-linear complex systems. Complex systems can be understood as networks that, depending on connection strength, behave linearly or nonlinearly, even discretely. There are many interesting technical links and equivalences between network models and the dominant latent variable approach in psychometrics, but conceptually they are very different. This will be discussed in the context of the measurement of cognitive functions and modeling of general intelligence. Secondly, I will present a new approach to educational measurement, motivated by the requirement of high frequent measurements in complex systems research. I will present the results of a web-based computerized adaptive training and monitor systems used by thousands of schools in the Netherlands, yielding over 1 billion item responses.

Title: Parameter Invariance and Skill Attribute Continuity in Diagnostic Classification Models: Bifactor MIRT as an Appealing Alternative

Speaker: Daniel Bolt, University of Wisconsin - Madison

Abstract: Psychometric models for diagnostic classification generally assume discrete skill attributes. Among concerns related to possible model misspecification is the likelihood of continuity in both the person skill attributes and the test item skill requirements. Using both simulated and real data, we explore the consequences of these forms of misspecification on the invariance of skill attribute mastery metrics across groups of different skill distributions, as might be applicable when studying skill acquisition over time. In this context, the bifactor MIRT model is presented as an appealing alternative. The applicability of the bifactor approach follows from the tendency for items measuring multiple conjunctively interacting skill attributes to primarily distinguish only with respect to the most difficult of the required skill attributes, especially when a strong higher order factor underlies the skill attributes. We illustrate these results by simulation and in application to frequently analyzed fraction subtraction datasets. Issues in the use of bifactor models for diagnostic assessment, including both advantages and disadvantages of this alternative approach, are considered.

Title: Discontinuation Rules in Ability Testing: New Results on Ignorability, Local Dependency, and Bias

Speaker: Matthias von Davier, National Board of Medical Examiners

Abstract: This presentation provides new results on a form of adaptive testing that is used frequently in intelligence testing. In these tests, items are presented in order of increasing difficulty. The presentation of items is adaptive in the sense that a session is discontinued once a test taker produces a certain number of incorrect responses in sequence, with the subsequent (not observed) responses commonly scored as wrong. The Stanford-Binet Intelligence Scales (SB5; Riverside Publishing Company, 2003) and the Kaufman Assessment Battery for Children (KABC-II; Kaufman & Kaufman, 2004) are examples that use this scoring rule. He & Wolfe (2012) compared different ability estimation methods in a simulation study for this discontinuation-based adaptation of test length. However there has been no study, to our knowledge, of the underlying distributional properties based on analytic arguments drawing on probability theory, of what these authors call stochastic censoring of responses. The study results obtained by He & Wolfe (2012) agree with results presented by DeAyala, Plake & Impara (2001) as well as Rose, von Davier & Xu (2010) and Rose, von Davier & Nagengast (2016) in that ability estimates are biased most when scoring the not observed responses as wrong. This scoring is used operationally, so more research is needed in order to improve practice in this field.

Title: Directions for Learning Research in Cognitive Diagnosis

Speaker: Jeff Douglas, University of Illinois Urbana-Champaign

Abstract: Online learning platforms offer great opportunities to model learning, which can in turn provide tools for item selection and guiding students more efficiently to mastery of a domain. Several aspects related to learning models are discussed. Selection of a measurement model and learning transition model are considered, along with techniques for assessing goodness of fit. The unresolved question of identifiability in learning models is discussed and compared with the same issue for static models. The notion of item selection to promote learning is proposed, along with the problem of mastery detection for the purpose of minimizing the expected time to learn a domain. Finally, utilization of response times is considered, and applications and benefits of using them are proposed. Real data illustrations are provided using an experiment of training spatial reasoning skills.

Title: Looking at Deep Neural Networks through IRT glasses

Speaker: Gunter Maris, ACTNext, by ACT

Abstract: Of late Deep Neural Networks have attracted quite a bit of attention and a range of interesting applications have been published. Training them however remains a resource intensive activity, which needs some level of fine tuning. Developing an understanding for what it is exactly, or even approximately, that a Deep Neural Network has learned seems most often impossible. Leveraging earlier work on the relationship between Multidimensional Item Response Theory models and Ising networks (Marsman et al, 2017) we can make progress in both areas. We show that Deep Neural Networks are nothing but Multidimensional IRT models with a particular mixture of multivariate normal distributions for the latent variables.

Marsman, Maarten & Borsboom, Denny & Kruis, Joost & Epskamp, Sacha & van Bork, Riet & Waldorp, Lourens & van der Maas, Han & Maris, Gunter. (2017). An introduction to network psychometrics: Relating Ising network models to item response theory models. Multivariate Behavioral Research.

Title: Real Time Performance Monitoring in Serious Gaming

Speaker: Jean-Paul Fox, University of Twente

Abstract: Serious Games (SGs) in the educational game market have been developed to support learning, where the primary focus is education rather than entertainment. The usefulness of SGs depend on the game characteristics, which need to enhance learning in players. To evaluate the effectiveness of an SG a randomized controlled trial (RCTs) can be used. Different RCTs have been carried out to review the effectiveness of SGs but this only provide information about the average effect on learning across players, and do not provide information about the performance improvement of single players.

For a single player, it would be of interest to evaluate changes in performance during a game session and also across game sessions. Accurate information about the players performance can be used to provide feedback during the game to improve their confidence and motivation. Personalized feedback should be given adapted to the speed of working and the level of performance to get the player more involved in the game. Given accurate information (e.g., response times and responses) about the performance it should be possible to identify learners disengagement. To enhance learning it is essential that disengagement is identified.

From the SGs, responses and response times for each task in the game can often be observed as indicators of the performance. The observations are considered to be a sequence of outcomes over time of a response process, which can be monitored using statistical process control (SPC) techniques. Then, across time changes in game performance of a player are tested to identify meaningful changes. Therefore, parameters of a response model are calibrated in the first time-interval. This response model serves as a baseline model, where the model with the estimated parameters describe the in-control situation. Boundary values of the estimated parameters can be used to specify a significant change in performance. A statistical test (e.g., likelihood ratio, Bayes factor) is used to test across time windows (a set of fixed time intervals) when a significant change is detected. A Bayes factor approach is also considered to quantify the accumulative evidence in favor of a hypothesis. The procedure can be applied to identify construct mastery.

Data from the SG Leos Pad (Kidaptive) were used to illustrate the method (Fox et al., 2017), where response and response time data from children were monitored. A (autoregressive) logistic regression model was used to identify significant changes in performances with respect to speed of working and ability. For different players, changes in performance was tested over time and control charts were used to monitor each learning process. To monitor jointly the performance of players with respect to speed and ability, a joint model for responses and response times is needed. It is shown how to monitor jointly and simultaneously assess changes in speed and ability.

Fox, J.-P., Steinrucke, J., van Steenbeek, C., and Verhagen, A.J. (2017). Change-point detection methods to identify increase of learning and disengagement. In preparation.

Title: Estimating the Cognitive Diagnosis Q Matrix with Expert Knowledge

Speaker: Steven Culpepper, University of Illinois Urbana-Champaign

Abstract: Cognitive diagnosis models (CDMs) are an important psychometric framework classifying students in terms of attribute or skill mastery. The Q matrix, which specifies the required attributes for each item, is central to implementing CDMs. The general unavailability of Q for most content areas and datasets poses a barrier to widespread applications of CDMs and recent research accordingly has developed fully exploratory methods for estimating Q. However, current methods do not always offer clear interpretations of the uncovered skills and existing exploratory methods do not use expert knowledge to estimate Q. We consider Bayesian estimation of Q using a prior based upon expert knowledge. The developed method can be used to validate which of the underlying attributes are predicted by experts and to identify residual attributes that remain unexplained by expert knowledge. We report Monte Carlo evidence about the accuracy of selecting active expert predictors and present an application using Tatsuoka's fraction-subtraction dataset.

Title: Do **I** Complete **Q**?

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

Abstract: A complete Q-matrix, which may or may not involve an identity matrix, is necessary for the identification of all attribute profiles. However, the completeness, or lack thereof, of a particular Q-matrix may vary from one cognitive diagnosis model (CDM) to another. A method that has been proposed to assess Q-matrix completeness is to compare the success probabilities across the items of the different attribute profiles. This method presupposes that the underlying CDMs are known, a condition that is difficult to meet in practice. The current work proposes a simulation-based approach to assess Q-matrix completeness. The proposed method involves determining the simplest CDMs empirically, and disentangling completeness from test reliability. A simulation study was conducted to evaluate the viability of the proposed method. Results show that the simulation-based method performs well under most conditions, but needs to be used with caution when the sample size is small and items are of inadequate quality.