# IMPS 2018

*Columbia University · New York City, New York, USA · July 9-13, 2018*

# Abstracts

## Talks

## Monday, July 9, 2018

### IMPS Registration: 8:30 AM – 1:30 PM

### Short Course: 9:30 AM – 5:00 PM

**Short Course: Continuous- and Discrete-Time Dynamic Modeling in R**
Sy-Miin Chow, The Pennsylvania State University; Michael Hunter, University of Oklahoma; Charles Driver, Max Planck Institute for Human Development; Peter Molenaar, The Pennsylvania State University

**Short Course: Computerized Adaptive Testing and Multistage Testing with R**
David Magis, KU Leuven; Duanli Yan, Educational Testing Service; Alina A. von Davier, ACTNext

**Short Course: New Matching for Causal Inference and Impact Evaluation**
Jose R. Zubizarreta, Harvard Medical School

## Tuesday, July 10, 2018 AM

### IMPS Registration: 8:00 AM – 5:00 PM

### Keynote Speaker: 9:15 AM – 10:15 AM

Chair: Anders Skrondal

**The Statistical Crisis in Science and What We Can Do About It**
Keynote Speaker: Andrew Gelman, Columbia University

### Refreshment Break: 10:30 AM – 11:00 AM

### Symposium 1: 11:00 AM – 12:00 PM

Chair: Maarten Marsman

**Symposium 1: Cognitive Development: Network Psychometric Approaches**

**Symposium 1 - Parallel Session: 1.1A: ASSESSMENT MEETS LEARNING: ON THE RELATION BETWEEN IRT AND BKT**
Benjamin Deonovic, ACTNext by ACT, Inc.

Learning and assessment deal with related topics but are separate concepts. Learning can be defined as the acquisition of knowledge, skills, values, beliefs, and habits through experience, study, or instruction. Assessments are instruments designed to observe behavior in a learner and produce data that can be used to draw inference about the knowledge, skills, values, beliefs, and habits that the learner has. Although learning and assessment serve the goals of education, which are to facilitate learning, the statistical models used to describe learning data and assessment data have significantly diverged and grown to leverage the salient features and distinct assumptions that are embodied in their respective data sets. The fields of educational data mining and learning analytics harness the dynamic, temporal, and large scale nature of learning data to construct models which can be used to predict learner performance, personalize and adapt instructional content, recommend intervention and curriculum changes, and provide information visualization to track progress. On the other hand, the same objectives are targeted by the field of psychometrics, using cross-sectional assessment data rather than longitudinal data. We explore the connections between Bayesian Knowledge Tracing (BKT) and Item Response Theory (IRT). BKT, a statistical model in educational data mining, is the most ubiquitous model used for

data obtained from intelligent tutoring systems, which are systems constructed to provide immediate and customized instruction to learners. IRT, a modeling framework developed in the field of psychometrics, was designed for constructing and analyzing assessments.

**Symposium 1 - Parallel Session: 1.1B: POLYA-YULE NETWORKS AND CONTAGION IN COGNITIVE DEVELOPMENT**

Maarten Marsman, University of Amsterdam; Lourens Waldorp, University of Amsterdam; Gunter Maris, ACTNext by ACT, Inc.

In network psychometrics, latent variable representations of network models that originate from statistical physics has been of primary interest. In particular, it has been shown that the notorious Ising network model can be expressed as a marginal IRT model; revealing a dual interpretation of observed correlations. A similar situation can be traced back to early models for the spread of disease, which can be seen to be either contagious (manifest interaction) or hereditary (latent variable). Greenwood & Yule (1920) proposed a model—independently rediscovered by Eggenberger & Polya (1923)—that can be derived from either perspective and thus giving a dual interpretation to the spread of disease. The mechanism of contagion can be used to model aspects of cognitive development, or a lack thereof. A case in point is reading at the start of primary education, where difficulties in acquiring the skills to read, or read well, eventually infects the performance on other topics in the curriculum, such as mathematics. Based on these ideas we introduce a novel network model—the Polya-Yule network—that shares several qualitative aspects with the Ising network model. In particular, the existence of a phase transition in more than one dimension. A convenient feature of the Polya-Yule model is that it is completely tractable, whereas an application of the full Ising model is computationally infeasible even for networks with a limited number of variables.

**Symposium 1 - Parallel Session: 1.1C: A NETWORK THEORY OF DEVELOPMENTAL INTELLIGENCE**

Alexander Savi, University of Amsterdam

The days of the genetic dominance in intelligence are numbered. Contemporary models, such as the multiplier effect model and mutualism model, not only convincingly capture explicit roles for the environment and internal cognitive processes, but also offer serious solutions to the IQ paradox. As the construct of general intelligence is deeply embedded in the complex dynamical system of cognitive development, the obvious next leap is to imagine a model of developmental intelligence. In this talk, we suggest a formal approach. We conceptualize intelligence as an evolving network. The static model, an extension of the Fortuin Kasteleyn model, provides a parsimonious explanation of the positive manifold and intelligence's hierarchical structure. On top of that, we suggest a dynamic growth mechanism that explains several developmental phenomena. Our approach offers a fundamental new lens on a century old construct, and may serve as a point of departure for studying gene-environment interactions, the influence of education, and various growth mechanisms to name just a few.

**Symposium 1 - Parallel Session: 1.1D: MECHANISMS AND DYNAMICS OF DEVELOPMENT**

Lourens Waldorp, University of Amsterdam; Maarten Marsman, University of Amsterdam

Mechanistic views on individual development are extremely important. They show us, for example, how it is the case that different cognitive performances are positively correlated and how children's cognitive performance become more diverse with age. A mechanistic model indicates what learning is. For instance, in a network model, learning could be perceived as the union of two or more concepts or knowledge chunks. This view elucidates how we should intervene with educational programs. Another more theoretical viewpoint is that of dynamical systems. Such a perspective provides insights into specific trajectories of individual development. It can also indicate what stages are encountered in learning and how and in what stage a person will end up in. For instance, with dynamical systems theory the question of whether learning is a gradual process or whether it is a sudden change can be addressed. In this talk we will try to combine both views on development. To do so we will show that a network can be approximated by stochastic differential or difference equations. These approximations allow for an analysis where the dynamics of a network can be investigated. The issue of ergodicity and of control can, for instance, be considered for such models.

## Symposium 2: 11:00 AM – 12:00 PM

Chairs: Yang Liu; Ji Seung Yang

**Symposium 2: Multi-stage Estimation and Inference in Measurement Modeling**

**Symposium 2 - Parallel Session: 1.2A: REPLENISHING THE GRIT SCALE USING RESTRICTED RECALIBRATION**

Monica Morell, University of Maryland, College Park; Ji Seung Yang, University of Maryland, College Park; Yang Liu, University of Maryland, College Park

Previously calibrated items are often re-used as anchor items when updating an item bank or expanding an existing instrument. If the expanded instrument that contains common items is administered to a sample that comes from a possibly different population than that of the original calibration dataset, it may be necessary to adjust the calibration model as item characteristics may change. The gold-standard approach, conducting multiple-group IRT analysis (Bock & Zimowski, 1997), is not possible when researchers only have access to published item parameter estimates and not the original response data. In such instances, parameters of the anchor items are often fixed at their calibrated values, conditional on which new item parameters are estimated. This method is henceforth termed restricted recalibration (RR; Liu, Yang, & Maydeu-Olivares, 2017). However, the naive estimate of the asymptotic covariance matrix (ACM) for model parameters under RR fails to account for the sampling variability carried over from original calibration and thus subsequent inferences may be misleading. The correct ACM should be obtained under the framework of pseudo maximum likelihood estimation (Parke, 1986). In the current study, two empirical datasets expanding the Grit short scale (Duckworth & Quinn, 2009) are used to compare the multiple-group analysis, RR with naive ACM, and RR with corrected ACM in terms of the standard errors for the new parameters, goodness of fit testing with the M2 statistic (Maydeu-Olivares & Joe, 2005), and detecting differential item functioning.

**Symposium 2 - Parallel Session: 1.2B: IRT ANALYSIS IN LONGITUDINAL RANDOMIZED CONTROLLED TRIALS**

Marian Strazzeri, University of Maryland, College Park; Ji Seung Yang, University of Maryland, College Park

A longitudinal randomized control trial is often used to estimate treatment effects on outcomes that are typically measured by administering the same set of items repeatedly across time. While this design is known for its better power to detect treatment effects by controlling for the between-subject variability, the one-stage full-information maximum likelihood (FIML) estimation of multiple group latent growth curve analysis with categorical responses data is challenging because high-dimensional integration is needed. Such a "curse of dimensionality" is exacerbated as the numbers of items and observed time points become large. The purpose of this study is to evaluate the comparative performance of five computationally more tractable methods to analyze longitudinal randomized trial data. Five approaches include plugging-in IRT scale scores (e.g., Expected A Priori scores), plausible values (e.g., Mislevy, Beaton, Kaplan, & Sheehan, 1992), doubly imputed plausible values (e.g., Yang, Hansen, & Cai, 2012), maximum likelihood estimation of mis-specified latent growth curve model that omits common item effects (e.g., Zheng & Yang, 2018), and limited information estimation of the fully specified latent growth curve model. Motivated by an empirical study (patient-reported treatment tolerability in cancer clinical trials) and literature (e.g., Glas, Geerlings, van de Laar, & Taal, 2009), a Monte Carlo simulation study is conducted to evaluate the Type I error rate and power to detect treatment effects.

**Symposium 2 - Parallel Session: 1.2C: CORRECTION FOR IRT THETA MEASUREMENT ERROR IN LINEAR MIXED MODELS**

Chun Wang, University of Minnesota; Gongjun Xu, University of Michigan; Xue Zhang, Northeast Normal University

When latent variables are used as outcomes in regression analysis, a common approach that is used to solve the measurement error issue is to take a multilevel perspective on item response modeling (IRT, e.g., Adams, Wilson, & Wu, 1997). Although recent computational advancement allow efficient and accurate estimation of multilevel IRT models, we argue that a two-stage divide-and-conquer strategy still

has its unique advantages. Within the two-stage framework, three methods that take into account heteroscedastic measurement errors of the dependent variable in stage II statistical analysis are introduced, they are closed-form marginalized MLE (CF-MMLE), Monte Carlo Expectation Maximization (MC-EM), and mean and covariance estimates methods. They are compared to the naïve two-stage estimation. A simulation study is conducted to compare the four methods in terms of model parameter recovery and their standard error estimation. The pros and cons of each method are also discussed to provide guidelines for practitioners. Finally, a real data example is given to illustrate the applications of various methods using the National Educational Longitudinal Survey data (NELS 88).

**Symposium 2 - Parallel Session: 1.2D: A TWO-STAGE PROCEDURE TO DETECT COMPROMISED ITEMS**
Xi Wang, Measured Progress; Yang Liu, University of Maryland, College Park

In continuous testing programs, some items are repeatedly used across test administrations to reduce item development cost. This, however, poses a threat to test security: examinees taking the test earlier may steal the items and then share them with future examinees. Therefore, it is of interest to evaluate whether items have become compromised over time due to examinees' preknowledge. In this study, we propose a method to detect compromised items when a test can be partitioned into two sets of items: secure items (T1) and possibly compromised items (T2). T1 items could be items that are used for the first time, whereas T2 items have been repeatedly used. Assuming item parameters are available, the proposed method detects compromised items in two stages: (1) Examinees' ability distribution is estimated from responses to T1 items; (2) The residual between observed proportion correct and model-implied value is calculated for each item in T2, where the model-implied proportion correct is calculated based on the estimated ability distribution in Stage 1. The standard error of the residual is derived analytically at the population level, and a computational more tractable sample-estimate is further derived. A simulation study has been conducted to evaluate the type-I error and power of the standardized residual statistic under different sample sizes, T1 test lengths, and compromise rates in T2 items.

**Symposium 2 - Parallel Session: 1.2E: IRT SCORE PROJECTION VIA CALIBRATION AND RESTRICTED RECALIBRATION**
Shuangshuang Xu, University of Maryland, College Park; Yang Liu, University of Maryland, College Park

When two tests measure distinct but related constructs, we could use multidimensional IRT models to predict scores of one test using responses to items on the other test. During this process, the sampling variability due to item parameter estimation is carried over to the predicted scores and thus must be accounted for. In the present work, we propose a multiple-imputation-based adjustment to the standard error of the projected scores. In particular, we consider the following two scenarios: (1) simultaneous estimation of item parameters and projected scores, and (2) restricted recalibration (Liu, Yang, & Maydeu-Olivares, 2017) and score projection based on published item parameter estimates. While the work of Yang, Cai, & Hansen (2012) is directly applicable to the fi rst scenario, the second scenario requires a simulation-based re-construction of the asymptotic covariance matrix for the model parameters. We apply the proposed score projection method to link the pediatric and adult anxiety scales in the Patient-Reported Outcomes Measurement Information System (PROMIS)—a problem originally studied by Thissen, Liu, Magnus, and Quinn (2015).

## Bayesian Statistical Inference: 11:00 AM – 12:00 PM

Chair: Jean-Paul Fox

**Bayesian Statistical Inference**

**Bayesian Statistical Inference - Parallel Session: 1.3A: A MULTIVARIATE PROBIT MODEL FOR LEARNING TRAJECTORIES WITH APPLICATION TO CLASSROOM ASSESSMENT**
Yinghan Chen, University of Nevada, Reno; Steven Culpepper, University of Illinois at Urbana-Champaign

Advances in educational technology provide teachers and schools with a wealth of information about student performance. A critical direction for educational research is to harvest the available longitudinal

data to provide teachers with real-time diagnoses about students' skill mastery. We propose a new dynamic cognitive diagnosis model that tracks the learning process over time and incorporates external covariates. In particular, we model the changes in skill profiles with a multivariate probit model and estimate the model parameters in a Bayesian approach. We also apply our method to an educational intervention study to provide a fine-grained assessment of the experimental intervention.

**Bayesian Statistical Inference - Parallel Session: 1.3B: BAYESIAN NONPARAMETRIC ORDERED LATENT CLASS MODELS**

Xiang Liu, Teachers College, Columbia University; Matthew Johnson, Teachers College Columbia University; Hui Soo Chae, Teachers College, Columbia University; Gary Natriello, Teachers College, Columbia University

Ordered latent class models (OLCM) impose inequality constraints on item response probabilities based on the ordering of the latent classes. These models have been widely used to analyze educational and psychological measurement data. OLCMs also have close connections to nonparametric item response theory models where the ordered discrete latent classes can be used to approximate the continuous latent trait. The traditional method of fitting OLCMs requires knowing the number of latent classes a priori. In determining the optimal number of classes, researchers often rely on comparative fit indices such as the Akaike information criterion (AIC) and Bayes information criterion (BIC). This approach can create a heavy computational burden, since models with different dimensions have to be estimated. Moreover, the subsequent inference conditional on the selected model ignores the uncertainty of the model selection process. In this paper we propose a modified Chinese restaurant process prior for constructing Bayesian nonparametric OLCMs with an unknown number of classes. In addition, we introduce a Gibbs sampling method for posterior computation across different dimensions. We conducted simulations to examine the effectiveness of the proposed method for recovering the model dimensions and approximating item response functions. Additionally, we analyzed real data sets in order to demonstrate the utility of this model.

**Bayesian Statistical Inference - Parallel Session: 1.3C: A HIERARCHICAL BAYESIAN COGNITIVE DIAGNOSTIC FACTOR MODEL FOR LEARNING TRAJECTORIES**

Albert Man, UIUC

With the increase in online and electronic learning, there is a need for tools to model the learning of students. One such interest is in classifying individuals into groups of similar learning trajectories. Cognitive diagnosis models (CDM) are diagnostic models used to classify respondents into groups based on latent skills estimated from item response data. However, CDMs traditionally have no capability for modelling dynamic, evolving skills. We propose a class of CDMs with an exploratory factor analysis (EFA) model to model latent learning trajectories over time, and to classify individuals into groups of similar learning trajectories. Parameter recovery of this model was evaluated through a Monte Carlo simulation study. The study results indicate that the proposed model provides good convergence rates and parameter recovery. The model was then fit on a N-back dataset, model fit was evaluated through posterior predictive model checking.

**Bayesian Statistical Inference - Parallel Session: 1.3D: A LATENT VARIABLE REPRESENTATION APPROACH FOR SCORE DISTRIBUTIONS IN TEST EQUATING**

Inés Varas, Department of Statistics - Pontificia Universidad Católica de Chile; Jorge González, Pontificia Universidad Católica de Chile; Fernando Quintana, Department of Statistics - Pontificia Universidad Católica de Chile

Comparability of measurements is an important practice in different fields. In educational measurement, the comparability of test scores is of crucial interest because scores are used to make important decisions. Equating methods have been developed to achieve the goal of having comparable scores from different test forms. All the approaches proposed in the equating literature make use of what is called an equating transformation which maps the scores on the scale of one test form, X, into their equivalents on the scale of another, Y. The equipercentile equating transformation is computed by composing the inverse cumulative distribution function of the scores in Y with the distribution of scores in X (Braun and Holland, 1982). Because test scores are usually integer numbers (e.g., total number of correct answers),

computing the equating transformation in this way is either difficult or simply impossible. This problem has usually been tackled using continuous approximations of the score distributions via a continuization step. All the continuization methods lead to parametric,semiparametric and nonparametric estimators of the equipercentile equating function. Because the continuization step can hide important distributional characteristics of the score distributions, our approach tries to avoid the continuization step. Considering scores as ordinal random variables, we propose a continuous latent variable formulation to perform an equipercentile-like equating based on a flexible Bayesian nonparametric model (Kottas et. al.,2005). The proposed model is applied to simulated and real data collected under an equivalent group design. Some methods to assess the performance of our model are also discussed.

**Bayesian Statistical Inference - Parallel Session: 1.3E: ADVANCEMENTS AND OPPORTUNITIES IN MARGINAL ITEM RESPONSE MODELING**
Jean-Paul Fox

In contrast to a conditional item response model, where local independence is modeled by including one or more latent variables, in a marginal item response model dependencies between item responses are directly modeled through a covariance structure. Complex dependencies implied by testlet effects, differential item functioning, or multidimensionality can be directly modeled and do not need the inclusion of additional latent variables. The marginal model is more parsimonious as the conditional counterpart, which leads to a more efficient MCMC estimation method for the marginal model. By modeling the dependencies between item responses through a covariance structure, the support for a positive (or non-zero) correlation can be directly tested through a Bayes factor and/or BIC. This makes it possible to identify for instance testlet effects, a multidimensional component, or measurement non-invariance. The marginal modeling framework is flexible and can include latent variables next to a covariance structure to model additional dependencies between item responses. Such hybrid marginal models can be used to expand more traditional latent variable models to account for additional dependencies, without changing the mean component of the model. The marginal model is also able to deal with different response types. Through simulation studies and real data examples, advancements and new opportunities in marginal item response modeling are discussed.

## Classification, Clustering, and Latent Class Analysis: 11:00 AM – 12:00 PM

Chair: Marie Wiberg

**Classification, Clustering, and Latent Class Analysis**

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 1.4A: PARTIAL IDENTIFIABILITY OF COGNITIVE DIAGNOSIS MODELS**
Yuqi Gu, University of Michigan; Gongjun Xu, University of Michigan

Cognitive Diagnosis Models (CDMs) are statistical modeling tools widely used in educational and psychological measurement. CDMs are a family of restricted latent class models, where pre-specified restrictions are imposed on the parameter space of latent class models through the Q-matrix. The Q-matrix-induced restrictions reflect practitioners' diagnostic assumptions about how the observed responses depend on the respondent's latent traits in the diagnostic test. Despite their popularity, CDMs suffer from nonidentifiability due to the models' discrete nature and complex restricted structure. This work addresses the fundamental identifiability issue of CDMs by developing a general framework for strict and partial identifiability of the model parameters. Developed conditions for identifiability only depend on the pre-specified Q-matrix structure and are easily checkable, which provide useful practical guidelines for designing statistically valid diagnostic tests. Furthermore, the theoretical framework is applied to establish, for the first time, the identifiability of several examples from cognitive diagnosis applications.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 1.4B: LATENT TRANSITION ANALYSIS WITH THE ORDINAL DINA MODEL AND COVARIATES**

Charles Iaconangelo, Pharmerit International; Daniel Serrano, Pharmerit International

Recently, longitudinal applications of cognitive diagnosis models have illustrated the utility of these models for detecting within-examinee change (see Kaya & Leite, 2017; Wang, Yang, Culpepper, & Douglas, 2018; Ye, Fellouris, Culpepper & Douglas, 2016). However, these techniques have not been applied to polytomous items. This work presents a comprehensive latent transition analysis model that incorporates covariates along with the deterministic input noisy "and" gate model for ordinal data (OR-DINA; Iaconangelo, 2018). This allows for the use of Likert scale item responses in the modeling of transitions across latent classes over time. It is applied in the context of identifying placebo effects in patient reported outcomes. Using a simulated example based on empirical data, the placebo responders are examined – that is, subjects assigned to the placebo group whose item responses indicated they achieved full recovery from the disorders (attributes) evaluated by the questionnaire. Placebo effects plague clinical trials, leading to low rates of detection of treatment efficacy for experimental drugs and procedures. In this example, a disproportionate number of the placebo responders are male. Estimating the transition matrix for each gender and testing for differences in the placebo responder profile via a bootstrap approach reveals a statistically significant gender effect. Identification of placebo responders may allow researchers to investigate the cause of the placebo effect and subsequently refine the clinical trial.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 1.4C: A DISJUNCTIVE REFORMULATION OF THE REDUCED REPARAMETERIZED UNIFIED MODEL**

Auburn Jimenez, University of Illinois at Urbana-Champaign; Steven Culpepper, University of Illinois at Urbana-Champaign

In the field of educational measurement and assessment, cognitive diagnosis models (CDMs) are useful methods for classifying individual ability into proficiency levels, based on the mastery of a specific set of skills, or attributes. The Generalized, Noisy Inputs, Deterministic "and" Gate (GNIDA) model, which is also known as the reduced reparameterized unified model (rRUM), is a widely accepted model for assessing performance on conjunctive tasks. We introduce the Generalized, Noisy Inputs, Deterministic "or" Gate (GNIDO) model as a disjunctive reformulation of the GNIDA model. We generalize the proof of Kohn & Chiu (2016) to establish duality between conjunctive and disjunctive versions of the rRUM. We show how to jointly model conjunctive and disjunctive items. We use conjunctive and disjunctive rRUMs to estimate the the Q of the Spielberg anxiety data.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 1.4D: APPLICATION OF GAUSSIAN COPULA DISTRIBUTION TO THE DINA MODEL**

Kevin Carl Santos, The University of Hong Kong; Mingchen Ren, University of Calgary; Jimmy de la Torre, The University of Hong Kong; Alexander R. de Leon, University of Calgary

Cognitive diagnosis models (CDMs) provide discrete multidimensional attribute profiles to examinees that can be valuable information for improving classroom instruction and learning. The deterministic input, noisy "and" gate (DINA) model is one of the widely known CDMs due to its simplicity and tractability. In this study, we reformulate the DINA model by using a copula in constructing the joint distribution of the latent variables to allow flexibility on the (conditional) marginal models for the test items. Unlike the traditional DINA model, which assumes conditional independence of test items, and the multivariate probit DINA (DINA-MP) model, which relies on the assumption that all latent variables underlying the test items are Gaussian, our new Gaussian copula DINA model (DINA-GC) model is flexible enough to account for conditional dependence between items, accommodate non-probit links, and allow the consideration of mixtures of probit, logit, robit, and other links. A computationally efficient parameter-expanded Monte Carlo Expectation-Maximization algorithm is outlined for maximum likelihood estimation of the parameters of DINA-GC. A simulation study is conducted to determine the performance of the proposed estimation procedure in terms of bias and efficiency. We include a re-analysis of the fraction-subtraction data to illustrate the model's application.

## Model Fit, Comparison, and Diagnostics: 11:00 AM – 12:00 PM

Chair: Ed Merkle

**Model Fit, Comparison and Diagnostics**

**Model Fit, Comparison and Diagnostics - Parallel Session: 1.5A: ADAPTING FIT INDICES FOR BAYESIAN SEM: COMPARISON TO MAXIMUM LIKELIHOOD**
Mauricio Garnier-Villarreal, Marquette University; Terrence D. Jorgensen, University of Amsterdam

In a frequentist framework, the overall fit of a Structural Equation Model (SEM) is typically evaluated with the $\chi 2$ test and at least one index of approximate fit, although it is common practice to report multiple fit indices. Bayesian SEM (BSEM) has increased in popularity in recent years, with a few programs providing user-friendly syntax for it's specification and estimation. Current BSEM software provides one measure of overall fit, Posterior Predictive p-value (PPP) which has shown to be unreliable. Due to noted limitations of PPP, common practice for evaluating Bayesian model fit focuses instead on model comparison, using information criteria (e.g., LOO, WAIC, DIC) or Bayes factors. Fit indices developed under maximum-likelihood estimation have not been incorporated into software for BSEM. However, a recently proposed Bayesian RMSEA showed promise for evaluating overall fit, but only in large samples. Our research presents an adaptation of the $\chi 2$ and seven commonly used fit indices (RMSEA, GammaHat, adjusted GammaHat, Mc, CFI, TLI, NFI) for BSEM. We propose adjusting the posterior distribution of $\chi 2$ to calculate the posterior distributions for the fit indices, which (according to our simulation results) makes the sampling variability of their posterior means equivalent to their frequentist counterparts across sample sizes, model types, and levels of misspecification. Also, showing the behavior of each fit index under the conditions of our simulation. The proposed fit indices therefore allow overall model-fit evaluation using familiar metrics of the original indices, while also quantifying their uncertainty with the posterior standard deviation and credible interval

**Model Fit, Comparison and Diagnostics - Parallel Session: 1.5B: POSTERIOR PREDICTIVE MODEL COMPARISON OF TWO PERFORMANCE ASSESSMENT SCORING METHODS**
Aaron Myers, James Madison University; Allison J. Ames, James Madison University; Brian C. Leventhal, James Madison University; Madison A. Holzman, James Madison University

Interpretations of scores from a performance assessment rely on the assumption that all raters use the scoring rubric as the rubric developers intended. Misalignment between raters' scoring processes and the rubric developers' intended scoring process may lead to invalid score inferences. An alternative to the rubric scoring method--the Diagnostic Rating System (DRS)--was developed to explicitly align raters' scoring processes with the rubric developers' intended scoring processes. Raters using the DRS are required to respond to a series of branching, selected-response statements resembling a decision tree. Each unique path through the tree leads to a unique score. These scores are modeled with an item response tree (IRTree) model specified to depict the decision-making process. In this study, independent groups of raters scored a set of essays using either the traditional rubric or the DRS. Each sample of responses was fit using a bifactor IRTree. The bifactor IRTree was specified to parse out the variability of scores due to raters from the variability of scores due to student ability. To understand the effects of the DRS rating method, model-data fit was compared between the two samples. Evaluation of IRTree model-data fit has been limited to information criteria indices and substantive interpretations. Posterior predictive model checking (PPMC) allows for graphical displays and numerical quantification of model-data misfit in a Bayesian framework. This study illustrates PPMC methods applied to a bifactor IRTree model independently fit to scores from raters using the DRS and scores from raters using a rubric.

**Model Fit, Comparison and Diagnostics - Parallel Session: 1.5C: HIGHER-ORDER ASYMPTOTICS AND ITS APPLICATION TO TEST FRAUD DETECTION**
Sandip Sinharay, Educational Testing Service

Detection of test fraud by individual examinees often involves a test of the hypothesis that the ability of an examinee is the same over two sets of items. For example, to detect fraudulent erasures, one could test the hypothesis that the ability of an examinee is the same over the erased and non-erased items; a

better performance on the erased items may indicate that the erasures are fraudulent (Sinharay, Duong, and Wood, 2017). Traditional frequentist approaches that are used in testing of this hypothesis include the Wald test, the likelihood ratio test, and the score test (e.g., Fischer, 2003; Finkelman, Weiss, & Kim-Kang, 2010; Glas & Dagohoy, 2007). However, approaches based on higher-order asymptotics (e.g., Barndorff-Nielsen & Cox, 1994; Ghosh, 1994) can be used to test the hypothesis of equal ability on two sets of items. We show how the modified signed likelihood ratio test (Barndorff-Nielsen, 1986) and the Lugannani-Rice approximation (Lugannani & Rice, 1980), which are based on higher-order asymptotics, can be applied to test the abovementioned hypothesis for common item response theory models for dichotomous and polytomous items. The two tests based on higher-order asymptotics are shown to improve over traditional hypothesis-testing approaches in a detailed simulation study. Two real data examples are also provided. The results indicate that though higher-order asymptotics have found a few applications in educational and psychological measurement (e.g., Bedrick, 1997; von Davier and Molenaar, 2003; Biehler, Holling, & Doebler, 2015), more such applications are warranted.

**Model Fit, Comparison and Diagnostics - Parallel Session: 1.5D: THE MODEL FIT OF CONDITIONALLY INDEPENDENT DYAD (CID) MODELS**

Yating Zheng, University of Maryland College Park; Tracy Sweet, University of Maryland, College Park; Qiwen Zheng, University of Maryland College Park

Conditionally independent dyad (CID) models, including stochastic blockmodels (Holland et al., 1983) and latent space models (Hoff et al., 2002), account for the network structure through latent variables. With the advantages of being easy to estimate and interpret, CID models are becoming more and more popular. Several R packages have been developed for CID models, such as statnet (Hancock et al., 2008) which includes latentnet (Krivitsky et al., 2008), CIDnetworks (Dabbs et al., 2014) and HLSM (Adhikari et al., 2014). However, a potential problem with CID models is that there is very little research on model fit; the only goodness of fit function included in latentnet mirrors the work completed by (Hunter et al., 2008) for a non-CID class of models. We explore goodness of fit methods for several CID models to understand which methods best assess model fit. The development of this work will improve our understanding of latent space models and help us select an optimal model among different social network models.

**Model Fit, Comparison and Diagnostics - Parallel Session: 1.5E: ADVANCEMENTS AND OPPORTUNITIES IN MARGINAL ITEM RESPONSE MODELING**

Jean-Paul Fox

In contrast to a conditional item response model, where local independence is modeled by including one or more latent variables, in a marginal item response model dependencies between item responses are directly modeled through a covariance structure. Complex dependencies implied by testlet effects, differential item functioning, or multidimensionality can be directly modeled and do not need the inclusion of additional latent variables. The marginal model is more parsimonious as the conditional counterpart, which leads to a more efficient MCMC estimation method for the marginal model. By modeling the dependencies between item responses through a covariance structure, the support for a positive (or non-zero) correlation can be directly tested through a Bayes factor and/or BIC. This makes it possible to identify for instance testlet effects, a multidimensional component, or measurement non-invariance. The marginal modeling framework is flexible and can include latent variables next to a covariance structure to model additional dependencies between item responses. Such hybrid marginal models can be used to expand more traditional latent variable models to account for additional dependencies, without changing the mean component of the model. The marginal model is also able to deal with different response types. Through simulation studies and real data examples, advancements and new opportunities in marginal item response modeling are discussed.

## Longitudinal Data Analysis: 11:00 AM – 12:00 PM

Chair: Charles Driver

**Longitudinal Data Analysis**

**Longitudinal Data Analysis - Parallel Session: 1.6A: MODEL SELECTION IN CONTINUOUS-TIME DYNAMICAL SYSTEMS MODELS WITH TIME-VARYING PARAMETERS**
Meng Chen, The Pennsylvania State University

Incorporating time-varying parameters into dynamical systems and related models offers researchers insights into whether and how the rules governing a process change over time (i.e. changes of change, as observed in many human behavioral systems). Applications involving time-varying parameters have been gaining traction in the statistical, social and behavioral sciences literature in the context of discrete-time models (e.g. Bringmann et al., 2017; Chow, Zu, Shifren & Zhang, 2011), but substantially less so in the continuous-time realm. Similarly lacking is an investigation into the performance of popular model comparison approaches to detect time-varying parameters in continuous-time models. We compare the utility of confidence intervals and popular information criterion measures such as the Akaike information criterion (AIC), Bayesian information criterion (BIC) and sample size adjusted BIC for model selection purposes in a Monte Carlo simulation study. Results from an empirical study of emotion regulation are further used to demonstrate how these approaches can be used to distinguish time-varying parameters from time-invariant ones.

**Longitudinal Data Analysis - Parallel Session: 1.6B: DISENTANGLING INDIVIDUAL DYNAMICS - PROBABILISTIC CLUSTERING OF LONGITUDINAL DATA**
Anja Ernst, University of Groningen; Marieke Timmerman, University of Groningen; Casper Albers, University of Groningen

Studying idiographic dynamics through time series models is becoming increasingly popular in the social sciences. Often, researchers are interested in generalizing to a population of individuals, rather than being interested in the single individuals per se. As dynamics can be rather heterogeneous across individuals, one needs smart ways to express their essential similarities and differences across individuals. A way to proceed is to identify subgroups of people who are characterized by qualitative differences in their dynamics. Recently, dynamic clustering methods have been proposed to discern groups of individuals who exhibit homogeneous dynamics. So far, these methods assume equal generating processes for individuals of a cluster. To avoid this, in empirical practice overly restrictive assumption, I will outline a probabilistic clustering approach based on the Gaussian finite mixture model that clusters on individuals' VAR coefficients, thereby allowing for individual deviations within clusters. I will contrast the proposed method to another time series clustering procedure drawing form the results of a simulation study and illustrating their performance on an empirical data set. The models are applied to N = 366 ecological momentary assessment data with external measures of depression and anxiety.

**Longitudinal Data Analysis - Parallel Session: 1.6C: MODELING INTRAINDIVIDUAL VARIABILITY USING A HIERARCHICAL MIXTURE LATENT MARKOV MODEL**
Tanja Lischetzke, University of Koblenz-Landau; Claudia Crayen, Freie Universität Berlin; Michael Eid, Free University of Berlin; Jeroen K. Vermunt, Tilburg University

Multilevel models are flexible modeling techniques that allow to analyze intraindividual dynamics in continuous variables. They are often applied to intensive longitudinal data (e.g., experience sampling/ambulatory assessment data). By contrast, models with categorical latent variables such as (Mixture) Latent Markov models have only rarely been applied to intensive longitudinal data. However, they may be useful to answer specific research questions on intraindividual variability and change. In particular, Mixture Latent Markov models might represent a meaningful alternative to autoregressive multilevel models for continuous variables if information on stability vs. change should be combined with information on the momentary level of the variable. As an example, we present an application from the domain of affect dynamics (fluctuations in pleasant-unpleasant mood). We show how a hierarchical mixture latent Markov model for experience sampling data can be defined so that within-days and

between-days transitions between mood states can be modelled and potential population heterogeneity in the change process can be acknowledged. In addition, we present results of a simulation study that was conducted to find out the minimal data requirements for this model with respect to the number of occasions within each day, the number of days, and the number of individuals.

**Longitudinal Data Analysis - Parallel Session: 1.6D: ESTIMATING STUDENT GROWTH PERCENTILE UNDER INEQUALITY RESTRICTIONS**
Ruitao Liu, ACT, Inc.

The Student Growth Percentile (SGP) calculation is based on estimated percentiles of the current year score distribution as a function of the previous year scores. The commonly used estimation method (Betebenner, 2009) has three limitations: (1) when sample size is not very large, the estimated percentiles may be out of score range, which can result in unacceptable estimated SGP values. (2) Intuitively, people expect that the median percentile of the current score distribution for students having a high previous score is greater than or equal to that for students having a low previous score. The current method can give an estimation that runs counter to this intuition. (3) Ideally, the estimated percentile functions should not cross each other. The existing method uses an ad hoc approach to achieve this goal. But it needs more clear explanation and supporting empirical evidence. In this project, two approaches are proposed to overcome the aforementioned limitations. The first approach translates the original quantile regression model to a constrained optimization problem. A set of inequality restrictions is constructed to simultaneously handle the three limitations. The second approach is a two-step procedure. Betebenner's (Betebenner, 2009) method is first used to get initial percentile estimations. Then a post-adjustment is applied to guarantee that adjusted percentile estimations satisfy the requirements coming from the three limitations. The performance of the two proposed methods and Betebenner's method are compared through an extensive simulation study which considers different sample size and functional relationships between pre-scores and current scores.

**Longitudinal Data Analysis - Parallel Session: 1.6E: RANDOM TIME ERRORS IN GROWTH CURVE MODELING**
Satoshi Usami, University of Tokyo; Kou Murayama, University of Reading

Growth curve modeling (GCM) has been one of the most popular statistical methods to examine participants' growth trajectories using longitudinal data. In spite of the popularity of GCM, little attention has been paid to the possible influence of random time errors — time specific random errors which influence all participants. In this presentation we demonstrate that the failure to take into account random time errors in GCM produces considerable inflation of Type-1 error rates in statistical tests of fixed effects (e.g., coefficients for the linear and quadratic terms). We propose a GCM that appropriately incorporates random time errors using mixed-effects models to rescue the problem. We also provide an applied example to illustrate that GCM with and without random time errors would lead to different substantive conclusions about the true growth trajectories. Comparisons with other models in longitudinal data analysis and potential issues of model misspecification are discussed.

## Structural Equation Modeling: 11:00 AM – 12:00 PM

Chair: Eva Ceulemans

**Structural Equation Modeling**

**Structural Equation Modeling - Parallel Session: 1.7A: EXTENDING THE 3-STEP APPROACH TO A MULTILEVEL STRUCTURAL EQUATION MODEL**
Yajing Zhu, London School of Economics and Political Science; Fiona Steele, London School of Economics and Political Science; Irini Moustaki, London School of Economics and Political Science

Latent class analysis (LCA) is widely used to derive categorical variables from multivariate data which are then included as predictors of a distal outcome. To correct for misclassification error and to avoid having outcomes influence the latent class membership (drawbacks of the modal-class and the 1-step approach, respectively), bias-correction 3-step approaches have been developed for situations where there is one

LC variable (e.g. Bolck et al., 2004; Vermunt, 2010; Asparouhov & Muthén, 2014; Bakk & Vermunt, 2014, 2016). Zhu et al. (2017) extended the maximum likelihood based approach for multiple associated LC variables and evaluated its performance under various types of model misspecification. In this study, we propose a general multilevel SEM that relates the latent class variables to multiple distal outcomes which are a mixture of time-to-event and categorical outcomes. The proposed method also corrects for potential endogeneity in covariates and informative dropout by specifying a factor structure for residual association using an individual-level continuous latent variable. In addition to accounting for correlations between within-individual responses, the residual association across equations also serves as a correction for potential violation of the local independence assumption that is implicit in modelling the relationship between LC variables and outcomes of interest. The proposed method is applied in an analysis of the effects of multiple latent dimensions of childhood socioeconomic situations on partnership stability in adulthood and later health, using data from the 1958 British birth cohort.

**Structural Equation Modeling - Parallel Session: 1.7B: MODELING LONGITUDINAL DYADIC DATA IN THE SEM FRAMEWORK**

Fien Gistelinck, University Ghent

In dyadic research, people are often interested in estimating the effect of one's own (i.e., actor effect) and one's partner (i.e., partner effect) predictor variable on an outcome variable. Due to the nonindependence between the two dyad members, statistical models such as the actor-partner interdependence model (APIM) have been developed to analyze the outcomes of the two dyad members simultaneously. When dyads are measured over time, not only the scores of the members are nested within a dyad, but the scores of a dyad member at different time points are also correlated. One way to incorporate both interdependencies is to extend the APIM to the longitudinal case: the over-time standard APIM (OSA). This model both accounts for between-dyad variation at level-2 (the so-called G-side), and for nonindependence of level-1 residuals in each dyad (the so-called R-side). The latter can take complex forms such as "UN@AR(1)", which allows for correlation within a dyad at a specific time point and a first-order autoregressive process for the measurements over time within each dyad member. While the implementation of such complex covariance structure is available in few multilevel modeling software (e.g., SAS), we show how it can be implemented in structural equation modeling (SEM) software, such as the R-package "Lavaan". Given the complexity of the code to model such advanced covariance structures in SEM, a Shiny-application was developed to enable applied dyadic researchers to fit an OSA on their longitudinal dyadic data within the SEM framework.

**Structural Equation Modeling - Parallel Session: 1.7C: MODELING INTERACTIONS BETWEEN LATENT VARIABLES**

Paul Lodder, Tilburg University; Wilco H.M. Emons, Tilburg University; Jelte M. Wicherts, Tilburg University

We conducted a simulation study to investigate bias and precision of four common methods to model interactions between latent variables: (1) the sum score multiplication approach; (2) the single indicator latent variable approach; (3) the matched pairs latent variable approach; (4) the Latent Moderated Structural Equations (LMS) approach. We simulated data for both continuous and ordinal items and varied sample sizes, latent skewness, and size of the interaction. Our main outcomes were the bias and precision of the estimated interaction. In case of continuous item scores and larger interactions, the sum score approach overestimated- and the single indicator approach underestimated the interaction effect. The matched pairs approach showed least bias, whereas the LMS approach was biased with small sample sizes and large skewness. The sum score method was most precise, followed by the matched pairs method. High skewness and smaller sample sizes resulted in less precise estimates, especially for the single indicator and LMS methods. The Type I error rates were approximately 5% for the single indicator and matched pairs methods. The sum score and LMS methods produced more false positive findings, especially when sample size was large and skewness was high. For ordinal items, all methods showed biased estimates of the interaction effect. For interactions based on continuous item scores, we recommend the matched pairs approach. We advise researchers to use continuous measures whenever possible, because modeling interactions based on ordinal rather than continuous item scores resulted in more bias, less precision, and more false positive findings.

**Structural Equation Modeling - Parallel Session: 1.7D: FORMATIVE MEASUREMENT THEORY: THREE CONCEPTUAL IMPEDIMENTS TO PROGRESS**
Keith Markus, John Jay College of Criminal Justice of The City University of New York

Bollen and colleagues have advocated the use of formative scales despite the fact that formative scales lack an adequate underlying theory to guide development or validation such as that which underlies reflective scales. Three conceptual impediments impede the development of such theory: the redefinition of measurement restricted to the context of model fitting, the inscrutable notion of conceptual unity, and a systematic conflation of item scores with attributes. Setting aside these impediments opens the door to progress in developing the needed theory to support formative scale use. A broader perspective facilitates consideration of standard scale-development concerns as applied to formative scales including scale development, item analysis, reliability and item bias. While formative scales require a different pattern of emphasis, all five of the traditional sources of validity evidence apply to formative scales. Responsible use of formative scales requires greater attention to developing the requisite underlying theory.

**Structural Equation Modeling - Parallel Session: 1.7E: USING GENERALIZED STRUCTURAL EQUATION MODELING IN AGREEMENT: A UNIFIED FRAMEWORK**
Jay Verkuilen, The Graduate Center, CUNY; Sydne McCluskey, CUNY Graduate Center; Aybolek Amanmyradova, CUNY Graduate Center

Assessment of coder and measurement agreement is a very common task across the sciences, ranging from chemistry and medicine to content and text analysis. Unsurprisingly, there has been a vast proliferation of agreement coefficients, most of which apply to particular data types (binary, nominal, ordinal, interval, etc.). These coefficients are useful but often difficult to interpret with a wide range of arbitrary cutoffs. Many users simply apply their literature's favored coefficient, often to transformed data without reference to the actual scale of their variables. The purpose of this work is to cast agreement models as a particular family of generalized structural equation models (GSEMs, e.g., Muthen, 2002). GSEMs provide statistical inference, gracefully handle missing data, non-Gaussian outcomes that are conditionally distributed in the exponential family, and inclusion of covariates for group comparisons or validity assessment via an explanatory modeling approach (e.g., de Boeck & Wilson, 2015). GSEMs can also be used to generate a semi-parametric distribution of random effects, important given the frequently non-Gaussian distribution created by sampling designs in agreement studies. Because agreement models are restricted GSEMs, testing of specific hypotheses is also possible. GSEMs can be estimated in either a MML or Bayesian fashion. Finally, detailed assessment of fit and identification of outlying cases can be accomplished using resampling or case deletion. The approach will be illustrated using a dataset taken from the agreement literature as well as a novel dataset comparing human and machine coders in a large automated international conflict event data processing environment.

## Attendee Lunch: On Your Own: 12:00 PM – 1:30 PM

# Tuesday, July 10, 2018 PM

## Symposium 3: 1:30 PM – 3:00 PM

Chair: Don van den Bergh (2.1A-C, 2.1E)

Chair: Eric-Jan Wagenmakers (2.1D)

**Symposium 3: Psychometric Developments in JASP**

**Symposium 3 - Parallel Session: 2.1A: NETWORK ANALYSIS IN JASP**
Adela-Maria Isvoranu, University of Amsterdam; Don van den Bergh, University of Amsterdam

In recent years, there has been an emergence in the estimation of network models from various sources of data, in numerous fields of psychology (e.g., clinical psychology, personality, attitudes). A large

number of empirical researchers who aim to estimate such network models are not familiar with programming environments and therefore often unable to perform these analyses. We introduce the network module for JASP, which provides a novel point and click suite for estimating these models. The network module allows for the estimation of Pairwise Markov Random Fields, such as the Gaussian graphical model and the Ising model, from binary, ordinal, categorical and continuous data. Furthermore, JASP allows for the computation of descriptive statistics, such as centrality indices, as well as bootstrapping methods to assess the stability and accuracy of these results. This talk will introduce the JASP network module, as well as discuss the estimation and interpretation of network models, the descriptive statistics that may be derived, and the bootstrap accuracy tests. We showcase this functionality in JASP by analyzing a dataset of patients diagnosed with a psychotic disorder.

**Symposium 3 - Parallel Session: 2.1B: BAYESIAN LATENT NORMAL INFERENCE FOR THE RANK SUM TEST, THE SIGNED RANK TEST, AND KENDALL'S TAU**
Johnny van Doorn, University of Amsterdam

Parametric assumptions are often violated under non-normality, outliers, or an ordinal measurement level. To overcome the impact of such violations one can use rank-based methods, such as rank correlations and the Wilcoxon tests. Bayesian inference for these procedures is straightforward when a latent normal distribution is introduced. This presentation outlines the general methodology and offers practical applications to tests of association (Kendall's tau) and difference of means (Wilcoxon's rank sum and signed rank tests).

**Symposium 3 - Parallel Session: 2.1C: BAYESIAN MULTINOMIAL TEST FOR INFORMED HYPOTHESES**
Alexandra Sarafoglou, University of Amsterdam; Maarten Marsman, University of Amsterdam; Quentin Gronau, University of Amsterdam; Eric-Jan Wagenmakers, University of Amsterdam

Researchers often approach data analysis of categorical data already expecting certain relations between the probabilities of the categorical events. For example, in replication research researchers seek to confirm whether the data of a replication study show the same trend as the original study (e.g., Verhagen & Wagenmakers, 2014). The key assumption of item response theory—latent monotonicity—is another example for ordinal expectations. Researchers are able to adequately capture their expectations and formalize them into testable hypotheses by stipulating ordinal constraints on the parameters of interest. The Bayesian framework includes several methods to evaluate these hypotheses using Bayes factors, i.e., the encompassing prior approach (Klugkist, Kato, & Hoijtink, 2005) or the conditioning method (Mulder et al., 2009). These methods, however, are potentially unstable and time-consuming if the number of categories increases. We introduce an accurate and fast alternative for evaluating ordinal constrained hypotheses in categorical data. Our approach is based on the bridge sampling method proposed by Bennet (1976) and Meng and Wong (1996). We also demonstrate how our method generalizes beyond the context of categorical data to any type of statistical model. Using the corresponding user-friendly JASP implementations, we will showcase straightforward testing of ordinal constrained hypotheses for the most commonly used statistical procedures used in empirical science.

**Symposium 3 - Parallel Session: 2.1D: USING CONTRASTS TO TEST INFORMED HYPOTHESES IN MULTIDIMENSIONAL IRT**
Don van den Bergh, University of Amsterdam; Maarten Marsman, University of Amsterdam; Timo Bechger, Cito

In the analysis of data from large-scale educational tests consistent patterns of results often emerge. As a motivating example, we consider here the analysis of data from the CITO end of primary school test in the Netherlands, which assesses the performance of over 150 000 pupils on different topics from the primary school curriculum. One consistent observation in the analysis of this CITO test data is the positive manifold; there is a dominant first eigenvalue with roughly equal loadings on all items. Another consistent observation is the strong contrast between mathematics-oriented items and language-oriented items; most pupils tend to be good in either mathematics or language and few are good in both. These patterns of observations emerge in the analysis of different versions of the CITO test data, i.e., different test, different pupils. Even though similar patterns can be observed in the analysis of similar test forms administered to similar populations in consecutive years, we often analyze our data as if we know

nothing about the problem at hand, wasting precious degrees of freedom along the way. Here, we introduce a method to incorporate prior information in multidimensional IRT models. The idea is to encode this information as eigenvectors or contrasts in the model. By replacing unknown eigenvectors with simple contrasts the number of free parameters can be significantly reduced, thus increasing statistical power. We demonstrate how to estimate Bayesian multidimensional IRT models with contrasts, and develop Bayes factor hypothesis tests for the inclusion (or exclusion) of these contrasts.

**Symposium 3 - Parallel Session: 2.1E: POSTERIOR DISTRIBUTIONS FOR RELIABILITY MEASURES**
Eric-Jan Wagenmakers, University of Amsterdam; Julius Pfadt, University of Amsterdam; Don van den Bergh, University of Amsterdam

Popular measures of a test's reliability include McDonald's omega, Guttman's lambda 6, the Greatest Lower Bound (GLB), and, mostly for historical reasons, Cronbach's alpha. This presentation first shows how these measures can be estimated within a Bayesian framework. Specifically, the posterior distribution for these measures can be obtained through Gibbs sampling -- for alpha, lambda 6, and the GLB one can sample the covariance matrix from a normal inverse Wishart distribution; for omega one samples the conditional posterior distributions from a single-factor CFA-model. Simulations show that -- with default priors-- the 95% Bayesian credible intervals are highly similar to the 95% frequentist bootstrapped confidence intervals, yielding a Bayesian validation for the frequentist bootstrap; in addition, the Bayesian posterior distribution can be used to address practically relevant questions, such as "what is the probability that the reliability of my test is between .7 and .9?", or, "how likely is it that the reliability of my test is higher than .8?". In general, the use of a posterior distribution attends users to the inherent uncertainty in the estimation of reliability measures.

## Symposium 4: 1:30 PM – 3:00 PM

Chair: Marieke E. Timmerman

**Symposium 4: Identifying Specific and Distinctive Factors: Exploratory Bi-factor Modeling and Beyond**

**Symposium 4 - Parallel Session: 2.2A: UNRAVELING THE MIX: FACTOR LOADING NON-INVARIANCE ACROSS MANY GROUPS**
Kim De Roover, Tilburg University; Eva Ceulemans, Katholieke Universiteit; Jeroen K. Vermunt, Tilburg University

Multigroup exploratory factor analysis (EFA) has gained popularity to address measurement invariance (MI), since it avoids overly restrictive zero loadings (Asparouhov & Muthén, 2009) as well as a sequence of model modifications bound to result in error fitting (Browne, 2001). MI is tested in a stepwise fashion, where the first step is to evaluate factor loading (weak) invariance. When this invariance is untenable, comparisons of group-specific loadings are warranted to identify the sources of non-invariance. Nowadays MI is often tested across a large number groups, where the number of between-group comparisons exponentially increases with the number of groups, elevating the chances of falsely detecting non-invariance (Rutkowski & Svetina, 2014). To increase insight, efficiency and specificity, an intuitive solution is performing a mixture clustering of the groups based specifically on EFA loadings. This drastically lowers the number of comparisons needed to pinpoint factor loading non-invariances and the clustering of the groups is an interesting result in itself. To this end, mixture simultaneous factor analysis (De Roover et al., 2018) is extended to accommodate a blend of cluster- and group-specific EFA parameters. Unique to existing approaches (Kim, Cao, Wang, & Nguyen, 2017), the proposed method avoids the stringent assumptions of confirmatory factor analysis, disentangles different levels of non-invariance and sets aside parameter differences that are irrelevant to the quest for MI.

**Symposium 4 - Parallel Session: 2.2B: IMPROVING BI-FACTOR EXPLORATORY MODELING: FACTOR LOADINGS DIFFERENCE-BASED EMPIRICAL TARGET ROTATION**
Eduardo Garcia-Garzon, Universidad Autónoma de Madrid

Target rotation has become the cornerstone of exploratory bi-factor analysis. However, determining how to empirically define a target matrix, in which researchers distinguish between expected salient and trivial

factor loadings, is still a matter of controversy. Traditional methods applying arbitrary cut-off points could lead to target misspecification and impair factor recovery, depending upon the appropriateness of the values chosen. Two alternatives for the determination of factor-specific, adaptive cut-off points were examined. Firstly, a cut-off estimation (SLiP) based on Promin (Lorenzo-Seva, 1999). Secondly, a cut-off derived from a strategy relying on finding the first relevant difference in the sorted normalized factor loading distribution (SLiD). Both alternatives were tested within the iterative target rotation based on Schmid-Leiman solution algorithm (SLi; Abad, Garcia-Garzon, Garrido & Barrada, 2017). Additionally, a non- iterative, totally specified bi-factor target rotation based on an oblique rotated solution (biFAD; Waller, 2017) was compared. The aforementioned methods were tested via a Monte Carlo simulation that manipulated sample size, number of specific factors, number of indicators, and cross-loading size. Arbitrary cut-off points ranging from .05 to .20 were evaluated for SLi. According to congruence coefficients, SLiD presented the best performance for all conditions except under presence of large cross-loadings, where BiFAD performed the best. However, the latter was suboptimal when recovering large structures. SLiP outperformed SLi, but its performance was unsatisfactory. Lastly, higher cut-offs (i.e., .20) for SLi provided worse results than smaller ones (i.e., .05). SLiP and SLi accuracy was diminished in the presence of factors with low average factor loadings.

**Symposium 4 - Parallel Session: 2.2C: EXPLORATORY ROTATION WITH (BLOCKWISE) BIFACTOR SIMPLIMAX**

Marieke Timmerman, University of Groningen

Exploratory bifactor analysis aims at identifying general and specific sources of common variance underlying an item set. Of the various available orthogonal rotation criteria, the two best performing ones in relevant conditions seem to be the iterative target rotation based on the Schmid-Leiman solution (SLi; Abad, Garcia-Garzon, Garrido & Barrada, 2017) and bi-quartimin (Jennrich & Bentler, 2011). In the present paper, we propose two alternative criteria, related to the Simplimax rotation criterion (Kiers, 1994) which directly aims for low and high loadings at particular positions in the loading matrix: Bifactor Simplimax aims at high loadings on the general factor, and exactly one high loading per item on the specific factors. Blockwise bifactor Simplimax has the same goal, with the distinction that the criterion steers towards an a priori known division of items (e.g., among subscales), such that items of the same subscale are loading on the same specific factor. The comparative performance of Bifactor Simplimax, Blockwise bifactor Simplimax, SLi and bi-quartimin is assessed by means of a simulation study under relevant conditions. The usefulness of the different methods is illustrated by a means of an empirical example, considering the structure underlying the SCL-90, a self-report questionnaire to evaluate a broad range of psycho psychological problems and symptoms of psychopathology.

**Symposium 4 - Parallel Session: 2.2D: RETRIEVING RELEVANT COMPONENTS IN DYADIC INTERATIONS**

Marlies Vervloet, Katholieke Universiteit Leuven; Katrijn Van Deun, Tilburg University; Eva Ceulemans, Katholieke Universiteit

Human behavior, feelings and thoughts continuously fluctuate across time, especially when interacting with other people. Studies on this topic mostly focus on interactions in dyads (such as romantic couples), and try to disentangle partner-induced fluctuations of the behavior from the within-person fluctuations. Because the currently available methods fall short when multiple variables are being measured, we developed a new method for unraveling the interpersonal dynamics of dyads, named DCovR, which is an extension of principal covariates regression (De Jong & Kiers, 1992). DCovR handles multiple variables by reducing the variables of each person separately to a limited number of summarizing variables, called components. Simultaneously, we regress the components of both persons on one another to capture mutual influences. Thus, components are found that are not only good summarizers of the variables, but that are relevant for predicting the component scores of the dyadic partner as well. Similarly to PCovR, DCovR also comes with a weighting parameter, with which a user can determine the degree of emphasis on reduction versus prediction when extracting components. In the present study, we evaluate the performance of DCovR by means of a simulation study. We check whether an underlying DCovR solution can be recovered in several challenging conditions, and we examine the role of the weighting parameter in this regard.

**Symposium 4 - Parallel Session: 2.2E: FINDING THE HIDDEN LINK: SPARSE COMMON AND DISTINCTIVE COMPONENT ANALYSIS**
Katrijn Van Deun, Tilburg University; Niek de Schipper, Tilburg University

Recent technological advances have made it possible to study human behavior by linking novel types of data to more traditional types of psychological data, for example linking psychological questionnaire data with genetic risk scores. Revealing the variables that are linked throughout these traditional and novel types of data gives crucial insight in the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of depression. Little or no theory is available on the link between such traditional and novel types of data, the latter usually consisting of a huge number of variables. The challenge is to select - in an automated way - those variables that are linked throughout the different blocks and this eludes current available methods for data analysis. To fill the methodological gap, we present here an extension of simultaneous component analysis. Constraints are introduced to impose a common and distinctive structure and to force automated selection of the relevant variables. We will present an efficient procedure that is scalable to the setting of a very large number of variables. Using simulated data and an empirical example, we will showcase the benefits of the proposed method and compare with various competing methods, including sparse PCA and rotation techniques.

## Bayesian Statistical Inference: 1:30 PM – 3:00 PM

Chair: Steven Culpepper

**Bayesian Statistical Inference**

**Bayesian Statistical Inference - Parallel Session: 2.3A: DIRICHLET PRIORS FOR LATENT DIRICHLET ANALYSIS OF CONSTRUCTED RESPONSE ITEMS**
Minho Kwak, University of Georgia; Seohyun Kim, University of Georgia; Jiawei Xiong, University of Georgia; Hye-Jeong Choi, University of Georgia

Latent Dirichlet Analysis (LDA) uses a Dirichlet distribution to model the number of latent topics in a corpus. Recent work on LDA has focused on its use for analysis of the text examinees provide in their responses to constructed response (CR) items (Kim, Kwak, et al., 2017). Dirichlet priors reported in the literature, however, were described for documents constructed for unconstrained responses such as twitter messages, web logs, or abstracts of scientific journals (Blei, 2012). Responses to CR items are relatively constrained in that the words are intended to answer specific questions on a test. As a result, CR responses are constrained to focus on the answers to CR items. In addition, if time limits are imposed, there is an additional constraint on the amount of time available to respond. As a result, the same Dirichlet priors used for analysis of less constrained kinds of writing may not be appropriate for analysis of CR items (Kwak, Kim & Cohen, 2017). Kwak et al. suggests that the number of latent topics may be relatively small compared to larger unconstrained kinds of documents. Non-informative priors have been suggested when the number of latent topics is small (Seo & Fokone, 2017). The use of non-informative priors, however, may be inappropriate depending on the purposes for which the texts were created (Contractor, Singla & Singla, 2016). The purpose of this study is to investigate Dirichlet priors for LDA for their usefulness in detecting latent topics in a corpus of documents of responses to CR items.

**Bayesian Statistical Inference - Parallel Session: 2.3B: ELABORATING ON ISSUES WITH BAYES FACTORS**
Jorge Tendeiro, University of Groningen; Henk Kiers, University of Groningen

Problems with frequentist statistics in general and null hypothesis significance testing in particular are a central issue in Psychology. Recent results related to the lack of reproducibility of statistical findings have urged researchers to change their way of doing science. One proposal that has received considerable attention in the literature is that of replacing p-values with Bayes factors. Bayes factors are often praised for offering a rational means of assessing evidence between two competing hypotheses or models. However, even though there are indeed reasons to prefer Bayes factors over p-values, Bayes factors are affected by a set of issues of their own. In this talk, we will highlight several standing issues with Bayes

factors, which include: Sensitivity to the choice of priors, lack of coherence with parameter estimation, lack of proper justification for the so-called default priors, questionable support towards (point) null hypothesis, and theoretical limitations of marginal likelihoods (on which Bayes factors are based). Our goal is two-fold: Further clarify what one can (and cannot) expect from Bayes factors and discuss alternative analytic approaches.

**Bayesian Statistical Inference - Parallel Session: 2.3C: BAYESIAN MODEL ASSESSMENT: USE OF CONDITIONAL VS MARGINAL LIKELIHOODS**

Edgar Merkle, University of Missouri; Daniel Furr, University of California, Berkeley; Sophia Rabe-Hesketh, University of California, Berkeley

Typical Bayesian methods for models with latent variables (or random effects) involve directly sampling the latent variables along with the model parameters. In high-level software code for model definitions (using, e.g., BUGS, JAGS, Stan), the likelihood is therefore specified as conditional on the latent variables. This can lead researchers to perform model comparisons via conditional likelihoods, where the latent variables are considered model parameters. In other settings, typical model comparisons involve marginal likelihoods where the latent variables are integrated out. This distinction is often overlooked despite the fact that it can have a large impact on the comparisons of interest. In this paper, we clarify and illustrate these issues, focusing on the comparison of conditional and marginal Deviance Information Criteria (DICs) and Watanabe-Akaike Information Criteria (WAICs) in psychometric modeling. The conditional/marginal distinction corresponds to whether the model should be predictive for the clusters that are in the data or for new clusters (where "clusters" typically correspond to higher-level units like people or schools). Correspondingly, we show that marginal WAIC corresponds to leave-one-cluster out (LOcO) cross-validation, whereas conditional WAIC corresponds to leave-one-unit (LOuO). These results lead to recommendations on the general application of these criteria to models with latent variables.

**Bayesian Statistical Inference - Parallel Session: 2.3D: USING PRIOR INFORMATION FOR MORE EFFICIENT NORM ESTIMATION**

Lieke Voncken, University of Groningen; Thomas Kneib, University of Göttingen; Casper Albers, University of Groningen; Marieke Timmerman, University of Groningen

A huge advantage of continuous norming with the Generalized Additive Models for Location, Scale, and Shape (GAMLSS; Rigby & Stasinopoulos, 2005) framework is that the characteristics of the raw score distribution can be estimated as a function of covariates (e.g., age). This means that you do not have to rely on the strict, unrealistic assumption that the test scores, conditional on the covariates, follow a normal distribution. The main downside of this flexibility is that a large normative sample is required to estimate the relationship between the distribution characteristics and covariates properly. That is why we propose to apply a Bayesian version of GAMLSS in the context of continuous norming. By using prior information in the creation of new norms, we hope to make norm estimation more efficient, which means that a smaller normative sample is required to estimate the norms with the same precision. We will investigate how and to what extent previously estimated norming models can be used as prior in the norm estimation of new tests. In a simulation study, we will investigate three issues. First, we will investigate the effect of chosen prior on the resulting estimated norms in a sensitivity analysis. Second, we will examine the quality of diagnostics that we will develop to determine how similar the existing and new normative samples are. Third, we will investigate how large the new normative sample minimally needs to be to obtain the minimum desired level of norm precision. The findings and implications will be discussed.

## Classification, Clustering, and Latent Class Analysis: 1:30 PM – 3:00 PM

Chair: Dylan Molenaar

**Classification, Clustering, and Latent Class Analysis**

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 2.4A: A GENERAL NONPARAMETRIC CD-CAT ALGORITHM FOR SMALL EDUCATIONAL PROGRAMS**
Yuan-Pei Chang, Rutgers, The State University of New Jersey; Chia-Yi Chiu, Rutgers, The State University of New Jersey

The computerized adaptive testing for cognitive diagnosis (CD-CAT) is aimed to assign examinees to the proficiency classes they belong to using individualized test items. Several parametric item selection methods for CD-CAT have been developed and in these algorithms, the best items are usually selected by optimizing different types of objective functions or taking into account the variability among latent classes. Although research has shown that these parametric methods perform well within the context of large-scale assessment, they require large samples to guarantee the reliable calibration of item parameters and accurate estimation of examinees' proficiency class membership. Such samples are simply not obtainable in small educational settings. In this study, a general nonparametric item selection (GNPS) method is proposed to overcome the obstacle. The proposed method uses the general nonparametric classification (GNPC; Chiu, Sun & Bian, Psychometrika, in press) method to classify examinees and for each examinee, the item with the highest discrimination ability is identified as the best item. The GNPS method can be used for data conforming to the general CDMs and therefore the models subsumed under them. Additionally, no model fitting is required for the proposed method, which makes it a superior approach for small samples. The simulations show that the proposed GNPS algorithm is more effective and efficient than some recently developed parametric methods when samples are small.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 2.4B: OVERLAPPING CLUSTERING: A FRAMEWORK, SOFTWARE, AND EMPIRICAL ANALYSIS**
Stephen France, Mississippi State University

Overlapping cluster analysis is a variant of cluster analysis where each item may be a member of more than one cluster. It has found particular use in marketing segmentation, where products may be members of more than one usage segment. Overlapping clustering methods have been developed from different clustering traditions. Additive decomposition methods, such as ADCLUS and INDCLUS, are discrete variants of continuous mapping methods. Fuzzy clustering methods can generate overlapping clustering solutions by setting thresholds for cluster membership. Partitioning clustering methods, such as k-means clustering, can be extended to overlapping clustering, by relaxing the cluster membership constraints. The R software package and associated framework described in this talk implements overlapping clustering methods from all of these traditions and also implements the generalized omega metric for cluster validation. Empirical work on both synthetic and real world datasets is described.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 2.4C: COMPARING CLASSIFICATION METHODS FOR RESPONSE DATA AND PROCESS DATA**
Xin Qiao, University of Maryland College Park; Hong Jiao, University of Maryland, College Park

Psychometric models for classification purposes, such as cognitive diagnostic models (CDMs) and Bayesian networks, have been developed rapidly nowadays with the development of complex assessment forms, such as simulation and game-based assessments. Data mining and machine learning techniques have also been utilized to analyze the process data collected from such assessments (He & von Davier, 2016). The classification accuracy of these statistical methods became a challenging research topic in the measurement field. Researchers have compared classification accuracy rates between CDM and support vector machine (SVM; Liu & Cheng, 2017) purely based on item responses. Process data, such as response time, has been shown to improve classification accuracy when integrated with regular response data in cognitive diagnostic models (Zhan et al., 2017). However, no study compared the performance of psychometric models and data mining methods when both response data and process data are incorporated. To fill this gap, current study conducts a simulation study to investigate the

classification accuracy of a hybrid Bayesian network model, JRT-DINA (Zhan et al., 2017) and SVM when the data consists of both regular response data and response time. This study evaluates the classification accuracy rates of the three methods under various simulation conditions. Discussions on the choice of these methods will be provided.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 2.4D: DETERMINE ATTRIBUTE PROFILE FOR DIAGNOSTIC COGNITIVE MEASUREMENT USING GENERATOR-EVALUATOR-NETWORK (GEN)**
Kang Xue, University of Georgia

Diagnostic Cognitive Models (DCMs) are the methods to diagnose students' attribute status (mastery or non-mastery) using their responses to the items which is directly designed. Most widely used DCMs (e.g. DINA, DINO, rRUM, GDM and LCDM) require specific mathematic relationships between responses and latent attribute profile and use responses to estimate both item and examinee parameters. In contrast, a novel generator-evaluator-network (GEN) is introduced in this framework to determine students' latent attribute profiles without specifying the diagnostic model. To achieve the latent attribute, the GEN consists two parts: 1) simple-structured-item based generative network (GN); 2) complex-structure-item based evaluation network (EN). The aim of GN is to map students' responses to latent attribute space just using simple structured items in one assessment, and EN is to evaluate the efficiency and accuracy of GN through the rest complex structured items. The contributions of the new method are three-folded: first, GEN is built based on general latent class theory and is not constricted by any specific models; second, since GN only uses simple structured items, the network fitting process could be more algorithmic efficient; third, unlike the traditional dimensional reduction methods, the GEN transfers the latent mapping process from unsupervised learning to supervised learning which could have more options for optimization.

## Item Response Theory: 1:30 PM – 3:00 PM

Chair: Alina A. von Davier

**Item Response Theory**

**Item Response Theory - Parallel Session: 2.5A: MODIFYING THE PROJECTIVE IRT MODEL TO REPRESENT A LATENT ABILITY COMPOSITE**
Terry Ackerman, University of Iowa; Edward Ip, Wake Forest School of Medicine; Shyh-Huei Chen, Wake Forest University

The projective IRT (PIRT) model (Ip, 2010) is designed to remove unwanted dimensionality from parameter estimation and to purify the interpretation of the reported score scale. That is, Ip (2010) and Ip and Chen (2012) provided the formulation of the local dependent PIRT model that is empirically indistinguishable from the multidimensional model. In their formulation they presented a two-dimensional scenario, in which the second latent ability being measured, $\theta_2$ represented a nuisance dimension and that practitioners may want to remove this invalid noise from the estimation process. By integrating out the nuisance dimension they demonstrated that the dependent PIRT model was empirically indistinguishable from two-dimensional IRT compensatory model. However, typically dimensionality in response data arises from the interaction of examinees with targeted content and cognitive levels of the test specifications. It many cases practitioners may not want to lose this richness and may prefer to identify a scale for their assessment that represents a composite of the manifest dimensionality. The purpose of this research is to create a composite PIRT (C-PIRT) model to allow testing practitioners to select a composite direction in the multidimensional latent space and obtain unidimensional item and ability parameter estimates. This research is part of an ongoing IES study. Results will include examples from real test data (e.g., PISA, ACT) and simulated data. Practical advantages of using the C-PIRT model will be discussed, including such issues as consistent interpretation of the score scale when difficulty and dimensionality are confounded and the mitigation of differential item functioning.

**Item Response Theory - Parallel Session: 2.5B: LATENT VARIABLE FREE MULTIDIMENSIONAL ITEM RESPONSE MODELS**

Mia Müller-Platz, RWTH Aachen University; Maria Kateri, RWTH Aachen University; Irini Moustaki, London School of Economics and Political Science

Large scale and complex data sets such as those collected in education or in household surveys require multidimensional factor models. In general, models with many factors and random effects can become infeasible to estimate in practice. We propose an alternative flexible way of analyzing multivariate responses that deals with the curse of dimensionality in factor models by modelling instead the conditional expectation of single items on other items responses and explanatory variables. This model formulation leads to composite likelihood estimation, which is feasible even for complex models. We establish the link of the new defined conditional specified models to standard multifactor item response models. The connection enables the estimation of latent variable response models from the composite likelihood estimates of the corresponding conditional specified response models. Simulation and empirical results will be presented.

**Item Response Theory - Parallel Session: 2.5C: A BAYESIAN MULTILEVEL TRIFACTOR IRT MODEL FOR SMALL SAMPLES**

Ken Fujimoto, Loyola University Chicago

Many research studies undertaken in education and psychology involve much smaller sample sizes (e.g., fewer than 500 participants) than those seen in large-scale studies. Yet, these smaller studies often use the same sampling and measurement processes that are used in large-scale studies, leading to complex multilevel multidimensional structures being represented in the item response data. Unfortunately, many of the item response theory (IRT) models that account for such structure (e.g., dual-dependent IRT models) have not been devised for use with small sample sizes. For this presentation, a Bayesian multilevel trifactor IRT model for small samples is introduced. This model incorporates prior distributions that allow complex multilevel multidimensional structures to be studied in small data sets without unduly influencing the results, yet the same priors can be used on larger samples. Also, the results of a simulation study examining the effectiveness of this model in capturing a data structure in which seven dimensions exist at the person level (layered in three tiers) and one dimension exists at the cluster level are reported. The findings show that, in sample sizes as small as 300 cases total (from 60 clusters), the model: (a) can accurately describe the cluster effects represented in the data (via the intraclass correlation); (b) can recover the model parameters; and (c) outperforms competing models (e.g., the multilevel cross-classified model and the bifactor model) with respect to predictive performance of the data. The presented model is also illustrated on empirical data from 317 persons nested within 123 family units.

**Item Response Theory - Parallel Session: 2.5D: COMPARING APPROACHES TO MODELING ZERO INFLATION IN OPEN-ENDED FREQUENCY DATA**

Brooke Magnus, Marquette University

This research compares three different methods of modeling zero inflation in open-ended symptom frequency data from the Patient Health Questionnaire, while also accounting for the digit preference that often manifests in self-report count data. The first method uses a single latent class to describe individuals who are "non-pathological" and do not endorse any of the symptoms on the questionnaire (Magnus & Thissen, 2017; Wall, Park, & Moustaki, 2015). The inclusion of a non-pathological class accounts for questionnaire-level zero inflation; however, it does not allow zero inflation to vary across items. This may be overly restrictive on psychopathology questionnaires comprising items of differing severity. For example, an item about suicide ideation is likely to exhibit a higher degree of zero inflation than an item about low energy levels, regardless of the size of the non-pathological class. An alternative method allows for zero inflation to vary across items by using a zero-inflated negative binomial (ZINB) IRT model, in which two sets of parameters are estimated: One models a zero process, the other models a count process (similar to Wang's (2010) IRT-ZIP model). The method proposed here combines these two approaches by including a non-pathological class to account for zero inflation at the questionnaire level while using ZINB IRT models to also capture item-level zero inflation that may not arise from a non-pathological class. Results suggest that accounting for zero inflation at the item level yields better fit and reduces the estimated size of the non-pathological class. Model interpretation is the focus.

**Item Response Theory - Parallel Session: 2.5E: A DYADIC ITEM RESPONSE THEORY MODEL**
Brian Gin, University of California, San Francisco; Nicholas Sim, University of California, Berkeley; Anders Skrondal, Norwegian Institute of Public Health & University of Oslo & University of California, Berkeley; Sophia Rabe-Hesketh, University of California, Berkeley

We propose a Dyadic Item Response Theory (dIRT) model for measuring interactions of pairs of individuals when the responses to items represent the actions (or behaviors, perceptions, etc.) of each individual (actor) made within the context of a dyad formed with another individual (partner). Examples of its use include the assessment of collaborative problem solving, or the evaluation of intra-team dynamics. The dIRT model generalizes both the Partial Credit Model, as well as the Social Relations Model developed by Kenny and colleagues. The responses of an actor when paired with a partner are modeled as a function of not only the actor's inclination to act and the partner's tendency to elicit that action, but also the unique relationship of the pair, represented by two directional, possibly correlated, interaction latent variables. Extensions of the dIRT model are discussed, such as accommodating triads or larger groups. Estimation of the dIRT model is performed using Markov-Chain Monte Carlo implemented in Stan, making it straightforward to extend the model in various ways. A simulation study shows that estimation performs well. We apply our approach to speed-dating data and find new evidence of pairwise interactions between participants, describing a mutual attraction that is inadequately characterized by individual properties alone. We show how the dIRT model can be extended to allow joint modeling of dyadic data and a "distal outcome" (of the pair mutually electing to date) by including a logistic regression of the distal outcome on the pairwise interaction and other latent variables.

Chair: Eric Schuler

**Item Response Theory**

**Item Response Theory - Parallel Session: 2.6A: MISSING VALUE IMPUTATION FOR BAYESIAN IRT MODEL**
Tianyang Zhang

The presence of missing data poses threats to the validity and reliability of IRT result. Evaluating effective ways of handling missing data, current simulations have only been done in traditional IRT framework. No simulation studies, however, have been done under Bayesian IRT framework. Rubin (1976) developed a new taxonomy based on the missingness nature, which are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The effects of different treatments for different missingness mechanisms are significantly different. This study looks into the effects of different parametric and nonparametric missing data treatments on the estimation of 3PL Bayesian IRT parameters under the large-scale educational assessment context for all missing scenarios. We simulated 600 datasets based on TIMSS 2015 Math. Each dataset contains 88 items and 1000 observations. Missing data were generated by three missing scenarios with 10%-50% missingness. Six methods were employed to deal with the missingness, including three nonparametric methods (ignore, incorrect, and correct) and three parametric methods (corrected item mean substitution, multiple imputation, and EM algorithm). Our simulation results indicated that, compared with the parametric methodologies, the nonparametric methods appear problematic in various ways. Rose, Davier, and Xu (2010) explained that "that procedure ignores the stochastic relation between the latent proficiency variable and the manifest item response" (p.4). Multiple imputation outperformed other methods in dealing with missingness with respect to the minimization of the mean square error and average absolute deviance.

**Item Response Theory - Parallel Session: 2.6B: T-TEST AND ANOVA FOR DATA WITH CEILING AND FLOOR EFFECTS**
Qimin Liu, University of Notre Dame; Lijuan Wang, University of Notre Dame

Psychology research commonly encounters ceiling/floor effects in the data. The current study examines the impact of and the methods for floor/ceiling effects in t test and ANOVA, two frequently used statistical methods. Based on our literature review, some researchers treated ceiling or floor data as if these data were true values while some other researchers simply discarded ceiling or floor data in conducting t-test and ANOVA. The present study examines the type I errors, statistical power, coverage probabilities, and effect size estimates from these conventional methods for t-test and ANOVA with

ceiling or floor data. In addition, censored regression was repurposed and investigated for its potential capacity of handling ceiling or floor data. Furthermore, novel methods that use properties of truncated normal distributions were proposed and evaluated for handling ceiling or floor data in t-test and ANOVA. Simulation studies were conducted to compare the performance of the methods. Overall, the proposed methods showed greater accuracy in effect size estimation and better controlled Type I error rates over other evaluated methods for t-test and ANOVA with ceiling/floor data. Discussion and future directions are also included.

**Item Response Theory - Parallel Session: 2.6C: MODELING NONIGNORABLE MISSING FOR NOT-REACHED ITEMS INCORPORATING ITEM RESPONSE TIMES**

Jing Lu, Northeast Normal University; Chun Wang, University of Minnesota; Jian Tao, Northeast Normal University

In educational and psychological assessments, tests are often administered with a fixed time limit, in which examinees always not reach the items at the end of the test. Glas and Pimentel (2008) proposed a modified Rasch model for not reached responses, however, their model only used item position as an indicator to explore the probability of response indicator. In practice, the not-reached items are not only related with item position, but more important rely on the elapsed time and remaining time of the test. In this study, we propose a multidimensional item response theory model to analyze the nonignorable missing data, specifically, one dimension for observed responses, one dimension for response times, and the other dimension for the missing data indicator. The missing data indicator is modeled using the response time information. Meanwhile, the correlations among ability parameter, speed parameter, and propensity of response can be investigated. The performance of the proposed model is demonstrated by a simulation study.

**Item Response Theory - Parallel Session: 2.6D: SEM APPROACH TO CCA WITH MISSING VALUES?**

Laura Lu, UGA; Fei Gu, University of Kansas

Canonical correlation analysis (CCA) is a generalization of multiple correlation that examines the relationship between two sets of variables. Traditional methods apply spectral decomposition to obtain canonical correlations and canonical weights. Anderson (2003) also provided the asymptotic distribution of the canonical weights under normality assumption. Lu and Gu (2016) proposed a new approach by using structural equation modeling (SEM) to canonical correlation analysis. Mathematical forms are presented to show the equivalence among these models. It is very flexible because it provides both canonical correlations and the covariances of canonical variates by using existing SEM software. However, Lu and Gu (2016) only focused on complete data and did not investigate the cases with missing values. In this article, we consider data sets with missing values under different mechanisms, missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Regarding estimation methods, maximum likelihood estimation are applied to obtain canonical correlations and weights. Bayesian estimation methods are also tried by employing Gibbs sampling and multiple imputations. Popular SEM software such as Mplus, EQS, CALIS, Lavaan and sem packages in R are demonstrated to illustrate the application. Related issues such as standard errors and comparison of estimation methods are discussed in the last section. Regarding these issues, it may require researchers writing their own computer programs.

**Item Response Theory - Parallel Session: 2.6E: PLANNED MISSING DATA APPROACH FOR MEASURE DEVELOPMENT: REDUCING PARTICIPANT BURDEN**

Eric Schuler, Uniformed Services University of the Health Sciences; Josh B. Kazman, Uniformed Services University of the Health Sciences; Kathleen G. Charters, Uniformed Services University of the Health Sciences; Stephen W. Krauss, Uniformed Services University of the Health Sciences; Patricia A. Deuster, Uniformed Services University of the Health Sciences

When developing and validating a new measure, it is necessary to start with a large item pool and include measures for construct validity (Clark & Watson, 1995). However, this approach could lead to lengthy surveys that create a high-level of burden for participants. Lengthy surveys can also degrade the quality of the data itself (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Kam & Meyer, 2015; Meade & Craig, 2012). One method to reduce participant burden is to reduce the number of items each person receives -

called planned missing data designs (PMD- Enders, 2010; Graham, Taylor, Olchowski, & Cumsille, 2006). By using PMD, no single participant is burdened with answering every test item. This results in a complete correlational matrix across participants, which can be used for subsequent analyses, such as exploratory factor analysis. In our talk, we will outline how to implement a PMD approach to measurement development and how to conduct a power analysis in the R statistical environment. We will discuss the decision-making process key factors for determining the appropriate methodology. These key factors include: (1) the number of items in the test item pool; and (2) time spent on the survey (as a proxy for participant burden). We will compare traditional measurement development strategies with a PMD approach on participant burden and the potential sample sizes needed for each. The talk will conclude with a discussion of the strengths and limitations of the PMD methodology.

## Validity and Reliability: 1:30 PM – 3:00 PM

Chair: Jelte Wicherts

**Validity and Reliability**

**Validity and Reliability - Parallel Session: 2.7A: REPRODUCIBILITY: THE ELEPHANT IN THE ROOM?**
Richard Artner, Katholieke Universiteit, Leuven

The replication crisis is one of the most well-researched topics of psychology in the last couple of years - and rightly so. An enormous amount of literature has been written about p-hacking, HARKing, publication bias, significance testing, replications, pre-registration, researchers degrees of freedom and consorts, however, an even more basic property of scientific research – reproducibility – has gotten less amount of attention. Given the raw data of a study, as well as, information about the fitted models, anyone should be able to come up with the same model parameter/coefficient estimations, test statistics and p-values. A failure to reproduce can be traced back to human error, since, in contrast to replications which necessarily deal with new data values and/or new data analysis methods, data handling and model fitting are the only sources of stochasticity. We investigate the reproducibility of over 40 articles published in 2012 in three different psychological journals. First, we identify the main results of each article, defined as a priori hypotheses of primary interest that are mentioned in the abstract, amounting to approximately 200 main results in total. Next, starting from the raw data, we closely follow the method section of those articles to aim to recalculate the main statistics of each main result. This talk will summarize the results and discuss the encountered difficulties, as well as, recommendations for improving reproducibility.

**Validity and Reliability - Parallel Session: 2.7B: PREREGISTRATION: COMPARING DREAM TO REALITY**
Aline Claesen, Katholieke Universiteit, Leuven

In the light of the replicability crisis, preregistration is one suggested method to gain more confidence in psychological research, since it freezes researcher degrees of freedom. Several journals implemented badges provided by the Open Science Framework, among which a preregistration badge, rewarding authors for open research practices. As a result, preregistered studies might inspire more confidence than studies without preregistration. However, this increased confidence is only warranted if the preregistration is accessible, has at least minimal amount of quality and if the article adheres to it. To investigate whether this increased confidence is just, we examined 23 articles with a preregistration badge, published in a high-impact journal between February 2015 and November 2017, using three consecutive criteria. First, we evaluated whether they provided a permanent, read-only, time-stamped document, which is publicly and non-ambiguously accessible on a third-party repository. If this was the case, we inspected whether a minimal amount of useful information was present. Finally, if both criteria were met, we assessed the adherence of the article to the preregistration. Results and implications will be discussed, which will hopefully guide us towards what can be considered as good or bad practices regarding preregistration.

**Validity and Reliability - Parallel Session: 2.7C: DOES LOW RELIABILITY LEAD TO A LOW REPLICATION RATE OF STUDY RESULTS?**
Shuo (Selena) Wang, Ohio State University

The replication crisis has recently incited the discussion on measurement quality. Although it may be evident that unreliable measurements hamper observation precision, the consequences of a low reliability coefficient are not yet well understood. In this study, we attempt to understand the reliability coefficient through its mathematical relationship with experimental effects. A multiple-group Structural Equation Model (SEM) was constructed to imitate an experiment allowing us to investigate reliability and experimental effects more closely. Both the reliability coefficient and effect size were computed from the model. The formulae show that reliability and Cohen's d move in the same direction upon the manipulation of the factor loadings and the residual variances of the manifest variable; yet move in the opposite direction upon the manipulation the (residual) variances of the latent variables. Well-fitting models (minor residual variances and sizable factor loadings) help us achieve favorable reliability and strong effect sizes. The negative relationship between effect size and the reliability coefficient suggests that pursuit of high reliability through increased sample variance will cause Cohen's d to decline given the same model. The cause of this counterintuitive finding is indicated by the Classical Test Theory assumption of independence between the true score and error. In future research, we seek to relax this assumption. Additionally, we suggest the use of reliability that is reflective of the experimental effect, calculated from meta-analysis, rather than the reliability calculated from the consistency of individual differences in individual study reporting.

## Symposium 5: 3:20 PM – 4:50 PM

Chair: Laura Bringmann

**Symposium 5: Modeling Intensive Longitudinal Data: Perks and Pitfalls**

**Symposium 5 - Parallel Session: 3.1A: INTERVENTIONISM AND WITHIN-PERSON CAUSES IN PSYCHOLOGY**
Markus Eronen, University of Groningen

A central goal in psychology is to find causal relationships that describe processes in individual persons. However, the gold standard for finding causes is randomized controlled trials, which result in population-level (between-person) causal relationships, and it is not clear to what extent and under which conditions they apply to individuals in the population. In this talk, I tackle this question from the perspective of the interventionist theory of causation (developed by, e.g., Judea Pearl and James Woodward). First, I will argue that extending population-level causal findings to individuals requires assuming causal homogeneity, that is, that all the individuals in the population have the same causal structure. However, due to the fact that individual differences are common in psychology, this assumption is unlikely to hold in most situations. Second, I will argue that making causal inferences directly at the individual level requires other strong assumptions, such as invariance (modularity) and causal sufficiency. Thus, finding within-person causes in psychology faces formidable theoretical and conceptual problems that need to addressed before methods of causal inference can be reliably applied to individuals.

**Symposium 5 - Parallel Session: 3.1B: SPECIFYING EXOGENIETY AND BILINEAR EFFECTS IN DATA DRIVEN MODEL SEARCHES**
Cara Arizmendi, University of North Carolina, Chapel Hill; Kathleen Gates, University of North Carolina, Chapel Hill; Aidan Wright, University of Pittsburgh; Barbara Fredrickson, University of North Carolina, Chapel Hill

In data driven model searches, all variables typically have the opportunity to be either exogenous and endogenous. Few methods allow for specification of exogenous variables prior to multivariate model searches. Allowing for specification of exogenous variables permits more realistic models (e.g. weather can predict emotions, but emotions cannot predict weather) and allows for the modeling of contextual change processes (e.g. beginning treatment versus no longer receiving treatment). We will explore the capabilities of allowing for the specification of exogenous variables in GIMME (Group Iterative Multiple

Model Estimation), a model search algorithm that allows for the modeling of idiographic individual level processes, as well as subgroup level and group level processes with intensive longitudinal data, using daily diary data in three examples. 1. Using data collected from individuals diagnosed with personality disorders, we show results where weather-related variables are specified as exogenous, and reports on affect and behavior are allowed to be both exogenous and endogenous. 2. We demonstrate the modeling of treatment effects in an intervention study looking at the effects of a 6-week meditation workshop on health behaviors in midlife adults. 3. Finally, we use the meditation intervention data to demonstrate modeling bilinear effects, where relationships between two endogenous variables are dependent on the current stage of the study for a given participant (i.e., currently attending meditation classes or not).

**Symposium 5 - Parallel Session: 3.1C: MULTILEVEL DYNAMIC TWIN MODELING**
Noémi Schuurman, Tilburg University

Recent developments in intensive longitudinal data collection and modeling have made it possible to fit dynamic statistical twin models. In these models the genetic and environmental components are dynamic processes, that vary over time within twins according to autoregressive process. The most well known dynamic twin model is arguably the simplex model, where the intraindividual differences of the components are modeled across twins on a population level. A more recent developed model is the Iface model, in which a similar model is fitted on the twin-pair level, based on more intensive longitudinal data for a single twin pair. We will discuss the missing link between these models - a multilevel extension of these models that allows for making both population level inferences, as well as twin pair level inferences. How to exactly interpret the intraindividual genetic and environmental components of these dynamic twin models, as well as the parameters of these models, is however non-trivial. We will examine the interpretation of these models, when gradually adding more random parameters (twin pair specific parameters) to the model.

**Symposium 5 - Parallel Session: 3.1D: NETWORK MODELS FOR CLINICAL PRACTICE?**
Laura Bringmann, University of Groningen

In an effort to bridge the gap between clinical research and practice, more and more clinicians and researchers are using some form of experience sampling method to study, for example, individuals with depression in real time. These kinds of intensive longitudinal data give a wealth of information, but also lead to many new methodological challenges. In the recently popular network approach, the focus is on the dynamics between symptoms: Studying networks of causally interacting symptoms will give information on which symptom one should intervene. However, others have argued that the network models used can get overly complicated and that the focus should be more on the symptom scores themselves. On this approach, the important symptoms are those with a high average score, and they can be found through, for example, visualisation techniques. In this talk, I will present and discuss both approaches. On the one hand, I will explain the newest kind of dynamic model (the time-varying vector autoregressive model) for inferring dynamical networks, and on the other hand, I will present new visualisation techniques that can highlight the importance of items in research and clinical practice. Additionally, I will discuss advantages and disadvantages of both sides, and discuss new ways of thinking about the relative importance of symptoms for interventions.

**Symposium 5 - Parallel Session: 3.1E: DIMENSION-REDUCTION YIELDS IMPROVED INSIGHT INTO AND PREDICTION OF NETWORK DYNAMICS**
Eva Ceulemans, Katholieke Universiteit; Kirsten Bulteel, Katholieke Universiteit Leuven; Annette Brose, Humboldt University Berlin; Francis Tuerlinckx, Katholieke Universiteit

To understand within-person psychological processes, a multivariate lag-one vector autoregressive (VAR(1)) model is often fitted to time series and the VAR(1) coefficients are displayed as a network. However, this approach is not without problems. First, we often expect substantial contemporaneous correlations between the variables, yielding multicollinearity issues. Moreover, as VAR(1) coefficients only capture unique direct effects, the potentially large shared effects are not included in the network. As a consequence, VAR(1) networks may be hard to interpret. Second, the highly parameterized VAR(1) model is prone to overfitting. Therefore, the predictive accuracy will be relatively low, as can be revealed

through cross-validation. To handle both problems, one may recur to penalized regression methods, such as VAR(1) with a lasso penalty. In this paper, we recommend a different solution, called principal component VAR(1) (PC-VAR(1)), in which the variables are first reduced to a few components; next the VAR(1) analysis is applied to the components. Reanalyzing data of a single participant of the COGITO study, we show that PC-VAR(1) has the better predictive performance, and that networks based on PC-VAR(1) give a more informative representation of both the lagged and the contemporaneous relations among the variables.

## Symposium 6: 3:20 PM – 4:50 PM

Chair: Willem J. Heiser; Antonio D'Ambrosio

**Symposium 6: Individual Differences in Rankings: Aggregation, Representation, Evolution & Prediction**

**Symposium 6 - Parallel Session: 3.2A: WEIGHTED AGGREGATION OF ORDINAL AND CARDINAL DATA**
Josè Luis Garcìa-Lapresta, University of Valladolid; Casilda Lasso De La Vega, Universidad del País Vasco (University of the Basque Country)

In some real problems analyzed in Welfare Economics, Sociology, Psychology, Marketing, Psychological Measurement, and other fields of research, some alternatives are evaluated in several criteria (or dimensions) with the purpose of rank order these alternatives. Alternatives can be countries, regions, products, services, or persons. The criteria on which the alternatives are evaluated can be of different nature. Some criteria may correspond to dichotomous (or binary) variables, as access to electricity, water or Internet, migration status, etc. Other criteria are assessed by linguistic terms of qualitative (or categorical or ordinal) scales, as health status, personal security, environmental quality, educational attainments, etc. In addition, others can be measured through numbers of a quantitative scale, as income, Gross National Product, unemployment rate, life expectancy, literacy rate, etc. Since criteria usually have different importance, frequently they have associated numerical weights. In this contribution some procedures that generate a reciprocal preference relation on the set of alternatives for each criterion are introduced. These kinds of preference relations capture the intensity of preference between each pair of alternatives. Since every weighted average of reciprocal preference relations is also a reciprocal preference relation, overall intensities of preference between pairs of alternatives are obtained, taking into account the weights assigned to the criteria. Then, for each fixed threshold, a ranking on the set of alternatives is attained.

**Symposium 6 - Parallel Session: 3.2B: ORDINAL UNFOLDING OF PREFERENCE RANKINGS USING THE KEMENY DISTANCE**
Antonio D'Ambrosio, University of Naples Federico II; Willem J. Heiser, Leiden University

Multidimensional unfolding can be seen as a special case of Multidimensional Scaling (MDS) with two sets of points, in which the within sets proximities are missing. Lack of the within-sets information is the major cause of well-known problems of degenerate solutions in analytical procedures, especially for ordinal (or non-metric) unfolding. Over the years, several approaches to avoid degenerate solutions in unfolding have been developed. Our approach belongs to the category of methods that aim to extend the unfolding data with information on the dissimilarities between rankings (Van Deun et al, 2007). By starting from a typical rectangular matrix, in which each row is a preference ranking, we propose a reconstruction strategy of the entire dissimilarity matrix based on the properties of the Kemeny distance (Kemeny and Snell, 1962) and the τ_X extended rank correlation coefficient (Emond and Mason, 2002). We show that our unfolding procedure can be used with any standard MDS program and produces non-degenerate solutions. These solutions are simple to compute, while comparable in quality with the ones returned by the PREFSCAL algorithm (Busing, Groenen and Heiser, 2005), currently the state-of-the-art method for avoiding degeneracies in unfolding.

**Symposium 6 - Parallel Session: 3.2C: RANKINGS VARYING IN TIME: A BAYESIAN APPROACH**
Elja Aryas, University of Helsinki

Rankings of individual persons, when based on repeated measurement of their performance, will generally vary in time. Moreover, some persons can be absent from the tests that are made either because of delayed entry into the study or early departure, or purely randomly. In this talk a Bayesian version of the well-known Mallows model for rank data is introduced, and then extended to situations in which the rankings can vary in time and be incomplete. The consequent missing data problems are handled by applying, within the considered MCMC, Bayesian data augmentation. The method is illustrated by analyzing some aspects of a data set describing the academic performance, measured by a series of tests, of a class of high school students over a period of four years.

**Symposium 6 - Parallel Session: 3.2D: DISTANCE-BASED DECISION TREES FOR RANKING DATA: THE ROLE OF THE WEIGHT SYSTEMS**
Mariangela Sciandra, University of Palermo; Antonella Plaia, University of Palermo

In everyday life ranking and classification are basic cognitive skills that people use in order to grade everything that they experience. Grouping and ordering a set of elements is considered easy and communicative; thus, rankings of sport-teams, universities, countries and so on are often observed. A particular case of ranking data is represented by preference data, where individuals show their preferences over a set of items. When individuals specific characteristics are available, an important issue concerns the identification of the profiles of respondents (or judges) giving the same/similar rankings. In order to incorporate respondent-specific covariates distance-based decision tree models (D'Ambrosio 2007, Lee and Yu 2010, Yu et al. 2010, D'Ambrosio and Heiser, 2016, Plaia and Sciandra, 2017) have been recently proposed. Actually, it can happen that one or some of the k items is more important than others, or, similarly, the top of the ordering can deserve more attention than the bottom. In these situations, changing the rank of very important items or changing the top of the ranking require different "weighting". In this contribution we want analyze the role of element and positional information (Kumar and Vassilvitskii 2010) when some distance measures for rankings are evaluated. Several weighting structures will be assumed for both positional and item weights, and we aim at identifying some particular behavior in the distance measures used. Analysis will be carried out both by simulation and by application to real dataset, especially in the framework of tree-based methods for rank data.

## Causal Inference and Mediation: 3:20 PM – 4:50 PM

Chair: Joost Van Ginkel

**Causal Inference and Mediation**

**Causal Inference and Mediation - Parallel Session: 3.3A: LOWER-LEVEL MEDIATION IN BINARY SETTINGS**
Haeike Josephy, Ghent University; Tom Loeys, Ghent University

In recent literature, researchers have put a lot of time and effort in expanding mediation to multilevel designs. Unfortunately, such extensions are often limited a continuous mediator and outcome, whereas research concerning multilevel mediation with a binary mediator and outcome remains rather sparse. Additionally, in lower-level mediation, the effect of the lower-level mediator on the outcome may oftentimes be confounded by an (un)measured upper-level variable. When such confounding is left unaddressed, the effect of the mediator, as well as the causal mediation effects, will be estimated with bias. In linear settings, bias due to unmeasured additive upper-level confounding is often remedied by separating the effect of the mediator into a within- and between-cluster component, but unfortunately, this solution no longer works when considering binary settings. To assess the severity of this transgression, we aim to tackle binary lower-level mediation from a counterfactual point of view (with a special focus on small clusters), by 1) providing non-parametrical identification assumptions of the direct and indirect effect, 2) parametrically identifying these effects based on multilevel logit-or probit-models, 3) considering estimation models for the mediator and the outcome, and 4) estimating the causal effects through an imputation algorithm that samples counterfactuals. Since steps three and four can be

completed in various ways, we compare the performance of three different estimation models: an uncentered and centred generalised linear mixed model, and a joint modelling approach. Employing simulations, we observe that the latter approach performs best.

**Causal Inference and Mediation - Parallel Session: 3.3B: FIXED VERSUS RANDOM EFFECTS MODELS FOR MULTILEVEL PROPENSITY SCORE ANALYSIS**

Jee-Seon Kim, University of Wisconsin-Madison; Youmi Suk, University of Wisconsin-Madison

Causal inference with observational data is challenging, as the assignment to treatment is often not random and people may have different reasons to receive or to be assigned to the treatment. Moreover, the analyst may not have access to all of the important variables, and this omission may lead to omitted variable bias as well as selection bias in nonexperimental studies. It is thus critical to account for the potential heterogeneity in selection processes and/or treatment effects as well as omitted variable bias in multilevel data. It is known that fixed effects models are robust against unobserved cluster variables while random effects models provide biased estimates of model parameters in the presence of omitted variables. This study further investigates the properties of fixed effects models as an alternative to the common random effects models for identifying and classifying subpopulations or "latent classes" when selection or outcome processes are heterogeneous. Our simulation study reveals that fixed-effects models outperform random-effects models in terms of the extraction of the correct number of latent classes and the classification of units, especially when the selection process is strong and cluster sizes are not too small. In discussion, it is shown that both fixed and random effects models are special cases of the generalized method of moments continuum with different assumptions on the endogeneity and exogeneity of the model predictors. The study concludes with recommendations for the proper use of fixed and random effects models for propensity score analysis with observational multilevel data.

**Causal Inference and Mediation - Parallel Session: 3.3C: DATA-BASED COVARIATE SELECTION FOR HIGH DIMENSION LOW SAMPLE SIZE DATA**

Rui Lu, Teachers College, Columbia University; Bryan Keller

With few exceptions, the propensity score literature has focused on estimating causal effects with moderate to large sample sizes. In the social and medical sciences, however, non-equivalent comparison group designs with small sample sizes are not atypical. Conditioning on many covariates in an attempt to satisfy the ignorability assumption may lead to $p > n$ estimation problems or inefficiency. In such cases, data-driven algorithms for selecting minimum covariate subsets may be useful. The primary aim of this study is to investigate the properties of three data-driven covariate selection techniques when used with small sample sizes under varying conditions. Stepwise logistic regression, Bayesian networks and random forests, are studied in a Monte Carlo simulation. In each scenario, we simulate small samples ranging from 50 to 500 with 90 noise covariates and 10 target covariates that have some association with either the propensity score or the potential outcomes. We generate data from several DAGs and implement de Luna, Waernbaum & Richardson' (2011) algorithms for covariate selection. Rosenbaum and Rubin's (1984; 2009) stepwise logistic regression approach is used as benchmark for comparison. The simulation results indicate random forest and Bayesian networks (using mutual information) outperform stepwise logistic regression by successfully reducing the dimension of the data set and including appropriate covariates suggested by the backdoor path criteria. Propensity scores based on selected covariate sets are used to assess bias and mean square error for each method. Results and implications for covariate selection with small samples are discussed.

**Causal Inference and Mediation - Parallel Session: 3.3D: IDENTIFYING RELEVANT PREDICTORS IN MESSY PSYCHOLOGICAL DATA**

Sierra Bainter, University of Miami

Selecting important predictors from some larger set can be a complicated problem in psychology, due in part to limited power and complex collinearity in the predictor set. This is especially true when the number of predictors is large relative to the sample size and there is limited information or theory to guide variable selection. As the number of predictors increases, it becomes intractable to consider models with all possible subsets of predictors and arbitrary to choose between them.

## Classification, Clustering, and Latent Class Analysis: 3:20 PM – 4:50 PM

Chair: Edward Ip

**Classification, Clustering, and Latent Class Analysis**

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 3.4A: A MIXED MEMBERSHIP RASCH MODEL**

Guoguo Zheng, University of Georgia; Hye-Jeong Choi, University of Georgia; Brian Bottge, University of Kentucky

When the population is a group of test examinees, mixed membership models assume that there are latent profiles within the population characterized by different response patterns across the test. Each examinee has a unique set of membership vector that indicates the probabilities of belonging to each profile. The Rasch model assumes that a response to a test item is determined by the examinee's ability level and item difficulty level, and this model holds for the entire population. In this study, we combined a mixed membership model with the Rasch model (MMR) and defined each profile by its unique set of item difficulty patterns across the test. In mixture Rasch models, each examinee can only belong to one of the subpopulations called latent classes. In contrast, the MMR allows examinees to have partial membership to all the profiles. We conducted a simulation study to investigate the MMR parameter recovery under practical testing conditions. We simulated 1000 examinees' responses to a 20-item test with three profiles. The simulation was repeated 30 times. The MMR was estimated using MCMC with 7000 iterations including a burn-in of 2000 iterations. We also examined the performance of DIC, AIC, BIC and AICM in model selection. Finally, we provided an empirical example that applied the MMR to a middle school fractions computation test. Two profiles were detected in the data. In profile 1, subtraction questions were more difficult than addition questions. In profile 2, questions with unlike denominators were more difficult than questions with the same denominators.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 3.4B: TOWARDS MULTIDIMENSIONAL LONGITUDINAL LEARNER PROFILES, A DYNAMIC BAYESIAN NETWORK APPROACH**

Josine Verhagen, Kidadaptive, Redwood City

Online educational products and educational games play an increasing role in students' educational journey. As all these products collect data about students, the question arises as to what extent it is possible to combine information collected from different educational environments to make inferences about a student's ability and progress over time.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 3.4C: IDENTIFYING DIVERGENT NONLINEAR GROWTH TRAJECTORIES USING RECURSIVE PARTITIONING**

Gabriela Stegmann, Arizona State University; Ross Jacobucci, University of Notre Dame; Sarfaraz Serang, University of Southern California; Kevin Grimm, Arizona State University

A common goal in longitudinal research is to model change and identify predictors of the between-person differences in the change trajectories. For instance, in educational studies researchers may be interested in measuring the rate of growth in achievement, and determine which primary skills are associated with the rate of growth. Nonlinear longitudinal recursive partitioning (NLRP) is a method that combines recursive partitioning (a data mining technique also known as regression trees) with the growth curve model (allowing for a linear or nonlinear model) in order to find variables that predict differences in the change trajectories. The data is recursively split into two nodes, as is done in regression trees. At each node, a growth curve model is estimated. The goal is to make the trajectories in the two nodes as homogeneous as possible. Therefore, every unique value of every predictor variable is evaluated as a potential split. The split that results in the highest homogeneity within the nodes is retained. This is done recursively on the resulting nodes until a stopping criteria (such as an extremely small gain in node homogeneity) is met. The NLRP method is presented and demonstrated with empirical data from the Early Childhood Longitudinal Study – Kindergarten Cohort, which is a longitudinal study that followed children from kindergarten to eighth grade. The study collected data on children's academic achievement

and demographic information. Using NLRP we explore various predictors of reading achievement. Benefits and limitations of NLRP are also discussed.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 3.4D: A HYBRID COGNITIVE DIAGNOSTIC MODEL**
Kazuhiro Yamaguchi, Graduate School of Education, the University of Tokyo; Kensuke Okada, Senshu University

Cognitive Diagnostic Models (CDMs) are useful methods to provide diagnostic information about the examinee's strengths and weaknesses in the learning process. Existing CDMs can be classified into two major categories, which are the compensatory and non-compensatory models. Compensatory models assume that low levels of one attribute in an examinee can be compensated for by high levels of another attribute. Conversely, non-compensatory models assume that an examinee must have all the required attributes to answer the items correctly. However, assumptions of both these models may be extreme ones. We believe that in most cases, actual diagnostic test items vary in degrees of their compensatory and non-compensatory properties. In order to implement this idea, we developed a new class of CDM in which the Item Response Function (IRF) is considered as a weighted combination of compensatory and non-compensatory model parts. More specifically, in the proposed CDM, the IRF is given as a mixture of the "Deterministic Input Noisy-Or gate" (DINO) and the "Deterministic Input Noisy-And gate" (DINA) IRFs. The proposed mixture model enables us to consider both compensatory and non-compensatory nature of items within one model. In other words, the proposed model is more flexible and less constrained than existing models in representing the examinees' problem solving process. Hamiltonian Monte Carlo algorithm, which is a variation of the Markov chain Monte Carlo method, was used to estimate the model parameters. Results of the simulation-based parameter recovery and empirical applications are presented.

## Item Response Theory: 3:20 PM – 4:50 PM

Chair: Qiwei He

**Item Response Theory**

**Item Response Theory - Parallel Session: 3.5A: COMPUTING EXPECTED FISHER INFORMATION FOR ADVANCED IRT MODELS**
Scott Monroe, University of Massachusetts Amherst

Advanced IRT models (e.g., multidimensional or multilevel IRT models) are becoming increasingly popular for modeling educational and psychological test data. Within a likelihood framework, numerous recent methodological advancements have facilitated estimation and inference for these models (e.g., Cai, 2010). One key quantity needed for numerous inferential procedures is the Fisher information matrix (FIM). In an IRT context, the FIM may be used to obtain parameter estimate standard errors, construct test statistics (e.g., Maydeu-Olivares, 2013), and facilitate test scoring (e.g., Yang, Hansen, & Cai, 2012). In general, for advanced IRT models, calculating the FIM is computationally challenging (due to high-dimensional integrals), and Monte Carlo methods are used. Typically, the observed FIM is approximated following Louis (1982). As an alternative, the current research proposes a new strategy to compute the expected FIM for such models. Unlike traditional approaches to calculating the expected FIM, the proposed strategy is computationally feasible, even for high-dimensional models with many items. In addition, the proposed strategy is computationally stable, and may be applied to other models (e.g., cognitive diagnostic models). Preliminary simulation work has been conducted using the proposed strategy to calculate expected FIM for tests with four correlated dimensions. Results show that standard errors are well-calibrated. Future simulation work will compare the performance of the observed FIM and proposed expected FIM.

**Item Response Theory - Parallel Session: 3.5B: ROBUST ESTIMATION FOR ITEM RESPONSE THEORY**
Maxwell Hong, University of Notre Dame

Self-report data is common in psychological and survey research. Unfortunately, many of these samples are plagued with careless responses due to unmotivated participants. The purpose of this study is to propose and evaluate a robust estimation method in order to detect careless, or unmotivated, responders while leveraging Item Response Theory (IRT) person fit statistics. First, we outline a general framework for robust estimation specific for IRT models. Subsequently, we conduct a simulation study covering multiple conditions to evaluate the performance of the proposed method. Ultimately, we show how robust estimation significantly improves detection rates for careless responders and reduce bias in item parameters across conditions. Furthermore, we apply our method to a real dataset to illustrate the utility of the proposed method. Our findings suggest that robust estimation coupled with person fit statistics offers a powerful procedure to identify careless respondents for further review, and to provide more accurate item parameter estimates in presence of careless responses.

**Item Response Theory - Parallel Session: 3.5C: SELECTING PRIORS UNDER THE 2PL AND 3PL ITEM RESPONSE MODELS**
Meina Bian, The University of Georgia; Victoria Tanaka, The University of Georgia; Seock-Ho Kim, University of Georgia

Accounting for auxiliary information and selecting the corresponding appropriate prior distribution is a problem encountered in Bayesian estimation under an item response theory (IRT) framework. The use of appropriate priors improves precision and stability of item parameter estimates (Mislevy, 1988). However, despite the importance of selecting a prior and defending this choice, there is not much transparency regarding the selection and use of priors in the literature. Based on previous work by Swaminathan and Gifford (1985, 1986), the role of priors in Bayesian estimation under the two- and three-parameter logistic (2PL and 3PL) models is investigated and presented in this paper. A literature review summarizes the variety of priors used in Bayesian estimation under the 2PL and 3PL models. A selection of these prior distributions is mathematically defined and further explored. The priors are evaluated using data simulated with OpenBUGS. Bayesian estimates obtained with informative priors are compared with estimates obtained with noninformative priors. A discussion of the implications of the use of priors, and Bayesian estimation, in IRT follows.

**Item Response Theory - Parallel Session: 3.5D: INVESTIGATING HYPERPRIORS FOR MODELING THE INTERTRAIT CORRELATIONS IN MIRT MODELS**
Meng-I Chang, Southern Illinois University Carbondale; Yanyan Sheng, Southern Illinois University Carbondale

Markov chain Monte Carlo (MCMC) algorithms have made the estimation of multidimensional item response theory (MIRT) models viable under a fully Bayesian framework. An important purpose of the MIRT models is to accurately estimate the interrelationship among multiple latent traits. In Bayesian hierarchical modeling, this is realized through modeling the covariance matrix, which is typically done via the use of an inverse-Wishart prior distribution due to its conjugacy property (Barnard et al., 2000). This prior, however, has problems because the marginal distribution for the variances has low density in a region near zero causing bias in Bayesian inferences (Gelman, 2006). To overcome these problems, other priors have been recommended, including the scaled inverse-Wishart (O'Malley & Zaslavsky, 2005), the hierarchical inverse-Wishart (Huang & Wand, 2013), and the LKJ priors (Lewandowski et al., 2009). In a study comparing these priors in a simple linear model, Alvarez et al. (2014) empirically showed that the inverse-Wishart prior performs poorly when the true variance is small even with large sample sizes. To date, no research, however, has investigated their comparisons under the MIRT context where latent variables are involved. Therefore, this study focuses on such models by comparing these four hyperprior specifications for the covariance matrix for the latent traits. Specifically, Monte Carlo simulations are carried out where sample sizes, test lengths, actual intertrait correlations, and true variances of the latent traits are manipulated. Findings from the study provide a set of guidelines on using these priors in estimating the Bayesian MIRT models.

**Item Response Theory - Parallel Session: 3.5E: SECOND-ORDER PROBABILITY MATCHING PRIORS FOR THE IRT PERSON PARAMETER**
Yang Liu, University of Maryland, College Park; Jan Hannig, The University of North Carolina at Chapel Hill; Abhishek Pal Majumder, Stockholm University

In applications of item response theory (IRT), it is often of interest to compute confidence intervals (CIs) for the person parameter that attain the desirable coverage of the true value in the frequentist sense. The ubiquitous use of short tests in social science research and practices calls for a refinement of the standard interval estimation procedures for the person parameter based on the first-order asymptotic theory, such as the Wald CI, which only works well when the test is sufficiently long. In the current paper, we propose a simple construction of second-order probability matching priors (PMPs) for the person parameter in unidimensional IRT models, which in turn yields CIs that maintain the prescribed coverage even in very short tests. The probability matching property is established based on an expansion of the posterior distribution and a shrinkage argument. The generic formulation of PMPs can be applied to a wide variety of unidimensional IRT models; Monte Carlo simulations are conducted to evaluate the performance of the proposed method under the two-parameter logistic, three-parameter logistic, graded, and nominal models.

## Multilevel/Hierarchical/Mixed Models: 3:20 PM – 4:50 PM

Chair: Javier Revuelta

**Multilevel/Hierarchical/Mixed Models**

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 3.6A: DOMINANCE ANALYSIS FOR DETERMINING PREDICTOR IMPORTANCE IN LONGITUDINAL MULTILEVEL MODELS**
Luciana Cançado, University of Wisconsin, Milwaukee; Razia Azen, University of Wisconsin, Milwaukee

The purpose of this study is to extend and evaluate Dominance Analysis (DA), a method used to determine the relative importance of predictors in various linear models (Budescu, 1993; Azen & Budescu, 2003; Azen, 2013), for use with longitudinal multilevel models. DA uses a measure of model fit to compute the additional contribution of each predictor to each possible subset model. A predictor is considered to dominate another if it consistently contributes more when added to each subset model or on average across models. The bootstrapping procedure is used to determine the consistency of the sample dominance results across other potential samples. This simulation study focused on predictors of outcomes measured on either 4 or 8 occasions for samples of either 30 or 200 individuals; thus measurement occasions (level-1 units) are considered to be nested within individuals (level-2 units). Additional factors varied in the simulations include different levels of model complexity (i.e., number of predictors at level-1 and level-2), size of predictor coefficients, predictor collinearity levels, and measures of model fit (i.e., R-squared analogues). The results are used to evaluate the accuracy of DA in rank-ordering the model predictors, the use of DA measures in inferential procedures testing whether one predictor significantly dominates another (including Type I error, power, and accuracy of estimation), and the reproducibility of the dominance results over repeated samples. This study will thus inform and provide recommendations to researchers who wish to determine the relative importance of predictors in longitudinal multilevel models.

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 3.6B: DYNAMIC GLMMS WITH CROSSED RANDOM EFFECTS: AN APPLICATION TO INTENSIVE BINARY TIME-SERIES EYE TRACKING DATA**
Sun-Joo Cho, Vanderbilt University; Sarah Brown-Schmidt, Vanderbilt University

As a method to ascertain person and item effects in psycholinguistics, a generalized linear mixed effect model (GLMM) with crossed random effects has met limitations in handing serial dependence across persons and items. This paper presents an autoregressive GLMM with crossed random effects, that accounts for variability in lag effects across persons and items. The model is shown to be applicable to intensive binary time series eye tracking data when researchers are interested in detecting experimental condition effects while controlling for previous responses. In addition, a simulation study shows that

ignoring lag effects can lead to biased estimates and underestimated standard errors for the experimental condition effects.

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 3.6C: ANALYSIS OF CHANGE WITH TWO TIME POINTS**
Ehri Ryu, Boston College

When the repeated measures are taken at two time points, research questions often involve the change between two time points, e.g., comparison of mean change between groups, association between the change and a covariate, or association between the changes in two or more series of repeated measures. There are four approaches to analyzing repeated measures at two time points (T1 and T2) without using difference scores. T2 measure can be analyzed with T1 measure controlled for. Within-subject or mixed analysis of variance (ANOVA) can be used with time as a within-subject factor with 2 levels. A multilevel model can be specified for within-individual and between-individual levels. A latent variable can be set up to represent the change in latent change score models. This study compares these approaches in terms of their applicability, flexibility, and performance, with particular focuses on the followings: comparing within-individual level random error in ANOVA and multilevel approaches, statistical inference with no within-individual level residual in multilevel model, direct representation of the change quantity in multilevel model and latent change score model, and parallel process model among two or more series in the latent change score approach.

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 3.6D: ESTIMATING RANDOM INTERCEPT MODELS WITH CLUSTER-ENDOGENOUS COVARIATES**
Nicholas Sim, University of California, Berkeley; Sophia Rabe-Hesketh, University of California, Berkeley; Anders Skrondal, Norwegian Institute of Public Health & University of Oslo & University of California, Berkeley

In estimating Random Intercept Models (RIMs), we assume that covariates at all levels do not co-vary with the random intercepts. A violation of this assumption (called cluster-level endogeneity) leads to inconsistent estimates when using standard estimation procedures. For two-level RIMs with such endogeneity, Hausman and Taylor (HT) devised an instrumental variable estimator (using only internal instruments) that produces consistent estimates. One can also explicitly model the endogeneity using a Structural Equation Model (SEM). Through simulation, we compare the HT and SEM estimators and evaluate their asymptotic and finite sample properties. We show that the SEM approach is also flexible enough to deal with different exchangeability assumptions for the covariates (e.g., whether the correlation between all units in a cluster are the same) and investigate how these exchangeability assumptions affect finite sample properties of the HT estimator. For the simulations, a new procedure is proposed for generating cluster- and unit-level covariates and random intercepts with a fully flexible covariance structure.

## Validity and Reliability: 3:20 PM – 4:50 PM

Chair: Patrick Shrout

**Validity and Reliability**

**Validity and Reliability - Parallel Session: 3.7A: A COMPARISON OF RELIABILITY GENERALIZATION METHODS**
Francesca Teora, The Graduate Center, CUNY

Reliability generalization (RG) involves the meta-analysis of reliability coefficients. RG is particularly relevant in the development of assessments where heterogeneity due to contextual factors such as language spoken, gender, culture, and age of examinees are important to assess. There are currently four primary RG models described in the literature, including (1) the fixed effects model (not generally recommended), (2) the Gaussian random effects model, (3) a Bayesian adaptation of the Gaussian random effects model (Brannick & Zhang, 2013), and (4) Bonett's (2010) varying coefficient model. In addition, because the sampling distribution of reliability coefficients is markedly non-Gaussian,

transformation of reliability coefficients is generally required. This study compares these methods with two semi-parametric methods: A Bayesian meta-analysis via the Dirichlet process prior (Jara et al., 2010) and Aitkin's non-parametric maximum likelihood (Aitkin, 1999; Skrondal & Rabe-Hesketh, 2004). These two methods are particularly useful when the distribution of reliability coefficients is non-Gaussian, which is frequently observed in meta-analysis more broadly but likely to occur in RG. We present results of a simulation study as well as apply the method to two datasets taken from the literature, one comparing reliabilities of the Maslach Burnout Inventory (MBI) subscales and another comparing Minnesota Multiphasic Personality Inventory (MMPI) subscales.

**Validity and Reliability - Parallel Session: 3.7B: DEPENDABILITY OF DIFFERENCE SCORES OF STUDENTS, CLASSES, AND SITES**

Rabia Karatoprak Ersen, University of Iowa; Won-Chan Lee, University of Iowa

Previous research on reliability of difference scores (e.g., Miller & Kane, 2001; Brennan, Yin, & Kane, 2003) has suggested that the error tolerance ratio or dependability coefficient of difference scores be reported, instead of classical reliability estimates. The purpose of this study is to examine dependability of difference scores using an empirical dataset. Difference scores, in this study, refer to the differences between a pre- and a post-survey which were administered to assess the effectiveness of an intervention. Multivariate generalizability theory (Brennan, 2001) will be used to analyze the data to evaluate dependability of difference scores at the levels of students, classes, and sites. The data collection design involved a nested structure such that students (p) were nested within classes (c) and classes were nested within sites (s). In addition, items (i) of the survey were clustered under different content (h) categories. G-study designs are unbalanced (p•) x (i•:h•) using students as the objects of measurement, (p•:c•) x (i•:h•) using classes as the objects of measurement, and (p•:c•:s•) x (i•:h•) using sites as the objects of measurement. The levels of the multivariate facet are pre survey and post survey, which are crossed with each facet.

**Validity and Reliability - Parallel Session: 3.7C: ESTIMATION OF IRT RELIABILITY WITH MULTIPLE LATENT OR OBSERVED GROUPS**

Björn Andersson, Centre for Educational Measurement, University of Oslo

Estimating the reliability of scores from a test provides important information regarding the precision of measurement. The reliability of scores on a test is however population dependent and it is reasonable to expect that the reliability is different for different subpopulations. In item response theory (IRT), the population dependence can be sourced to differential item functioning or to differences in the distributions of the latent variable in different observed or latent groups. In this study, we illustrate the impact on reliability coefficients from these two sources in IRT using multiple group and mixture IRT models and show how reliability for different subgroups can be estimated from a single sample. Furthermore, we outline how to estimate confidence intervals for the reliability using asymptotic expansions and the bootstrap. Implications and recommendations for applied work are outlined.

**Validity and Reliability - Parallel Session: 3.7D: CROSS-CULTURAL SCALE COMPARABILITY USING PARTIAL-BY-ITEM&COUNTRY INVARIANCE ANALYSIS**

Hynek Cigler, Masaryk University, Czech Republic; Agnes Stancel-Piątak, IEA-Hamburg; Minge Chen, IEA-Hamburg

To conduct cross-cultural comparisons it is crucial to establish scales that are equivalent among cultures. The construction of respective measurement instruments has proved to be challenging in the context of Large Scale Assessment (LSA), in which the number of participating countries is large. This study uses Multiple Group Confirmatory Factor Analysis (MGCFA) to compare the results from the traditional measurement invariance analysis with two partial measurement invariance approaches. The data used in this study comes from the TALIS Starting Strong Survey, which is a unique international dataset of staff and leader characteristics in Early Childhood Education (ECEC). Traditional measurement invariance approach requires the parameters of all items for all countries the same. Due to the system diversity, it might be challenging to achieve cross-cultural comparability of latent traits in the ECEC context. The traditional partial invariance (partial by item) method allows some items to vary across all countries, whereas other items are still invariant. This study adopted a more flexible strategy by allowing some

items to be freely estimated for some countries to create comparable constructs, an alternative approach we named as partial-by-item&country invariance. We expect that the partial-by-item&country measurement invariance will outperform the traditional invariance analysis as well as the partial-by-item invariance analysis with respect to the model fit as well as to the number of scalar invariant scales. This approach allows enhancing the cross-country comparability in future LSA without violating statistical assumptions while simultaneously considering cultural differences.

## State of the Art: 5:00 PM – 5:50 PM

Chair: Matthew Johnson

**Diagnosing Diagnostic Models: From von Neumann's Elephant to Model Equivalencies and Network Psychometrics**
State of the Art: Matthias von Davier, National Board of Medical Examiners

Chair: Willem Heiser

**Watching Children Grow Taught Me All I Know**
State of the Art: James Ramsay, McGill University

## Dissertation Prize: Sacha Epskamp: 5:50 PM – 6:30 PM

Chair: Terry Ackerman

**Network Psychometrics: Current State and Future Directions**
Sacha Epskamp

## Welcome Reception and Poster Session: 6:30 PM – 8:30 PM

# Wednesday, July 11, 2018 AM

## IMPS Registration: 8:00 AM – 3:15 PM

## Symposium 7: 8:30 AM – 10:00 AM

Chair: Silvia Bianconcini

**Symposium 7: Recent Advances in the Analysis of Complex Data Structures**

**Symposium 7 - Parallel Session: 4.1A: FINITE MIXTURES OF REGRESSION MODELS FOR LONGITUDINAL RESPONSES AND OMITTED COVARIATES BIAS**
Marco Alfò, Sapienza University of Rome

Individual-specific, time-constant, random effects are often used to model dependence and/or to account for omitted covariates in regression models for longitudinal responses, especially when short individual sequences are observed. Longitudinal studies have known a huge and widespread use in the last few years as they allow to distinguishing between so-called age and cohort effects; these relate to differences that can be observed at the beginning of the study and stay persistent through time, and changes due to the temporal dynamics in the observed covariates. While there is a clear and general agreement on this purpose, the random effect approach has been frequently criticized for not being robust to the presence of correlation between the observed (i.e. covariates) and the unobserved (i.e. random effects) heterogeneity. Often, this is felt as a reason to choose the fixed effect estimator instead, with the Hausman test being advocated for this purpose. Starting from the so-called correlated effect approach, we argue that the random effect approach may be parameterized to account for potential correlation between observables and unobservable. Specifically, when the random effect distribution is estimated

non-parametrically using a discrete distribution on K <=n locations, a further, more general, solution could be adopted. This is illustrated via a large scale simulation study and a benchmark data example.

**Symposium 7 - Parallel Session: 4.1B: INFERENCE BASED ON DIMENSION-WISE QUADRATURE FOR THE ANALYSIS OF MULTIDIMENSIONAL LONGITUDINAL DATA**

Silvia Bianconcini, Department of Statistical Sciences - University of Bologna; Silvia Cagnone, Department of Statistical Sciences - University of Bologna

We consider approximate methods for likelihood inference to longitudinal and multidimensional data within the context of health science studies. The complexity of these data necessitates the use of sophisticated statistical models that can pose significant challenges for model fitting in terms of computational speed, memory storage, and accuracy of the estimates. Our methodology is motivated by a study that examines the temporal evolution of the mental status of the US elderly population between 2006 and 2010. We propose modeling the individual mental status as a latent process also accounting for the effects of individual specific characteristics, such as gender, age, and years of educational attainment. We describe the specification of such a model within the generalized linear latent variable framework, and its efficient estimation using a recent technique, called dimension-wise quadrature. The latter allows a fast and streamlined analytical approximate inference for complex models, with better or no degradation in accuracy compared with the standard techniques, such as Laplace approximation and adaptive quadrature. The model and the method are applied in the analysis of cognitive assessment data from the Health and Retirement Study combined with the Asset and Health Dynamic study.

**Symposium 7 - Parallel Session: 4.1C: BI-FACTOR MIRT OBSERVED-SCORE EQUATING UNDER THE NEAT DESIGN**

Valentina Sansivieri, Department of Statistical Sciences, University of Bologna; Matteucci Mariagiulia, Department of Statistical Sciences, University of Bologna

Traditional Item Response Theory (IRT) equating procedures are based on unidimensionally scored test forms. Multidimensional Item Response Theory (MIRT) test equating has been studied in few works. Among others, Brossman & Lee (2013) developed MIRT observed and true-score equating with a multidimensional Two-Parameter Logistic (2PL) model. Wang et al. (2014) focused on the applicability of MIRT methods in vertical equating. Lee & Lee (2016) developed a bi-factor MIRT observed-score equating procedure for mixed-format tests under the equivalent group (EG) design based on the 2PL model. In this work, we propose an observed-score equating procedure based on the bi-factor extension of the Three-Parameter Logistic (3PL) model under the nonequivalent groups with anchor test (NEAT) design. The bi-factor 3PL model is chosen because it allows for the presence of overall and specific traits and the guessing parameter which are especially important in educational assessment. The NEAT design is chosen because, when an anchor test is available, it allows to work with nonequivalent groups. The first results obtained by using simulated data show that, in presence of bidimensionality, the proposed equating procedure is more efficient than the unidimensional observed-score equating. An empirical study is also presented based on data coming from the Italian national student assessments conducted by the Italian National Institute for the Evaluation of the Education System (INVALSI).

**Symposium 7 - Parallel Session: 4.1D: MODEL-BASED AND FUZZY CLUSTERING ALGORITHMS: A COMPARATIVE ASSESSMENT**

Paolo Giordani, Sapienza University of Rome; Marco Alfò, Sapienza University of Rome; Maria Brigida Ferraro, Sapienza University of Rome; Luca Scrucca, University of Perugia; Alessio Serafini, Sapienza University of Rome

Model-based and fuzzy clustering (algorithms) are widely used clustering methods. In both cases, observation units are assigned to clusters via a soft allocation rule. In the former approach, it is assumed that the data are generated by a mixture of probability distributions (usually multivariate Gaussian) in which each component represents a different group or cluster. Each observation unit is ex-post assigned to the clusters according to the so-called MAP principle, i.e. based on the posterior probabilities of component membership. In the latter case, no probabilistic assumptions are made and each observation unit belongs to the clusters with the so-called fuzzy membership degrees, taking values in [0,1], based on the distances between the observation units and the cluster prototypes. Therefore, it is quite obvious that

the posterior probability of component membership may play a role similar to the membership degree. The aim is at comparing the performance of both approaches by means of a simulation study. In detail, in the model-based context, finite mixtures of either Gaussian or t distributions are investigated, whilst, in the fuzzy context, the standard fuzzy k-means algorithm and the Gustafson-Kessel variant for ellipsoidal clusters are considered.

## Symposium 8: 8:30 AM – 10:00 AM

Chair: Gunter Maris

**Symposium 8: Non-cognitive Psychometric Theory and Assessment**

**Symposium 8 - Parallel Session: 4.2A: THE ISING MODEL OF ATTITUDES**
Han L.J. van der, University of Amsterdam Maas; Jonas Dalege, University of Amsterdam

We discuss the psychometric properties of the Ising representation of the Causal Attitude Network (CAN) model. The CAN model conceptualizes attitudes as networks consisting of evaluative reactions and interactions between these reactions, conforming to a small-world structure. By adopting the Ising model paradigm we can formalize several key assumptions underlying the study of attitudes, derive original predictions, and establish new links between existing psychometric models and attitude modeling. By introducing Hebbian learning in the attitude network model, dynamic properties of attitudes, such as ambivalence, gradual versus abrupt attitude change, and implicit versus explicit measurements of attitudes, can be explained.

**Symposium 8 - Parallel Session: 4.2B: A FREE ENERGY THEORY OF CREATIVE THINKING**
Gunter Maris, ACTNext by ACT, Inc.; Lu Ou, ACTNext by ACT, Inc.; Vanessa Simmering, ACT, Inc.; Vanessa Simmering, ACT, Inc.; Benjamin Deonovic, ACTNext by ACT, Inc.; Maria Bolsinova, ACTNext by ACT, Inc.

The measurement of creative thinking, like all of the other 21-th century skills, is, to say the least, poorly developed. Based on the Free Energy principle, the dominant high level theory of neuroscience, we develop an account of how the creative thinking process could work. The theory is developed to the level that it is both in accord with empirical evidence, and allows for mapping out its measurement and assessment consequences. What sets this line of research apart from much of the literature on the topic is its focus on the thinking process.

**Symposium 8 - Parallel Session: 4.2C: THE MATHEMATICS OF COLLABORATION: INSIGHTS FROM GAME THEORY AND MATHEMATICAL BIOLOGY**
Lu Ou, ACTNext by ACT, Inc.; Vanessa Simmering, ACT, Inc.; Gunter Maris, ACTNext by ACT, Inc.

Collaboration is a complex dynamic process, where the subjects who participate in it constantly update their own goals and strategies. Depending on the extent to which the goals and strategies are shared, the collective behavior can easily alternate between a collaborative mode and a competitive mode. Different modes of the group behavior can lead to distinct outcomes. In order to gain insights on how to facilitate optimal outcomes, we represent the collaborative or competitive process as mathematical models, as developed in game theory and mathematical biology, and illustrate the model behavior using mathematical derivations and simulations. We will cast mathematical conclusions on solutions to real-world problems, and offer practical advice on how to facilitate collaboration in work and education.

**Symposium 8 - Parallel Session: 4.2D: A CROSS-DISCIPLINARY LOOK AT NONCOGNITIVE ASSESSMENTS**
Vanessa Simmering, ACT, Inc.; Lu Ou, ACTNext by ACT, Inc.; Gunter Maris, ACTNext by ACT, Inc.

Educators and employers seek to evaluate noncognitive skills, such as diligence, integrity, and collaboration, because of the intuitive role they play in educational and workplace success. However, exactly how these skills are defined – and therefore assessed – varies across research disciplines. This talk will briefly review the way these skills have been conceptualized and measured across disciplines including cognitive science, developmental psychology, economics, and education. Synthesizing across

the commonalities, differences, and limitations in these various approaches will have important implications for the development and interpretation of noncognitive assessments.

## Computer-Based Testing: 8:30 AM – 10:00 AM

Chair: John Donoghue

**Computer-Based Testing**

**Computer-Based Testing - Parallel Session: 4.3A: USE OF THE NONPARAMETRIC ISOTONIC MODEL IN CAT**

Mario Luzardo, University of the Republic, Uruguay

Xu & Douglas (2006) studied the possibility of using Ramsay nonparametric model in a computerized adaptive test (CAT). They explore the possible methods of item selection but they did not use the Fisher maximum information method because the derivatives of the item characteristics curves (ICC) may not be estimated well. Here we use a computerized adaptive test where the item bank follow the nonparametric isotone model proposed by Luzardo & Rodriguez (2015). The model is based on the estimation of the inverse of the ICC and it use a two-stage process. The first stage uses a nonparametric estimator of the ICC by means of nonparametric kernel regression; the second uses the above result to estimate the density function of the inverse ICC. By integrating the density function, and then symmetrized it, we obtain the ICC estimator. We can consider estimating its derivatives from its formula and to use the Fisher information to select the item to be administrated. This work focuses on comparing four methods of selecting items in the CAT; the random selection, the maximum Fisher information criterion, the Kullback-Leibler information criterion and a procedure based on the expected Shannon entropy. We show the results of a simulation study which compare the procedures using to measure the adequacy, the root mean squared error, the bias and the item exposure rate.

**Computer-Based Testing - Parallel Session: 4.3B: EXTENDING THE DIFFUSION–IRT MODEL TO FORCED-CHOICE RESPONSE TIME DATA**

Kyosuke Bunji, University of Tokyo; Kensuke Okada, Senshu University

Ipsative measurement, such as multidimensional forced-choice tasks, is a promising solution to overcome the problem of response biases such as faking. However, there exist several analytical and practical problems in statistically analyzing ipsative data. Brown & Maydeu-Olivares (2013) demonstrated that most of these problems can be resolved by applying the item response theory (IRT) framework. Meanwhile, Tuerlinckx & De Boeck (2005) developed a response-time-incorporated IRT model by extending the diffusion model, which is one of the well-known models that represent underlying cognitive processes. By combining these two approaches, in this study, we proposed a novel extension of the diffusion–IRT model to forced-choice responses and response times. The proposed model inherits the merits from both these previously proposed models: it can extract more abundant information from response time, and it resolves the problems of ipsative measurement. From a simulation study, we found that the proposed model enables us to estimate parameters correctly even in a variety of different conditions related to the number of dimensions, number of items, and factor correlations. Furthermore, we applied the proposed model to real personality data collected from two forced-choice questionnaires. Findings revealed that the proposed model allows us to obtain additional information from the response times, while it resolves the problems of ipsative measurement. These results indicate that the proposed approach may be used as a less-biased psychometric assessment based on forced-choice responses and response time measurement.

**Computer-Based Testing - Parallel Session: 4.3C: RELIABILITY OF AN ADAPTIVE ASSESSMENT BASED ON KNOWLEDGE SPACE THEORY**

Christopher Doble, McGraw-Hill Education; Jeffrey Matayoshi, McGraw-Hill Education; Eric Cosyn, McGraw-Hill Education; Hasan Uzun, McGraw-Hill Education; Arash Karami, McGraw-Hill Education

A large-scale simulation study of the assessment effectiveness of a particular instantiation of knowledge space theory is described. In this study, data from more than 700,000 actual assessments in mathematics

using the ALEKS (Assessment and LEarning in Knowledge Spaces) software were used to determine response probabilities for the same number of simulated assessments, for the purpose of examining reliability. The results indicate a reliability comparable to that of assessments having mathematics content overlapping with that of the ALEKS assessment, though these assessments are more limited in scope than the ALEKS assessment. Some consequences and future directions will be discussed.

**Computer-Based Testing - Parallel Session: 4.3D: APPLYING HIGHER ORDER MODELS TO IMPROVE CAT ADAPTIVE LEARNING ALGORITHMS**
Ruitao Liu, ACT, Inc.; Terry Ackerman, University of Iowa; Yu-Lan Su, ACT, Inc.

This research aims to apply higher order CDM and IRT models to improve CAT adaptive learning algorithms. This study has two basic parts: first to confirm the feasibility of adopting standards as attributes in Q-matrix and second to facilitate the implementation of adaptive learning algorithms. The data include Math and Reading operational data from a large scale paper-and-pencil testing program, and corresponding UIRT item parameter estimates used to develop a CAT item pool for the same assessment. First, the response data will be calibrated using both the HO-DINA and the HO-IRT models. The estimated item parameters, proficiency profiles, item fit and model fit will be reviewed. The estimated skill proficiency patterns and the reliabilities for item standards estimated from the CDM model will be reviewed and discussed with content experts to evaluate the appropriateness and reasonability of adopting the tests' common core standard coding into the CDM model's Q-matrix. The estimated theta scales from HO-DINA and HO-IRT will be linked to the CAT pool theta scales. Examinee response data will be simulated for the CAT pool, and be estimated using HO-DINA and HO-IRT to obtain the estimated mastery profiles, higher-order thetas, and domain thetas. The study will examine differences between item selection procedures for adaptive learning using examinees' estimated mastery skill patterns versus estimated theta and domain theta ranges. The results will be evaluated using criteria such as item selection accuracy and efficiency of feeding in practice items for test-takers and then score and route to the next item.

## Measurement Invariance and DIF: 8:30 AM – 10:00 AM

Chair: Matthias von Davier

**Measurement Invariance and DIF**

**Measurement Invariance and DIF - Parallel Session: 4.4A: USING NLMM TO DETECT THE NON-UNIFORM AND UNIFORM DIFS**
Guan-Yu Chen, Beijing Normal University; Ping Chen, Beijing Normal University

More and more educational and psychological assessments have started to adopt testlet design, which constructs a bundle of items that share a common stimulus (Wang & Wilson, 2005). However, the use of testlets in a test can lead to local item dependence (LID), which will affect the following IRT-based analyses including the differential item functioning (DIF) detection. Both generalized linear mixed models (GLMM) and nonlinear mixed models (NLMM), which extend the IRT to explanatory IRT, can model LID and DIF directly (De Boeck & Wilson, 2004). Although GLMM and bi-factor models (also can be considered as NLMM) have been employed to detect the uniform DIF under the condition of LID (Beretvas & Walker, 2012; Fukuhara & Kamata, 2011; Ravand, 2015), the how to detect the non-uniform DIF under the condition has not been studied yet which greatly limits the application of explanatory IRT in DIF detection. This research tried to detect the non-uniform and uniform DIFs using NLMM under the LID context. The real data of the Chinese eighth grade students from the National Basic Education Assessment was used, and the results from SAS PROC NLMIXED will be compared with those from the IRT Likelihood Ratio method (Thissen, Steinberg, & Wainer, 1988). The preliminary results showed the advantages of the NLMM-based method in Type I error, suggesting the new method can be used to deal with DIF under complex scenarios.

**Measurement Invariance and DIF - Parallel Session: 4.4B: DETECTION OF DIFFERENTIAL ITEM FUNCTIONING WITH DIFNLR PACKAGE**

Adéla Drabinová, Faculty of Mathematics and Physics, Charles University, and Institute of Computer Science of the Czech Academy of Sciences; Patrícia Martinková, Faculty of Education, Charles University, and Institute of Computer Science of the Czech Academy of Sciences

The R package difNLR (Drabinová, Martinková & Zvára, 2018) has been developed for detection of Differential Item Functioning (DIF), based on extensions of logistic regression model. These include guessing and non-attention parameters which can differ for different groups. For dichotomous data, eleven predefined models have been implemented, however, user can constraint some parameters to be the same for different groups and hence create wide range of models that can be seen as proxies for item response theory models. The difNLR package offers various methods for estimation of parameters and DIF detection procedure. It also covers procedures in DIF identification such as item purification or corrections for multiple comparisons. Moreover, simulation studies suggest good properties even in smaller samples (Drabinová & Martinková, 2017), and thus the family of models offered by the difNLR library seems to be promising in DIF detection.

**Measurement Invariance and DIF - Parallel Session: 4.4C: ACQUIESCENCE AND PERSON DIFFERENTIAL FUNCTIONING: SOLVING PDIF WITH BALANCED SCALES**

Ricardo Primi, Universidade São Francisco; Daniel Santos, Universidade de São Paulo and EduLab21, Ayrton Senna Institute; Filip De Fruyt, Ghent University and EduLab21, Ayrton Senna Institute; Oliver P. John, University of California, Berkeley and EduLab21, Ayrton Senna Institute

Likert-type self-report scales are frequently used in large-scale educational assessment of social-emotional skills. Self-report questionnaires rely on the assumption that their items elicit information only about the trait they are supposed to measure. However, different response biases may threat this assumption. Specifically, in children, the response style of acquiescence is an important source of systematic error. Balanced scales have been proposed as a solution to control for acquiescence, but the reasons why this design feature worked from the perspective of modern psychometric models have been underexplored. Three methods for controlling for acquiescence are compared: (a) Classical ipsatization by partialling out the mean, (b) an Item Response Theory method to measure Person Differential Functioning (PDIF), and (c) random intercept item factor analysis. Comparative analyses are conducted on self-ratings on a fully balanced 30-item scale assessing Conscientious Self-management provided by 40,649 students (aged 11 to 18). Acquiescence bias was explained as person differential item functioning and it was demonstrated that: (a) the acquiescence index is equivalent to PDIF, (b) balanced scales resolve PDIF, and (c) that random intercept factors are equivalent to PDIF and the substantive factor is controlled for PDIF

**Measurement Invariance and DIF - Parallel Session: 4.4D: LINKING SCALES WITHOUT COMMON ITEMS**

Ya Zhang, Western Michigan University

Mixture item response theory (IRT) models have been shown to improve the identification of latent group structure and facilitate the estimation of model parameters when covariates are incorporated or the Bayesian estimation method is employed. However, the efficiency of mixture IRT models in differential item functioning (DIF) analysis has not been systematically studied due to the challenges of identifying DIF with a relatively complex model. The present study explored the effect of covariate and estimation method on the detection of latent DIF under the mixture IRT framework. A Monte Carlo simulation study was performed by manipulating the magnitude of DIF, type of DIF, proportion of DIF items, group impact, and relationship between the covariate and the latent group membership. The generated response data were analyzed using the mixture 2PL IRT model by manipulating the inclusion of covariate and the estimation method (maximum likelihood and Bayesian estimation). The estimation results were evaluated in terms of the recovery of the latent group structure, recovery of the model parameters, and detection of DIF. The study provided insights and suggestions on the use of mixture IRT models in the analysis of DIF.

**Measurement Invariance and DIF - Parallel Session: 4.4E: LINKING SCALES WITHOUT COMMON ITEMS**
Michael Hunter, University of Oklahoma; David E. Bard, University of Oklahoma Health Sciences Center

The classical psychometric way of linking scales relies on the scales having common or overlapping items (e.g. Bauer & Hussong, 2009). We develop a method to link scales together without any common items by instead relying on the smoothness of the change in the population characteristics of the latent variable(s) as a function of some measured covariate. The approach we take is threefold. First, through simulation studies we develop and validate a method that, under certain circumstances, can link together versions of a measure in the extreme case where there is no item overlap. Second, we apply the method developed in the simulations to Ages and Stages Questionnaire (ASQ) data from a large community sample to place the various versions of the ASQ on the same scale. Third, we map the common scale onto the raw sum scores of the ASQ subscales, thus creating transformations of the raw scores that are on the same scale.

## Item Response Theory: 8:30 AM – 10:00 AM

Chair: Thorsten Meiser

**Item Response Theory**

**Item Response Theory - Parallel Session: 4.5A: A MODEL-BASED EMPIRICAL COMPARISON OF DISCRETE OPTION AND TRADITIONAL MULTIPLE-CHOICE ITEMS**
Daniel Bolt, University of Wisconsin, Madison; Nana Kim, University of Wisconsin, Madison; Yiqin Pan, University of Wisconsin, Madison; Carol Eckerly, Alpine Testing Solutions; John Sowles, Ericsson, Inc.

Discrete option multiple choice (DOMC) items differ from traditional multiple-choice (MC) items in that the presentation of response options is done sequentially until either one (or more) correct options (keys) are selected to score the item as correct or one of the distractors is selected to score the item as incorrect. This delivery format can be appealing in computer-based test administrations for various reasons. Using empirical data collected for the same (equivalent) items administered in both DOMC and MC formats, we study the effect of key location (order of option presentation) on DOMC items in terms of its variability across items and persons, focusing on both response correctness and response time. Beyond informing about item format effects, we suggest the comparison lends insight into how students solve individual multiple-choice items. The resulting analysis can inform the application of process models (and the potential need for varying process models) across items within a multiple-choice test.

**Item Response Theory - Parallel Session: 4.5B: TWO MODIFIED STATISTICS FOR S-X^2**
Zhuangzhuang Han, Teachers College Columbia University; Xiang Liu, Teachers College, Columbia University

S-X^2, a squared-residual based goodness-of-fit statistic, has gained prominence in item-fit fit analysis for dichotomous item response models (e.g., IRT and CDM) due to its stable performance in terms of the type-I error and power that is shown in past comparison studies. However the early statistical article has shown that this type of statistics constructed on the basis of grouped residuals using the MLEs obtained from ungrouped data will have a non-chi-square limiting distribution. Using the chi-square approximation will lead to inflated type-I error. S-X^2 is modified based on two correction ideas borrowed from the mathematical statistical studies to follow the chi-square distribution. Simulations on the type-I error and power are conducted to substantiate the modification.

**Item Response Theory - Parallel Session: 4.5C: A MODIFIED S-X2 FOR DICHOTOMOUS IRT MODELS WITH MISSING DATA**
Xue Zhang, Northeast Normal University; Chun Wang, University of Minnesota; Jian Tao, Northeast Normal University

Item-level fit analysis not only serves as a complementary check to global fit analysis, it is also essential in scale development because the fit result will guide item revision/deletion (Liu &Maydeu-Olivares, 2014).

During data collection, missing response data may likely happen due to various reasons. Chi-square-based item fit indices are the most widely used statistics to assess item-level fit, yet none of them can perform well or can be directly applicable to deal with missing data, as they rely on parameter estimates (e.g., Yen's Q1 and McKinley and Mill's G2) or total scores (Orlando and Thissen's S-X2 and S-G2). To this end, we propose modified versions of S-X2 and S-G2 , which can be used to evaluate item-level fit when response data is incomplete, denoted as M-X2 and M-G2. They are the generalization of Orlando and Thissen (2000)'s indices. Instead of using observed total score for grouping, the new indices rely on correct response proportion. The new indices are equivalent to the original Orlando and Thissen (2000)'s indices when response data is complete.Their performances are evaluated via simulation studies, the manipulated factors include test length, sources of misfit, misfit proportion, missing proportion and missing type. The results from simulation studies are consistent with Orlando and Thissen's results (2000, 2003), and the performances of the proposed indices are much better than the default index in flexMIRT (i.e., S-X2). Keywords: item fit, missing data, S-X2 , M-X2, M-G2

**Item Response Theory - Parallel Session: 4.5D: THE IMPACT OF BIB SPIRALING ON LOCAL DEPENDENCE DETECTION STATISTICS**

Katherine Castellano, Educational Testing Service; Yue (Helena) Jia, ETS

Scenario-based tasks (SBTs) are becoming increasingly common in K-12 assessments, particularly in tests aligned to science and technology standards. However, because items within SBTs may elicit more similar responses than between SBTs, SBTs may violate the key assumption in standard IRT models of local item independence. Ignoring such dependency would potentially result in biased item, test, and person parameters, which can, in turn, affect inferences about subgroup performance. SBTs may be more susceptible to local dependence (LD) compared to reading passage item sets given their novelty. For testing programs, matrix sampling of items can be utilized to reduce testing time burden for individual students while still providing schools/districts with rich information about overall performance. Established testing programs that solely focus on group-level reporting, such as NAEP and PISA, already have a history of relying on matrix sampling and have begun to include SBTs within their assessments. Although matrix sampling results in item responses missing at random, it may have implications for effectively identifying LD among item pairs. Little research has studied the effect of matrix sampling, specifically balanced-incomplete-block (BIB) spiraling, on the performance of common LD detection statistics, such as Yen's $Q_3$, Mantel-Haenszel, or chi-squared statistics. This study fills that gap by conducting a simulation informed by NAEP test characteristics, finding that all such statistics have deflated power and inflated Type I error rates under BIB spiraling of items to students, particularly when SBTs form entire blocks. It also investigates the extent that ignoring LD affects subgroup inferences.

**Item Response Theory - Parallel Session: 4.5E: METRIC RECOVERY IN IRT SIMULATION STUDIES**

Leah Feuerstahler, Fordham University

In item response theory (IRT), Monte Carlo simulation studies are commonly used to evaluate the adequacy of new methods or the effects of model misspecification. The results of these studies are typically quantified in terms of item parameter accuracy or item response function accuracy (cf. Luecht & Ackerman, 2018). However, these measures are only indirectly related to the latent trait metric on which examinee scores are computed and reported, even though scores on the latent metric are often a primary outcome of IRT analyses. Moreover, when estimated item parameters are treated as fixed to estimate examinee scores, these scores are estimated on the recovered metric, which is not necessarily close the data-generating metric (Zhang, 2005; Zhang & Lu, 2007). In this talk, I propose a preliminary definition of a metric as a variable that makes a certain set of predictions on the response probability space. Metric recovery, therefore, is concerned with the extent to which the recovered metric makes the same predictions as the data-generating metric. Through a series of examples, I demonstrate how metric recovery can be used to establish minimum data requirements, the practical effects of model misspecification, and to determine ranges of the latent trait metric for which a calibrated item bank is well-determined. I argue that metric recovery is an important criterion that should be directly investigated in IRT simulation studies.

## Multilevel/Hierarchical/Mixed Models: 8:30 AM – 10:00 AM

Chair: Karl Schweizer

**Multilevel/Hierarchical/Mixed Models**

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 4.6A: TESTING VARIANCE COMPONENTS IN LINEAR MIXED MODELING USING PERMUTATION**
Han Du, University of California, Los Angeles; Lijuan Wang, University of Notre Dame

Linear mixed modeling (LMM) is widely used to deal with repeated measures, clustered subjects, or both in practice. In LMM, inference of variance components provides evidence of heterogeneity between individuals or clusters. When only nonnegative variances are allowed (constrained estimation), there is a boundary (i.e., 0) in the variances' parameter space. With the boundary issue, regular statistical procedures for inferring such a parameter could be problematic. We aim to introduce a practically feasible permutation method to make inferences about variance components while considering the boundary issue in linear mixed modeling. The permutation tests with different settings (i.e., constrained estimation vs. unconstrained estimation, specific test vs. generalized test, different ways of calculating p-values, and different ways of permutation) were examined with both normal data and non-normal data. In addition, the permutation tests were compared with the likelihood ratio test with mixtures of chi-squared distributions as reference distributions. In testing a subset of the variance components and testing all the variance components, the permutation tests and the likelihood ratio tests have their specific strengths and limitation respectively in different scenarios in terms of Type I error rates, statistical power, and availability of the methods. An example about the development of verbal IQ and performance IQ for 204 children based on the Wechesler Intelligence Scale for Children is used to illustrate the application of the permutation tests.

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 4.6B: A BETTER CORRECTION FOR UNRELIABILITY FOR META-ANALYSES**
Zijun Ke, Sun Yat-sen University; Xin Tong, University of Virginia

As a powerful tool for synthesizing information from multiple studies, meta-analysis has enjoyed high popularity in many disciplines. Conclusions stemming from meta-analyses are often used to direct theory development, calibrate sample size planning, and guide critical decision making or policy making. However, meta-analyses can be conflicted, misleading, and irreproducible. One of the reasons for meta-analyses to be misleading is the improper handling of unreliability. We will show that when reliabilities of the two focal variables are correlated, existing meta-analysis procedures with or without corrections for unreliability can frequently detect nonexistant effects, and provide biased estimates and intervals with coverage rates far below the intended level. A better approach to correcting for unreliability was proposed and evaluated via a simulation study.

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 4.6C: QUANTIFYING EXPLAINED VARIANCE IN MULTILEVEL MODELS: AN INTEGRATIVE FRAMEWORK OF R-SQUARED MEASURES**
Jason Rights, Vanderbilt University; Sonya Sterba, Vanderbilt University

Researchers often mention the utility and need for R-squared measures of explained variance for multilevel models (MLMs). Although this topic has been addressed by methodologists, the MLM R-squared literature suffers from several shortcomings: (1) analytic relationships among existing measures have not been established so measures equivalent in the population have been re-developed 2 or 3 times; (2) a completely full partitioning of variance has not been used to create measures, leading to gaps in the availability of measures to address key substantive questions; (3) a unifying approach to interpreting and choosing among measures has not been provided, leading to researchers' difficulty with implementation; and (4) software has inconsistently and infrequently incorporated available measures. We address these issues with the following contributions. We develop an integrative framework of R-squared measures for MLMs with random intercepts and/or slopes based on a completely full decomposition of variance. We analytically relate 10 existing measures from different disciplines as special cases of 5 measures from our framework. We show how our framework fills gaps by supplying additional measures that answer new

substantive research questions. To facilitate interpretation, we provide a novel and integrative graphical representation of all the measures in the framework; we use it to demonstrate limitations of current reporting practices for MLM R-squareds, as well as benefits of considering multiple measures from the framework in juxtaposition. We supply and empirically illustrate an R function, r2MLM, that computes all measures in our framework to help researchers in considering effect size and conveying practical significance.

**Multilevel/Hierarchical/Mixed Models - Parallel Session: 4.6D: COMPARING LINEAR MIXED MODELS WITH NON-NESTED RANDOM EFFECTS**

Ting Wang, The American Board of Anesthesiology; Huaping Sun, The American Board of Anesthesiology; Ann Harman, The American Board of Anesthesiology

In practice, it is a major challenge for researchers to compare linear mixed models (multilevel models) with non-nested random effects. This is especially true in social sciences, because the clustering situation is naturally complicated. For example, students are nested in schools, but also in neighborhoods. Comparing models with either one of the clustering or both clustering (cross random effect) is a non-trivial statistical question since the usual likelihood ratio statistic does not extend to non-nested models, and the information criteria such as AIC and BIC is prone to large variability at the sample size typically used in social sciences' linear mixed model contexts. In this presentation, we extend Vuong's likelihood ratio tests of non-nested models to linear mixed models via the framework of traditional regression model with correlated error terms. This approach enables formal comparisons of linear mixed models with non-nested random effects by utilizing case-wise log likelihood we recently developed in an R package. We also demonstrate the tests' performance under different modeling situations through simulations. The tests offer researchers a useful statistical tool for linear mixed models comparisons with various random effects specifications.

## Model Fit, Comparison and Diagnostics: 8:30 AM – 10:00 AM

Chair: Yanhoung Bian

**Model Fit, Comparison and Diagnostics**

**Model Fit, Comparison and Diagnostics - Parallel Session: 4.7A: REVISITING THE STATISTICAL ASSUMPTIONS IN TUCKER EQUATING**

Jorge González, Pontificia Universidad Católica de Chile; Marie Wiberg, Umeå University

When equating is performed using score data collected under the nonequivalent groups with anchor test design (NEAT), different assumptions are needed to estimate the parameters used to compute the equating transformation. The Tucker equating method makes two types of assumptions: i) both conditional expectation and conditional variances of tests scores given the anchor scores are assumed to be the same in both populations, ii) the conditional expectations are assumed to be linear functions while the conditional variances are assumed to be constants, pretty much like in ordinary linear regression analysis. Using data from two SAT administrations, Braun and Holland (1982) showed how it is possible to assess the second type of assumptions by comparing the sample conditional means of scores with those obtained by a linear regression model. These authors, however, did not show any assessment on the conditional variances assumption. In this paper we propose to use estimators of the conditional expectations that are smoother than the sample conditional means, and show how to assess the assumption of constant variance. A comprehensive simulation study is performed to compare the results of using sample conditional means, smoother nonparametric estimators of conditional expectations and linear regressions. The comparison and assessment of assumptions in Tucker equating is also performed on real data. The paper ends with some practical recommendations.

**Model Fit, Comparison and Diagnostics - Parallel Session: 4.7B: VALIDATING TEST SCORES AGAINST A FALLIBLE CRITERION**

Paul Jewsbury, Educational Testing Service

Criterion-related validation studies of diagnostic test scores are often complicated by the use of a reference test, such as clinical diagnosis of the construct of interest, rather than the construct of interest directly. Mostly commonly, the method of Known Group Validation is used that assumes the reference test is an infallible measure of the construct of interest, an assumption unlikely to be true. Methods for estimating the criterion-related validity of diagnostic test scores with a fallible reference test for the construct of interest are reviewed, including introducing new models and expanding and adapting existing models for the purpose of test validation. A type of latent class model, Mixed Group Validation, allows for infallible reference tests (Frederick, 2000). However, many constructs of interest do not conform to the assumptions of a latent class, and several Mixed Group Validation studies have been shown to be biased as a consequence (Jewsbury & Bowden, 2013). The Neighborhood model is adapted for diagnostic test score validation as a method that allows for infallible reference tests but does not assume the construct is a latent class. As both Mixed Group Validation and the Neighborhood model make strong assumptions that may be true only for special cases, the Method of Bounds for Test Score Validation was developed as an interval-based method for more general use. Simulations were conducted that identified the relative strength of the alternate methods across a range of plausible research study designs, and the results were summarized.

**Model Fit, Comparison and Diagnostics - Parallel Session: 4.7C: MEASUREMENT ERROR, RELIABILITY, AND PREDICTIVE ACCURACY OF HIGHEST SCORES**

Dongmei Li, ACT, Inc.; Deborah J. Harris, The University of Iowa

Academic achievement test programs often report a composite score as well as scores on each subject test. For students sending multiple sets of test scores for admissions, an institution might choose to use the highest reported composite scores or even the "super scores" obtained by combining the highest subject test scores from each test event for decision making. Previous research has revealed some interesting but seemingly contradictory results regarding these scores. For example, while they were shown to be biased estimates of student performance, highest and super composite scores were also shown to be more reliable (as defined by the squared correlation between true and observed scores) (Li, 2017) and slightly more accurate in prediction of future performance (Mattern, Radunzel, Bertling, & Ho, 2017) than using the most recent scores. This study is intended to provide explanations or resolutions to some of the seemingly contradictory results revealed in previous research by exploring the following research questions:

1. With different ways to calculate score reliability, which ones are appropriate for highest or super scores?
2. In what situations do biased estimates have higher reliability?
3. How are measurement error and reliability related to predictive validity?

These questions will be addressed both theoretically and empirically using simulated data. Results from this study will shed light on the relationships between measurement error, reliability, and predictive validity of highest or super composite scores to inform decisions about appropriate uses of these scores, informing college admission practices and other high stakes scenarios.

**Model Fit, Comparison and Diagnostics - Parallel Session: 4.7D: HOW TEST BLUEPRINTS AND MEASUREMENT MODELS IMPACT SUBSCORE VALUE**

Tyler Matta, University of Oslo; Kondwani Kajera Mughogho, University of Oslo

It is common for educational tests within a single content area or subject to produce both subject/content area scores and underlying domain-level scores (subscores), even if the test blueprint does not account for such a domain structure. Interest in domain-level scores has spawned two areas of psychometric research, a) models for estimating overall and domain-level scores (e.g., de la Torre & Song, 2009; Rijmen, et al., 2014; Yao, 2010; Zhang & Stout, 1999) and b) subscore value (e.g., Haberman 2008; Haberman & Sinharay, 2010). With that, there has been little research on how a test blueprint and measurement model interact to impact subscore value. This paper draws on simulated data to understand the unique and interactive impact of a test blueprint and measurement model on subscore

value. Specifically, the simulation will focus on a single subject-level trait comprised of three domain traits. Three domain-specific item banks will be used to build two tests, one using a unidimensional blueprint (focused on the subject level trait) and the other using a multidimensional blueprint (focused on the domain-level traits). Item parameters, subject-level traits, and domain-level traits will be estimated using a) consecutive UIRT, b) composite UIRT, c) MIRT with maximum information overall scores (Yao & Schwarz 2006), and d) higher-order IRT. Subscore value will be estimated for each test-measurement model combination using PRMSE (Haberman & Sinharay, 2010). In addition to PRMSE, subject-level scores, domain-level scores, and item parameters will be compared.

**Model Fit, Comparison and Diagnostics - Parallel Session: 4.7E: CLASSIFICATION ACCURACY FOR NONPARAMETRIC CLASSIFICATION APPROACHES IN COGNITIVE DIAGNOSIS**
Yanhong Bian, Rutgers, the State University of New Jersey; Yan Sun, Rutgers, the State University of New Jersey

Cognitive diagnosis models (CDMs) and several nonparametric classification methods (Chiu, Sun, & Bian, in press; de la Torre, 2011; Henson, Templin, & Willse, 2009; DiBello, Roussos, & Stout, 2007) have been developed to classify examinees to the proficiency classes they belong based on the cognitive attributes they prosses. There are many classification consistency and accuracy indices developed in the literature, and some approaches are also proposed to measure the consistency and accuracy under CDM (Cui, Gierl, & Chang, 2012; Templin & Bradshaw, 2013; Lee, 2010). However, the classification consistency and accuracy for nonparametric classification approach in cognitive diagnosis has seldom been proposed. This study introduces the classification accuracy indices under the general nonparametric classification method (GNPC; Chiu, et al., in press) which can be applied to the saturated CDMs. One method measures the distance between the observed item response pattern and the final estimated ideal response pattern over all latent classes. Another method utilizes the estimated weight, which is essentially the proportion of correct responses for each latent class, as well as the estimated latent class distribution and converts them into a pseudo-posterior measure. The performance of two indices will be investigated using simulation study by comparing with model-based indices (Cui et al., 2012; Templin & Bradshaw, 2013).

## Refreshment Break: 10:00 AM – 10:15 AM

## State of the Art: 10:15 AM – 11:00 AM

Chair: Daniel Bolt

**Equivalent Dynamic Models**
State of the Art: Peter Molenaar

Chair: Ulf Bockenholt

**Response Styles and Dispersion in Regression and Item Response Models**
Invited Speaker: Gerhard Tutz

## Career Award for Lifetime Achievement: Kenneth A. Bollen: 11:00 AM – 12:00 PM

Chair: Sophia Rabe-Hesketh

**Specify Globally, Estimate and Test Locally: A MIIV Approach to Structural Equation Models**
Kenneth A. Bollen, University of North Carolina, Chapel Hill

## Meeting of Members: 12:00 PM – 12:30 PM

## Attendee Lunch: On Your Own: 12:30 PM – 1:30 PM

# Wednesday, July 11, 2018 PM

## Keynote Speaker: 1:30 PM – 2:30 PM

Chair: Klaas Sijtsma

**Large-Scale Probabilistic Modeling**
Keynote Speaker: David Blei, Columbia University

## Refreshment Break: 2:30 PM – 3:00 PM

## Symposium 9: 3:00 PM – 4:30 PM

Chairs: Eva Ceulemans; Janne Adolf

**Symposium 9: Timely Perspectives on Dynamic Models for Time Series and Panels**

**Symposium 9 - Parallel Session: 5.1A: UNDERSTANDING THE TIME COURSE OF INTERVENTIONS VIA CONTINUOUS TIME MODELING**
Charles Driver, Max Planck Institute for Human Development

How long does a treatment take to reach maximum effect? Is the effect maintained, does it dissipate, or perhaps even reverse? Do certain sorts of people respond faster or stronger than others? Is the treatment more effective in the long run for those that respond quickly? I describe a continuous time dynamic modelling approach for considering the potentially complex shape of intervention effects over time, as well as mediation and individual differences in such a context, with examples using the R software ctsem.

**Symposium 9 - Parallel Session: 5.1B: GAUSSIAN PROCESS PANEL MODELING R PACKAGE GPPM**
Julian Karch, Leiden University

Recently, we have introduced Gaussian Process Panel (GPPM) as a powerful modeling framework for longitudinal panel data. The main advantage of GPPM is its flexibility. Most popular modeling approaches for time-series and longitudinal panel data can be considered special cases of GPPM. This not only includes structural equation modeling and hierarchical linear modeling but also state-space modeling in its time-discrete and its time-continuous variant, as well as generalized additive models. Consequently, GPPM is perfectly suited for investigating the dynamics of individual change. In this talk, we will present the new R package gppm that makes GPPM easily accessible to applied researchers. During the presentation, we will focus on the gppm package's capabilities for modeling the dynamics of individual change.

**Symposium 9 - Parallel Session: 5.1C: TAILORING KERNEL CHANGE POINT DETECTION TO CAPTURE ABRUPT AUTOCORRELATION SHIFTS**

Jedelyn Cabrieto, Katholieke Universiteit Leuven; Francis Tuerlinckx, Katholieke Universiteit; Peter Kuppens, Katholieke Universiteit, Leuven; Janne Adolf, Katholieke Universiteit Leuven; Eva Ceulemans, Katholieke Universiteit

Change point detection methods capture abrupt changes in a time series. Non-parametric variants of this approach such as Kernel Change Point (KCP) detection are attractive especially when information on the underlying distribution of the data is limited. Yet, these methods can signal any parameter change (mean, variance, autocorrelations, etc.), leaving the user clueless as to what parameter has changed when a change point is signaled. This is an important drawback from a substantive perspective where changes in a pre-specified parameter are of interest. We will demonstrate that KCP can be adapted to detect change points in specific parameters by implementing it on the running statistics rather than the raw data. These running statistics are obtained by sliding a window across the time series, and in each window, extracting the statistic value. In this talk, we will focus on the AR(1) parameter, which in emotion research, has been linked to emotional inertia, the extent to which feelings carry over from one moment to the next. Recent cross-sectional findings demonstrated that emotional inertia is related to psychopathology. Therefore, it makes sense to investigate at the within-person level whether the onset of mental health problems is preceded by a change in emotional inertia. Through a simulation study, we will assess the performance of our proposed approach and compare it to a parametric regime switching model. We also provide an illustrative example where we reanalyze the data of Wichers and Groot (2016).

**Symposium 9 - Parallel Session: 5.1D: SCALING TIME IN AUTOREGRESSIVE MODELS**

Janne Adolf, Katholieke Universiteit Leuven; Francis Tuerlinckx, Katholieke Universiteit; Eva Ceulemans, Katholieke Universiteit

Theories and models that conceive of psychological phenomena as dynamic processes are gaining ground, and the time scale, at which these unfold, becomes a topic of interest. Scaling time in studying psychological dynamics is a multifaceted problem, but can be addressed as a statistical one in the context of a formal dynamic model. The popular vector autoregressive (VAR) model makes clear assumptions about how the parameterized dynamics depend on the time scale. Recently available continuous-time formulations explicate these dependencies and thus promise the translation of results obtained at a specific time scale to any other. We probe this promise by investigating how well VAR model parameters can be estimated from time series data sampled at different rates. To this end, we take an information-theoretic perspective and quantify the amount of information present in the data with respect to the parameters by the Fisher information matrix (FIM). The FIM serves as a gradual measure of practical model identifiability, that is, the estimability of a per se identified model given data. By analytic and simulation results, we show that slowing down the sampling rate leads to VAR model parameters that become less and less identifiable at a given time scale of interest. While a model with less identifiable parameters may still be estimable, solutions can become highly inaccurate and variable. Additionally, distinct parameters can become highly correlated. This affects statistical power and hampers the interpretability of modeling solutions. We discuss possibilities to prevent such problems by optimizing study design.

## Symposium 10: 3:00 PM – 4:30 PM

Chair: Yunxiao Chen

**Symposium 10: Advances in Cognitive Diagnostic Assessment**

**Symposium 10 - Parallel Session: 5.2A: TWO-TIER LATENT TRAIT MODELS FOR COGNITIVE DIAGNOSIS**

Hyeon-Ah Kang, University of Texas at Austin

Over the past several decades, research has led to sophisticated models that provide cognitive profiles of mastery and nonmastery on a set of predefined attributes. These models, commonly referred to as cognitive diagnostic models (CDMs), characterize an individual's proficiency on the basis of attributes

assessed by test items. The major focus of the existing CDMs has been on classifying examinees into several latent classes. The possibility of simultaneously evaluating both the overall test performance and specific sub-skill mastery from a single test administration has seldom been explored. The present study proposes new statistical modeling that provides information about both the generic ability levels and specific cognitive profiles. The model assumes that the ability and attribute profiles preserve an equal standing in explaining the covariation among the observable indicators. Because all latent variables have immediate influences on the item performance, the model is named a two-tier CDM. The present modeling allows for direct links between the generic ability parameters and item indicators, and hence, one can achieve fine-grained estimation of ability parameters by capitalizing on a large number of observable variables. The study provides a contrasting perspective from the higher-order CDM (de la Torre & Douglas, 2004) where the generic ability parameters are connected only through the attribute profiles. The study presents simulation studies as well as empirical data analysis to validate the proposed framework. All experimentation is implemented in comparison with the higher-order CDM.

**Symposium 10 - Parallel Session: 5.2B: JOINT MAXIMUM LIKELIHOOD ESTIMATION FOR HIGH-DIMENSIONAL EXPLORATORY ITEM RESPONSE ANALYSIS**

Xiaoou Li, University of Minnesota

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. In this paper, 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 data sets. The proposed method is applied to a large-scale personality assessment data set from the Synthetic Aperture Personality Assessment (SAPA) project. Simulation studies are conducted to evaluate the proposed method.

**Symposium 10 - Parallel Session: 5.2C: A REINFORCEMENT LEARNING APPROACH TO PERSONALIZED LEARNING RECOMMENDATION SYSTEM**

Xueying Tang, Columbia University; Yunxiao Chen, Emory University; Xiaoou Li, University of Minnesota; Jingchen Liu, Columbia University; Zhiliang Ying, Columbia University

Personalized learning refers to instruction in which the pace of learning and the instructional approach are optimized for the needs of each learner. With the latest advances in information technology and data science, personalized learning is becoming possible for anyone with a personal computer, supported by a data-driven recommendation system that automatically schedules the learning sequence. The engine of such a recommendation system is a recommendation strategy that, based on data from other learners and the performance of the current learner, recommends suitable learning materials to optimize certain learning outcomes. A powerful engine achieves a balance between making the best possible recommendations based on the current knowledge and exploring new learning trajectories that may potentially pay off. Building such an engine is a challenging task. We formulate this problem under the Markov decision framework and propose a reinforcement learning approach to solving the problem.

**Symposium 10 - Parallel Session: 5.2D: IDENTIFIABILITY OF RESTRICTED LATENT CLASS MODELS**

Gongjun Xu, University of Michigan; Yuqi Gu, University of Michigan

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.

**Symposium 10 - Parallel Session: 5.2E: A JOINT MIXTURE LEARNING MODEL FOR RESPONSES AND RESPONSE TIMES**
Susu Zhang, University of Illinois at Urbana-Champaign; Shiyu Wang, University of Georgia

The increased popularity of computer-based testing has enabled researchers to collect various types of process data, including test takers' reaction time to assessment items, also known as response times. Extensive research has been conducted on the joint modeling of response accuracy and response times, which can improve the estimation accuracy of item parameters and examinees' latent traits or latent classes, further our understanding of individuals' test-taking behavior and the test items' characteristics, and help us differentiate examinees using different test-taking strategies. Recent researches in Diagnostic Classification Models begin to explore the relationship between speed and accuracy to understand students' fluency of applying the mastered skills, in addition to mastery information, in a learning environment. This can be achieved by modeling the changes in response accuracy and response times throughout the learning process. We propose a mixture hidden Markov Diagnostic Classification Model framework for learning with response times and response accuracy. Such a model accounts for the heterogeneities in learning styles among students by modeling the different learning and response behaviors among subgroups, and it may provide instructors with valuable information that can be used to design individualized instructions. A Bayesian modeling framework is developed for parameter estimation, and the proposed model is evaluated through a simulation study and is fitted to a real dataset collected from a computer-based learning system for spatial rotation skills.

## Computer-Based Testing: 3:00 PM – 4:30 PM

Chair: David Magis

**Computer-Based Testing**

**Computer-Based Testing - Parallel Session: 5.3A: ASYMMETRY IN FIXED-PRECISION M-CAT: MULTIDIMENSIONAL SELECTION VERSUS MARGINAL STOPPING**
Johan Braeken, CEMO, University of Oslo; Muirne Paap, University of Groningen

Standard implementations of a Multidimensional Computerized Adaptive Testing (M-CAT) algorithm have item selection rules that are searching for items that optimize the Fisher information volume. A variable-length M-CAT would usually include a stopping rule requiring all dimensions being measured with a fixed minimum precision. In contrast to the inherently multidimensional selection rule, this stopping rule is defined at the marginal levels of the latent traits distribution: standard error smaller than a pre-determined threshold value for each dimension. This asymmetry between selection rule and stopping rule leads to side-effects that might not always be anticipated at first glance. We will first revisit and discuss the issue from a distribution and practical perspective, subsequently propose some work-arounds in the form of alternative selection rules, and elaborate on their effectivity to tackle the issue in practice.

**Computer-Based Testing - Parallel Session: 5.3B: INVESTIGATING THE ITEM SELECTION METHODS IN VARIABLE-LENGTH CD-CAT**
Ya-Hui Su, National Chung Cheng University; Hua-Hua Chang, University of Illinois, Urbana-Champaign

Cognitive diagnostic computerized adaptive testing (CD-CAT) not only obtains useful cognitive diagnostic information measured in psychological or educational assessments, but also has great efficiency brought by computerized adaptive testing. At present, there are only a limited numbers of previous studies examining how to optimally construct cognitive diagnostic test. The cognitive discrimination index (CDI; Henson & Douglas, 2005) and attribute-level discrimination index (ADI; Henson, Roussos, Douglas, & He, 2008) has been proposed to assemble cognitive diagnostic tests. The CDI measures an item's overall discrimination power whereas the ADI measures an item's discrimination power for a specific attribute. It is challenging, while constructing assessments in CD-CAT, to meet various constraints simultaneously. The priority index approach (Cheng & Chang, 2009; Cheng, Chang, Douglas, & Guo, 2009) was proposed to manage many constraints simultaneously in CAT. This approach could be not only implemented easily but also computed efficiently. Su and Chang (2017) firstly

integrated this approach with the posterior-weighted CDI and ADI, constraint-weighted posterior-weighted CDI (CW-PWCDI) and constraint-weighted posterior-weighted ADI (CW-PWADI), for test construction in fixed-length CD-CAT, and they found examinees yielded different precision. In reality, if the same precision of test results is required for all the examinees, some examinees need to take more items and some need to take fewer items than others do. Therefore, this study was to investigate the performance of the CW-PWCDI and CW-PWADI in variable-length CD-CAT through simulations.

**Computer-Based Testing - Parallel Session: 5.3C: EXPLORING A NEW RESCORING METHOD FOR FLAWED ITEMS IN CAT**

Chunxin (Ann) Wang, ACT, Inc.; Yi He, ACT, Inc.; Jie Li, ACT, Inc.; Jin Zhang, ACT, Inc.

Occasionally in operation, due to some unexpected reasons, an item that was intended for scoring may be identified as flawed in administration; consequently, rescoring is needed. The commonly used rescoring methods in paper & pencil tests are removing the item from scoring and rescoring the item as correct. In computer adaptive tests (CAT), these rescoring methods were still found to work well (Potenza and Stocking, 1997). To identify which method produced more accurate rescoring results, Wang et al. (2016) conducted simulations under a Scripted CAT (McKinley et al., 2014; Lee et al., 2014, Wang et al., 2017), in which included the two commonly used rescoring methods and the method of rescoring the item as wrong. They examined the interactions with the factors of item's position, IRT statistics and examinees' ability levels. They found that the rescoring method of removing the item from scoring yielded the smallest errors. Across all factors, they found the larger impact of rescoring was for examinees at the high and the low ability levels. With the purpose of making the re-scoring method more accurate and being fair to students from all ability levels, a new rescoring method is proposed, called ability-matching method. With this method, the flawed item is given a score of 0 or 1 based on whether the examinee's estimated ability is less than or greater than the flawed item's IRT-b. Simulations will be conducted to compare the proposed method and other rescoring methods with different factors under the Scripted CAT.

**Computer-Based Testing - Parallel Session: 5.3D: DERIVING STOPPING RULES FOR ADAPTIVE RATER MONITORING**

Zhuoran Wang, University of Minnesota, Twin Cities

Adaptive rater monitoring was developed to efficiently evaluate and monitor rater performance before and during operational rating. The previous studies focus on the fixed-length tests. Due to the limited number of items in the item bank, the estimate accuracy of raters whose rater parameters widely spreading out are compromised. Therefore, instead of stopping the test with a predetermined fixed test length, the authors use a more informative stopping criterion that is directly related to measurement accuracy. Specifically, this research derives seven stopping rules. Some of them quantify the measurement precision of the rater parameter vector (i.e., minimum determinant rule [D-rule], minimum eigenvalue rule [E-rule], and maximum trace rule [T-rule]), some others quantify the amount of possible change that can be caused by each item (i.e., biggest change of rater parameters [Cb-rule] and average change of rater parameter[Ca-rule]), the rest quantify the accuracy of rater effect estimate (i.e., variance of leniency effect estimate [Vl-rule] and variance of centrality effect estimate [Vc-rule] ). The simulation results showed that all seven stopping rules successfully terminated the test when the mean squared error of rater parameter estimation is within a desired range, regardless of rater parameter values. It was found that when using the Vc-rule, raters with large parameter spread tended to have tests that were twice as long as the tests received by raters with small parameter spread. However, the test length difference with the other rules is not very dramatic.

## Measurement Invariance and DIF: 3:00 PM – 4:30 PM

Chair: David Hessen

**Measurement Invariance and DIF**

**Measurement Invariance and DIF - Parallel Session: 5.4A: THE MANY FLAVORS OF SCORE-BASED TESTS FOR DIFFERENTIAL ITEM FUNCTIONING**
Rudolf Debelak, University of Zurich; Lennart Schneider, University of Tuebingen; Achim Zeileis, University of Innsbruck; Carolin Strobl, University of Zurich

The invariance of the item parameters across the population is a central assumption of most item response theory (IRT) models, and the detection of its violation (differential item functioning, DIF) is of high practical relevance. Several recent papers have suggested score-based tests to detect DIF effects in the context of IRT. These tests are a data-driven approach for detecting parameter invariance along person covariates of interest (e.g. gender or age), and do not require the definition of focal or reference groups like many alternative methods. These tests are adapted to IRT models with several parameters, giving emphasis to the two-parameter logistic (2PL) model. The modified tests allow to specify which model parameters, e.g., the difficulty and/or the discrimination parameters, are assessed to detect violations of measurement invariance along which person covariates. Using simulation studies, the modified tests are evaluated in the context of the 2PL model. The results of this evaluation show that the tests can detect a wide range of DIF types and that the proposed tests are sensitive against the particular DIF types of interest. The consequences for the tests' application in practice are discussed based on these insights.

**Measurement Invariance and DIF - Parallel Session: 5.4B: A COMPARISON OF ANCHOR METHODS FOR DETECTING DIF IN MULTIPLE GROUPS**
Thorben Huelmann, University of Zurich; Carolin Strobl, University of Zurich

In order to test individual items of a psychological test for Differential Item Functioning (DIF), a common strategy is to first select a set of anchor items. Ideally these items are DIF-free and can be used to align the scales of the two groups. If the anchor items are not handpicked by content experts, some kind of heuristic or statistical test is needed to select them. Kopf, Zeileis & Strobl (2015) describe a variety of approaches for selecting anchor items for settings with two groups, that all depend on some form of ranking the candidate anchor items. However, these approaches cannot be directly applied in multiple-group scenarios, where the number of summary statistics determining the ranking of each item depends on the number of focal groups. In order to apply the existing anchoring approaches, some kind of aggregation rule needs to be applied. In this talk three aggregation rules will be presented and their impact on the anchor selection process as well as on the final DIF analysis will be illustrated and discussed.

**Measurement Invariance and DIF - Parallel Session: 5.4C: COMPARING DIFFERENTIAL DISTRACTOR FUNCTIONING DETECTION METHODS UNDER THE NESTED LOGIT MODELING FRAMEWORK**
Armi Lantano, Center for Educational Measurement, Inc.; Kevin Carl Santos, The University of Hong Kong

Differential distractor functioning (DDF) analyses are employed to determine whether different distractors attract various groups disproportionately. Although incorrect answers were not the focus before, the results of DDF analyses can be used in identifying distractors that could possible cause differential item functioning (DIF), thus, providing valuable information for potential item revisions or design of new tests for multiple choice items. Under the nested logit modeling framework, this study compares three DDF detection methods, namely, the odds ratio approach (Terzi and Suh, 2015), standardization approach (Dorans et al., 1992) and the log-linear approach (Green et al., 1989). These performance of these approaches are evaluated and compared using a simulation study. Test length, sample size, significance level, and DDF/DIF magnitude and pattern are the factors manipulated in the simulation study.

**Measurement Invariance and DIF - Parallel Session: 5.4D: LOCALIZED ITEM RESPONSE THEORY (LIRT): DETECTING REGIONAL DIFFERENCES IN ITEM FUNCTIONALITY**
Samantha Robinson, University of Arkansas

Mappings of spatially-varying Item Response Theory (IRT) parameters are proposed, allowing for investigation of potential uniform and non-uniform Differential Item Functioning (DIF) based upon geographic location without need for pre-specified groupings and prior to confirmatory DIF testing. This localized approach to IRT modeling provides a flexible framework, with current emphasis being on 1PL/Rasch and 2PL models. Applications to both simulated examination data and empirical survey data are presented to demonstrate the method, illustrate its benefits, and advocate for the use of local IRT modeling in a variety of contexts such as in the analysis of International Large-Scale Assessment (ILSA) data for education or in the analysis of Patient Reported Outcome Measure (PROM) data for use in Mobile Health Applications (MHA). There is not only practical value with this method but also visual appeal when initial attempts to consider measurement invariance are made across national, state, or other political boundaries. The approach and the visualization it affords have great potential for affecting policy; regional disparities and latent spatial trends in item functionality can be identified and used in a beneficial manner regardless of geographic location (e.g., to increase educational opportunity or to increase access to quality health care while reducing associated costs of that health care).

## Item Response Theory: 3:00 PM – 4:30 PM

Chair: Lale Khorramdel-Ameri

**Item Response Theory**

**Item Response Theory - Parallel Session: 5.5A: A MODEL FOR TRUE-FALSE EXAMS BASED ON SIGNAL DETECTION THEORY**
Lawrence DeCarlo, Teachers College, Columbia University

Exams can be viewed as being a signal detection task. For example, in a true-false test, the examinee's task is to detect whether an item is true (signal) or false (noise); in a multiple choice test, the task is to detect the correct answer (signal) among distractors (noise). Signal detection theory (SDT) views these tasks as consisting of two basic psychological components: a perceptual component – for example, the perceived plausibility of an item on a true-false test – and a decision component – for example, the use of a decision threshold that delineates a response of true versus false. The item-response model that follows directly from this conceptualization is derived. An interesting result is that the SDT-IRT model is closely related to IRT models, and in particular the 2-parameter logistic model (2PL). However, there are also important distinctions, in that SDT-IRT views 'difficulty' and 'guessing' as arising from the same process – item bias – and so a single parameter accounts for both, rather than two parameters as in the 3PL. The SDT-IRT models are fit using Bayesian estimation and are compared to IRT models both in simulations and in real-world data.

**Item Response Theory - Parallel Session: 5.5B: AN IRT TREE MODEL WITH NOT-ALL-DISTINCT LEAVES FOR NON-RESPONSE MODELING**
Yu-Wei Chang, Feng Chia University; Nan-Jung Hsu, National Tsing-Hua University, Taiwan

An IRT tree model with 4 leaves of not-all-distinct response categories is proposed for non-response modeling. In particular, the model in Knott et al. (1990) is further extended to have more than one path of branches resulting in responses belonging to the same response category. Potential applications of the proposed model will be given in our talk. The model does not fall within the multi-dimensional item response model framework in Jeon & de Boeck (2016), and its estimation is an issue. Penalized quasi-likelihood estimation procedure is suggested. Simulations are conducted to demonstrate the validation of the estimation procedure and the advantage of the proposed model. The model is further applied to an entrance examination data set for illustration.

**Item Response Theory - Parallel Session: 5.5C: A MULTILEVEL TESTLET MODEL FOR RESPONSE AND RESPONSE TIME**

Evan Olson, University of Maryland, College Park

With the widespread use of computer-based assessment (CBA), process data, including item response time, may be readily obtained. Such process data may be combined with product data, item responses, to inform respondent speed and accuracy. Klein Entink, Fox, & van der Linden (2009) described a joint multilevel approach for modeling response and response time. This multilevel approach addressed the person clustering often employed in large-scale assessment such as multinational studies. For their model, three levels were specified. Jiao, Kamata, Wang, & Jin (2012) discussed a multilevel dual dependence model supporting the benefits of representing item clustering in addition to person clustering for item responses. The model included four levels. In a proposed new model, a dual person and item clustering approach for item responses is jointly modeled with item response time. The model is also a four-level multilevel model. Including the item clusters, known as testlets, has been shown to improve parameter estimation in item response models. The new model affords the ability to characterize and accommodate for the violation of item independence assumed in traditional item response models. To demonstrate the plausibility of the proposed model, a simulation using Bayesian Markov Chain Monte Carlo (MCMC) estimation will be performed to evaluate parameter accuracy and compare alternative models. An application using a 2015 dataset from the Programme for International Student Assessment (PISA) with estimation of a two-parameter logistic measurement model will also be presented.

**Item Response Theory - Parallel Session: 5.5D: PSYCHOMETRIC MODELS OF SMALL GROUP COLLABORATIONS: TOWARDS A GENERAL THEORY**

Peter Halpin, New York University; Peter van Rijn, ETS

Social combination theory represents group performance using (a) a model for the performance of individual group members, and (b) a decision function (linear map) from individual performance to group performance. Halpin and Bergner (in press) treated individual performance using a standard item response theory (IRT) model and showed how to estimate parameters of a restricted class of decision functions. In this presentation we work towards more general results on the decision function. We consider conditions under which the parameters of the decision function are identified at the group level, and at the level of individual group members. We also consider model restrictions that ensure latent monotonicity of the group item response functions, and analyze the item information function of the decision parameters to provide insights about optimal test design. The results provide a theoretical basis for maximum likelihood estimation of a relatively large class of social combination IRT models, the utility of which is illustrated with real-data examples involving dyads.

**Item Response Theory - Parallel Session: 5.5E: A TAXONOMY OF ITEM RESPONSE MODELS IN PSYCHOMETRIKA**

Seock-Ho Kim, University of Georgia

Articles on Psychometrika for last 30 years are sorted based on two classification frameworks by Thissen & Steinberg (1986) and van der Linden (2016a) (cf. Nering & Ostini, 2010). Articles are also further sorted by the parameter estimation methods (e.g., Baker & Kim, 2004; de Ayala, 2009) as well as the computer programs used to implement the estimation methods (e.g., Hambleton, Swaminathan, & Rogers, 1991, pp. 159-160; van der Linden, 2016b). Articles from other journals in education and psychology are also reviewed to explore practical use of various item response models. Guiding principles for the taxonomy (cf. Bloom, 1956) are discussed.

## Multivariate Analysis: 3:00 PM – 4:30 PM

Chair: Steffi Pohl

**Multivariate Analysis**

**Multivariate Analysis - Parallel Session: 5.6A: ONE DIRECTION? ON THE MODELING OF CIRCULAR DATA IN PSYCHOLOGY**
Jolien Cremers, Utrecht University; Irene Klugkist, Utrecht University

In psychology, there are numerous examples of data that can be regarded and analysed as circular data. Research areas from psychology in which circular variables arise include: personality measurement, cognitive maps, visual perception of space, visual working memory and more. Despite there being numerous examples of circular data being collected in different areas of psychology, the knowledge of this type of data is not well-spread and literature in which these types of data are analysed with more complex methods than a one-way ANOVA for circular data is scarce. More complex models for circular data however do exist. Among these models are a projected normal regression and mixed-effects model. The interpretation of effects from these model is not straightforward and we introduce new tools that alleviate this problem. Additionally, an R-package was created that allows applied researchers to use the new interpretation tools and fit both projected normal regression models and mixed-effects models. By means of step-by-step analyses of example data from the field of cognitive psychology we outline the use of the tools from the package and their usefulness to applied researchers.

**Multivariate Analysis - Parallel Session: 5.6B: SUPERVISED CLASSIFICATION WITH MATRIX SKETCHING**
Roberta Falcone, University of Bologna; Laura Anderlucci, University of Bologna; Angela Montanari, University of Bologna

Matrix sketching is a data compression technique that has been recently developed in the computer science community. An input matrix A is efficiently approximated with a smaller matrix B, so that B preserves most of the properties of A up to some guaranteed approximation ratio. In so doing numerical operations on big data sets become faster. Sketching algorithms generally use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. The statistical properties of sketched regression algorithms have been widely studied in Woodruff (2014) and in Ahfock, Astle and Richardson (2017). In this work, we study the performances of sketching algorithms in the supervised classification context, both in terms of misclassification rate and of boundary approximation, as the degree of sketching increases. We also address, through sketching, the issue of unbalanced classes, which hampers most of the common classification methods.

**Multivariate Analysis - Parallel Session: 5.6C: POWER ANALYSIS FOR ANCOVA DESIGNS**
Gwowen Shieh, National Chiao Tung University

The analysis of covariance (ANCOVA) provides a useful approach for combining the advantages of two widely recognized procedures of multiple regression and analysis of variance (ANOVA). There are numerous published sources that address statistical theory and applications of power analysis for linear regression and ANOVA. However, relatively little research has attempted to address the corresponding issues for ANCOVA. This article discusses the general distributions and power calculations of the overall test of treatment effects under the two frameworks of fixed and random covariates. Currently available formula relies exclusively on the simple result of ANOVA with a deflated variance to take into account the relationship between the response and covariate variables. Under the assumptions of a priori specified covariate values and multinormal distributed covariate variables, the exact power functions of the omnibus test are derived. The obtained power functions under fixed and random modeling structures reveal that the existing procedure suffers the essential problem of omitting the influence of covariate properties. According to the analytic justification and empirical assessment, the suggested approaches have a clear advantage of accurate power estimation over the approximate method. Computer codes are also presented to implement the recommended power calculation and sample size determination in planning ANCOVA studies.

## Applications: 3:00 PM – 4:30 PM

Chair: Gunter Maris

**Applications: Applications**

**Applications - Parallel Session: 5.7A: MEASURING STUDENT'S PROFICIENCY IN MOOCS: MULTIPLE ATTEMPTS EXTENSIONS FOR THE RASCH MODEL**
Dmitry Abbakumov, Katholieke Universiteit, Leuven; Wim Van den Noortgate, Katholieke Universiteit, Leuven; Piet Desmet, Katholieke Universiteit, Leuven

Popularity of massive open online courses (MOOCs) has been growing intensively. Students, professors, and universities have an interest in accurate measures of students' proficiency in MOOCs. However, these measurements face several challenges: (a) assessments are dynamic: items can be added, removed or replaced by a course author at any time; (b) students may be allowed to make several attempts within one assessment; (c) assessments may include an insufficient number of items for accurate individual-level conclusions. Therefore, common psychometric models and techniques of CTT and IRT do not serve perfectly to measure proficiency. In this study we try to cover this gap and propose cross-classification multilevel logistic extensions of the common IRT model, the Rasch model, aimed at improving the assessment of the student's proficiency by modeling the effect of attempts and by involving non-assessment data such as student's interaction with video lectures and practical tasks. We illustrate these extensions on the logged data from one MOOC and check the quality using a cross-validation procedure on three MOOCs. We found that (a) change in the performance over attempts depends on both students and items; (b) student's activity with video lectures and practical tasks are significant predictors of response correctness; (c) overall accuracy of prediction of student's item responses using the extensions is 6% higher than using the traditional Rasch model. In sum, our results show that the approach is an improvement in assessment procedures in MOOCs and could serve as an additional source for accurate conclusions on student's proficiency.

**Applications - Parallel Session: 5.7B: USING CPA-BASED WEIGHTED RESIDUAL METHOD TO DETECT CARELESSNESS**
Xiaofeng Yu, University of Notre Dame

Careless or inattentive responding is a frequently observed aberrant response behavior in research based on questionnaires or surveys, which jeopardizes test validity and the 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), especially in some online survey data. Change point analysis (CPA), which is a widely used statistical process control method, can be applied to detect aberrant behaviors. In this research, we propose a CPA-based weighted residual test to detect BRR behavior. The performance of this CPA-based method was evaluated in a comprehensive simulation study, as well as in an empirical study. The critical values used in the studies were based on a Monte Carlo simulation. Type I error rates and detection rates were evaluated. Results showed that the critical values have a small fluctuation with the increase of test length with the condition of known item parameters. When item parameters were known, the CPA method could obtain a relative high detection rates with type I error rates close to the corresponding nominal level. When item parameters were unknown, the careless responses can have negative effect on parameter estimates, and so as to the detection rates. The data cleansing process can improve the detection rates. Take 120-item test with 10% BRR prevalence as an example, under different BRR severities, the detection rates for the known-item-parameter, unknown-item-parameter without and with data cleansing are as least as .86, .77, and .90, respectively.

**Applications - Parallel Session: 5.7C: RAPID GUESSING IN TECHNOLOGY-ENHANCED ITEM**
Rong Jin, Houghton Mifflin Harcourt; Johnny Denbleyker, Houghton Mifflin Harcourt; JP Kim, Houghton Mifflin Harcourt

The rapid guessing (RG) is a common behavior of test takers when they face somewhat difficult items. Then, examinees usually give responses in very short time without interacting with items. Such disengaged responses not only provide no valuable information about the examinees' achievement

levels, but also decrease the precision of item parameter estimates. How to identify RG behavior and how to improve item calibration with emerge of RG responses are very meaningful research questions. The present study focuses on technology-enhanced items (TEI) and explores these research areas. Several RG identification methods have been proposed but how they perform in TEI has not been evaluated. In addition, previous studies have shown that response time (RT) modeling was helpful to improve item calibrations when aberrant behavior occurs. Therefore, two goals are targeted in the present study. First is to explore two RG identification methods in TEI and they are visual inspection of RT distribution and common K-second threshold. The second goal is to evaluate two calibration approaches with emerge of RG in TEI. In approach 1, item response theory (IRT) modeling is applied after identified RG responses are excluded. In approach 2, RT modeling proposed by van der Linden (2007) is used with all responses. The data used for analyses come from a large field test event for a K-11 mathematics test. The results of this research will provide insights into RG in TEI, which is helpful to TEI's future application in operational psychometric work.

**Applications - Parallel Session: 5.7D: AN EVALUATION OF WORDING EFFECTS MODELING UNDER THE ESEM FRAMEWORK**

Luis Eduardo Garrido, University of Virginia; Hudson Golino, University of Virginia; María Dolores Nieto, Universidad Autónoma de Madrid; Kiero Guerra Peña, Pontificia Universidad Católica Madre y Maestra; Agustín Martínez Molina, Universidad de Zaragoza

The combination of positive and negative polarity items in rating scales often produces artifactual systematic variance that can have a strong influence on model fit and factor structure. This wording variance can be modeled by incorporating additional "method factors" that are uncorrelated with the substantive factors. However, under the framework of exploratory structural equation modeling (ESEM), systematic evaluations of the most promising techniques are currently lacking. Thus, the objective of the current study was to assess the performance of the correlated trait-correlated (method minus one) model (CT-C[M-1]) and random intercept item factor analysis (RIIFA) with ESEM models. A Monte Carlo study was conducted for polarity balanced binary datasets with three underlying factors that included the manipulation of four relevant variables: percentage of cases with wording bias (0%, 10%, 20%, 30%, 40%), sample size (300, 500, 1000), factor loadings (0.50, 0.60, 0.70), and factor correlations (0.00, 0.30, 0.50). The results indicated that wording bias above 10% strongly impacted on model fit and the recovery of the uncontaminated population structure. Both models were able to mostly explain the sample data with wording bias, but multiple method factors were needed when the factor correlations were low. The RIIFA technique provided the best recovery of the uncontaminated population structure, as CT-C(M-1) tended to strongly underestimate the substantial loadings of the items specified to load on the method(s) factor(s). In conclusion, wording bias may be modeled effectively using the RIIFA technique, but it will produce non-negligible levels of improper solutions for certain data conditions.

## Symposium 11: 4:30 PM – 6:00 PM

Chair: Willem Heiser

**Symposium 11: The History of Psychometrics**

**Symposium 11 - Parallel Session: 6.1A: THE HISTORY OF PSYCHOMETRICS: AN ACADEMIC GENEALOGY**

Lisa D. Wijsen, University of Amsterdam; Denny Borsboom, University of Amsterdam; Tiago Cabaço, Humboldt-University, Germany; Willem J. Heiser, Leiden University

Psychological and educational testing has become widespread in both society and science, and psychometrics has therefore become a scientific discipline of central importance. The history of psychometrics, however, is a topic rarely touched upon in either historical or psychometric research. In this talk, I will present the evolution of psychometrics using an academic genealogy of past presidents of the Psychometric Society. Although genealogical trees were originally developed as a tool for displaying pedigrees, they have also proved useful for visualizing advisor-student relations and tracking patterns of intellectual descent in academia. The genealogies show that most of the presidents are descendants of Wilhelm Wundt, James Angell, William James, Albert Michotte or Carl Gauss. Interestingly, many

historical figures often associated with the history of psychometrics in the literature, such as Francis Galton, Charles Spearman, and Stanley Smith Stevens, do not occur in the tree or play a marginal role. The individual genealogies, rooted in different sections of psychology, show that psychometrics is inherently multidisciplinary. Moreover, they show that the discipline has become increasingly diverse in terms of formal training, nationality and gender. Altogether, the genealogy exposes and preserves the rich and multidisciplinary background of psychometrics.

**Symposium 11 - Parallel Session: 6.1B: HISTORY AND FUTURE OF PSYCHOMETRICS IN THE NETHERLANDS**
Klaas Sijtsma, Tilburg University

This presentation is a follow-up of an earlier article (Van der Heijden & Sijtsma, 1996) that addressed the development of methodology and statistics in the Dutch social and behavioral sciences since 1945. Accidently, it appears that this article falls in between two era's in several ways. The presentation briefly discusses the development of Dutch psychometrics between 1945 and 1995, and then addresses the differences between this development and the development we have seen since 1995. The topics I address are: national versus international orientation, publishing little and in Dutch outlets versus publishing much in international outlets, no or little computing power versus immense computing power (and statistics that is analytically tractable versus computational statistics), and small data sets (experiments, tests, surveys) versus huge data sets (wearables, brain data, internet data). I draw some conclusions and make predictions about future developments that are not entirely Dutch.

**Symposium 11 - Parallel Session: 6.1C: L.L. THURSTONE'S "PSYCHOMETRIC LABORATORY"**
David Thissen, University of North Carolina at Chapel Hill

In approximately 1930, L.L. Thurstone tacked a sign that said "Psychometric Laboratory" on the door of his workroom in the Social Science Research Building at the University of Chicago. That sign remained until Thurstone retired from Chicago in 1952 and moved to the University of North Carolina at Chapel Hill, where the Psychometric Laboratory became an official research center under Thurstone's direction. The center was re-named the "L.L. Thurstone Psychometric Laboratory" in 1968, and has continued to function under that name. This presentation recounts some of the products and describes some of the facilities of the Lab, with biographical notes about some of the many dozens of students, postdoctoral visitors, and members of the faculties at both Chicago and UNC. The history of the Lab is closely tied to that of the Psychometric Society: More than a dozen of the past presidents of the Society were or had been at one time faculty or students at the Lab, and a few more had been postdoctoral visitors. The annual meeting of the Society was hosted by the Lab in Chapel Hill in 1968 and 1981, and on the occasions of the Lab's 25th and 50th anniversaries in 1978 and 2002. On another front, computing, the Lab brought the first electronic computer to UNC, housed one of the first facilities in the world for computerized collection of psychological data, and has been a significant source of psychometric software. The Lab's history is a view of part of the history of psychometrics.

**Symposium 11 - Parallel Session: 6.1D: FROM PSYCHOLOGICAL THEORIES TO PSYCHOMETRIC TOOLS**
Denny Borsboom, University of Amsterdam

Psychometrics and mathematical psychology form an interface between the fields of general statistical modeling and of substantive psychological research. Previous work by Gigerenzer (1991) has shown that statistical models (e.g. the analysis of variance model) are sometimes used as a heuristic of discovery for generating psychological theories (e.g., the theory of attribution). In the current paper, I argue that the reverse also happens: in the history of psychology, substantive psychological theories have repeatedly been transformed into general statistical models. Two examples are used to establish this thesis. First, I show how the theory of general intelligence or g, which was originally proposed as a substantive explanation for the positive manifold by Spearman (1904), was gradually stripped of its substantive content and developed into the general statistical tool of factor analysis. Second, I show how the co-activation theory of learning ("what fires together wires together"), originally proposed by Hebb (1949), was transformed into the general statistical tool of neural network modeling. In both cases, similar mechanisms are at work, and I will propose a general characterization of these processes in terms of

recent theories on the transmission of cognitive goods between fields. Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. Psychological Review, 98, 254-267. Hebb, D.O. (1949). The Organization of Behavior. New York: Wiley. Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. The American Journal of Psychology, 15, 201-292.

## Symposium 12: 4:30 PM – 6:00 PM

Chair: Jelte M. Wicherts

**Symposium 12: Novel Approaches to Dealing with Heterogeneity in Meta-analysis**

**Symposium 12 - Parallel Session: 6.2A: RANDOM-EFFECTS META-ANALYTIC STRUCTURAL EQUATION MODELING WITH MAXIMUM LIKELIHOOD ESTIMATION**
Suzanne Jak, University of Amsterdam

Meta-analytic structural equation modeling (MASEM) is a statistical technique to fit hypothesized models on the combined summary data of multiple independent studies. Existing methods typically consist of two stages (Viswesvaran & Ones, 1995). In Stage 1, correlation matrices from different studies are combined to estimate a pooled correlation matrix with maximum likelihood estimation. The pooled correlation matrix can be estimated under a fixed-effect (equal effect) or a random-effects model. In Stage 2, a structural equation model, such as a path model or factor model, is fitted to the pooled correlation matrix using weighted least squares estimation (Cheung & Chan, 2005; Cheung, 2014). In will present and evaluate an alternative (one-stage) random-effects MASEM method that uses maximum likelihood estimation throughout. This new method can be seen as the random-effects version of ML MASEM (Oort & Jak, 2015). It is expected that the new method will outperform existing methods in terms of parameter estimation and the evaluation of model fit. Specifically, the comparative fit index (CFI) is known to behave differently with weighted least squares estimation in comparison with maximum likelihood estimation (Yuan and Chan, 2005). Moreover, the new method is the only correlation-based MASEM method that facilitates the inclusion of continuous moderator variables in MASEM.

**Symposium 12 - Parallel Session: 6.2B: REPRODUCIBILITY OF PSYCHOLOGICAL META-ANALYSES: CODING ERRORS, OUTLIERS, AND HETEROGENEITY**
Esther Maassen, Tilburg University

Various studies have assessed the prevalence of reporting errors and inaccuracy of computations in psychological articles. These problems related to errors in primary studies extend to the level of meta-analyses in biomedical and health literature. The goal of this study was to systematically assess the reproducibility of psychological meta-analyses. To this end, we randomly selected meta-analyses from the psychological literature, and included 33 meta-analyses that reported effect sizes at the study level. We subsequently determined the prevalence of reporting or computational errors in 500 of the primary study effect sizes, and checked whether corrections of these effect sizes altered overall meta-analytic effect sizes, confidence intervals, and heterogeneity estimates. We documented how often we were unable to reproduce primary study effect sizes and the main meta-analytic outcomes. Additionally, we documented how meta-analysts dealt with issues related to heterogeneity, outlying primary studies, signs of publication bias, and possibly dependent effect sizes. Common issues in the reproducibility of meta-analyses pertain to the omission of necessary information on effect size computations and meta-analytic approaches. We present error rates and highlight the importance of using meta-analytic reporting standards and other practices that might help improve reproducibility of meta-analyses.

**Symposium 12 - Parallel Session: 6.2C: NOVEL APPROACHES TO DEALING WITH HETEROGENEITY IN META-ANALYSES**
Jelte M. Wicherts, Tilburg University

Meta-analyses are increasingly being used to collate evidence from research lines in psychology and other fields. The goals of these meta-analyses are to estimate mean effects, heterogeneity, and potential moderation of effects due to study-level characteristics. This symposium consists of four talks that bear on the key issue of heterogeneity in meta-analyses. Each talk approaches heterogeneity in meta-analysis

from different perspectives, varying from methodological issues (coding errors), to statistical approaches to modelling publication bias, structural equation models, and analyses of multiple moderators. Specifically, the talks deal with heterogeneity that is possibly due to coding errors and outliers (Maassen), consider heterogeneity as implemented in the novel tool p-uniform (van Aert), deal with heterogeneity when meta-analyzing structural equation models (meta-analytic SEM; Jak), and explain heterogeneity with multiple moderators in meta-analysis (Li, Dusseldorp, & Meulman). In the first talk, Esther Maassen selected a random sample of meta-analyses from the psychological literature and re-computed the effect sizes to study reproducibility of results, to see how meta-analysts deal with heterogeneity, and to determine whether coding errors and outliers affect heterogeneity estimates. In the second talk, Robbie van Aert will present a new random effects version of the method of p-uniform, that was developed to correct for publication bias under heterogeneity. In the third talk, Suzanne Jak will present a new maximum likelihood based method to deal with heterogeneity of meta-analytic structural equation models. In the fourth talk, Xinru Lin and her colleagues will present a new flexible R package called metaCART for meta-analysis that deals with multiple moderators and potential interactions between these moderators. We will end with a general discussion opened by Wicherts.

**Symposium 12 - Parallel Session: 6.2D: CORRECTING FOR PUBLICATION BIAS IN A META-ANALYSIS WITH P-UNIFORM***

Robbie van Aert, Tilburg University; Jelte M. Wicherts, Tilburg University; Marcel A.L.M. van Assen, Tilburg University, Utrecht University

Meta-analysis is now seen as the "gold standard" for synthesizing evidence from multiple studies. However, a major threat to the validity of a meta-analysis is publication bias that refers to situations where the published literature is not a representative reflection of the population of completed studies. In its most extreme case this implies that studies with statistically significant results get published and studies with statistically nonsignificant results do not get published. A consequence of publication bias is that the meta-analytic effect size is overestimated. The p-uniform method is a meta-analysis method that corrects estimates for publication bias, but the method overestimates average effect size in the presence of heterogeneity in primary study's true effect sizes (i.e., between-study variance). We propose an extended and improvement of the p-uniform method called p-uniform*. This new p-uniform* method is an improvement in three important ways, because it (i) is a more efficient estimator, (ii) eliminates the overestimation of effect size in case of between-study variance in true effect sizes, (iii) enables estimating and testing for the presence of the between-study variance in true effect sizes. We will explain the p-uniform* method and discuss the results of an analytical study and Monte-Carlo simulation study where p-uniform* was compared to a selection model approach to correct for publication bias. We offer recommendations for correcting meta-analyses for publication bias in practice, and a R package as well as an easy-to-use web application for applying p-uniform*.

**Symposium 12 - Parallel Session: 6.2E: R-PACKAGE METACART: A FLEXIBLE TOOL FOR META-ANALYSIS WITH MULTIPLE MODERATORS**

Xinru Li, Leiden University; Elise Dusseldorp, Leiden University; Jacqueline J. Meulman, Leiden University

In meta-analysis, heterogeneity often exists between studies. Knowledge about study features (i.e., moderators) that can explain the heterogeneity in effect sizes can be useful for researchers to assess the effectiveness of existing interventions and design new potentially effective interventions. When there are multiple moderators, they may amplify or attenuate each other's effect on treatment effectiveness. In this situation, we say that there are interaction effects between the moderators. Usually, interaction effects are neglected in meta-analytic studies. One reason for this is the lack of appropriate methods that are able to identify interactions between multiple moderators in situations without a priori hypotheses. To overcome this problem, a new approach called meta-CART was proposed with the advantage of dealing with many moderators and identifying interaction effects between them (Li et al., 2017). The method follows the paradigm of classification and regression trees (CART) to partition studies into more homogeneous subgroups by influential moderators, and simultaneously tests the subgroup meta-analysis results. In our presentation, we will introduce the R-package metacart, which provides user-friendly functions to perform meta-CART analysis for various types of moderators (i.e., continuous, ordinal, and nominal variables), with fixed- or random-effects model assumptions and various options to

control the partitioning process. Practical application of the package will be illustrated on real-world meta-analytic data sets.

## Computer-Based Testing: 4:30 PM – 6:00 PM

Chair: Johan Braeken

**Computer-Based Testing**

**Computer-Based Testing - Parallel Session: 6.3A: COMPARISON OF TWO ITEM PREKNOWLEDGE DETECTION METHODS USING RESPONSE TIME**
Chunyan Liu, National Board of Medical Examiners

Security for high stakes assessments is always a concern for test developers. Breached or compromised items tend to become easier both in terms of difficulty and in the time needed to respond for the test takers with preknowledge compared to those without preknowledge. This is detrimental to the validity of test scores since the ability of these test takers is overestimated. Therefore, it is crucial for the test developers to attempt to flag compromised items and detect the test takers with preknowledge to ensure test validity. Response time (RT) recorded during a computer-based testing (CBT) or computerized adaptive testing (CAT) have been demonstrated to be effective in identifying compromised items and test takers with item preknowledge (Meijer & Sotaridona, 2006; Qian, et al, 2016). Meijer and Sotaridona (2006) proposed the effective response time method (ERT) to identify item preknowledge, in which the RTs of the correctly answered items for the able examinees were used to estimate the slowness parameter of person and time intensity of item. Qian et al (2016) detected the item preknowledge using MCMC method based on the lognormal response time model (van der Linden, 2006). The purpose of this study is to compare the performance of these two methods in terms of Type I and Type II errors, and power. The following factors will be considered: the percentage of compromised items, the percentage of the RT reduction of the compromised items, and the percentage of test takers with preknowledge.

**Computer-Based Testing - Parallel Session: 6.3B: BAYESIAN MASTERY PROMOTION OF MULTIPLE ATTRIBUTES WITH RESPONSE TIMES**
Sangbeak Ye, University of Missouri, Kansas City

Implementing computerized systems to teach multiple skills may consist of a sequential process of administering a set of items for each skill and advancing to the subsequent skill when mastery is achieved. With the common availability of the response time data, the more data-driven adaptation of this two-fold learning process can be developed if response time distributions of masters and non-masters can be characterized. In this study, instructional effectiveness of items were first identified as transition probabilities driving a state of non-mastery to mastery jointly under cognitive diagnosis model. The skills vector of each examinee is expressed as posterior probabilities that reflect the progress of the assessment and likelihood of transition, and they were utilized to select the subsequent item as well. Through the Bayesian item selection, the goal is to promote learning to hasten mastery and to minimize the size of the assessment where the byproduct benefit was to decrease the item exposure rate. Simulation studies were conducted under restrictive conditions with misspecified response time distributions and estimates of transition probabilities. This approach can be adopted in assessments in e-learning environments where the transitions of all targeted attributes are presumed.

**Computer-Based Testing - Parallel Session: 6.3C: STATISTICALLY EFFICIENT MASTERY-BASED LEARNING**
Georgios Fellouris, University of Illinois, Urbana-Champaign; Yanglei Song, University of Illinois, Urbana-Champaign

The widespread use of computerized e-learning environments in education calls for sophisticated and rigorous statistical methods for skill acquisition. The fundamental goal in this framework is to teach each individual student a certain skill using a given pool of items. Some of these items may have high instructional value, whereas other may be more informative for testing purposes. The problem then is to

decide how to administer these items, in real time, in order to (i) minimize the time needed for mastery, (ii) detect this mastery quickly as soon as it happens, (iii) control the probability of a false detection. Therefore, the role of an item assignment rule in this setup is dual; first, to accelerate the learning process; second, to contribute to the quick and accurate detection of the latent time of skill acquisition. We propose an on-line item assignment rule that is easy to implement and is inspired by mastery-learning theory (Bloom,1968). The proposed scheme controls explicitly the probability of false detection and, more importantly, it is asymptotically efficient. Specifically, it requires the smallest possible expected number of items to a first-order asymptotic approximation as the probability of false detection goes to 0. This asymptotic optimality property is established under a general learning model, where there is no optimal rule. For a simple, "stationary" learning model, an optimal (computationally intensive) rule is obtained based on dynamic programing, and a simulation study reveals that the performance loss of the proposed approach in this case is minimal.

**Computer-Based Testing - Parallel Session: 6.3D: INCREASING COMPARATIVE JUDGMENT EFFICIENCY: A REFERENCE-BASED ALGORITHM FROM CAT RESEARCH**

San Verhavert, University of Antwerp; Antony Furlong, International Baccalaureate; Renske Bouwer, University of Antwerp; Vincent Donche, University of Antwerp; Sven De Maeyer, University of Antwerp

Over the last five years several studies have proven that Comparative Judgement (CJ) is a reliable and valid assessment method for a variety of competences, expert assessment and peer assessment. The method is based on Thurstone's law of comparative judgement (1927). Since its early days, CJ has suffered from low efficiency (Bramley, Bell & Pollitt, 1998). Therefore one of the most important questions in CJ today is "how can the method be made more efficient without affecting the reliability"? Pollitt (2012) proposed an adaptive algorithm based on Computer Adaptive Testing (CAT). This algorithm however artificially inflates the reliability of the estimates (Scale Separation Reliability, SSR; Bramley & Vitello, 2018). A solution is to use a fixed reference set of representations against which new representations are compared. Such an algorithm was developed and implemented in D-PAC (www.d-pac.be) and tested in an experiment together with the International Baccalaureate (www.ibo.org). The main question in this experiment is, how does the reliability of the reference set affect the results and the efficiency? A two-step approach was taken. First, 160 essays were compared by 15 assessors reaching a SSR of .92. Second, 20 essays were selected from the 160 to be re-judged using the adaptive algorithm. For this, four reference sets with SSR's of .50, .70, .80 and .92 were created by selecting subsets of the original data. The reference set with SSR .80 showed the best balance between RMSE and number of comparisons. These results are replicated in a post-hoc simulation study.

**Computer-Based Testing - Parallel Session: 6.3E: COMPARING SELF-REPORTED PERSEVERANCE WITH BEHAVIORS INFERRED FROM PROCESS DATA**

Xiaying Zheng, American Institutes for Research; Young Yee Kim, American Institutes for Research; Markus Broer, American Institutes for Research

In recent years, noncognitive constructs have received increasing attention from researchers, educators, and policymakers as they have been found to be relevant to academic and workforce success. One such construct newly introduced in the 2017 National Assessment of Educational Progress (NAEP) is Perseverance, measured by eight survey items. However, as a new index, its validity has not yet been evaluated in relation to other criteria that should be theoretically related to Perseverance. Process data (log or event data) from the 2017 NAEP digitally-based assessment (DBA) provides a unique opportunity to examine whether examinees demonstrated persevering behaviors when taking the cognitive examinations. As a first step in our exploratory research on the Perseverance index, this research uses the 2017 NAEP DBA process data to investigate the relationship of the Perseverance index with a set of action variables derived from the process data that are expected to be related to perseverance, including but not limited to, times spent on each cognitive examination item, number of revisits to items, omitting/skipping behaviors, and editing behaviors when answering constructive response items. For this purpose, the partial correlations between the Perseverance index and each of the action variables are estimated, with student abilities and item characteristics (e.g., item format and difficulty) taken into account to control for potential confounding effects. The results of this research could provide important insights on the inferential usefulness of the Perseverance index and inform future development of noncognitive indices.

## Estimation and Computation Methods: 4:30 PM – 6:00 PM

Chair: Eric-Jan Wagenmakers

**Estimation and Computation Methods**

**Estimation and Computation Methods - Parallel Session: 6.4A: A RESIDUAL EQUIVALENCE TEST FOR MIIV-2SLS ESTIMATION**
Ian Campbell, University of Notre Dame

We propose a residual equivalence test to replace the Sargan test in model-implied instrumental variable two-stage least squares (MIIV-2SLS) estimation of systems of equations. MIIV-2SLS estimation is an alternative to maximum likelihood (ML) for estimating parameters in models that are systems of equations, such as factor, mediation, growth, and other common psychology models. 2SLS estimation makes fewer assumptions than ML but requires instrumental variables, which can be difficult to find. Bollen (2001) outlined a procedure for using the proposed model structure itself to identify instrumental variables. A key part of this procedure relies on using the Sargan (1958) test to assess if the instrumental variables are uncorrelated with the composite error terms, as the model implies they should be. This Sargan test is a test of exact model fit with a null hypothesis stating the correlation equals zero. Tests of exact model fit like this are problematic, as failing to reject the null hypothesis does not actually support the null hypothesis. We propose a residual equivalence test that can replace the Sargan test and allows researchers to directly support the model's claim of zero association. The advantages of this equivalence test procedure over the Sargan test are discussed. A power analysis is performed to demonstrate the impact of the threshold level on the statistical power of the residual equivalence test, and a Monte Carlo simulation is conducted to compare four different multiple comparison procedures for testing multiple MIIVs jointly, ultimately recommending the Nyholt correction.

**Estimation and Computation Methods - Parallel Session: 6.4B: THE UNCERTAINTY OF NUISANCE PARAMETER IN POWER AND SAMPLE SIZE CALCULATION**
Chuchu Cheng, Boston College; Hao Wu, Boston College

In this paper we propose a new method of power and sample size calculation in dependent t test and independent t test. In practice, power and sample size are commonly calculated using textbook formulas that involve both effect size and some nuisance parameters. For example, the nuisance parameter can be the correlation between two measures in dependent t test, and variance ratio of the two groups in independent t test. While researchers can specify the size of effect to be detected, the nuisance parameter is usually estimated from a previous study or a pilot study. Thus, there is randomness in the nuisance parameter estimate. In most studies, the point estimate of nuisance parameter is used, which ignored its uncertainty. To improve the estimation accuracy and computation efficiency, we construct the confidence intervals of power or sample size from the confidence interval of the nuisance parameter to account for its randomness. Simulation studies were conducted to compare different bootstrap methods, maximum likelihood and Delta method in constructing the confidence intervals.

**Estimation and Computation Methods - Parallel Session: 6.4C: HIGHER-ORDER ACCURATE APPROXIMATIONS FOR LATENT VARIABLE MODELS**
Shaobo Jin, Uppsala University

Marginal maximum likelihood is commonly used in various latent variable models, in which latent variables are often integrated out. Since such integrals often do not have closed form solutions, various numerical methods have been suggested in the literature such as Laplace approximation, adaptive Gauss-Hermite quadrature, and full exponential Laplace approximation with first-order Laplace approximation. In this study, we first show that full exponential Laplace approximation with qth-order Laplace approximation has the error rate of order q+1. Hence, the second-order Laplace approximations in the literature can be modified to achieved the third-order accuracy. We then show that the gradient from the full exponential type approximation is equivalent to the properly calculated gradient of approximated observed log-likelihood. At last, we also discuss how to achieve a higher-order approximation from the commonly used EM algorithm with adaptive Gauss-Hermite quadrature approximation.

**Estimation and Computation Methods - Parallel Session: 6.4D: PRECISION OF POWER COMPUTATIONS FOR PSEUDO EXACT OR CONDITIONAL TESTS**
Jan Philipp Nolte, UMIT - Health and Life Sciences University; Clemens Draxler, UMIT - Health and Life Sciences University

Draxler & Zessin (2015) derived the power function for a class of exact or conditional tests of assumptions of the (binary) Rasch model and suggested an MCMC approach developed by Verhelst (2008) for the numerical approximation of the power of the tests. In this contribution, the Verhelst approach is compared with an exact sampling procedure proposed by Miller & Harrison (2013) for which the discrete probability distribution to be sampled from is exactly known. Of interest is in particular the variance and thus the precision of the power computations in the context of the analysis of differential item functioning in the case of smaller sample sizes (< 150) and item numbers (< 10). Tentative results show no substantial differences between the two numerical procedures. Regarding the question of computation time the Verhelst approach will have to be considered much more efficient.

## Item Response Theory: 4:30 PM – 6:00 PM

Chair: Peter Halpin

**Item Response Theory**

**Item Response Theory - Parallel Session: 6.5A: SPEED-ACCURACY-DIFFICULTY INTERACTION IN JOINT MODELING OF RESPONSES AND RESPONSE TIME**
Dandan Liao, University of Maryland, College Park; Hong Jiao, University of Maryland, College Park; Matthias von Davier, National Board of Medical Examiners

With the rapid development of information technology, computer-based tests have made available the collection of auxiliary data, including response times (RTs). A commonly adopted assumption in joint modeling of RTs and item responses is that item responses and RTs are conditionally independent given a person's speed and ability (e.g., Thissen, 1983; van der Linden, 2007). However, researchers have found that the conditional independence assumption between item responses and RTs is likely to be violated in various ways (e.g., De Boeck, Chen, & Davison, 2017; Meng, Tao, & Chang, 2015), and that item difficulty is associated with the direction of conditional dependence between item responses and RTs (e.g., Bolsinova, De Boeck, & Tijmstra, 2016; Goldhammer, Naumann, & Greiff, 2015). However, such an interaction has not been explicitly explored in jointly modeling of RT and response accuracy. In the present study, various approaches for joint modeling of RT and response accuracy are proposed to account for the conditional dependence between responses and RTs due to the interaction among speed, accuracy, and item difficulty. Three simulation studies are carried out to compare the proposed models with existing models that do not take into account the conditional dependence with respect to model fit and parameter recovery. The consequence of ignoring the conditional dependence between RT and item responses on parameter estimation is explored. Further, empirical data analyses are conducted to investigate the potential violations of the conditional independence assumption between item responses and RTs and obtain a more fundamental understanding of examinees' test-taking behaviors.

**Item Response Theory - Parallel Session: 6.5B: THE MIXTURE RESPONSE TIME MODEL FOR DETECTING UNEXPECTED RAPID-GUESSING BEHAVIORS**
Chia-Ling Hsu, Faculty of Education, The University of Hong Kong; Kuan-Yu Jin, The University of Hong Kong; Shu-Ying Chen, National Chung Cheng University

Rapid guessing, which means that respondents tend to respond to an item quickly by guessing the answers, is commonly observed in ability tests. In particular, rapid guessing is not a consistent status for a respondent; instead, a respondent can rapidly guess on any item due to non-motivation, speededness, responding strategy, or whatever reasons. By the aids of the development of computer-based testing, timing information can be a supportive evidence for identifying if a respondent endorses a response thoughtfully or dismissively. In this study, a mixture item response model incorporating response time is proposed to detect unexpected rapid-guessing behaviors during testing. We conducted a simulation study to evaluate the parameter recovery of the new model by using the Marcov chain Monte Carlo

estimation with Gibbs sampling. Results showed that the parameters in the new model can be recovered fairly well; ignoring rapid-guessing behaviors by fitting a standard item response model would overestimate item difficulties and underestimate item item-intensities seriously. Finally, a computer-based mathematical test was analyzed as an example for demonstration.

**Item Response Theory - Parallel Session: 6.5C: DETECTING ABERRANT RATING SCALE RESPONSES USING CHANGE POINT ANALYSIS**

Vivian Chan, Multi-Health Systems, Inc.; Gregory Gunn, Multi-Health Systems, Inc; Gill Sitarenios, Multi-Health Systems, Inc

Data for all sorts of assessments—ability, personality, and others—are impacted by inattentive, careless, or rushed responding. Change point analysis, an approach that can detect unusual changes in latent mean scores in sequences of responses, is one promising approach to capture these kinds of aberrant responding. Although change point analysis has been examined in ability test scores (e.g., Lee & von Davier, 2013; Shao, Li & Cheng, 2015; Sinharay, 2017), it has not been examined in a behavioral rating scale. This study explores change points in rating scale data collected online from a demographically representative sample of 1,145 U.S. adults. The validity of change points was examined in two ways. One way involved comparing change point patterns with 8 other data quality indicators (i.e., archival, direct and statistical indicators). Another way involved comparing patterns of change points with the order of the valence of the items (i.e., positively worded phrases or negatively worded phrases). Findings reveal that change points were associated with other indicators in detecting aberrant responses. Change points were typically absent whenever a subsequent item had a different valence, which meant this property of the items may not be a factor in change points detected. The utility of change point analysis in rating scales will be discussed.

**Item Response Theory - Parallel Session: 6.5D: CHANGE POINT ANALYSIS FOR DETECTION OF BACK RANDOM RESPONDING**

Alex Brodersen, University of Notre Dame; Ying "Alison" Cheng, University of Notre Dame

Aberrant responses in psychological survey data have been demonstrated to be a realistic concern in several prior studies such as Meade & Craig (2012), Baer et al. (1997), and Woods (2006). Back random responding (Clark, Gironda, & Young, 2003), a form of aberrant responding that is particularly challenging to detect in which respondents respond normally during initial phases of completing a survey and then respond in model deviant ways (random, straight line, or pattern responding) after some unknown point in the assessment. Change point analysis (CPA) is a method of detecting changes in a sequence of data frequently used in the field of statistical quality control. This method has been previously applied to psychological data in detecting compromised items in computerized adaptive testing (CAT) and speededness behavior in high stakes educational testing. Previous implementations of CPA for use in item response data (Shao, Li, & Cheng, 2015) are not suitable for psychological assessment due to assumptions that are typically only valid in educational testing data. The present study derives necessary modifications to CPA required for use in detecting back random responding. A simulation study was conducted to demonstrate the utility of this method. Results are presented with a brief discussion of conclusions and considerations for further study.

**Item Response Theory - Parallel Session: 6.5E: SEPARATING TRAIT AND VARYING THRESHOLDS IN RESPONSE STYLE IRT MODELS**

Mirka Henninger, University of Mannheim

Varying threshold models are promising candidates to account for heterogeneous use of response scales such as tendencies towards extreme or middle categories. However, some approaches constrain the variance-covariance matrix of trait and response styles to a diagonal matrix for identification. Besides the risk that this assumption is violated in empirical data, it limits the interpretation of varying threshold effects and their relation to the content trait. I therefore propose a new identification constraint for a varying threshold model accounting for response styles that allows for the full estimation of the covariance matrix. By using a sum-to-zero constraint for varying thresholds, I fix the location of the trait on the latent continuum and therewith separate trait and response styles. Results of two simulation studies supported the ability of the new approach to capture and correctly account for response style

effects when there exists covariance between trait and threshold variations in the generated data. In the latter case, the new model using a sum-to-zero constraint showed a superior parameter recovery than a model assuming unrelated threshold variations and traits. In an analysis of empirical survey data, model fit increased when covariances were freed for estimation. Furthermore, estimated covariances between varying thresholds were substantially high. Altogether, the results indicate the relevance of the new identification constraint for psychometricians and applied researchers: it increases the informative value of the varying threshold model without impeding its flexibility in modeling response scale use with minimal a priori assumptions.

## Network Analysis: 4:30 PM – 6:00 PM

Chair: Han L.J. van der Maas

**Network Analysis**

**Network Analysis - Parallel Session: 6.6A: A NETWORK APPROACH TO BINARY TIME SERIES DATA**
Nadja Bodner, Katholieke Universiteit, Leuven; Peter Kuppens, Katholieke Universiteit, Leuven; Nicholas B. Allen, University of Oregon; Lisa B. Sheeber, Oregon Research Institute; Eva Ceulemans, Katholieke Universiteit

Family processes have been identified as one of the important factors that contribute to affect (dys)regulation during adolescence. Numerous studies have investigated affective family interactions zooming in on single aspects of the interaction: frequency and duration of affective behavior, reactions of children to parental behavior or vice versa, or synchronicity of affect. In the present study, we integrated several of these aspects into affective family networks. To this end, we calculated the strength of ties between each combination of affective behaviors of different family members using a Jaccard similarity index (Jaccard,1901). This measure enables us to quantify the co-occurrence of affect as well as its temporal sequencing, depending on whether non-lagged or lagged data are used. Concurrent expressions of affect, the temporal sequencing of affective behaviors among family members and frequencies of different affects were finally depicted in networks to capture the interaction dynamics in easy-interpretable pictures. We applied this method to binary time series data resulting from a micro-coded problem solving family interaction between adolescents and their parents. The interaction was coded for the presence and absence of anger, sadness and happiness in second- to-second intervals. Comparing the affective networks of families of depressed adolescents with those of families of non-depressed adolescents revealed a anger-driven family dynamic that may contribute to vulnerability to, or maintenance of, adolescent depressive disorders. Our findings underline the importance of investigating affective family interactions in a holistic way.

**Network Analysis - Parallel Session: 6.6B: A LONGITUDINAL NETWORK: ASSESSING STUDY TIME ALLOCATION USING EMA**
Iris Yocarini, Erasmus University Rotterdam; Joran Jongerling, Erasmus University Rotterdam; Samantha Bouwmeester, Erasmus University Rotterdam

Network models are very popular at the moment, with most applications found in the field of clinical psychology. However, there are also other interesting applications. In this study we are interested in evaluated student's study time allocation during an Introduction to Psychometrics course and a network model approach offers a new way to improve the validity of this type of research. Previous studies into study time allocation and its relation to study success often have taken a retrospective approach or the use of time-consuming paper diaries, therefore suffering from memory biases and high dropout rates. With recent developments of mobile applications for the use of ecological momentary assessment (EMA) a new approach is available to measure study time allocation more efficiently in real time. In addition to the relations between study time allocation, perceived stress, and study success, and differences in these relations between students at a specific point in time, we are also interested in changes over time. Expectantly, students' time invested in studying varies during a course and increases as the test date approaches. Consequently, we take a dynamic network modelling approach. As these models have several requirements concerning the data, this study is a first step towards developing Bayesian multilevel time-

varying vector autoregressive models and functions as a first exploration in which we evaluate the practical side of using these models. In this pilot approximately 58 students were included.

**Network Analysis - Parallel Session: 6.6C: A LATENT SPACE MODEL FOR SOCIAL INFLUENCE**
Tracy Sweet, University of Maryland, College Park

One aspect of social interaction is social influence, the idea that beliefs or behaviors change as a result of one's social network. The purpose of this article is to introduce a new model for social influence based on a latent space model, which employs latent positions in a social space to predict ties. Our model instead predicts a node-level variable and uses these latent positions to model social influence; that is, individuals are influenced most by those who are ``closest'' to them in this latent space. We describe this model along with some of the contexts in which it can be used and explore some of the operating characteristics using a series of small simulation studies. We conclude with an example of teacher advice-seeking networks to show that changes in beliefs about teaching mathematics can be attributed to network influence.

**Network Analysis - Parallel Session: 6.6D: A STRONG TRUE-SCORE MODEL FOR MULTIDIMENSIONAL TESTS**
Stella Kim, University of Iowa; Won-Chan Lee, University of Iowa

A strong true-score theory defines a mathematical relationship between observed scores and true scores, as it relies on strong assumptions about the distributions of true and error scores. Keats and Lord (1962) used a two-parameter beta distribution and a binomial distribution for true and error scores, respectively. The resulting observed-score distribution is a negative hypergeometric distribution. Lord (1965) suggested a four-parameter beta distribution for true scores and either a binomial or compound binomial distribution for the errors. Lord (1969) also proposed a model that allows more general true score distributions. The current strong true-score models are limited in that they can deal only with "unidimensional" data using a univariate beta distribution for true scores. However, there has been an emerging need for a psychometric model applicable to "multidimensional" data such as a test consisting of multiple sub-content areas each of which is associated with a unique true score distribution. As the first step towards general multivariate situations, this study extends the conventional univariate strong true-score models to bivariate data employing a bivariate beta distribution (Arnold & Ng, 2011) for true scores. The bivariate strong true score model is applied to smoothing observed score distribution, estimating reliability and conditional standard errors of measurement, and computing classification consistency and accuracy.

## Applications: 4:30 PM – 6:00 PM

Chair: Selena Wang

**Applications**

**Applications - Parallel Session: 6.7A: THE INS AND OUTS OF EFFECT SIZES: EFFECT SIZE CLASSIFICATION**
Sheri Kim, Australian National University

The Replication Crisis has revived discussion on the need to improve the research methods used in psychological research. Along with this discussion, there has been a re-emphasis on the benefits of using effect sizes. However, the use of effect sizes appear to be a "checking the box" process, with researchers simply opting for well-known effect sizes such as Cohen's d and eta-squared (Fritz, Morris & Richler, 2012) and a general lack of interpretation alongside the effect size reported (Sun, Pang & Wang, 2010). Researchers' theoretical understanding of effect sizes remain underdeveloped (Kelley & Preacher, 2012). Researchers need clear guidance on which effect size to use, and what each effect size measure is capable of, as incorrect use of effect sizes may result in inaccurate or confusing information (e.g., McGrath & Meyer, 2006; Smithson & Shou, 2016). One way that methodology researchers attempt to gain and communicate clarity about effect size indices (ESIs) is to classify the indices into groups. These classifications systems are useful tools for understanding effect sizes, and can also provide guidance for

selecting between ESIs. However, they also have limitations. Classification systems are not exhaustive (they do not include all ESIs), the categorisations often conflict, and classification systems do not fully inform the researcher about the ESIs (they ignore some characteristics of effect sizes that may be important to consider). This article suggests a unified framework to guide researchers through the zoo of ESIs to choose, report and interpret appropriate effect sizes for their research.

**Applications - Parallel Session: 6.7B: QUANTIFYING AND EVALUATING INTERVENTION EFFECTS IN SINGLE-CASE EXPERIMENTAL DESIGNS**

Hsiu-Ting Yu, National Chengchi University; Jungkyu Park, McGill University

Single-case experimental designs (SCED) are critically important in the fields of psychology where only a small sample is available. In typical SCED studies, the intervention effects are usually examined by visual inspection and basic descriptive statistics. Computational methods have been proposed for evaluating the magnitude and direction intervention effectiveness for SCED including difference effect sizes (Busk & Serlin, 1992), regression-based effect sizes (Allison & Gorman, 1993), and the percentage of nonoverlapping data (PND; Struggs, Mastropieri, & Casto, 1987). The first two methods assume the data are independent; however, this assumption is not tenable since the data in SCED are repeated measures and correlated. The PND is not compromised by the serial dependency in the data but it faces problems due to the characteristics of SCED such as variability and ceiling or floor effect in the baseline condition and the presence of trends in the data. Indices have been proposed in the literature: the percentage of data exceeding the median, the improvement rate difference, the percentage of all nonoverlaping data, the pairwise data overlap squared, the percentage of data exceeding a median trend, Tau-U, and nonoverlap of all pairs. This research systematically investigates the psychometric properties of these proposed methods for quantifying and evaluating intervention effect in SCED studies. The performance of these indices and methods are compared under varies experimental designs and scenarios using simulation approaches to investigate the effects on Level, Trend, and Variability. Empirical data are also used to demonstrate the usages these indices in empirical SCED studies.

**Applications - Parallel Session: 6.7C: USING MULTIPLE IMPUTATION TO BALANCE UNBALANCED DESIGNS FOR TWO-WAY ANALYSIS OF VARIANCE**

Joost Van Ginkel, Leiden University; Pieter M. Kroonenberg, Leiden University

Two-way Analysis of Variance (Two-way ANOVA) is a widely used statistical technique in both experimental and non-experimental research. The technique tests whether two categorical independent variables (factors) plus their interaction make a significant contribution to the prediction of a numerical outcome variable. In order for the interpretation of the results of Two-way ANOVA to be unambiguous, it is important that the design is balanced. However, in practice unbalanced designs occur frequently. One way to resolve the problem of unbalancedness is to consider unbalancedness a missing-data problem, and to use multiple imputation to balance the design. In the current study, simulations are carried out in which the multiple-imputation approach is compared with the two most commonly used methods for dealing with unbalanced designs, namely the Type-I and Type-III sum of squares.

# Thursday, July 12, 2018 AM

## IMPS Registration: 8:00 AM – 11:30 AM

## Keynote Speaker: 8:30 AM – 9:30 AM

Chair: Susan Embretson

**An Urgent Assessment Question and a Proposed Answer, with an Eye toward Bayesian Inference**
Keynote Speaker: Robert Mislevy

## Refreshment Break: 9:30 AM – 9:45 AM

## Symposium 13: 9:45 AM – 11:15 AM

Chair: Denny Borsboom

**Symposium 13: Network Psychometrics 1: Advances in Network Structure Estimation**

**Symposium 13 - Parallel Session: 7.1A: GGM ESTIMATION USING THE BAYESIAN LASSO**
Joran Jongerling, Erasmus University Rotterdam

Recent years have seen an emergence in the use of Gaussian Graphical Models (GGM) – network models of partial correlation coefficients. Due to the large number of parameters in GGMs, the estimation of such models usually involves some form of regularization to protect against overfitting. Currently, a popular form of regularized GGM estimation is by using the graphical LASSO (GLASSO), which limits the sum of the absolute edge weights, and therefore shrinks small estimates toward zero. It has recently been demonstrated that results from GLASSO estimation may be unstable in light of sampling variation, and that, in particular, bootstrapping methods fail to form confidence intervals around descriptive statistics based on the estimated network model, such as centrality indices (Epskamp, Borsboom, & Fried, 2017). In addition, the GLASSO poorly handles missing data. In the current project we evaluate a Bayesian version of the graphical LASSO (Wang, 2012), which may remediate both problems by allowing the formation of Bayesian credibility intervals as well as sampling missing data from the posterior distribution. In simulation studies, we compare different model selection rules, and compare the performance of the Bayesian GLASSO, to both GLASSO estimation using the Extended Bayesian Information Criteria for model selection, and Bayesian estimation with a Wishart-prior for the variance-covariance matrix. Finally, we have implemented the Bayesian GLASSO in the R package BayesGGM.

**Symposium 13 - Parallel Session: 7.1B: ON TUNING PARAMETER SELECTION FOR ESTIMATING NETWORK MODELS**
Anna Wysocki, University of California, Davis

Network models are gaining popularity as a way to estimate direct effects among psychological variables and investigate the structure of constructs. A key feature of network estimation is determining which edges are likely to be non-zero. In psychology, this is commonly achieved through the GLASSO regularization method that estimates a precision matrix of Gaussian variables using an l1-penalty to push small values to zero. A tuning parameter, $\lambda$, controls the sparsity of the network. There are many methods to select $\lambda$, each of which can lead to vastly different graphs. The most common approach is to minimize the extended Bayesian Information Criterion (EBIC), but the consistency of this method for model selection has primarily been examined in high dimensional settings (i.e., $p > N$) that are uncommon in psychology. Further, for common situations in psychology, there is some evidence that alternative selection methods may have superior performance. Here, with simulation, we compare different methods for selecting $\lambda$, including the stability approach to regularization selection (STARS), k-fold cross-validation, the rotation information criterion (RIC), and the extended Bayesian information criterion (EBIC). We assessed performance using specificity (i.e., true negative rate) and sensitivity (i.e.,

true positive rate) of identifying non-zero edges, in addition to parameter bias and expected out-of-sample prediction. Our results demonstrate that tuning parameter selection should be made based on data characteristics and the inferential goal. We end with recommendations for selecting the tuning parameter when using GLASSO to estimate psychological networks.

**Symposium 13 - Parallel Session: 7.1C: COMPARING CONSTRAINT-BASED CAUSAL DISCOVERY ALGORITHMS**

Nitin Bhushan, University of Groningen; Laura Bringmann, University of Groningen; Casper Albers, University of Groningen

Substantive theories in psychology describe causal relationships between variables and graphical causal models are a useful representation of such relationships. Historically, the gold standard to test such causal hypothesis and estimate the effect of interventions have been randomized controlled trials (RCTs). But oftentimes, RCTs are not feasible due to various regulatory, ethical, or practical constraints. When RCTs are not feasible and substantive theories yet to be developed, causal discovery algorithms can discover probabilistic causal relationships between variables of interest from observational data. In this talk, we assess three such procedures which use conditional independence as a constraint to infer underlying causal structures; the PC algorithm (Spirtes et al., 2000), Linear Non-Gaussian acyclic models (LinGaM; Shimizu et al., 2006), and the Fast Causal Inference algorithm (FCI ; Spirtes et al., 1995; Zhang, 2008). The PC algorithm assumes a linear model with Gaussian errors and no latent confounders. The LinGaM algorithm relaxes the Gaussian error assumption and retains assumptions of linearity and absence of latent confounders. The FCI algorithm allows for latent confounders while retaining linear Gaussian assumptions. To validate these procedures in scenarios typical to psychology, we perform a simulation study varying the sample size, number of variables, the effect of latent confounders, non-normality of the error distribution, and graph density. We score these procedures using three metrics. We discuss the results of our study and further discuss implications of such procedures for theory development in psychology.

**Symposium 13 - Parallel Session: 7.1D: ESTIMATING PSYCHOLOGICAL NETWORKS WITH THE BAYESIAN BOOTSTRAP**

Donald Williams, University of California, Davis

An important goal for psychological science is developing methods to characterize relationships between variables. A common approach uses structural equation models to connect latent factors on a structural level to a number of observed measurements. More recently, network models have been developed that provide an alternative approach for characterizing conditional relationships among variables with the precision matrix. Whereas classical (i.e., frequentist) methods such as GLASSO are commonly used in psychology, Bayesian methods remain relatively uncommon in practice and methodological literatures. Here we propose a Bayesian method that uses probability weights drawn from a Dirichlet distribution over the input data to estimate the precision matrix. These probability weights provide a Bayesian bootstrap estimate for the posterior, which is computationally cheap compared to Markov chain Monte Carlo sampling. Specifically, we introduce two models based on a non-regularized and regularized approach, both of which rely on directional posterior probabilities for determining non-zero relationships. The latter allows for estimating network models when there are more variables (p) than observations (n). With numerical experiments, we demonstrate that performance often exceeds that of classical methods with respect to correctly identifying conditional relationships. In addition, both models show similar results for frequentist risk measured with Kullback–Leibler divergence. We discuss implications for psychology, as well as the Bayesian literature on the topic of estimation in high-dimensional settings.

**Symposium 13 - Parallel Session: 7.1E: MODERATED NETWORK MODELS**

Jonas Haslbeck, University of Amsterdam

Pairwise network models such as Gaussian Graphical Models (GGMs), Ising models and Mixed Graphical Models (MGMs) have become a popular tool to analyze dependencies among observed variables. We extend these models by allowing each pairwise interaction between variables Xi and Xj to be a linear combination of all remaining variables. These moderation effects are plausible in many situations, for instance one would expect that the positive impact of activity on mood depends on how fatigued an

individual is. We show how to estimate moderated network models with a nodewise estimation approach, report the sample complexity required to recover moderation effects, present a free and open-source implementation in R, and discuss open problems for future research.

## Symposium 14: 9:45 AM – 11:15 AM

Chair: Jelte M. Wicherts

**Symposium 14: Meta-science and Responsible Research Practices**

**Invited Symposium 14 - Parallel Session: 7.2A: RECOVERING FROM A CRISIS OF CONFIDENCE: DESIGNING AND CONDUCTING EFFECTIVE REPLICATION STUDIES**
Samantha Anderson, University of Notre Dame; Ken Kelley, University of Notre Dame; Scott E. Maxwell, University of Notre Dame

The "replication crisis" has garnered increasing attention in psychology, following several failed attempts to replicate high-profile findings. Although the replication crisis is due to a variety of factors, issues related to the statistical power of the replication study and the original study are important considerations. First, we describe series of simulations that compared various approaches to sample–size planning for replication studies for a variety of designs in psychology. Approaches were compared in terms of the average statistical power and assurance achieved, where assurance is the proportion of times the sample–size planning approach achieves its intended level of statistical power. Results indicated that common sample–size planning procedures yielded replication study statistical power levels well below 80% in most cases, particularly when the original study was not adequately powered. Thus, the statistical power of the original study can be a limiting factor in the statistical power of the replication study. Alternatively, a likelihood–based approach that takes both publication bias and uncertainty in the original study's sample effect size into account yielded improved statistical power and assurance. Second, we illustrate software to implement this approach in practice. Finally, we emphasize that researchers may have different goals for the replication study, and that it is important to tailor sample-size planning to the specific goal. For example, very different sample sizes may be required when the goal is to detect a non–null effect, evidence a null effect, or achieve a large degree of precision.

**Invited Symposium 14 - Parallel Session: 7.2B: CONCEPTUAL EFFECT SIZE ANALYSIS AND THE DESIGN OF SMALL N STUDIES WITH ADEQUATE POWER**
Patrick Shrout, New York University

In recent years there have been renewed calls for designing studies with sufficient power. Although power is a function of effect size, study design and sample size, some commentators have emphasized sample size only. In fact, some journals desk-reject studies that do not have at least 90 subjects per group. When special populations are of substantive interest, or methods are costly or time intensive, it is often difficult or impossible to carry out studies with recommended fixed large sample sizes. In this talk, I discuss the benefits of focusing on effect size rather than sample size to design studies with adequate power. I describe an approach that is inspired by generalizability theory that models effect size as a function of characteristics of the persons, interventions, measures, and contexts. As an initial step in a research program where large samples are infeasible, I recommend that investigators specify a combination of subject type, intervention strength, precise measures and potent context where the expected effect is very large. Once the proof-of-concept effect is established, further studies can explore conditions where the effect sizes are weaker. In the discussion, I explore the skepticism that some methodologists have expressed about empirical reports with large effects, and propose that such skeptical concerns can be addressed by pre-registration of a formal effect-size conceptual analysis.

**Invited Symposium 14 - Parallel Session: 7.2C: RESEARCHER DEGREES OF FREEDOM: PROBLEMS AND SOLUTIONS**
Jelte Wicherts

The designing, collecting, analyzing, and reporting of psychological studies entail many choices that are often arbitrary. The opportunistic use of these so-called researcher degrees of freedom aimed at

obtaining statistically significant results is problematic because it enhances the chances of false positive results and may inflate effect size estimates. In this talk, I present different degrees of freedom that researchers have in formulating hypotheses, and in designing, running, analyzing, and reporting of psychological research. I discuss research highlighting the prevalence and effects of the opportunistic use of these researcher degrees of freedom. Subsequently, I argue that pre-registration might help counter many of these effects, but also present recent results highlighting that many current pre-registrations are insufficiently specific to disallow p-hacking. I discuss the value of peer reviewing pre-registrations in the registered report format.

**Invited Symposium 14 - Parallel Session: 7.2D: ABANDON STATISTICAL SIGNIFICANCE**
Blakeley McShane, Northwestern University; David Gal, University of Illinois at Chicago; Andrew Gelman, Columbia University; Christian Robert, Universitè Paris-Dauphine; Jennifer L. Tackett, Northwestern University

We discuss problems the null hypothesis significance testing (NHST) paradigm poses for replication and more broadly in the biomedical and social sciences as well as how these problems remain unresolved by proposals involving modified thresholds, confidence intervals, and Bayes factors. We then discuss our own proposal, which is to abandon statistical significance. In particular, we recommend dropping the NHST paradigm—and the p-value thresholds associated with it—as the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences. We propose that the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with currently neglected factors (e.g., prior and related evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain) as just one among many pieces of evidence. We have no desire to "ban" p-values. Instead, we offer two concrete recommendations for how our proposal can be implemented in practice.

## Computer Based Testing: 9:45 AM – 11:15 AM

Chair: Dmitry Belov

**Computer Based Testing**

**Computer Based Testing - Parallel Session: 7.3A: UTILIZING RESPONSE TIME IN ON-THE-FLY MULTISTAGE ADAPTIVE TESTING**
Yang Du, University of Illinois, Urbana-Champaign; Anqi Li, University of Illinois, Urbana-Champaign; Hua-Hua Chang, University of Illinois, Urbana-Champaign

With the booming of adaptive testing development, multiple adaptive testing designs have been universally studied in both academia and testing industry. Among these adaptive designs, computerized adaptive testing (CAT) and multistage testing (MST) (Yan et al., 2014) have gained great popularity. While CAT flexibly tailors item difficulty to examinees' ability levels, MST mitigates early-stage theta estimation inaccuracy and allows for within-stage item revisiting (Yan et al., 2014; Zheng & Chang, 2015). On-the-fly multistage adaptive testing (OMST), as an integration of traditional CAT and MST, shares both of their merits and further makes quality control easier and faster (Zheng et al., 2014; Zheng & Chang, 2015). But in order to ensure ability estimation accuracy, test security, and satisfaction of other statistical and nonstatistical constraints, OMST still entails cautious item selection procedures. While some research, such as Choe et al. (2017), have already investigated the role of Response Time (RT) in item selection procedures in CAT, very few studies have yet examined RT's role in the domain of OMST. Building upon previous research on RT-included item selection procedures and OMST, we proposed several RT-included item selection methods in OMST. Through simulation studies, our research purports to investigate the performances of these methods, thereby unveiling how RT can facilitate item selection efficiency and estimation accuracy in OMST.

**Computer Based Testing - Parallel Session: 7.3B: HEURISTIC ASSEMBLY OF A CLASSIFICATION MULTISTAGE TEST WITH TESTLETS**
Zhuoran Wang, University of Minnesota, Twin Cities

In addition to the advantages of shortening test and balancing item bank usage, multistage testing (MST) has its unique merit of incorporating testlets. Testlet refers to a group of items sharing the same piece of stimulus. As MST can include an entire testlet in one module, fewer stimuli are required than items. On the other hand, CAT selects item one by one, thus excludes the possibility of several items sharing the same stimulus (Boyd, Dodd, & Fitzpatrick, 2013). In this way, testlets in MST save the stimuli processing time and facilitate ability estimate (Zheng, Chang, & Chang, 2013). In order to utilize the advantages brings by testlet, a classification MST was designed to upgrade an operational listening test. A heuristic module top-down assembly procedure incorporating testlet was developed based on the modified NWADH (Luecht, 1998; Zheng, et al, 2012). A three-stage classification MST with 1-3-6 panel design was assembled to classify examinees into seven levels. A real data-based simulation study was conducted to compare the performance of the classification MST and the operational linear test in terms of ability recovery and classification accuracy. The bi-factor model was used in item parameter calibration and examinee scoring. Results show the 30-item MST had a similar performance as the 44-item linear test with prior knowledge of examinee ability and outperformed the 44-item linear test without prior information, in both ability recovery and classification accuracy. In conclusion, the classification MST can shorten the test while keep a good accuracy.

**Computer Based Testing - Parallel Session: 7.3C: CONCERNING THE ANALYSIS OF MULTI-STAGE DATA**
Timo Bechger, Cito

The idea of a multi-stage test as a useful compromise between a CAT and a linear test is quite old but we have only recently gained experience with the use of multi-stage testing in a large and high stakes project. What we have learned is that a multistage test is something other than a CAT or a linear test and requires a special treatment: from the construction of the test up to the analysis of the data and the reporting of the results. The purpose of this talk is to share some of our experiences. For those interested in technical matters; we have used IRT using CML in multi-stage testing for the first time.

**Computer Based Testing - Parallel Session: 7.3D: A TWO-LEVEL ADAPTIVE TEST BATTERY**
Qi Diao, ETS; Wim J. van der Linden, ACT, Inc.

Running a test in an adaptive mode is much more efficient than as a fixed form. As many testing programs consist of multiple subtests rather than a single test, measurement efficiency can be further increased by using the information in already administered subtests as collateral information when selecting the next subtest. An appropriate framework for implementing the approach is multilevel item response theory in combination with Bayesian updating of the examinee's ability parameters. In the study reported in this presentation, both the choice of the next subtest and the items within each subtest are adaptive. Theoretical and empirical examples are presented to demonstrate use of the algorithm. The results of the study show substantial advantages of two-level adaptation in a test battery relative to within-test adaptation only.

## Estimation and Computational Methods: 9:45 AM – 11:15 AM

Chair: Samantha Bouwmeester

**Estimation and Computational Methods**

**Estimation and Computational Methods - Parallel Session: 7.4A: FACTOR ANALYSIS OF ORDINAL DATA VIA RIDGE METHOD**
Ge Jiang, University of Notre Dame; Ke-Hai Yuan, University of Notre Dame

This 'large p small N' problem is increasingly common in psychology and has attracted attention from researchers in various areas. It describes the difficulty of data analysis when there is a large number of

variables but a limited number of samples. In addition, ordinal data using Likert-type scales are frequently collected in psychology, posing another layer of complexity to data analysis. In such a case, existing estimation methods of the factor analytic model may not produce accurate parameter estimates and reliable model fit statistics. In this talk, I will present two studies to address the issues of parameter estimates and model fit statistics. In the first study, we developed an estimation method, called ridge generalized least squares (rGLS), to improve the efficiency of parameter estimates. It inherits the stability of an existing estimation method and the optimality of an asymptotically efficient estimation method, and thus yields more efficient parameter estimates with large p and small N. In the second study, we capitalize on the parameter estimates from rGLS and correct the likelihood ratio statistic for more reliable model fit. Although many corrections have been proposed to this statistic, most of them are ad hoc rather than principled. We developed empirical corrections to this statistic so that the new fit statistics yield improved Type I error control than the likelihood ratio statistic with large p and small N.

**Estimation and Computational Methods - Parallel Session: 7.4B: SPARSE MULTIGROUP FACTOR ANALYSIS**

Elena Geminiani, University of Bologna

Sparse factor analysis is an efficient method to estimate a penalized model where small loadings are automatically set to zero by a lasso penalty. Thanks to the variable selection property of the lasso, the resulting loading matrix possesses an optimal "simple structure" and rotation techniques and subjective loading thresholding are no longer necessary. We present an extension of sparse factor analysis to the case in which the observations belong to distinct groups (e.g. nations, cultural or socio-economical groups). The proposed methodology allows to estimate a sparse factor model within each group while inducing equal loadings across groups. This is accomplished with a fused lasso penalization, which simultaneously shrinks loadings within each group as well as their differences across groups. Sparse multigroup factor analysis is then more flexible than ordinary multigroup exploratory factor analysis in that it can identify relevant indicators across groups as well as group-specific indicators. We will illustrate the theoretical details of this technique and discuss its performances.

**Estimation and Computational Methods - Parallel Session: 7.4C: EXPLORATORY FACTOR ANALYSIS WITH SPARSE CORRELATION MATRICES**

Minami Hattori, University of Notre Dame; Elizabeth Daly, University of North Carolina at Wilmington; Guangjian Zhang, University of Notre Dame

The paper is concerned with exploratory factor analysis with sparse correlation matrices. A number of questionnaires often exist for measuring relatively new psychological constructs such as narcissism. Studies of such constructs often involve a few of such questionnaires. Factor analyzing the results of multiple such studies simultaneously can contribute to the development of psychological theory. A large correlation matrix synthesizing multiple studies would contain many 'holes', because each study contributes observations on, and correlations among, different sets of items. We adapt the iterated principal factor method for estimating exploratory factor analysis models with sparse correlation matrices. We conduct a Monte Carlo study to assess statistical properties of the proposed method.

**Estimation and Computational Methods - Parallel Session: 7.4D: CONSTRAINED FACTOR MODELS WITH APPLICATIONS**

Henghsiu Tsai, Institute of Statistical Science, Academia Sinica

This paper focuses on factor analysis of a data matrix. We propose statistical methods that enable analysts to leverage their prior knowledge or substantive information to sharpen the estimation of common factors. Specifically, we consider a doubly constrained factor model that enables analysts to specify both row and column constraints of the data matrix to improve the estimation of common factors. The row constraints may represent classifications of individual subjects whereas the column constraints may show the categories of variables. We derive both the maximum likelihood and least squares estimates of the proposed doubly constrained factor model and use simulations to study the performance of the analysis in finite samples. The Akaike information criterion (AIC) is used for model selection. Real data are used to demonstrate the application of the proposed model.

**Estimation and Computational Methods - Parallel Session: 7.4E: A NEW FACTOR ANALYSIS PROCEDURE CONSTRAINED BY LAYERED SIMPLE STRUCTURE**
Naoto Yamashita, Graduate School of Human Sciences, Osaka University; Kohei Adachi, Graduate School of Human Sciences, Osaka University

In factor analysis, loading matrix is referred for interpreting extracted common factors, usually followed by factor rotation. In this research, we propose a new procedure, as a new alternative to factor rotation, in order to obtain easily-interpreted loading matrix. It is called Layered Factor Analysis (LFA), as the loading matrix is constrained to be a sum of matrices called layers which has a specific simple structure. LFA allows to obtain loading matrix with simple structure without post-hoc transformation, as well as special structure such as bifactor structure. The effectiveness of LFA is demonstrated by real data examples. As an extension of LFA to other procedures for multivariate analysis is also discussed.

## Item Response Theory: 9:45 AM – 11:15 AM

Chair: Yunxiao Chen

**Item Response Theory**

**Item Response Theory - Parallel Session: 7.5A: GAUSSIAN VARIATIONAL ESTIMATION FOR MIRT**

April Cho, University of Michigan

Multidimensional Item Response Theory (MIRT) is widely used in assessment and evaluation of educational and psychological tests. It models the interaction between individuals' latent traits and their responses. The challenge in parameter estimation in MIRT is that likelihood involves multidimensional integrals due to latent variable structure. Various methods have been proposed which involve direct numerical approximations to the integrals or approximations based on Monte Carlo simulation. However, these methods are known to be computationally demanding in high dimensions and dependent on sampling data points from a posterior distribution. We propose a Gaussian Variational EM algorithm which adopts a variational inference approach from machine learning that approximates probability densities through optimization. Its basic strategy is to approximate intractable likelihood by a computationally feasible lower bound. Simulation studies show that the proposed algorithm is computationally more efficient and achieves better performance in parameter estimation than existing approaches. Moreover, the proposed algorithm provides an effective way to estimate dimension of latent traits and to identify the item-trait relationship via Lasso-based regularization.

**Item Response Theory - Parallel Session: 7.5B: A PRINCIPAL COMPONENT SOLUTION OF A GENERALIZED PARTIAL CREDIT MODEL**
David Hessen, Utrecht University

A multidimensional generalized partial credit model in which latent variables are assumed to have a multivariate normal distribution given any fixed response pattern, is derived. It is shown that if the number of latent variables is set equal to the number of items, the model can be fitted to data as a regular log-linear model whose maximum likelihood estimates can be used to obtain a principal component solution. Generalizations of the model are discussed, which can each be used as alternative hypothesis model in a generalized likelihood ratio test. In an example, the application of the principal component solution of the model to empirical data, is illustrated.

**Item Response Theory - Parallel Session: 7.5C: EXPLANATORY ITEM RESPONSE THEORY MODELS: IMPACT ON VALIDITY AND TEST DEVELOPMENT?**
Susan Embretson, Georgia Institute of Technology

Many explanatory item response theory models have been developed since Fischer's (1973) linear logistic test model was published. The current Handbook of Item Response Theory (van der Linden, 2016) contains chapters on many of these models. However, despite the existence of these models for several decades and their applicability to typical test data, actual impact on test development and validation has

been very limited. The purpose of the current paper is to explicate the importance of these models in validation and test development. A framework that explicates the interrelationships of the five aspects of validity (Embretson, 2017) is elaborated. It will be shown that the response processes aspect of validity impacts other aspects of validity. The Standards for Educational and Psychological Testing (2014) describes methods on response process that require special data collection methods, such as eyetracker movements or verbal protocols. However, the applicability of explanatory IRT model results, which can be obtained from traditional test data are not described. Examples are elaborated in this presentation to illustrate the relevancy of such results. Similarly, test development activities, such as item design, test design and testing procedures typically do not use information from explanatory IRT modeling results. Examples are presented to illustrate the potential impact on test interpretation. In conclusion, the requirements for stronger impact of explanatory IRT model results on testing practices are described.

**Item Response Theory - Parallel Session: 7.5D: TO Θ, OR NOT TO Θ: A SIMULATION STUDY ON THE VALIDITY OF IRT**

Jonathan Park, California State University, Fullerton; Netasha Pizano, California State University, Fullerton; Kathleen Preston, California State University, Fullerton

Psychological and educational research heavily rely upon survey methods to arrive at approximations of true, underlying constructs that are difficult to directly assess (i.e., depression, anxiety, extraversion). In situations where high stakes conclusions are being drawn from the data (i.e., acceptance to gifted programs, diagnostic situations, etc.), misclassification of a single participant may result in serious ramifications (i.e., missed educational opportunities, medical costs, etc.). As such, it is prudent that researchers be certain that scales, tests, and surveys assess the constructs of interest as accurately as possible to avoid issues with potentially costly misclassification. Various means for approximating latent constructs have been developed such as unweighted summative scores (SS; Crocker & Algina, 1986), alpha adjustment (α; Cronbach, 1951), Confirmatory Factor Analysis (CFA; Brown, 2014), and Item Response Theory (IRT; Lord, 1980, Embretson & Reise, 2013). Prior research comparing the various methodologies have revealed little differences in ability estimation under instances of ideal item functioning (see Reise & Waller, 2009). However, little to no previous research has assessed how these methods perform in identifying ability when poorly functioning items are present. To ameliorate this gap, a simulation study was conducted to see how the various testing methodologies performed in estimating a latent trait under realistic testing environments. In addition, the simulation also sought to assess how each method preserved a predictive relationship between two measures. A novel algorithm was developed for the purposes of identifying optimal 2PL Item Response Models. Results and implications of the simulations are discussed.

**Item Response Theory - Parallel Session: 7.5E: THE EFFECT OF USING PRINCIPAL COMPONENTS TO CREATE PLAUSIBLE VALUES**

Tom Benton, Cambridge Assessment

In all large scale educational surveys such as PISA and TIMSS the distribution of student abilities is estimated using the method of plausible values. This method treats student abilities within each country as missing variables that should be imputed based upon both student responses to cognitive items and a conditioning model using background information from questionnaires. Previous research has shown that, in contrast to creating specific estimates of ability for each individual student, this technique will lead to unbiased population parameters in any subsequent analyses provided the conditioning model is correctly specified (Wu, 2005). More recent research has shown that, even if the conditioning model is incorrectly specified, the approach will provide a good approximation to population parameters as long as sufficient cognitive items are answered by each student (Marsman et al, 2016). However, given the very large amount of background information collected in studies such as PISA, background variables are not all individually included in the conditioning model and a smaller number of principal components are used instead. Furthermore, since no individual student answers cognitive items from every dimension of ability, we cannot rely on sufficient items having been answered to ignore possible resulting misspecification in the conditioning model. This presentation will first use simple simulations to show how relying upon principal components within the conditioning model can potentially lead to bias in later estimates. A real example of this issue will also be provided based upon analysis of regional differences in performance in PISA 2015 within the UK.

## Principal Components and Correspondence Analysis: 9:45 AM – 11:15 AM

Chair: Victoria Savalei

**Principal Components and Correspondence Analysis**

**Principal Components and Correspondence Analysis - Parallel Session: 7.6A: ASSESSING INFERENCE ACCURACY FROM SAMPLED CORRESPONDENCE ANALYSIS SOLUTIONS**
Joseph Grochowalski, The College Board

Sampled contingency tables result in correspondence analysis (CA) solutions that can vary widely from sample to sample, making it difficult to make inference about categorical associations in a population. Existing literature attempts to define confidence regions for sampled CA coordinates, but the various authors conflict on the best approach. This study aims to provide a comprehensive overview of the sources of variability in CA solutions (e.g., inertia distribution across axes, associations between categories, sample sizes, etc.), and how different methods for creating confidence regions perform based on these sources of variability. In this large simulation, I created population tables with varying numbers of row and column categories, sample sizes, associations between categories, and strengths of associations between categories. I then sampled from these population tables and applied various methods of calculating confidence regions for CA coordinates, and evaluated their performance (i.e., whether the confidence regions provided correct inference about the relationships in the population) based on the features of the population and the resulting sample. In general, sampled CA solutions provided poor inference about population associations, and often confidence regions either did not include population coordinates, or—when sample solutions were to constrained to have partial bootstrap regions—did not estimate associations in the sampled data that were consistent with population associations. The results of this simulation provide information about the sources of the instability in CA samples and how to identify sample CA results that have a greater chance of misestimating true population associations.

**Principal Components and Correspondence Analysis - Parallel Session: 7.6B: REGULARIZED SIMULTANEOUS COMPONENT ANALYSIS FOR JOINT ANALYSIS OF MULTIBLOCK DATA**
Zhengguo Gu, Tilburg University; Katrijn Van Deun, Tilburg University

Thanks to recent technological developments, often multiple blocks of data are available on the same participants in psychological research. As a result, psychological researchers face a situation where the data they are accustomed to, for example, questionnaire or experimental data are supplemented with more novel sorts of data including sentiments expressed online, genetic scores, and neural activity. The joint analysis of linked blocks of traditional and novel data with respect to the same subjects, also called multi-block data, holds promising prospects. We present a regularized simultaneous component analysis (RSCA) model for joint analysis of multi-block data. The RSCA model combines the traditional simultaneous component analysis model (e.g., Timmerman & Kiers, 2003), a method that is similar to principal component analysis (PCA), with the Lasso and the Group Lasso penalties, resulting in a flexible framework for joint analysis of multiblock data. The current work is a significant improvement of the RSCA model we proposed previously (Gu & Van Deun, 2016). We introduce an alternative, yet much simpler approach to estimating the model that replies on coordinate descent instead of a Majorization Minimization procedure and present its implementation in a user-friendly R package. Further, by means of a simulation study, we examine model selection techniques for the RSCA model, including repeated double cross validation, four different types information criteria, and an index for sparseness. Finally, we present an empirical example where we analyze three-block survey data on parent-child relationships

**Principal Components and Correspondence Analysis - Parallel Session: 7.6C: 3WAYPACK: A MENU-DRIVEN PROGRAM SUITE FOR THREE-WAY ANALYSIS**
Pieter Kroonenberg, Leiden University

The program suite contains the following procedures/model: Tucker2, Tucker3 (eigenvalue-based, regression-based and operating on multi-mode matrices), Tucker4, Parafac, Procrustes analysis , Three-mode mixture method clustering, Three-way correspondence analysis, Simultaneous component

analysis, Three-mode hierarchical classes (Hiclas) It also contains extensive preprocessing and missing data handling, Splitting and subsetting three-way data sets. Output handling such rotations, joint biplots, residual analysis for the entire data set and subsets. Moreover, with a relatively straightforward extension of the Delphi interface it is possible to call external independently working programs - Ceulemans' three-mode Hiclas program is an example. An extensive manual exists for earlier versions, but it might not yet be available for the upcoming IMPS2018 The interface is written in Delphi but the main analysis programs are in Fortran90 compiled with the NAG compiler. These programs can operate within any Interface made for them, they can also be used in a stand-alone mode and as such used in simulation studies. The programs write basic ASCII files as output, HTML code for inspecting the output in a browser and they generate plotting code for GnuPlot so that publication quality plots can be made which also can be save in EMF format for further editing in presentation programs such as Micorsoft PowerPoint. Unfortunately there is no direct link with R or MatLab, apart from the fact that it should be possible to build an interface in those languages to call the programs and handle the output.

**Principal Components and Correspondence Analysis - Parallel Session: 7.6D: DIMENSION NUMBER FOR VARIANTS OF THREE-WAY CORRESPONDENCE ANALYSIS**

Rosaria Lombardo, University of Campania "Luigi Vanvitelli"; Michel van de Velden, Erasmus University Rotterdam; Eric J. Beh, University of Newcastle

Variants of three-way correspondence analysis (CA) have been considered for analysing bivariate and trivariate associations in three-way contingency tables with regard to the nominal variables (Carlier & Kroonenberg, 1996; Kroonenberg, 2008, chap. 17; Beh & Lombardo, 2014, chap. 11) and to the ordinal and mixed variables (Lombardo, Beh & Kroonenberg, 2016; Lombardo & Beh, 2017). The association in such three-way contingency tables has been modelled using correspondence analysis based on the components from a three-mode component analysis, such as Tucker3 or PARAFAC (Kroonenberg, 2008), and/or the polynomial components from the trivariate moment decomposition (Lombardo & Beh, 2017). Detecting the optimal number of components is of fundamental importance for visualising association in contingency tables. Here, common rules of thumb similar to those used in principal component analysis and factorial analysis are reviewed (Cattell, 1966) and new computational proposals based on inverse correspondence analysis (Groenen & van de Velden, 2004) are pursued for choosing the best dimension number of a three-way CA model.

**Principal Components and Correspondence Analysis - Parallel Session: 7.6E: JOINT DIMENSION REDUCTION AND CLUSTER ANALYSIS OF MIXED DATA**

Michel van de Velden, Erasmus University Rotterdam; Alfonso Iodice D'Enza, Università di Cassino e del Lazio Meridionale; Angelos Markos, Democritus University of Thrace

Methods for joint dimension reduction and cluster analysis have been around for quite some time. Several proposals followed by extensions, generalizations and modifications have been proposed over time. In addition, researchers appraised different proposals both theoretically and empirically by studying their performances both on simulated and observed data. In this paper, we review existing methods and proposals for joint dimension reduction and cluster analysis of mixed data; that is, categorical and numerical data. We propose a general method that encompasses several of the existing methods. Moreover, our general method gives rise to alternative options for joint dimension reduction and cluster analysis of categorical variables, and provides several options for joint dimension reduction and cluster analysis of mixed variables.

# Thursday, July 12, 2018 PM

## Symposium 15: 11:15 AM – 12:45 PM

Chair: Hyo Jeong Shin; Lale Khorramdel

**Symposium 15: Modeling Response Times in Large-Scale Assessments**

**Symposium 15 - Parallel Session: 8.1A: RELATING RESPONSE TIMES AND RESPONSE STYLES USING MIXTURE IRTREE MODELS**

Lale Khorramdel, Educational Testing Service; Matthias von Davier, National Board of Medical Examiners; Artur Pokropek, European Commission, Joint Research Centre; Ulf Böckenholt, Northwestern University; Thorsten Meiser, University of Mannheim

Response style (RS) bias linked to rating and Likert-type scales jeopardizes the validity and cross-cultural equivalence of non-cognitive constructs in comparative large-scale assessments. The purpose of the study is to examine the relationship between response time and RS for a better understanding of response behavior and for developing more sophisticated models for RS. A multi-process IRTree approach to detect RS (Böckenholt, 2012) and extensions of it were successfully applied to empirical data (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) and evaluated using simulated data (Pokropek, Khorramdel & von Davier, under review). Findings show that the approach can be a successful tool to control for RS bias but also that the measurement of RS is not straightforward. Not all respondents show RS and the ones who do may not show it to the same extent or in the same direction. Therefore, the current study combines a multidimensional IRTree approach with mixture IRT models to differentiate between groups of respondents who give valid responses versus RS in cross cultural surveys (PIAAC and PISA). The resulting latent classes are used to identify RS, and are related to response times as covariates. Including response time information may hint to potential causes of RS such as low test-taking motivation, low reading ability, fatigue effects, or item content not targeting parts of the sample. As a consequence, item and assessment development as well as test administration could be improved to reduce the proportion of invalid responses in future test applications.

**Symposium 15 - Parallel Session: 8.1B: MIXTURE MODELS FOR RESPONSE ACCURACY AND CATEGORIZED RESPONSE TIMES**

Hyo Jeong Shin, ETS; Matthias von Davier, National Board of Medical Examiners

Many international large-scale assessments have started administering computer-based assessments that collect RT data along with RA data. In the context of low-stakes, large-scale assessments, RT data can be useful in evaluating the validity of responses and investigating how test takers manage their time (i.e., Lee & Jia, 2014; Lee & Chen, 2011; Weeks, von Davier, & Yamamoto, 2016). Although there is increasing interest in utilizing RT data now found in publicly available data, RT data generally has not been incorporated in analytic procedures used in international large-scale assessments. This may lead to biased estimates in the secondary analyses that are using or attempting to make inferences about RT (e.g., correlations between RA and RT). Therefore, the current study investigates the possibility of joint modeling of RT and RA in the item calibration stage to extend the currently used operational approach. We use the multidimensional mixture IRT model, a special case of the general diagnostic models (von Davier, 2008) that incorporates RT and RA as separate dimensions: (a) a model that assumes zero correlation between RT and RA, and (b) a model that allows different correlations between them depending on the estimated latent classes. In preparation for the analyses, RT data are categorized into stanine values to allow improved handling of non-normal distributions with outliers and improved handling of missing data stemming from complex rotated booklet design. We investigate the comparability of this approach across multiple countries with different levels of performance using the PISA 2015 data.

**Symposium 15 - Parallel Session: 8.1C: RESPONSE MIXTURE MODELING: ACCOUNTING FOR HETEROGENEITY IN ITEM CHARACTERISTICS ACROSS RESPONSE TIMES**

Dylan Molenaar, University of Amsterdam; Paul de Boeck, Ohio State University

In item response theory modeling of responses and response times, it is commonly assumed that the item responses have the same characteristics across the response times. However, heterogeneity might arise in the data if subjects resort to different response processes when solving the test items. These differences may be within-subject effects, that is, a subject might use a certain process on some of the items and a different process with different item characteristics on the other items. If the probability of using one process over the other process depends on the subject's response time, within-subject heterogeneity of the item characteristics across the response times arises. In this talk, the method of response mixture modeling is presented to account for such heterogeneity. Contrary to traditional mixture modeling where the full response vectors are classified, response mixture modeling involves classification of the individual elements in the response vector. In a simulation study, the response mixture model is shown to be viable in terms of parameter recovery. In addition, the response mixture model is applied to a large scale arithmetic test in the Netherlands to illustrate its use in practice.

**Symposium 15 - Parallel Session: 8.1D: DISENTANGLING MISSINGNESS DUE TO LACK OF SPEED FROM MISSINGNESS DUE TO QUITTING**

Steffi Pohl, Freie Universität Berlin; Esther Ulitzsch, Freie Universität Berlin; Matthias von Davier, National Board of Medical Examiners

Missing values at the end of a test can occur for a variety of reasons: On the one hand, examinees may not reach the end of a test due to time limits and a lack of speed. On the other hand, examinees may not attempt all items and end the test early due to, e.g., fatigue or a lack of motivation. We use response times retrieved from computerized testing to distinguish missing data due to a lack of speed from missingness due to quitting. On the basis of this information, we present an approach that allows to disentangle, simultaneously model, and account for different missing data mechanisms underlying not reached items. The proposed model combines research on missing data and research on response times. In doing so, the model a) supports a more fine-grained understanding of the processes underlying not reached items and b) allows to obtain less biased and more efficient ability estimates. In a simulation study we evaluate the proposed model and compare its performance to current state of the art models for not reached items. In an empirical study we show which insights can be gained from this model on test taking behavior and the missing process.

**Symposium 15 - Parallel Session: 8.1E: MARGINAL JOINT RESPONSE MODELING OF RESPONSE ACCURACY AND RESPONSE TIMES**

Konrad Klotzke, University of Twente

Large-scale testing programs in educational measurement often use response accuracy (RA) and response time (RT) data to make inferences about test takers' ability and speed, respectively. Computer-based testing offers the possibility to collect item response time information by recording the total time spent on each item. Together with RA data, this kind of information can be used in test design to make more profound inferences about response behavior of the candidates. When following the popular modeling framework of van der Linden (2007), Klein Entink, Fox, and van der Linden (2009), Fox, Klein Entink, and van der Linden (2007), and van der Linden and Fox (2016), joint models are constructed by connecting an item response theory (IRT) model with an RT model, thereby defining a relationship between the person and item parameters. Although these joint models have been successfully applied in educational measurement, they are based on rather strict assumptions. For instance, it is assumed that respondents work at a constant speed and with a constant speed-accuracy trade-off throughout the test. In practice, different test strategies and heterogeneous time-management qualities most likely lead to variable working speed and a variable speed-accuracy trade-off. To extend the psychometric inferences for response accuracy and response times, a marginal joint modeling approach is proposed in which test takers' latent variables for ability and speed are integrated out. This approach will be advantageous in

identifying changes in ability, speed, and their relationship, and evaluating statistical hypotheses using the Bayes factor. Examples are given to illustrate the new modeling framework.

## Symposium 16: 11:15 AM – 12:45 PM

Chair: Kadriye Ercikan

**Symposium 16: Advanced Psychometric Modelling and Applications of Process Data Analysis in Educational Assessments**

**Invited Symposium 16 - Parallel Session: 8.2A: PROCESS DATA IN ADVANCING MEASUREMENTS AT ETS**
Kadriye Ercikan, Educational Testing Service (ETS)

Digitally based assessments allow us to capture data about examinee behaviors related to their engagement with the test referred to as response process data. Such data have the potential to advance the science and practice of measurement in significant ways. This presentation will contrast claims and sources of data that support validity of claims in paper based versus digitally based assessments that utilize process data, highlight three key uses of process data and provide an overview of the four presentations.

**Invited Symposium 16 - Parallel Session: 8.2B: PROCESS DATA IN LARGE-SCALE K-12 ASSESSMENTS: PERSPECTIVE FROM NAEP**
Christopher R. Agard, ETS; Lan Shuai, ETS; Mo Zhang, ETS; Paul Deane, ETS; Jiangang Hao, ETS; Mengxiao Zhu, Educational Testing Service; Fred Yan, ETS

As K-12 tests move to digitally based assessment (DBA) platforms, they provide the new affordance of capturing detailed logs of student interactions with assessment tasks, also known as the process data. Examples of process data range from students changing responses in a multiple-choice item to complex action sequences while conducting experiments in a physics simulation. Process data have been an integral part of the National Assessment of Educational Progress (NAEP) since the beginning of its digital transition. We will share some of the practices and lessons learned from NAEP. The presentation begins with an overview of NAEP process data: what they are, why they are collected, and how they are being used. Process data have been implemented in the NAEP 2017 and 2018 operational tests for 4th, 8th, and 12th grades. We will share process data results from subjects such as reading, mathematics, science, and writing to illustrate uses of process data in the NAEP assessment program, namely a) informing item design, b) providing validity support for psychometric reports, and c) shedding new light on strategies and cognitive processes students use in problem-solving. We will conclude with a discussion on lessons learned in NAEP, including a) the importance of cognitive theories of response processes in designing process data-based evidence; b) the need for interdisciplinary approach to process data analysis; c) ways in which process data may be used to investigate validity claims of an assessment program; and d) how insights from process data may shed new light on learning.

**Invited Symposium 16 - Parallel Session: 8.2C: PROCESS DATA IN ESSAY WRITING**
Hongwen Guo, Educational Testing Service

Keystroke logging (KL) offers the opportunity of analyzing the text production process by using stochastic processes. Analyses based on KL may help researchers understand the strategies that different writers execute during composition, providing more evidence of writing beyond the final production. We focus on modeling of writing processes in terms of writing states (such as long pause, editing, and text producing) and their duration time. Keystroke timing data and features are used to automatically classified keystroke events into a sequence of writing states; and then Markov models and Semi-Markov models are fit to these sequences of states and their duration time. Our results show that the more general semi-Markov models fit the data better. We also investigate subgroup differences in their writing processes given the same writing proficiency.

**Invited Symposium 16 - Parallel Session: 8.2D: ANALYSIS OF TIMING DATA IN A SIMULATION-BASED INTERACTIVE TASK**
Yi-Hsuan Lee, ETS

There has been an increasing need in the measurement field to develop new capabilities for new task formats and assessment types. It is believed that the next generation of assessments should measure such skills as problem-solving, critical thinking, etc., which may not be readily assessed by traditional educational assessments with multiple-choice and constructed-response items. To address this need, efforts have been made at ETS in developing simulation-based and interactive tasks in recent years. This study focused on the timing data collected from 480 test takers who participated in one of such tasks, and investigated how those test takers interacted with the task. In this talk, I will present the statistical properties of the timing data and how they relate to the test takers' task performance, item responses, demographics, and personality. Given that both the task under study and the test-taker population are very different from the typical cases in educational measurement, discussion will be made to connect the findings from this study to general results from traditional educational assessments in the literature on timing data.

**Invited Symposium 16 - Parallel Session: 8.2E: PREDICTIVE FEATURE GENERATION AND SELECTION FROM PROCESS DATA IN LARGE SCALE ASSESSMENTS**
Qiwei He, Educational Testing Service

A variety of timing and process data such as action sequences can be recorded in log files accompanying test performance data when students respond to items in computer-based assessments. This talk draws on process data collected from problem solving items in technology-rich environment in two large scale assessments, the Programme for International Assessment of Adult Competencies (PIAAC) and the Programme for International Student Assessment (PISA) to address how sequences of actions are related to task performance and how to generate features from the process data. The studies were motivated by methods in text mining and natural language processing, for which the target language shares similar structure as action sequences in process data. We generated features via two approaches: disassembling action sequences into mini-sequences with n-grams, and creating features that reflect solving strategies, goal-directed behaviors, and latency information. Two case studies in PISA and PIAAC were used to illustrate a combined strategy in process data analysis, which integrated bottom-up, i.e., exploring the test takers' behavioral features from the observed process data, and top-down approaches, i.e., defining test takers' behavioral features from guiding theories and hypothesis in one systematic framework. We conclude that the methods developed in our study hold promise in process data analysis, and are expected to be applied in scenario-based items with similar settings.

## Classification, Clustering, and Latent Class Analysis: 11:15 AM – 12:45 PM

Chair: Maria Bolsinova

**Classification, Clustering, and Latent Class Analysis: Classification, Clustering, and Latent Class Analysis**

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 8.3A: BEYOND THE Q-MATRIX: A GENERAL APPROACH TO COGNITIVE DIAGNOSTIC MODELS**
Yinyin Chen, University of Illinois at Urbana Champaign

Restricted latent class models (RLCMs) are latent variable models developed to infer latent skills, knowledge, or personalities that underlie responses to educational, psychological, and social science tests and measures. Recent research focused on theory and methods for using RLCMs in an exploratory fashion to infer the latent processes and structure underlying responses. We report new theoretical results about sufficient conditions for generic identifiability of RLCM model parameters. An important contribution for practice is that our new generic identifiability conditions are more likely to be satisfied in empirical applications than existing conditions that ensure strict identifiability. Learning the underlying latent structure can be formulated as a variable selection problem. We develop a new Bayesian variable selection algorithm that explicitly enforces generic identifiability conditions and monotonicity of item

response functions to ensure valid posterior inference. We present Monte Carlo simulation results to support accurate inferences and discuss the implications of our findings for future RLCM research and educational testing.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 8.3B: DO I COMPLETE Q?**
Jimmy de la Torre, The University of Hong Kong; Wenchao Ma, The University of Alabama

A central component for most cognitive diagnosis models (CDMs) is an item and attribute association matrix (Q-matrix; Tatsuoka, 1983), which specifies whether an attribute is measured by each item. 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 CDM to another. A method that has been proposed by Kohn & Chiu (2017) 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 is 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. A set of real data is also analyzed to examine the proposed procedure.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 8.3C: A MULTIPLE CATEGORY COGNITIVE DIAGNOSTIC MODEL FOR SIMULTANEOUSLY IDENTIFYING SKILLS AND MISCONCEPTIONS**
Bor-Chen Kuo, National Taichuang University of Education; Chun-Hua Chen, National Taichuang University of Education

Diagnosing skills and misconceptions are both important goals of diagnosing tests for education. Some cognitive diagnostic models have been proposed for simultaneously identifying them from multiple choice items (GDCM-MC, DiBello, Henson & Stout, 2015; SISM, Kuo, Chen, & de la Torre, 2017). In addition, the test data of constructed response items can provide more useful diagnostic information than multiple choice items, for examples, problem solving process, operational sequences in computerized item. But GDCM-MC and SISM did not mention how to cooperate these more complex information. Actually, some cognitive diagnostic models have tried to apply the complex information provided by constructed response items (Kuo, Chen, & Mok, 2015; Ma, W., & de la Torre, J., 2016). However, these models only focus on identifying skills but not misconceptions. In this study, a new model that can simultaneously identify skills and misconceptions by using the complex information provided by constructed response items will be introduced. An expectation-maximum (EM) algorithm is also applied to estimate the model parameters. Finally, results of simulation studies is presented and discussed.
Keywords: multiple category model, EM algorithm, skills, misconceptions.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 8.3D: STUDY OF CHARACTERISTICS OF TEST COLLUSION GRAPH**
Dmitry Belov, Law School Admission Council

Test collusion (TC) is sharing of test materials or answers to test questions before or during the exam. Potential sources of shared information include teachers, test preparation entities, the Internet, or even examinees collaborating during the exam. Because TC poses a serious threat to the validity of score interpretations, accurately identifying individuals involved in collusion is important. Using graph theory, it is possible to build a graph in which each examinee is defined by a vertex, and every significant response similarity index (RSI) produces an edge connecting the vertices for those two corresponding examinees. This proposal will study how changes in the RSI significance level affect different characteristics of the TC-graph. Among the outcome variables considered are measures of the graph's sparseness and propensity for developing clusters, the number of individuals involved in the largest connected component, the total number of connected components, the size of the largest clique (subset in which all vertices are interconnected), and the distribution of clique size. If TC is not present, each significant RSI represents a false positive; hence, the corresponding TC-graph is expected to consist of entirely random

edges. Characteristics of a TC-graph based on real data that are unusual for a random graph may provide an indication of TC. Using real and simulated datasets with varying amounts and severities of collusion, we will systematically examine changes in graph characteristics as a function of RSI significance level to develop guidelines for optimizing our ability to detect groups of examinees involved in TC.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 8.3E: A COPULA MODEL FOR RESIDUAL DEPENDENCY IN COGNITIVE DIAGNOSTIC MODELS**
Zhihui Fu, Shenyang Normal University and Northeast Normal University; Ya-Hui Su, National Chung Cheng University

Cognitive diagnosis models (CDMs) have been received the increasing attention by educational and psychological assessment. In practice, most CDMs are not robust to violations of local item dependence (LID). Many approaches have been proposed to deal with the LID, such as conditioning on other responses and additional random effects (Mark Patrick Hansen 2013); however, these have some drawbacks, such as non-reproducibility of marginal probabilities and interpretation problem. Braeken, Tuerlinckx, and De Boeck (2007) introduced a new class of marginal models that makes use of copula functions to capture the residual dependence in item response models. In this paper, we applied the copula methodology to model the residual dependencies in CDMs. It is shown that the proposed copula model could overcome some of the dependency problems in CDMs, and the estimated model parameters recovered well through simulations. Furthermore, we develop a free R package to fit the proposed copula CDMs.

## Estimation and Computational Methods: 11:15 AM – 12:45 PM

Chair: Samantha Anderson

**Estimation and Computational Methods**

**Estimation and Computational Methods - Parallel Session: 8.4A: A NONLINEAR DYNAMICAL MODEL OF AFFECT**
Tim Loossens, Katholieke Universiteit, Leuven

Emotions play a prominent role in guiding our actions and determining our well-being. In an attempt to better understand affective dynamics, many computational models have been put forward in psychology and neurobiology. Often, these models suffer from some limitations, which makes them unfit for practical uses. Moreover, there is a large degree of isolation between neurobiological and psychological models. Inspired by the Ising Decision Maker (IDM), which was introduced in studies concerning Choice Response Time, we propose a new affective model. It combines principles from neurobiology and theoretical physics to work its way up from a microscopic description of pooled populations of stochastic binary neurons to a psychological model described by a nonlinear equation of motion of the affective state. Essentially, affective dynamics are described as the emerging collective behavior of pools of stochastic binary neurons, which interact with one another by means of excitation or inhibition. Given its nonlinear nature, it can account for several features, such as multimodalities, metastable states and sudden transitions, features which are observed in affective data samples and which are generally not captured by current models. In this talk, we will explain the model and how it relates to different experimental paradigms, focusing particularly on time series data obtained through Experience Sampling Methods. We will dwell on the statistical inference tools that have been developed, how they allow accurate fitting of the model and how they enable model validation.

**Estimation and Computational Methods - Parallel Session: 8.4B: THE THIRD-ORDER POLYNOMIAL TRANSFORMATIONS OF STUDENT'S T-DISTRIBUTION**
Mohan Pant, University of Texas at Arlington

The third-order polynomial transformations of standard normal distribution, proposed by Fleishman (1978), are widely popular for fitting data and simulating univariate and multivariate non-normal distributions with user specified values of skew, kurtosis, and Pearson correlation. Although Fleishman's polynomial transformations can produce non-normal distributions with a wide range of skew and kurtosis

values (Headrick, 2010), they may not be capable of fitting some empirical and theoretical distributions with values of skew and kurtosis lying within the valid skew versus kurtosis boundary graph. In order to obviate the aforementioned problem, a new family of non-normal distributions based on third-order polynomial transformation of Student's t-distribution is proposed. A system of equations for the product moment-based indices of mean, variance, skew, and kurtosis associated with this new family is derived. Also demonstrated is a methodology to solve for the shape parameters of this new family of distributions. Also included are the examples of fitting this new family of distributions to real-world data arising from some educational and psychological phenomena. Further, a methodology is demonstrated for simulating correlated non-normal distributions with user specified Pearson correlation matrix. The proposed methodology can be applied in a variety of contexts such as fitting data and Monte Carlo simulation studies. Results of a Monte Carlo simulation example indicate that the proposed methodology yields estimates of skew, kurtosis, and Pearson correlation in close proximity of their corresponding parameters.

**Estimation and Computational Methods - Parallel Session: 8.4C: SIMULATION BASED ESTIMATION AND INFERENCE FOR GLLVM**

Maria-Pia Victoria-Feser, Research Center for Statistics, University of Geneva; Guillaume Blanc, University of Geneva; Irini Moustaki, London School of Economics and Political Science

An important challenge frequently encountered in practice, is the computational aspect of the estimation procedure in models with latent variables. Outside the normal setting, getting the MLE becomes computationally very challenging, not only because of the very large number of parameters that are estimated, but also because of the marginalization of the latent variables from the likelihood function. The numerical approximation of these integrals, includes adaptive quadratures ( Rabe-Hesketh et al., 2002) or the Laplace approximation (see e.g. Huber et al., 2004) which is equivalent to an H-likelihood estimation (Wu and Bentler, 2010). Even with approximated likelihood or composite likelihood (Lindsay, 1988), model estimation remains numerically quite challenging in large settings (Varin et al., 2011 and Katsikatsou et al., 2012). Estimation methods that overcome the computational burden in large scale problems (large number of observed and latent variables) are increasingly needed in many disciplines such as educational testing. Moreover, the simultaneously estimation of factor loadings and factor scores with desired estimator properties is still an open research topic with some recent contributions by Unkel and Trendafilov (2010) and Stegeman (2016) and Chen and Zhang (2017). We capitalize on Guerrier et al. (2017) and propose an indirect estimator for the GLLVM with Binary manifest variables, based on an auxiliary estimator for the GLLVM in the normal case. The new estimator is consistent and fast to compute, even in high dimensional settings. Moreover, simulations can be extended, at a low computational cost, to provide finite sample inference for the model's parameters.

**Estimation and Computational Methods - Parallel Session: 8.4D: HIGHLY CORRELATED FACTORS: DIMENSIONALITY ASSESSMENT USING CANONICAL CORRELATION ANALYSIS**

Ginette Lafit, Quantitative Psychology and Individual Differences, Katholieke Universiteit, Leuven; Eva Ceulemans, Katholieke Universiteit

Determining the dimensionality of a set of variables plays a critical role in psychological theory building. Several methods have been proposed to determine the number of factors in a set of variables. Parallel analysis (PA) and multiple average partial procedure (MAP) have shown a good performance when the sample size is large and the factors are minimally correlated. However, when factors are more than minimally correlated or when there is a small number of variables per factor, PA and MAP tend to underestimate the number of dimensions. In the present article, we propose an agglomerative hierarchical clustering method using canonical correlation analysis (CCA) to determine the number of factors and the allocation of each item to a unique factor. Our proposal differs from previous work since we use CCA to estimate the dimensionality taking into account the linear association among variables. We conduct an extensive simulation study and we show that our procedure performs comparable to the popular factor analysis techniques and can satisfactorily determine the number of dimensions when the factors are highly correlated and when there is small number of variables per factor.

## Item Response Theory: 11:15 AM – 12:45 PM

Chair: Carl Francis Falk

**Item Response Theory**

**Item Response Theory - Parallel Session: 8.5A: ON A GENERALIZATION OF LOCAL INDEPENDENCE IN ITEM RESPONSE THEORY**

Stefano Noventa, University of Tübingen; Andrea Spoto, University of Padova; Jürgen Heller, University of Tübingen; Augustin Kelava, University of Tübingen

Knowledge Space Theory (KST) structures are introduced within Item Response Theory (IRT) as a possible way to model local dependence (LD) between items. The aim is threefold: Firstly, to generalize the usual characterization of local independence without introducing new parameters; secondly, to merge the information provided by the IRT and KST perspectives; and thirdly, to contribute to the literature that bridges continuous and discrete theories of assessment. In detail, connections are established between the KST Simple Learning Model (SLM) and the IRT General Graded Response Model (GRM), and between the KST Basic Local Independence model (BLIM) and IRT models in general. As a consequence, local independence is generalized to account for the existence of prerequisite relations between the items, IRT models become a subset of KST models, IRT likelihood functions can be generalized to broader families, and the issues of local dependence and dimensionality are partially disentangled. Models are discussed for both dichotomous and polytomous items and considerations are drawn on their interpretation (e.g., relevance of the parameters, definition of polytomous items as knowledge structures of dichotomous ones, interpretation of Rasch model as a probabilistic version of Guttman's scale). Considerations on possible consequences in terms of model identifiability and estimation procedures are also provided.

**Item Response Theory - Parallel Session: 8.5B: ADAPTIVE PAIRWISE COMPARISON IN EDUCATION**

Elise Crompvoets, Tilburg University/Cito; Anton Béguin, Cito; Klaas Sijtsma, Tilburg University

Pairwise comparison is becoming an increasingly popular assessment method in educational contexts. Pairwise comparison is a method where several raters compare persons in pairs on a trait in order to obtain a rank order of these persons on this trait. An advantage of this method is the holistic approach to traits (i.e., viewing traits as a whole), which is useful for skills that are difficult to measure analytically (e.g., creative thinking). However, pairwise comparison asks a lot from the raters/teachers because many pairwise comparisons are needed for reliable measurement. Pollitt (2012) proposed an adaptive approach to reduce the number of required comparisons. The challenge is to efficiently select person pairs to be compared while the person parameters are still estimated/uncertain. Unfortunately, current algorithms do not take the uncertainty of the person parameters into account. We developed an Adaptive Selection Algorithm (ASA) that maximizes the information from the comparisons while accounting for uncertainty of the person parameters. We investigated the performance of the ASA in a simulation study in terms of reliability, rank order accuracy and the uncertainty of the person parameters in comparison with a Semi-random Selection Algorithm (SSA) under various numbers of persons to be compared and proportions of total comparisons. The commonly-used Bradley-Terry-Luce model was used to analyze the data. The reliability and rank order accuracy did not improve, but the uncertainty of the person parameters decreased using the ASA. In addition, the reliability estimate used for pairwise comparison designs strongly overestimated the true reliability.

**Item Response Theory - Parallel Session: 8.5C: VUONG TESTS FOR MODEL SELECTION OF MIRT MODELS**

Lennart Schneider, University of Tuebingen; Phil Chalmers, University of Georgia; Rudolf Debelak, University of Zurich

Multidimensional Item Response Theory (MIRT) allows for flexible modeling of test data, being capable of reflecting more complex psychological constructs. However, with the application of MIRT models having become more common, the task of selecting the best fitting model has become more difficult. In this talk, we introduce Vuong's (1989) general approach for model selection of both nested and non-nested

models and use it to conduct model selection of MIRT models, relying on the R packages mirt (Chalmers, 2012) and nonnest2 (Merkle & You, 2017). Results of a simulation study focusing on compensatory MIRT models are presented. Comparing the Vuong tests to the traditional likelihood ratio test as well as information criteria and model fit indices, we show that the Vuong tests are a helpful tool to determine the number of dimensions required to adequately model the data, outperforming the traditional likelihood ratio test, which performed surprisingly poorly in the scenarios examined.

**Item Response Theory - Parallel Session: 8.5D: TESTGARDENER: A PROGRAM FOR OPTIMAL SCORING AND GRAPHICAL ANALYSIS**

Juan Li, McGill University; James Ramsay, McGill University; Marie Wiberg, Umeå University

Since 1998, TestGraf has been widely used by researchers and test designers. As its successor, TestGardener is designed to aid the development, evaluation, and use of multiple choice examinations, psychological scales, questionnaires, and similar types of data. This software implements the optimal scoring of binary data (Ramsay & Wiberg 2017) and multi-option data (publication in preparation). Optimal scoring is proposed to be used as a supplement of sum scoring, since it explores the interaction between test taker performance and item impact. Spline smoothing and functional data analysis are also used in fitting the data and generating ICCs. Using TestGardner does not require any programming skill or formal statistical knowledge beyond what would be provided by a first course in statistics in a social science department. More statistically sophisticated users will also find information that they may find helpful and have more control of the analysis. Most of the output from TestGardener is in graphical form that is designed to be self-explanatory, and the program is used interactively. In this talk, I will demonstrate how to use TestGardener to analysis real testing data with various item types and explain some main displays.

**Item Response Theory - Parallel Session: 8.5E: MODEL SELECTION FOR MONOTONIC POLYNOMIAL ITEM RESPONSE MODELS**

Carl Francis Falk, McGill University

One recent semi-parametric approach for item response modeling involves use of a monotonic polynomial in place of the linear predictor for popular response functions. This approach has been applied to allow more flexibility to two-parameter and three-parameter logistic models, and the generalized partial credit model. In addition, use of maximum marginal likelihood allows the approach to be used with planned missing data designs, which are common in conjunction with a long test. However, since polynomial order may vary across items, selection of polynomial order becomes difficult as test length increases. For polynomial orders greater than one, the number of possible order combinations increases exponentially with test length. In this talk, I reframe this issue as a combinatorial optimization problem, and apply a metaheuristic algorithm known as simulated annealing to aid in finding a suitable model. In some ways, simulated annealing resembles Metropolis-Hastings. A random perturbation of polynomial order for some item is generated. Acceptance of this candidate model depends on the change in model fit and the current algorithm state. In simulations, this approach was found to be a computational feasible way to select a model that fits better according to standard information criterion. Results on the recovery of response functions and respondent scores were mixed, with the performance of this approach versus alternatives depending on the studied conditions.

## Resampling and Simulation Techniques: 11:15 AM – 12:45 PM

Chair: Duanli Yan

**Resampling and Simulation Techniques**

**Resampling and Simulation Techniques - Parallel Session: 8.6A: A RESAMPLED-BASED METHOD FOR OUTLYING VARIABLE DETECTION**
Sopiko Gvaladze, Katholieke Universiteit, Leuven; Kim De Roover, Tilburg University; Eva Ceulemans, Katholieke Universiteit

In psychology, many studies yield multivariate multi-block data. To deal with the high-dimensionality, researchers often reduce the variables to factors or components (depending on the framework that is used). Before one can compare the scores on these components across the blocks (e.g., mean comparison) one should first make sure that those components measure the same construct across the blocks. To this end, the variables that hamper equivalent interpretation of the components across blocks, which we call outlying variables, should be removed. Moreover, detecting outlying variables can, for example, be used to test theories about which variables are invariant and which ones have different meanings across cultures. In this paper, we first scrutinize the lower bound congruence method (LBCM) that was recently proposed for outlying variable detection. LBCM is based on clusterwise SCA-P and uses Tucker's congruence to measure similarity across the cluster loading matrices. We demonstrate that LBCM has the tendency to output false positives for both real and simulated data. In order to address this issue, we present two new outlying variable detection heuristics: the test-based method and the hybrid method. Both methods use a resampling technique to obtain a sampling distribution for the congruence coefficient, given that no outlying variable is present. In a simulation study, we show that the hybrid method outperforms the test-based method and LBCM, and we demonstrate how to apply the methods to two real-data applications.

**Resampling and Simulation Techniques - Parallel Session: 8.6B: TEST-TAKING MOTIVATION IN INTERNATIONAL SURVEYS: AN IRT APPROACH**
Denise Reis Costa, CEMO; Hanna Eklöf, Umeå University

In this work we propose an IRT approach to model the self-reported test-taking motivation in PISA and TIMSS Advanced studies. Both assessments are used to evaluate educational quality and student proficiency in an international context. Using six items included in the student's questionnaire as a national option in the Swedish context, we created an IRT motivation scale for each assessment. For the PISA, we evaluated changes in test-taking motivation between 2012 and 2015, whereas the 2008 and 2015 cycles were used for the TIMSS Advanced. Both scales were created using a unidimensional Generalized Partial Credit model. Differential item functioning analysis was also incorporated into the analysis, due to a clear difference on the self-reported motivation over the time.

**Resampling and Simulation Techniques - Parallel Session: 8.6C: IMPACT OF ITEM POSITION EFFECTS ON RESPONSE TIMES IN SURVEYS**
Olga Kunina-Habenicht, University of Education Karlsruhe; Frank Goldhammer, German Institute for International Educational Research (DIPF), Centre for International Student Assessment (ZIB); Ralph Carstens, International Association for the Evaluation of Educational Achievement (IEA) Data Processing and Research Center (DPC); Philipp Koehme, International Association for the Evaluation of Educational Achievement (IEA) Data Processing and Research Center (DPC)

In international large-scale assessments often incomplete booklet designs are applied. For achievement tests the booklet or item position effects on item difficulty were frequently investigated, showing that item difficulty can either increase at the end of the test due to fatigue or decrease due to test-wiseness. We address the impact of item position effects on response times in surveys. Time spent on an item can partly indicate the extent to which cognitive resources are allocated to processing the questions and responding. To our knowledge, for surveys the impact of item position effects on response times has not yet been investigated systematically. Our study focuses on how the time allocation for questionnaire items changes while proceeding through the questionnaire. Similarly to cognitive assessments, we

assumed that with increasing questionnaire length test takers spend less time on single items due to fatigue. We used data from approximately 19.000 teachers and 40 different countries from the field-test of the international computer-based TALIS 2018 survey. It won't be possible to identify the individual countries. In TALIS a rotation design with three booklets was applied, whereby the position of selected questionnaire scales was a-priori systematically varied twofold. We compared response times between two different item positions for two selected questionnaire scales. For both scales, ANOVAs and structural equation models showed a small significant fixed effect of the scale position on response time with lower times at the latter position. Further, ANOVAs showed significant random effects for country indicating that response times differ between countries.

# Friday, July 13, 2018 AM

## IMPS Registration: 8:00 AM – 2:00 PM

## Symposium 17: 8:30 AM – 10:00 AM

Chair: Sacha Epskamp

**Symposium 17: Network Psychometrics II: Psychometric Extensions to Network Modeling**

**Symposium 17 - Parallel Session: 9.1A: BOUNCING BACK FROM ADVERSITY: A DYNAMICAL CONCEPTUALIZATION OF PSYCHOLOGICAL RESILIENCE**
Gabriela Lunansky, University of Amsterdam

The question of why some people develop psychopathology following adversity while others do not, has been extensively investigated in the psychological literature. A pressing yet unanswered question is what makes the latter group of people resilient in that they are seemingly able to withstand the 'attack' that adverse events launch on their health and well-being. Research seems to be predominantly about rather static indicators (e.g., childhood maltreatment) while the concept of resilience is deeply dynamical. Moreover, while other theoretical definitions sound dynamic, quite a few operationalizations are static in nature. To date, there is no consensus on what resilience exactly is conceptually, which seriously hampers progress as various definitions and operationalizations abound. Resilience is a concept that has been extensively studied in dynamical systems such as the growth of a bacteria population and the turbidity of a lake. A relatively new modeling approach conceptualizes mental disorders as a dynamical system in which the dynamics are driven by a network that consists of symptoms of a disorder that stand in certain direct relations to one another. Based on such a dynamical systems perspective, this presentation will discuss an alternative framework of psychological resilience. It will be shown how the architecture of psychopathological networks can be related to resilience indicators (e.g., connectivity of the network is related to robustness). Hereby, we will open the door to study resilience as a dynamical process, while simultaneously taking into account possible static factors that play a role in this dynamical process.

**Symposium 17 - Parallel Session: 9.1B: A FUSED LATENT AND GRAPHICAL COX MODEL FOR MULTITYPE RECURRENT EVENTS**
Yunxiao Chen, Emory University

Latent factor models have been the statistical foundation for the modern measurement theory, receiving wide applications in educational testing, personality assessment, and mental health diagnosis. With the advances in technology, innovative computer-based instruments, such as simulation-based and game-based assessments, are developed to uncover the underlying characteristics of individuals (e.g., complex problem-solving ability, cognitive ability, and personality). Such new measurement tools collect individuals' entire process of solving one or multiple problems, producing data that cannot be handled by the classical latent factor models. In this paper, we propose a model for the factor analysis of multivariate recurrent event processes, a type of data that is commonly collected in the innovative computer-based assessments. This model contains a latent variable component and a sparse graphical component, where the former captures the underlying factors shared by the counting processes of multiple event types and the latter accounts for ad-hoc event-type specific dependence. The proposed model can be viewed as an

extension of the Cox proportional hazards model and is constructed by imposing structures on the intensity functions of a multivariate counting process. Statistical procedures and a computational algorithm are proposed for model selection and parameter estimation. The proposed method is motivated by the analysis of problem-solving process data recorded in computer log-files from the Programme for International Student Assessment (PISA) 2012. Simulation studies are conducted to further evaluate the performance of the proposed method.

**Symposium 17 - Parallel Session: 9.1C: NETWORK-BASED ADAPTIVE ASSESSMENT**
Sacha Epskamp

The network perspective to psychology conceptualizes observed variables (e.g., attitudes, symptoms, and moods) as causal agents in a complex interplay of psychological, biological, sociological and other components. Aiming to map out this interplay, recent psychometric developments have led to methodology for estimating network models from psychological datasets (network psychometrics). These methods have since grown popular in diverse fields of research. This presentation will discuss how network psychometrics may be used in an innovative adaptive measurement system, leading to solutions for diverse problems such as diagnostic tools (e.g., diagnosing which symptoms of the entire DSM a person endorses), patient monitoring (e.g., assessing changes in severity of select symptoms over time), online applications (e.g., voting recommendation apps), and the analysis and gathering of large epidemiological datasets. Data gathering in network psychometrics is currently static: first a dataset is collected, and then a statistical model is estimated. Participants are required to respond to many questions, possibly several times per day over a period of weeks. In network-based adaptive assessment, only informative items will be administered that best predict other responses (adaptive administration), and a network model will be updated as new data becomes available (adaptive model estimation). The presentation will show first results in which the network structure of items from a popular online application (N = 5,998) is used to derive an optimal adaptive item-order. Furthermore, future directions and research topics will be discussed.

**Symposium 17 - Parallel Session: 9.1D: DIMENSIONALITY ASSESSMENT ON INTENSIVE LONGITUDINAL DATA USING EXPLORATORY GRAPHICAL VAR**
Hudson Golino, University of Virginia; Eva Ceulemans, Katholieke Universiteit; Marlies Vervloet, Katholieke Universiteit Leuven

Assessing the number of dimensions (latent variables) in cross sectional data can be done by several classical procedures, from Kaiser-Guttman rule (eigenvalue greater than one) to parallel analysis. Recently, a new approach termed Exploratory Graph Analysis (EGA) was proposed, where a network is estimated using regularized partial correlations, with the level of sparsity controlled via graphical lasso. A community detection algorithm for weighted networks can be used to assess the number of clusters. It was empirically demonstrated, via a simulation study, that the number of latent variables simulated can be correctly recovered using EGA, even when the correlation between factors is high (.7). In the current presentation, a similar approach will be followed to assess the number of dimensions (latent variables) in intensive longitudinal data using a graphical vector autoregressive model (order 1). The number of latent variables can be assessed using community detection algorithms in the partial contemporaneous correlation matrix (PCC). The PCC is the correlation between two variables at the same point in time after removing the linear effects of the other variables at the same point in time and all variables at previous times. A simulation study will be presented to show the accuracy of the proposed approach. The presentation will end with empirical examples of intensive (intraindividual) longitudinal research in the field of psychological disorders.

**Symposium 17 - Parallel Session: 9.1E: SPARSE NETWORK AND COMPONENT (SNAC) MODEL FOR INTEGRATING MULTI-SOURCE DATA**
Pia Tio, Tilburg University; Lourens Waldorp, University of Amsterdam; Katrijn Van Deun, Tilburg University

Network analysis has successfully been applied to many different types of psychological data, including personality, cognitive performance, and clinical symptoms. While investigating these different areas in isolation of the other ones is useful, a better understanding of their structure requires an integrated analysis. Investigating such cross-source relationships requires (possibly) large data sets containing

information about individuals from multiple sources (big data). Such data are becoming more and more commonplace. However, estimating a network using big data is not without its challenges. The dimension of the dataset, often containing more variables than observations, hinders accurate estimation of relations, even when some form of regularisation (e.g. lasso penalty) is used. Reducing the number of variables to those involved in cross-source relationships would be a straightforward way to remove (or at least reduce) this problem, except that we do not yet know which variables are involved in cross-source relationships. An additional challenge is that big data contains data from different sources that inherently may have different characteristics. For example, indicators of cognitive performance are expected to correlate much higher with one another than indicators of gene expression. Applying network analysis to such data without taking this difference into account again leads to inaccurate estimation of relationships. We propose the Sparse Network and Component (SNAC) model, which combines regularized simultaneous component analysis with the network framework. In a simulation study SNAC analysis outperforms sparse network analysis when estimating cross-source relationships from multi-source data. Here we present an empirical application of SNAC using R.

## Measurement Invariance and DIF: 8:30 AM – 10:00 AM

Chair: Anton Beguin

**Measurement Invariance and DIF**

**Measurement Invariance and DIF - Parallel Session: 9.2A: DETECTING DIFFERENTIAL ITEM FUNCTIONING FROM THE MULTI-PROCESS APPROACH**
Hui-Fang Chen, City University of Hong Kong; Kuan-Yu Jin, The University of Hong Kong

Multi-process IRT or (IR)tree models use a tree-like approach to describe the process of reaching a response category in Likert-type data and perform well in detecting extreme response style (ERS). These approaches usually divide the process into three steps: (1) indifference; (2) direction; and (3) intensity. A binary pseudo item (BPI) is created in each step, and such decomposition would inevitably result in non-imputable missingness in the second and third steps. Then, these BPIs are then examined with simple-structure multidimensional IRT models. Up to date, however, none of studies has investigated differential item functioning (DIF) under the framework of the tree approach. The present study examined how the logistic regression (LR) and the odds ratio (OR) methods performed in DIF detections through a series of simulations. Results showed that the OR method was rather robust across all DIF conditions and yielded satisfactory false positive rates (FPRs) and true positive rates (TPRs), whereas the LR method yielded inflated FPRs, especially when fitting to the BPIs derived from the second and third steps.

**Measurement Invariance and DIF - Parallel Session: 9.2B: SECULAR CHANGES IN FLUID INTELLIGENCE OF NORWEGIAN ADOLESCENTS: A RANDOM DIF APPROACH**
Fredrik Helland-Riise, Centre for Educational Measurement at the University of Oslo (CEMO); Ole Christian Lang-Ree, The Norwegian Defence University College; Johan Braeken, CEMO, University of Oslo

Studies on the military entrance tests in Norway have contributed to insights in secular changes in intelligence occurring in the developed world, also known as the Flynn-effect. By using simple sum scores, previous studies implicitly assumed measurement invariance across time, taking for granted that there is no cohort-wise differential item functioning (DIF) and that fair comparisons across cohorts were possible. This study revisits this issue from the item level perspective for the figural matrices part of the military entrance test. As these tests are modernized through automatic item generation (AIG) or administered using for instance computer-adaptive testing (CAT), it is important to establish strong theory and validity evidence on the construct - what features affect the difficulty of an item (radicals) and what features are merely cosmetic (incidentals). Theory on figural matrices points towards the number of elements and rules, as well as the type of rule and perceptual ambiguity in an item as being radical features, but are they stable across time? We will address this question by investigating differential feature functioning (DFF) between cohorts. Between 2011 and 2017, a figural matrices test of 36 items was administered to cohorts of Norwegian adolescents. DIF and DFF between cohort groups will be analyzed within an explanatory item response modelling framework. We expect to shed light on secular changes occurring in fluid intelligence in Norway, as well as strengthening the construct representation

evidence for figural matrices tests in general, potentially also pointing at which features contribute to the expected secular changes.

**Measurement Invariance and DIF - Parallel Session: 9.2C: COMPARISON OF DIF DETECTION METHODS USING THE C-RUM MODEL**

Kevin Krost, Virginia Polytechnic Institute and State University; Gary Skaggs, Virginia Polytechnic Institute and State University

Cognitive diagnostic models are a growing focus within psychometrics due to their many positive features. Most research has focused on their methodological features and model developments, with empirical studies increasing in popularity recently. Many models have been developed, but the noncompensatory deterministic, input noisy and gate (DINA) model remains one of the most common. Several methodological issues have been evaluated based on this model, including differential item functioning (DIF) (Hou et al., 2014; Li & Wang, 2015). Another model is the compensatory reparameterized unified model (C-RUM), which makes different assumptions about item parameters. Thus far, DIF has not been evaluated for the C-RUM model. This simulation study evaluated the power and Type I error rates of the Wald and likelihood ratio tests to detect DIF using the C-RUM model. Similar conditions to previous studies were replicated, including a test length of 30 items and a five attribute Q-matrix. Additional fully-crossed factors were evaluated, including the comparison of p-values and Bonferroni-corrected p-values, sample size, Q-matrix complexity, DIF type, and magnitude. There were many findings about power, Type I error rates, and the impact of factors. Power and Type I error rates were higher with the unadjusted p-value than the Bonferroni p-value. In most scenarios, the likelihood ratio test had higher power and lower Type I error rates than the Wald test. Power was higher for detecting uniform DIF than non-uniform DIF, and Type I error rates were lower. Power increased as both sample size and DIF magnitude increased.

**Measurement Invariance and DIF - Parallel Session: 9.2D: DIF EFFECT SIZE MEASURES**

Daniella Reboucas, University of Notre Dame

Statistical significance in DIF detection is commonly paired with consideration of its practical significance. In large samples, an estimated DIF effect may be statistically significant but negligible in practice. Using DIF effect size measures along with a DIF assessment method has been shown to reduce inflated type I error rates (Hidalgo, Gómez-Benito & Zumbo, 2014). Some commonly used effect size measures are the Delta scale in the Mantel-Haenszel method (Holland & Thayer, 1988), R2 in the logistic regression approach (Zumbo & Thomas, 1997), and, under the item response theory framework, the signed area between the item characteristic curves of the focal and reference groups (Raju, 1988) and standardized indices based on the probability difference between groups (Wainer, 1993). Some measures take into account the underlying ability distribution of reference/focal groups, while others do not perform well when the ability distributions of focal and reference groups differ (Jin, Myers, Ahn & Penfield, 2012). For an appropriate effect size measure, it should demonstrate certain desirable traits at the population level: sensitivity to the size of model misspecification, and insensitivity to sample characteristics (e.g., size and distribution) and estimation method. In this study, we will evaluate existing effect size measures against these criteria, and propose an alternative effect size measure that has desirable characteristics on these attributes.

**Measurement Invariance and DIF - Parallel Session: 9.2E: THE EFFECT OF DIF ON MINI AND MIDI ANCHOR TEST**

Çiğdem Akın Arıkan, Hacettepe University; Hatice İnal, Mehmet Akif Ersoy University

The non-equivalent groups anchor test (NEAT) design is the most popular equating design in literature. In this design, the anchor test is used to try to separate difficulty differences from ability differences (Kolen & Brennan, 2004). The items within the anchor test should be a representative sample (refer to mini test) of the test (Budescu, 1985; von Davier, Holland, & Thayer, 2004; Kolen & Brennan, 2004). However, Sinharay and Holland (2006) report that there is no proof that the spread of the difficulty of the anchor test should be representative of the total test. Sinharay and Holland (2006, 2007) propose using midi tests that is matched to the full test in terms of content and mean difficulty, but with less variance in item difficulty. For this study, another issue that should be considered is the effect of differential item

functioning in equating. The purpose of this study is to examine the effect of DIF in mini or midi anchors on the group invariance in IRT equating. Within this scope, the factors chosen to investigate the population invariance in equating are types of anchors, the frequency of anchor items displaying DIF, direction and magnitude of DIF in the anchor test. Data analysis will be conducted in two steps. At the first step, for each condition, equated scores will be obtained by using IRT equating method (mean-mean and mean standard deviation). At the second step, population invariance indexes will be calculated for each condition.

## Validity and Reliability: 8:30 AM – 10:00 AM

Chair: Elise Dusseldorp

**Validity and Reliability**

**Validity and Reliability - Parallel Session: 9.3A: NEW METHOD FOR DETECTION OF TEST COLLUSION AMONG MULTIPLE EXAMINEES**

Hongling Wang, ACT, Inc.; Chi-Yu Huang, ACT, Inc.

Test collusions among multiple examinees such as illicit coaching by teachers and examinees obtaining answers from item harvesters greatly jeopardize test validity. However, the detection methods related to this field of test security are still lacking. This study explores a simple two-step process to detect test collusions among multiple examinees. First, the suspected pairs among all the possible examinee pairs who have strong response similarity are flagged. Index B (Angoff, 1974) and many other indices in literature can achieve this purpose. Then, the suspected examinees who are involved in the suspected pairs are clustered into suspected collusion groups. The suspected examinees who are related directly through a suspected pair or indirectly through more than one suspected pairs are grouped as a suspected collusion group. Responses of the examinees in each group are then analyzed to determine whether they share some types of answers or not. In this study, we will simulate a test of 50 items by 5000 examinees. Four factors will be considered in the simulation: numbers of collusion groups (1, 2, and 4 groups), sizes of collusion groups (10, 20, and 40 examinees), numbers of colluded items (20, 30, and 50 items), and percentages of item response agreement within a collusion group (80% and 100%). We will explore the criteria that are used to detect collusion groups. In each condition, we will compute type I error and type II error of the new method in flagging collusions.

**Validity and Reliability - Parallel Session: 9.3B: GAUGING UNCERTAINTY IN TEST-TO-CURRICULUM ALIGNMENT INDICES**

Anne Traynor, Purdue University; Shuqi Zhou, Purdue University

During the development of large-scale school achievement tests, recruited panels of independent Subject-Matter Experts (SMEs) use systematic judgmental methods to rate the correspondence between a given test's items and the objective statements in a particular curricular standards document. The individual experts' ratings may then be used to compute a variety of mean "alignment" indices, which are compared to suggested criterion values to determine how well the given test matches its target item domain. Existing studies of variability in alignment index values have focused on whether alignment between a particular test and its corresponding curriculum is significantly greater than would be expected by chance — which seems to be a low standard for test content quality. The magnitude of alignment index variability across experts within a panel, and across randomly-sampled expert panels, is largely unknown. Using rater-by-item level data from more than 40 alignment reviews for US states' Grade K-12 achievement tests, we will compute the observed standard deviations of individual experts' index values, and estimate bootstrap standard error, for the alignment indices suggested by Frisbie (2003) and Webb (2007). The results will permit us to characterize "typical" and "extreme" rater and sampling error levels, which may have implications for the interpretation of alignment evidence during validation. Then, using mixed-effects generalizability theory models, we will predict the number of SME raters that would be required for these alignment indices to attain acceptable precision.

**Validity and Reliability - Parallel Session: 9.3C: EXAMINATION QUALITY: MERGING RELIABILITY AND VALIDITY**
James Penny, Castle Worldwide, Inc.; Robert L. Johnson, University of South Carolina

After a brief introduction to serve as an advance organizer for the session, we will provide the new NCCA Standard regarding reliability in sufficient detail to frame the next few slides describing common forms of reliability and when they might be used. We will then repeat those steps for validity: the new NCCA Standard regarding validity, common forms of validity, and when we might encounter and use them. We then will focus the discussion on content validity because its use is so common in certification, and then present an easily implemented methodology to compute a numerical index with values between 0 and 1 that will measure content validity in a manner reflective of other indices of reliability and validity. We will finish the brief descriptions of reliability and validity with a synopsis of the previous slides to cement the current frameworks of reliability and validity in the thinking of the audience as we segue into the next section of the session where we present a new framework to create a single measure of examination quality from extant measures of reliability and validity. Although the mathematical complexity of the quality framework is on par with that of the Pythagorean Theorem, we do not intend to delve into the mathematics more than to show a very few equations and how they work (mostly to demonstrate they exist), followed by a few 3-dimensional graphics to further illustrate the interplay of reliability and validity with quality.

**Validity and Reliability - Parallel Session: 9.3D: EFFECTS OF RELIABILITY ON INCREMENTAL VALIDITY: A STATISTICAL LEARNING PERSPECTIVE**
Bunga Citra Pratiwi, Leiden University

Incremental predictive validity (IV) research is vital for various purposes such as to guide decisions about the addition of a new psychological test to a selection system (Schmidt & Hunter, 1998). The common approach to assess IV is based on regression techniques within the null-hypothesis testing framework. Finding a significant increase in explained variance is the usual evaluation criterion. We argue that when adding a test for the purpose of prediction, a different definition of IV should be employed. The field of statistical learning provides an appropriate definition of prediction, that is, the predictive accuracy of a prediction rule which is evaluated on out-of-sample data. In the context of IV, there will be two prediction rules that we must test, one that contains only data from previous test(s) and in the second rule we add data from the test to be validated. The increase in out-of-sample predictive accuracy (decrease in prediction error) is our definition of IV. There are two main goals of this study: 1) to investigate the relationship between the reliability of a test and its IV given this new definition and 2) to compare the classical least squares method with correction methods such as the simulation-extrapolation algorithm (Cook & Stefanski 1994), and a shrinkage method (Hoerl & Kennard, 1970) on their assessment of IV. Based on simulation experiments, we show that the relationship between reliability and IV depends on a wide range of circumstances and that in most cases, a shrinkage method helps to optimize IV.

## Estimation and Computational Methods: 8:30 AM – 10:00 AM

Chair: Maarten Marsman

**Estimation and Computational Methods**

**Estimation and Computational Methods - Parallel Session: 9.4A: NEW EFFICIENT AND PRACTICABLE ADAPTIVE DESIGNS FOR CALIBRATING ITEMS ONLINE**
Ping Chen, Beijing Normal University; Yinhong He, Beijing Normal University; Yong Li, Beijing Normal University

In the online calibration scenario, the calibration framework proposed by van der Linden and Ren (2015) (referred to as VR framework) is practically feasible, because it assigns the optimal new item to the current examinee who reaches a seeding location according to some criterion (e.g., D-optimality). Under the VR framework, if a new item is assigned to the current examinee, it means that the current examinee is more helpful in accurately calibrating this new item than in calibrating any other new items. However,

compared to all examinees in the examinee pool, the current examinee may not be the optimal design point for calibrating the selected new item. In this regard, the VR framework may be less efficient than the traditional optimal designs with static examinee pool. To improve the calibration efficiency of VR framework and evaluate the importance of the current examinee relative to the optimal design point for each new item, two new calibration designs (i.e., maximin efficiency design and maximean efficiency design) are proposed, depending on whether the "worst" or "mean" value from a small neighborhood of the ability estimates is used to handle the uncertainty inherent in ability estimates. Simulation studies were conducted under a variety of conditions to compare the two new designs with the D-VR-optimal design (i.e., D-optimal design under the VR framework) and random design in terms of calibration efficiency and precision. Results showed that maximean efficiency design performed the best, followed by maximin efficiency design, D-VR-optimal design, and random design.

**Estimation and Computational Methods - Parallel Session: 9.4B: PREPAID PARAMETER ESTIMATION WITHOUT LIKELIHOODS**

Merijn Mestdagh, Katholieke Universiteit, Leuven

In various fields, many models of interest are analytically intractable. As a result, statistical inference is greatly hampered by computational constraints. However, given a model, different users with different data are likely to perform similar computations. Computations done by one user are potentially useful for other users with different data sets. We propose a pooling of resources across researchers and datasets to capitalize on this. More specifically, we preemptively chart out the entire space of possible model outcomes in a prepaid database. Using advanced interpolation techniques, any individual estimation problem can now be solved on the spot. We created prepaid databases for three challenging models and demonstrate how they can be distributed through an online parameter estimation service. Our method outperforms state-of-the-art estimation techniques in both speed (with a 23,000 to 100,000-fold speed up) and accuracy, and is able to handle previously inestimable models.

**Estimation and Computational Methods - Parallel Session: 9.4C: ESTIMATING CROSS-CLASSIFIED ITEM RESPONSE MODELS USING THE MH-RM ALGORITHM**

Nicholas Rockwood, The Ohio State University

Item response models with crossed random effects are notoriously difficult to estimate using Maximum Likelihood (ML) due to the high dimensional integrals in the marginal likelihood function, for which there is no closed-form solution. For high dimensional IRT models without crossed random effects, the Metropolis-Hastings Robbins-Monro algorithm has shown to be relatively accurate and efficient (Cai, 2010a, 2010b). The MH-RM algorithm works by iteratively using Monte Carlo (MC) to simulate the random effects from their conditional posterior distributions and updating the model parameters using a Robbins-Monro step (Robbins & Monro, 1951) which, over time, filters out the random noise induced by the MC step. Despite its utility for traditional item response models, the MH-RM algorithm has not yet been applied to models involving crossed random effects, which is the focus of the present research. I focus on a particular example in which multiple teachers provide ratings of multiple students on a series of items, resulting in item responses that are cross-classified within teachers and students. The model of interest includes 4 factors at the student level and 2 factors at the teacher level. Comparing the estimates obtained using the MH-RM algorithm to those obtained using Bayesian estimation with diffuse priors results in no substantive differences, but the MH-RM algorithm was noticeably faster. Following, the estimates were used as population parameters in a small-scale simulation study using 100 replications. The MH-RM algorithm recovered the true parameters well, demonstrating its potential use for ML estimation of similar cross-classified item response models.

**Estimation and Computational Methods - Parallel Session: 9.4D: CONTROLLING THE IMPACT OF RESPONSE BIASES: AN FA-BASED APPROACH**

David Navarro-González, Rovira i Virgili University; Andreu Vigil-Colet, Rovira i Virgili University; Pere Joan Ferrando, Rovira i Virgili University; Urbano Lorenzo-Seva, Rovira i Virgili University

The responses to self-reports are susceptible to response bias, which is a systematic tendency to answer the items on some other basis than the specific item content. A review of the literature on response biases indicates that Acquiescence (AC) and Social Desirability (SD) can impact both the individual scores

and the factorial structure of typical response measures such as personality traits. Given these findings, test developers often use some type of procedure to control or minimize the effect of AC and SD when designing questionnaires. However, most of these procedures are purely descriptive and have some shortcomings due to the ad hoc approaches inherent in those methods. For overcoming these limitations, Ferrando, Lorenzo-Seva & Chico (2009) developed a restricted FA model to assess simultaneously the effects of Acquiescence and Social Desirability, thus allowing these biases to be modelled as additional factors that can be distinguished from the content factors. The procedure has been considered of interest in applied research and it has been successfully used to develop different questionnaires. We also have developed a stand-alone program that enables the implementation of the procedure, providing an easy way of performing factor analysis by controlling the effect of AC and SD. In this communication, we are going to expose the findings of our investigations, showing that AC is the main source of the distortions on the factor loading matrices in the questionnaires tested so far.

## Item Response Theory: 8:30 AM – 10:00 AM

Chair: Bernard Veldkamp

**Item Response Theory**

**Item Response Theory - Parallel Session: 9.5A: MODELING EXTREME RESPONSE STYLES IN BEHAVIORAL GENETICS USING IRTREES**

Jack DiTrapani, The Ohio State University; Nicholas Rockwood, The Ohio State University; Minjeong Jeon, UCLA

Behavioral genetics studies often employ the ACE model to decompose the variance of a latent trait into variances due to genetic and environmental influences. The manifest variables for these types of analyses are often categorical in nature (e.g., item responses). In such circumstances, an item response theory (IRT) model can be incorporated into the ACE model as the measurement model. Standard IRT models do not typically account for different response style tendencies, such as extreme response style (ERS), which is a given respondent's propensity to endorse item categories at a scale's endpoints regardless of the underlying latent trait being measured. Work utilizing item response trees (IRTrees; De Boeck & Partchev, 2012) to model ERS tendencies has shown that ignoring ERS can alter conclusions reached about the latent trait under investigation (Böckenholt, 2017). However, ERS has yet to be explored using IRTrees in the context of ACE models, which is the focus of the present research. Specifically, we use an IRTree to model the latent trait of interest and a latent ERS trait. The variances of these latent variables are decomposed into genetic and environmental factors. The utility of this research is twofold. First, we obtain estimates of the proportions of variance in ERS and the substantive latent trait due to genetic and environmental influences. Second, we compare the estimated genetic and environmental influences in the trait of interest when the IRTree model is used, relative to the model that does not control for ERS.

**Item Response Theory - Parallel Session: 9.5B: A THRESHOLD FOR DETECTING INATTENTIVE RESPONSE BEHAVIOR USING MIXTURE IRT**

Juyeon Lee, University of Georgia; Allan S Cohen, University of Georgia; Seock-Ho Kim, University of Georgia

Mixture IRT has been used for detecting inattentive response behavior on questionnaire data. Mixture IRT permits distinguishing normal responses from inattentive responses by assuming individual differences in usage of the response categories. This approach is not always straightforward, however, since characterizing different latent classes is typically an exploratory approach. A confirmatory mixture IRT has been proposed to deal with this issue (Böckenholt & Meiser, 2017; Jin et al., 2017). In this confirmatory model, the number of latent classes and the probability of the response categories are specified first. These two studies have demonstrated good performance of a mixture polytomous IRT model with constraints on the threshold to capture the inattentive response behavior. There is yet no clear consensus, however, about how to impose constraints on thresholds. In this study, we focus on examining constraints on the threshold, comparing use of pre-knowledge based on response style (RS; Jin et al., 2017) and the linearly restricted threshold distance across latent classes by (Böckenholt & Meiser, 2017). The present study assumes two types of the inattentive responses: RS (e.g., an individual

tendency toward a particular response category, such as extreme RS or midpoint RS, and a random guessing group assumed to reflect carelessness or unwillingness to respond). An empirical data set will be analyzed to illustrate issues involved in comparing thresholds and a simulation study to investigate different thresholds and priors. MCMC estimation will be used to estimate model parameters.

**Item Response Theory - Parallel Session: 9.5C: A COMPARISON OF MODELS FOR CONTROL OF RESPONSE STYLES IN ORDINAL ITEMS**

Walter Leite, University of Florida; Jue Zhou, University of Florida; Anne Corinne Huggins-Manley, University of Florida; Ricardo Primi, Universidade São Francisco

This paper aims to describe and evaluate a new psychometric model to adjust estimates of latent traits for the effects of response styles to scales with ordinal items, and to compare the new method with existing methods. These goals are accomplished with a Monte Carlo simulation study and an analysis of items from the Programme for International Student Assessment (PISA, 2012). Response styles are individual tendencies to respond to ordinal items with certain response categories in consistent ways that are unrelated to the individual's level on the latent trait being measured, such as extreme response style (ERS; tendency to select the extreme categories) and middle-category response style (MRS; tendency to select the middle category). It has been shown that response styles result in biased latent trait estimates when the trait is correlated with the response style, because score variability due to response styles is incorrectly attributed to variation in the trait. The new psychometric model is a mixture generalized partial credit model with constraints on ordering of item thresholds. It will be compared with Falk and Cai's (2016) multidimensional nominal response model (MNRM), and Bolt et al. (2014) MNRM with anchoring vignettes. This paper will provide guidelines about which models work best under a variety of testing conditions such as the number of scale items, number of respondents, and number of response categories. These guidelines are valuable to researchers who need to select a method to control for response styles in their assessments.

**Item Response Theory - Parallel Session: 9.5D: CONFIRMATORY IRT(REE) MODELS FOR ANALYZING THE DYNAMICS OF RESPONSE-STYLE EFFECTS**

Thorsten Meiser, University of Mannheim; Mirka Henninger, University of Mannheim; Hansjörg Plieninger, University of Mannheim

Multidimensional IRT models and multi-process IRTree models have been proposed for analyzing and controlling response styles in rating data, such as a general tendency to choose extreme or midpoint categories. Here we extend IRT and IRTree models to accommodate and test moderating effects of item position and item complexity on the strength of response-style effects. For that purpose, individual response styles are considered as latent factors in multidimensional IRT, random-threshold IRT or multi-process IRTree models, and the model parameters are specified as functions of item position and complexity. Modeling the loading weights of response-style factors as functions of position and complexity (i.e., fixed links models) allows one to analyze sources of variation in the impact of individual response styles on the choice of rating categories over items. Modeling the difficulty parameters of pseudo-items in IRTrees as functions of position and complexity (i.e., LLTMs) affords testing sources of heterogeneity between items in prompting certain kinds of response categories. This talk outlines the formal foundations of confirmatory IRT and IRTree models for the analysis of variations in the strength of response-style effects, including model equations and implementation, parameter estimation and model selection. The extended IRT and IRTree models are illustrated with applications to empirical data sets.

**Item Response Theory - Parallel Session: 9.5E: TESTING WITHIN-CLASS DISTRIBUTIONS IN MIXTURES OF RESPONSES AND RESPONSE TIME**

Renske Kuijpers, University of Amsterdam; Ingmar Visser, University of Amsterdam; Dylan Molenaar, University of Amsterdam

In the past, mixture models have been applied to the responses of psychometric test administrations to reveal differences in the response processes across subjects (e.g., Mislevy & Verhelst, 1987; Schmittmann, Dolan, & Van der Maas, 2005). More recently, interest has grown in adding the item response times to this mixture analysis to improve the detection of differences in response processes across subjects (e.g., Van der Maas & Jansen, 2003; Molenaar, Tuerlinckx, & Van der Maas, 2015) and to

enable tests on differences within-subjects (Wang & Xu, 2015; Molenaar, Oberski, Vermunt, De Boeck, 2016; Schnipke, & Scrams, 1997). In order to enable mixture modeling of both the item response times and the responses, a distributional assumption is needed for the within-class response time distribution. Since violations of the assumed response time distribution will bias the modeling results (e.g., Bauer, & Curran, 2003), choosing an appropriate within-class distribution is important. However, testing this distributional assumption is challenging as the within-class response time distribution is by definition different from the marginal distribution (i.e., aggregated over classes). Therefore, existing tests on the observed (marginal) response time distribution cannot be used. In this paper, we propose a statistical test on the within-class response time distribution in the dynamical mixture modeling framework for responses and response times by Schnipke and Scrams (1997), Wang and Xu (2015), and Molenaar et al. (2016). We investigate the viability of the newly proposed test in a simulation study, and we apply the test to a real dataset.

## Structural Equation Modeling: 8:30 AM – 10:00 AM

Chair: Irini Moustaki

**Structural Equation Modeling**

**Structural Equation Modeling - Parallel Session: 9.6A: GENERATION OF MULTIVARIATE NON-NORMAL RANDOM NUMBERS WITH SPECIFIED MULTIVARIATE MEASURES**

Wen Qu, University of Notre Dame; Haiyan Liu, University of Notre Dame; Zhiyong Zhang, University of Notre Dame

In real-world scenarios, data are not always normally distributed, which can invalidate the statistical hypothesis testing and yield unreliable results when using methods developed for normal data. The existing methods of generating multivariate non-normal data create data mostly according to specific univariate marginal measures such as the univariate skewness and kurtosis (i.e. Vale & Maurelli, 1983), but not multivariate measures such as Mardia's skewness and kurtosis. Multivariate measures may have an impact on statistical model and analysis, which motivates this research. The purpose of this research is to generate desired multivariate random numbers. In this study, we propose a new method of generating multivariate non-normal data with specific multivariate skewness and kurtosis based on non-elliptical data framework. This approach allows researchers to better control their simulation designs by evaluating the effect of multivariate non-normality. A simulation study was conducted by varying conditions including sample size, covariance, number of variables, and different combinations of multivariate skewness and kurtosis. The results of the simulation confirm that the generating method can produce stable results under different settings and covariance does not play a crucial role in multivariate skewness and kurtosis for our method, and therefore, variance-covariance matrix only affects the marginal measures rather than multivariate measures.

**Structural Equation Modeling - Parallel Session: 9.6B: DIAGNOSTIC ACCURACY OF BARTLETT SCORE ESTIMATES FROM A MISSPECIFIED SEM**

Esther Beierl, University of Oxford; André Beauducel, University of Bonn; Markus Buehner, Ludwig Maximilian University Munich; Moritz Heene, Ludwig Maximilian University Munich

A one-factorial compared to a correct two-factorial model represents a typical misspecification in the structural part of an SEM. in contrast to sensitivity studies of measures of model fit, effects of this type of misspecification on individual diagnoses of factor score estimates have not yet been investigated. we conducted a population simulation study based on predefined factor scores and predefined psychological diagnoses and investigated the extent to which the validity of individual diagnoses would be impaired by making dichotomous diagnoses of Bartlett factor score estimates from a misspecified one-factorial rather than a correct two-factorial model. furthermore, we investigated diagnostic accuracy from overall sum scores. We used different sizes of factor correlations and a balanced versus imbalanced number of indicators per factor to create the population models. we also varied the size of factor loadings and used different base rates for giving dichotomous diagnoses of Bartlett factor score estimates and from overall sum scores. concerning accuracy of diagnoses of factor score estimates from the two-factorial model, the size of factor loadings had the highest impact, in terms of accuracy of diagnoses of factor score estimates

from the one-factorial model and the overall sum scores, the base rates had. the size of factor correlation and the number of indicators per factor also had an impact on the diagnostic accuracy of the factor score estimates from the one-factorial model and the sum scores, especially in interaction with different sizes of factor loadings. implications for psychometrics are discussed.

**Structural Equation Modeling - Parallel Session: 9.6C: POST-SELECTION INFERENCE IN STRUCTURAL EQUATION MODELING**
Po-Hsien Huang, National Cheng Kung University

Most statistical inference methods are established under the assumption that the fitted model is known in advance. In practice, however, researcher often modify their initially specified model by some data-driven selection process. The selection process makes the finally fitted model to be random and also influences the sampling distribution of the estimator. Therefore, implementing naive inference methods may result in wrong conclusions, which is possibly a cause of the current crisis in reproducibility in psychological science. Recently, some advances in valid post-selection inference are made. One promising approach is the polyhedra method proposed by Taylor and his colleagues (Lee, Sun, Sun, & Taylor, 2016; Taylor & Tibshirani, 2017). The method considers the sampling distribution of post-selection estimator, conditioned on the selection done by L1 penalization (lasso). According to polyhedra lemma, the conditional distribution is truncated normal. In this study, we extend the polyhedra method to the case of L1-penalized structural equation modeling (SEM). The proposed method is also compared with two other simple but conservative alternatives: data splitting and Scheffe's correction.

**Structural Equation Modeling - Parallel Session: 9.6D: EFFECT OF MEASUREMENT RELIABILITY ON PRETEST-POSTTEST DESIGN ANALYSIS**
Yasuo Miyazaki, Virginia Tech

The key interest in this investigation is on how much bias occurs when the random assignment is imperfect and therefore when there was a preexisting difference between treatment and control groups in terms of the pretests, and which direction the bias goes. Attenuation bias in correlation coefficient or regression coefficient is well known in behavioral and social research. Compared to the regression coefficient on the variable that a measurement error, the effect of measurement error in covariate on the regression coefficient of treatment grouping variable is not well known. From the general conclusion in two-predictors multiple regression, it can be inferred that when the measurement error exists in the pretest and when there is a preexisting difference between two groups the bias occurs. Because this bias can go either direction, i.e., positive or negative, it is actually a more serious problem than the attenuation. Because of this seriousness of the problem, the issue was quite emphasized in the literature in 1980s (e.g., Cohen & Cohen, 1983) and was referred to as the weakest point in linear models (Darlington, 1990). In the present study, not only the direction and the size of bias, but also other statistical properties such as variance, mean squared error, power, and coverage are studied through a Monte Caro simulation by comparing the performance of the model that takes into account the measurement error through structural equation model and one that ignore the measurement error. Implications of the results are discussed in the context of program evaluation.

## Applications: 8:30 AM – 10:00 AM

Chair: Denny Borsboom

**Applications**

**Applications - Parallel Session: 9.7A: EXPLORING RELATIONSHIP BETWEEN PROBLEM SOLVING SEQUENCE PATTERNS AND BACKGROUND VARIABLES**
Qiwei He, Educational Testing Service; Hok Kan Ling, Columbia University; Jingchen Liu, Columbia University; Zhiliang Ying, Columbia University

Technical advances in computer-based testing provide possibility in recording a variety of timing and process data such as sequences of actions in log files accompanying test performance data when test takers respond to items. With the help of process data, sequence patterns could be identified in test

takers' problem solving behaviors by different performance groups. Background variables are of importance to further exam consistency of test takers' behavior patterns across items and countries in large scale assessment. The primary objectives of this study are to (1) investigate the effects in adding background variables as additional features in identifying malleable factors associated with test takers problem-solving proficiency, and (2) explore the relationship between sequence patterns and background questionnaires. This empirical study used log file data of 8,599 test takers from six countries as well as their background and cognitive dataset collected in Programme for International Assessment of Adult Competencies (PIAAC). We employed different regression analyses (i.e., linear regression, logistic regression and Poisson regression) to investigate the relationship between different variables of interest according to the type of the response variables. The results show that background variables play an important role in identifying subgroups that need special supports when mapping with process data. Consistent problem solving patterns were found across items and countries in subgroups, which help generalize our findings in a broader scope.

**Applications - Parallel Session: 9.7B: A BAYESIAN HIDDEN TOPIC MARKOV MODEL FOR LATENT LINGUISTIC STRUCTURE**

Kenneth Wilcox, University of Notre Dame

Topic models are a probabilistic approach to extracting structural information from discrete data and have been used with great success for text mining. The topic model is a mixture model that represents a collection of words (i.e., a document) as a mixture of topic-specific distributions over the words. A recent development in topic modeling, the Hidden Topic Markov Model (HTMM), incorporates hidden Markov models to model contiguous latent topics and better handle word sense disambiguation. This is a departure from the standard "bag-of-words" assumption of the seminal Latent Dirichlet Allocation (LDA) model and has been shown to have better predictive accuracy for new documents and qualitatively more interpretable topics. We detail the state space model used by the HTMM and propose a Gibbs sampling algorithm for Bayesian inference. The performance of the Gibbs sampler and an EM algorithm are compared in a simulation study. The HTMM is demonstrated on an empirical data set and compared to LDA. Unlike LDA, the HTMM is capable of modeling dependency among the latent topics in a corpus of documents. The topic distributions can be used for a variety of goals including identification of similar documents, network analysis, and identification of key topics in a corpus. The HTMM and related topic models offer a fully probabilistic statistical framework for modeling textual data that often results from open-ended questionnaires, diaries, interviews, and other information-rich sources of data that cannot be easily analyzed with traditional statistical methods.

**Applications - Parallel Session: 9.7C: USING RESPONSE TIME DATA TO MODEL COGNITION AND COGNITIVE DECLINE**

Lynne Schofield, Swarthmore College; Seth Sanders, Duke University; William Dale, City of Hope National Medical Center; Henrik Olsson, Sante Fe Institute; L. Philip Schumm, University of Chicago; Linda Waite, University of Chicago

Social science researchers are increasingly interested in cognitive aging and its relationship to other life events. Most large longitudinal datasets have no or limited direct measures of cognition. Using the National Social, Health and Aging Project (NSHAP), which contains two measures per respondent of the Montreal Cognitive Assessment (MoCA) taken five years apart, we show that the time it takes to answer standard questions measuring cognition is highly correlated with both current measured levels and later declines in cognition. These response time measures are also highly correlated with 5-year mortality. We also find that the item-to-item variability in response time appears to be related to cognitive scores. Data on the time to answer questions is routinely captured with time-stamping in computer assisted interviewing, yet it is rarely used by the social science research community. Our paper shows the usefulness of collecting item-by-item, time-stamped response time data and using these data to better understand respondent's cognitive capacity and behavior. Our results suggest a large amount of useful information about cognition is likely contained within most social science surveys with time-stamping that has to date gone unused and may be useful for modeling the aging process. In addition, the results have implications for clinicians who may be able to use the response time data to supplement the cognitive score and to better understand current cognitive performance and later cognitive decline.

**Applications - Parallel Session: 9.7D: AN APPLICATION OF A TOPIC MODEL TO CONSTRUCTED RESPONSES IN FORMATIVE ASSESSMENTS**
Hye-Jeong Choi, University of Georgia; Minho Kwak, University of Georgia; Jiawei Xiong, University of Georgia; Seohyun Kim, University of Georgia

Topic models have been used in a variety of contexts to extract latent topics from text data. Paul and Dredze (2011) used the topic model latent Dirichlet allocation (LDA; Blei et al., 2003) to analyze Twitter messages for health-related topics. Phan et al. (2008) examined medical texts and Lauderdale and Clark (2014) combined LDA with a multidimensional IRT model to study political preferences. Little research has been reported, however, on analysis of the text of students' responses to constructed response (CR) items. In this study, we will apply a topic model to a formative assessment for English Language Arts for grades 6-8. We will use the information from the topic model with the rubric-based score for constructing a better report to educators and parents. The following research questions will be investigated: 1. Do students' responses to items within different genre reveal the same topics? If not, what are the differences across genres? (genre specific) 2. Do responses of students in different grades to items within the same genre reveal the same topic? If not, what are the differences across grade? (grade specific) 3. How does each topic relate to the rubric-based scores?

## Refreshment Break: 10:00 AM – 10:15 AM

## State of the Art: 10:15 AM – 11:00 AM

Chair: Sophia Rabe-Hesketh

**Optimal Designs and Statistical Power Analysis of Studies with Multilevel Data**
State of the Art Speaker: Miriam Moerbeek, University of Utrecht

Chair: David Thissen

**An Overview of Recent Developments and Applications of Bayesian Model Averaging**
Invited Speaker: David Kaplan, University of Wisconsin, Madison

## Symposium 18: 11:00 AM – 12:30 PM

Chair: Maria Bolsinova

**Symposium 18: Modeling Response Times and Response Processes**

**Symposium 18 - Parallel Session: 10.1A: MODELS FOR HINT USE IN ADAPTIVE LEARNING SYSTEMS**
Maria Bolsinova, ACTNext by ACT, Inc.; Benjamin Deonovic, ACTNext by ACT, Inc.; Burr Settles, Duolingo; Masato Hagiwara, Duolingo; Meirav Attali, ACTNext by ACT, Inc. & Fordham University; Gunter Maris, ACTNext by ACT, Inc.

Adaptive learning systems aim at supporting students in acquiring knowledge and skills in a particular domain. The students' progress is monitored through them continuously solving items which match their level and aim at specific learning goals. Scaffolding and providing learners with hints are powerful tools in helping the learning process. One way of introducing hints is providing a hint to the student after an incorrect response and giving her a second attempt for solving the item. Another way of introducing hints is to make hint usage then student's choice. When the student is certain of her response, she answers without hints, but if she is not certain or simply does not know how to approach the item she requests a hint. In this presentation we develop measurement models for these different ways of providing hints. In the case when using a hint is a learner's choice, it can be treated as a random variable along with response time and accuracy and included in the scoring rule which serves as the basic for the measurement model. This scoring rule is an extension of the scoring rule in the SRT model (Maris & van der Maas, 2012): Learners get a penalty for using the hint(s) when the response is correct, and for not using the hint(s) when the response is incorrect. In this presentation the properties of the different models for hint usage

and their implications are discussed. An application to data from Duolingo, a language learning company, is presented.

**Symposium 18 - Parallel Session: 10.1B: ITEM RESPONSE PROCESSES IN THE PAIRWISE EXCHANGE MODEL**

Joost Kruis, University of Amsterdam; Dylan Molenaar, University of Amsterdam; Maarten Marsman, University of Amsterdam; Gunter Maris, ACTNext by ACT, Inc.

The derivation and application of traditional psychometric measurement models like the Rasch model (Rasch, 1960) is primarily based on desirable statistical properties or measurement-theoretic assumptions (van der Maas et al., 2011) and have a limited connection to the psychological process that generated the item responses. As such, establishing validity of psychological measurement instruments is hampered as it is unclear what process transmitted the variation into the item scores (Borsboom et al., 2003). Developing process measurement models is thus a worthwhile effort as they enable psychometric modelling in terms of explicit psychological processes instead of arbitrarily defined statistical relations. In this talk, we adopt a connectionist network psychometric approach and present an urn model where responses to psychometric test items are the product of pairwise comparisons of the response alternatives. We introduce the resulting response process model for two-, multiple-, and open-choice items, and discuss the key concepts underlying the statistical and theoretical framework of this process model. Specifically, first, we derive the mathematical properties of the models and show that the marginal distributions of the responses are instances of well known IRT models. Next, we demonstrate that the expected response times follow a log-normal distribution which is commonly used in practice. In addition, we discuss possible approaches for the transition from a network structure representing knowledge to a particular urn configuration in the process model framework. Finally, we discuss how learning can be incorporated in the model by a response feedback loop which updates the network structure.

**Symposium 18 - Parallel Session: 10.1C: RELATIONSHIP BETWEEN PLANNING AND EXECUTION: WHAT HAPPENS ON RESPONSE LEVEL?**

Zhaojun Li, The Ohio State University; Paul de Boeck, The Ohio State University

Traditional latent variable model research focuses on the relationship among latent variables and the relationship between latent variables and manifest variables. Whereas there may be important additional information in the item residuals. In this study, planning time and execution time during a game-based assessment with ten items were recorded. The relationship between planning and execution was explored on both latent variable level and manifest variable level. With factor analysis, the latent variables for planning and execution were found to be positively correlated. In contrast, there were negative item-wise residual correlations between planning and execution times. The positive correlation implies that people who plan more slowly also spend more time on executing, while the negative dependency indicates that planning may have a negative effect on execution. It seems to pay off for faster execution to spend more time on planning. Among other methods we used GLMM (generalized linear mixed models) for further investigation of the relationship. Consistent with the factor analysis results, a positive correlation between planning and execution and a negative effect of planning on execution were found. Moreover, the negative effect varied among items, which means that how much execution benefits from planning depends on the items. This study suggests that analyses on different variable levels may show different relationships. Item residuals should be paid more attention in future latent variable model research, especially when parallel data are collected regarding the same items (e.g., planning and execution time of the same item).

**Symposium 18 - Parallel Session: 10.1D: AN URN-SCHEME TO TRACK ACCURACY AND RESPONSE TIMES IN LEARNING ENVIRONMENTS**

Abe Hofman; Maria Bolsinova, ACTNext by ACT, Inc.; Han van der Maas, ACTNext by ACT, Inc.; Gunter Maris. ACTNext by ACT, Inc.

An observed response can be conceptualized as an outcome of a match (with a certain duration) between an item and a player. In modelling these observed responses, it is usually assumed that person abilities are not changing -- at least not while making a test. If the skills of players or the difficulty of items

possibly change over time (even after every response), this classical assumption of a constant ability fails. This is, for example, the case in learning environments with instructions or feedback. In this talk we introduce an urning algorithm that is designed to track changing parameters. The skill of each player and the difficulty of each item is represented with an urn with a certain probability (configuration of balls) and a certain size. This results in binomial distributed random variables with known standard errors. Second, we illustrate the urnings algorithm by analysing data from a large online adaptive learning environment (Klinkenberg, Straatemeier & Van der Maas, 2011) including both accuracy and response time data. We focus on the (developmental) relation between estimates based on speed and accuracy. Third, we show that this system allows for an adaptive selection of items to the skills of players, without any rating inflation.

## Item Response Theory: 11:00 AM – 12:30 PM

Chair: Silvia Cagnone

**Item Response Theory**

**Item Response Theory - Parallel Session: 10.2A: USING PROJECTIVE IRT AS A METHOD FOR VERTICAL SCALING**

Tyler Strachan, The University of North Carolina at Greensboro; Edward Ip, Wake Forest School of Medicine; Terry Ackerman, University of Iowa

The Projective Item Response Theory (PIRT, Ip & Chen, 2012) is designed to allow comparison of a latent trait of interest across tests that may contain different mixes of multiple latent abilities in their items. The idea underlying PIRT is to integrate out the secondary dimensions of a test such that the projected model provides both item parameters and ability estimates for the primary dimension. Typically, in PIRT a 2-dimensional IRT is first used to approximate the latent ability structure of the test and the stronger dimension is identified as the primary dimension. Subsequently the projection is conducted for the identified dimension. This study proposes the use of PIRT as a method for vertical scaling – placing tests from a series of different grade levels on a common scale. The scenario of interest is when there exists a common ability of interest measured between tests at each grade level, but there also may exist idiosyncratic abilities measured within tests at each grade level. In other words, the latent ability structure may vary in both the dimensionality as well as how the items load on the different dimensions across grades. The PIRT is used to remove "contamination" from these idiosyncratic abilities, allowing the placement of all grades on a common "purified" scale. A simulation experiment will be conducted to evaluate the effectiveness of the PIRT as a method for vertical scaling.

**Item Response Theory - Parallel Session: 10.2C: WHAT DO YOU GET WHEN MIXING PROFILE ANALYSIS AND LOCAL EQUATING?**

Richard Feinberg, National Board of Medical Examiners; Matthias von Davier, National Board of Medical Examiners

Profile analysis (PA, Verhelst, 2012) is an approach that looks at sub-scales defined by item categorizations that co-exist with the notion of a unidimensional overall proficiency scale. PA can be understood as an attempt to resolve the discrepancy between unidimensional modeling and customer requests to provide meaningful reports on strengths and weaknesses of test takers. Local equating (LE, e.g. van der Linden, 2012) is an approach that aims at producing comparable scores using an ability estimate (or a proxy) and adjust the equated score by the conditional distribution of observed scores given the ability estimate. In our talk, we combine the ideas of PA and LE and use these together with a well known approach that produces exact distributions of expected scores given ability in order to construct a method for evaluating and reporting sub scores.

**Item Response Theory - Parallel Session: 10.2D: USING ITEM DIFFICULTY LEVEL TO LINK ASSESSMENTS OVER TIME**
Anton Béguin, Cito

In educational measurement results of different test versions cannot be directly compared since results can be influenced by both the difficulty level of the items and the proficiency level of the candidates. With high-stakes educational test often a linking design is used creating common items between some of the test forms in the design and allowing for the application of linking or equating procedures. Due to security issues some examination systems have to rely on assumptions of random equivalent groups and judgments about the difficulty level by subject experts. In the current paper the validity of the assumption of random equivalent groups is evaluated based on an assumption of random difficulty of the items between examinations in different years. Obviously examinations in a subject fluctuate in difficulty between the years, but often the fluctuation can be assumed to be random. If we aggregate difficulty level over the exams of different subjects, the fluctuation will cancel out and differences in average item difficulty over examinations can be interpreted as a difference in the overall proficiency of the population. This overall proficiency change is estimated based on both facility values and average IRT difficulty parameters. This is improved by weighting the differences in item difficulty based on stability of item difficulty over the years prior to year n. In subjects with historically relatively low variance over time the impact of an observed difference in facility as an indicator of change in population proficiency is higher.

## Classification, Clustering, and Latent Class Analysis: 11:00 AM – 12:30 PM

Chair: Jimmy de la Torre

**Classification, Clustering, and Latent Class Analysis**

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 10.3A: INCORPORATING ITEM FEATURES INTO COGNITIVE DIAGNOSIS MODELING WITH RESPONSE TIMES**
Manqian Liao, University of Maryland, College Park; Hong Jiao, University of Maryland, College Park

A number of cognitive diagnosis models (CDMs) have been proposed to model both response and response times (e.g., Minchen, de la Torre, & Liu, 2017; Zhan, Jiao, & Liao, 2017) because response times, as process data, provide rich information about respondents' mastery profiles. However, factors that affect item psychometric characteristics (e.g., guessing, slipping, time-intensity) in cognitive diagnosis assessment (CDA) remain unknown. Cognitive models and theoretical literature have identified certain item features that can affect students' performance through influencing working memory load (e.g., Gierl & Haladyna, 2012; Leighton & Gokiert, 2008). Therefore, this study proposes using either observed (e.g., length, item type) or latent (e.g., content, genre) item features to explain the item parameters in the joint models of response and response times for cognitive diagnosis. Specifically, the proposed model is a combination of the linear logistic test model (LLTM; Fischer, 1973) and the CDM with response times (Zhan et al., 2017). The item features are extracted using computer through text mining techniques. The PISA 2015 mathematics data are used for empirical data analysis. A simulation study is followed where model parameters are estimated using the Bayesian Markov chain Monte Carlo (MCMC) method and parameter recovery and classification accuracy are evaluated under different simulation conditions. The findings of this study would reveal some features that are closely related to item psychometric characteristics and potentially provide recommendations to item writing and item revision in CDA.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 10.3B: CUT-SCORES THAT MINIMIZE ULTIMATE CLASSIFICATION ERROR IN TEST BATTERIES**
Irina Grabovsky, NBME

Certification in a profession is often granted to candidates upon successfully completing several assessments testing different skills. We consider an examination consisting of more than one test. The accuracy of a selection method for the battery can be characterized by combination of two possible errors: a false positive - choosing someone who is undeserving, and a false negative - rejecting someone who is truly deserving. (Grabovsky & Wainer, 2017) introduced a graphical method to select cut-scores minimizing classification error for a single test. The current research develops further the methodology of

selection of optimal cut scores for assessments containing multiple measures. For example, for a test battery containing two test components with given cut-scores, the total probability of classification error is the sum of probabilities of false positive and false negative classification of passing both components of the battery . We calculate the classification error for the battery consisting of the two tests combined, and determine the vector of cut-scores that deliver that minimum for the battery. Data used in this project are simulated based on two highly reliable licensure examinations testing related traits. The method of selecting cut scores by minimizing ultimate classification error for a battery will provide valuable guidance to examination program administrators in the real life standard setting processes.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 10.3C: ADDITIVE TREES FOR FITTING THREE-WAY (MULTIPLE SOURCE) PROXIMITY DATA**
Hans Friedrich Koehn, University of Illinois, Urbana-Champaign

Additive trees are graph-theoretic network models for proximity data collected on a set of objects. Each object is represented as a node in a connected graph, so that the length of the paths connecting the nodes reflects the proximities observed among objects. For additive trees, Carroll, Clark, and DeSarbo (1984) developed the INDTREES algorithm to accommodate three-way two-mode data in explicitly modeling individual differences that might underlie the input proximity judgments. The path lengths of the individual trees are estimated using a conjugate gradient routine for minimizing a least-squares loss function that is augmented by a penalty term to account for violations of the constraints imposed by the four-point condition (that determines an additive tree structure). This study presents an alternative method for fitting additive trees to three-way two-mode proximity data that does not rely on gradient-based optimization nor on penalty terms. Instead, the path lengths of the trees are estimated by an iterative projection algorithm minimizing a constrained least-squares loss function. The constraints are defined by the four-point condition. Simulations are reported for evaluating the performance of the proposed method in comparison with that of the INDTREES algorithm. A real-world data set is analyzed for illustration.

**Classification, Clustering, and Latent Class Analysis - Parallel Session: 10.3D: A LONGITUDINAL AND POLYTOMOUS DIAGNOSTIC CLASSIFICATION MODEL**
Matthew Madison, University of California, Los Angeles; Yu Bao, University of Georgia

The transition diagnostic classification model (TDCM; Madison & Bradshaw, in press) is a general, longitudinal DCM designed to simultaneously classify examinees into attribute mastery profiles and model examinee transitions between different attribute mastery statuses over time. Previous studies employing longitudinal DCMs have only modeled dichotomous attributes (i.e., mastery/non-mastery). We view this as a significant limitation, as latent traits are often operationalized and interpreted polytomously (i.e., basic, proficient, advanced). In this study, we capitalize on the generality of the TDCM by specifying the polytomous DCM (PDCM; Bao & Bradshaw, 2017) as the measurement model within the TDCM. This specification allows for the modeling of examinee changes in polytomous attributes over time and a more fine grained measure of examinee change. Via simulation, we examine model classification accuracy and reliability under different conditions to assess data requirements. Additionally, we utilize this novel model extension to analyze pre-test/post-test data from a diagnostic mathematics assessment.

## Model Fit, Comparison and Diagnostics: 11:00 AM – 12:30 PM

Chair: Wicher Bergsma

**Model Fit, Comparison and Diagnostics**

**Model Fit, Comparison and Diagnostics - Parallel Session: 10.4A: INTERPRETATION OF GENERAL AND SPECIFIC FACTORS IN THE BIFACTOR MODEL**
Wes Bonifay, University of Missouri

Recent measurement research has demonstrated that the bifactor model has a heightened propensity to fit well to any possible item response pattern. Because this model tends to fit well, researchers often

conclude that the general and specific factors are necessarily useful and interpretable. However, this is not the case; the general and/or the specific factors could represent unwanted and uninterpretable noise in the data. Through an extensive simulation study and an empirical example, the present work provides guidelines for proper interpretation of the bifactor model. In the simulation study, many "true" data sets were generated from a two-factor structure with varying inter-item correlations. One-factor, two-factor (uncorrelated), and bifactor models were fit to 1,000 replications of each data condition and compared via several common fit indices. In all instances, the bifactor model unsurprisingly provided superior fit to the data. However, the ratio of within- to between-factor inter-item correlations was found to provide meaningful bounds on bifactor interpretability. If this ratio is below a certain threshold, then a one-factor model is sufficient and bifactor specific factors should not be interpreted; if this ratio is above a certain threshold, then an uncorrelated two-factor model is sufficient and the bifactor general factor should not be interpreted. It is only between these bounds that both the general and specific factors should be interpreted. These results were confirmed using a simple empirical example in which the bifactor model fits exceedingly well, but includes a general factor that is undeniably difficulty to interpret.

**Model Fit, Comparison and Diagnostics - Parallel Session: 10.4B: LIMITED INFORMATION GOODNESS-OF-FIT TESTING OF DCMS FOR POLYTOMOUS ITEMS**

Seungwon Chung, University of California, Los Angeles/CRESST

Diagnostic classification models (DCMs) are psychometric models developed with the aim of classifying respondents into latent classes based on item responses. While research on global absolute fit in DCMs is still in its infancy, limited-information fit statistics such as M2 (Maydeu-Olivares & Joe, 2005) have been proposed as one possible approach. Prior research on application of the M2 statistic has been limited to dichotomous items. Recently, a number of DCMs to accommodate polytomously scored items have been proposed, as constructed-response items and Likert-type scales are widely used in educational and psychological measurement. Hence, this study investigates the application of overall limited goodness-of-fit statistics in DCMs, extending it to polytomously scored data. Specifically, along with the M2 statistic, I examine Cai and Hansen's (2013) M2* statistic and Cai and Monroe's (2014) C2 statistic, which are additional M2-inspired limited-information test statistics proposed to address the sparseness issue with polytomous items. Furthermore, I show that the M2-type of statistics provide sample size independent fit indices such as the RMSEA and TLI. In particular, I illustrate a potential usage of TLI developed for discrete data (Cai, Chung, & Lee, to be submitted). Simulation studies will be conducted to assess whether the three M2-type statistics remain applicable in polytomous DCMs and whether they can detect model misspecification. Various conditions of model misspecification will be used to reflect mixed results regarding sensitivity to misspecification in previous research. This study will establish goodness-of-fit evaluation for polytomous DCMs and provide guidance in assessing model fit.

**Model Fit, Comparison and Diagnostics - Parallel Session: 10.4C: DETECTING PERSON MISFIT AND ABILITY ESTIMATES USING PERSON FIT STATISTIC**

Aminat Egberongbe, Joint Admissions and Matriculation Board (JAMB); Bashir Galadanchi, Bayero University; Kunmi Popoola, Joint Admissions and Matriculation Board, Bwari, Abuja, Nigeria; Patrick Onyeneho, Joint Admissions and Matriculation Board

In person-fit analysis, aberrant item score patterns are detected by means of statistics that signal whether an individual's item score are consistent with expected results or not. When the discrepancy between the observed and the expected item score pattern is large, this indicates person misfit. The rationale behind this study emanated from the assertion that responses of examinees with aberrant pattern may not provide useful and valid measure of their ability. The purpose of this study therefore is to explore the use of lz index in accurately identifying unusual responses in the 2018 Mock test in the Use of English (UOE). Log-Likelihood Index is used in the identification of misfitting persons in this study which employed the ex-post facto research design. The data comprises of a random sample of 1,500 examinees. The 3-parameter IRT logistic model was used in the calibration process. Ability estimation was determined using the Maximum Likelihood Estimation (MLE) method. The lz index is standardized, so that a value of 0.0 reflects a perfectly typical response string, while values greater than 2.0 indicate unexpectedly good fit (overfit). Values below -2.0 indicate poor fit (noise). Examinees' responses identified as exhibiting unexpected responses pattern were removed and the data recalibrated. When the two estimates were compared using a paired sample t-test, there was an increase in ability estimate and

reduction in infit and outfit values. When persons whose answers to questionable items are not removed from the set of items used, the validity of the measurement may be uncertain.

**Model Fit, Comparison and Diagnostics - Parallel Session: 10.4D: TESTING THE CONDITIONAL INDEPENDENCE AND DIMENSIONALITY OF NONPARAMETRIC COGNITIVE DIAGNOSIS MODELS**
Youn Seon Lim, Zucker School of Medicine at Hofstra University

Nonparametric cognitive diagnostic models are useful in cognitive diagnosis modeling for calibration efficiency, especially when sample size is small or large, or the latent attributes are more complex. This article proposes the Mantel-Hanszel chi-squared statistic as an index for detecting the misspecification of latent attributes as well as testlet effects in nonparametric cognitive diagnosis models. The proposed theoretical considerations are augmented by simulation studies conducted to assess the performance of the Mantel-Hanszel statistic under various conditions within the nonparametric diagnosis framework, with a special focus on situations were the set of latent abilities assumed to underlie the data was underspecified.

## Longitudinal Data Analysis: 11:00 AM – 12:30 PM

Chair: Josine Verhagen

**Longitudinal Data Analysis**

**Longitudinal Data Analysis - Parallel Session: 10.5A: POWER AND TYPE I ERROR RATES IN SINGLE CASE DESIGNS**
Samantha Bouwmeester, Erasmus University Rotterdam; Joran Jongerling, Erasmus University Rotterdam; Iris Yocarini, Erasmus University Rotterdam

In clinical and educational settings sample size is often too small to evaluate an intervention using the common statistical tools. Single case designs (SCD) may offer an appropriate alternative when the number of measurements within the baseline and treatment conditions is sufficient but the number of participants is small. In SCD the effect of an intervention is most often evaluated in a qualitative or exploratory way. However, a permutation test allows researchers to statistically evaluate the effect of an intervention. For this permutation test the number of measurements within a condition is permutated for one or a few participants. In our study we evaluated the power and type I error rate of the permutation test for SCD by varying 1) the number of measurements, 2) the means, 3) the standard deviations, and 4) the number of participants. The preliminary results show that the power of the n=1 design is low as long as the effect size is not huge. The type 1 error rate hardly ever reached the assumed alpha level. For n>1 designs, the preliminary results are more promising. We discuss the do's and don'ts of the SCD's given the power and type I results and show an app in which the SCD analyses as well as the power and type I error analyses can easily be performed.

**Longitudinal Data Analysis - Parallel Session: 10.5B: PLANNED MISSING DESIGNS: EFFECTS ON LATENT GROWTH CURVE MODELS**
Maria de Fátima Salgueiro, Instituto Universitário de Lisboa (ISCTE-IUL);, Paula C.R. Vicente, Business Research Unit (BRU-IUL), ISCTE-IUL, Lisboa; Catarina Marques, Instituto Universitário de Lisboa (ISCTE-IUL)

Missing data is one of the most frequent problems to be addressed in longitudinal data analysis. Missing data in a longitudinal study is often due to attrition, unit non response or item non response. However, omissions can also be a consequence of the design of the study: in a planned missing design part of the missingness is due to an option made by the researcher to avoid the burden on the respondent and, hence, increase the quality of the data that are available (C.K. Enders, 2010). The statistical analysis of longitudinal data can be done using latent growth curve models (LGCM), which allow to capture information about interindividual differences in intraindividual change over time. The patterns of change are summarized in relatively few parameters: the means and variances of the random effects, as well as the covariance between intercept and slope (Bollen & Curran, 2006). This talk presents the main results and conclusions from a Monte Carlo simulation study conducted to investigate the effect of non-response

due to a planned missing design on parameter estimates, standard errors and fit measures. LGCMs with unconditional linear growth (and three or four time points) are considered. Sample sizes of 100, 250 and 500 observations are used. The impacts of different patterns and percentages of missingness are discussed.

**Longitudinal Data Analysis - Parallel Session: 10.5C: BIVARIATE MIXED-EFFECTS MODELS FOR SEGMENTED TRAJECTORIES**
Yadira Peralta, University of Minnesota; Nidhi Kohli, University of Minnesota; Eric Lock, University of Minnesota; Mark Davison, University of Minnesota

Developmental processes rarely occur in isolation. Interrelations between two or more processes as they unfold over time are often of interest. Bivariate mixed-effects models (BMEMs) constitute a versatile statistical framework to jointly model trajectories over time of two developmental processes while focusing on revealing the interrelations among them. Furthermore, frequently, longitudinal trajectories do not portray a steady pattern of change. Different stages or segments of development are present in the data. Piecewise linear-linear mixed-effects models are a flexible tool that allows two or more linear segments corresponding to distinct developmental phases to intersect at the knot or change point within the same overall trajectory. However, an abrupt transition between the two linear segments might not be realistic. The bent-cable model accommodates this behavior by defining a transition interval where the change from one linear phase to the other occurs. In the present study, we extend and compare a piecewise linear-linear BMEM (PL-BMEM) and a bent cable BMEM (BC-BMEM) using a Bayesian inference approach via Monte Carlo simulation study. Both, the PL-BMEM and the BC-BMEM are adaptable methods for modeling segmented trajectories jointly and investigating their associations over time. This study aims to (1) extend the framework of PL-BMEM and BC-BMEM to allow for the estimation of flexible covariance structures based on a given theoretical relationship between two outcomes, (2) assess the robustness of the two models via simulation study, and (3) develop a software program that makes these methods accessible to a broad audience.

## Structural Equation Modeling: 11:00 AM – 12:30 PM

Chair: Esther Beierl

**Structural Equation Modeling**

**Structural Equation Modeling - Parallel Session: 10.6A: EQUIVALENCE TESTING FOR FACTOR INVARIANCE ASSESSMENT WITH CATEGORICAL INDICATORS**
Holmes Finch, Ball State University; Brian F. French, Washington State University

Factorial invariance assessment is central in the development of educational and psychological instruments. Establishing factor structure invariance is key for building a strong validity argument, and establishing the fairness of score use. Fit indices and guidelines for judging a lack of invariance is an ever-developing line of research. Two equivalence testing approaches to invariance assessment, based on the RMSEA and CFI indices have been introduced (e.g., Yuan & Chan, 2016). Simulation work (Finch & French, in press) demonstrated that this approach is effective for identifying loading and intercept noninvariance under a variety of conditions, when indicator variables are continuous and normally distributed. However, in many applications indicators are categorical (e.g., ordinal items). As Yuan and Chan noted, equivalence testing based on the RMSEA and CFI indices must be adjusted to account for the presence of ordinal data to ensure accuracy of the procedures. The purpose of this simulation study is to investigate the performance of three alternatives for making such adjustments, based on work by Yuan, Chan, and Bentley (2000), Yuan (2008), and Maydeu-Olivares and Joe (2006). Equivalence testing procedures based on RMSEA and CFI using these adjustments are investigated, and their performance compared with one another. Manipulated factors include sample size, magnitude of noninvariance, proportion of noninvariant indicators, model parameter (loading or intercept), and number of indicators. Outcomes of interest include Type I error and power rates. A total of 1000 replications for each combination of conditions are conducted. Results and discussion focus on the accuracy of the adjustments.

**Structural Equation Modeling - Parallel Session: 10.6B: ESTIMATING STANDARDIZED SEM PARAMETERS GIVEN INCORRECT MODEL AND NONNORMAL DATA**
Keke Lai, University of California, Merced

When both model misspecifications and nonnormal data are present, it is unknown how trustworthy various point estimates, standard errors (SEs), and confidence intervals (CIs) are for standardized SEM parameters. We conducted simulations to evaluate maximum likelihood (ML), conventional robust SE estimator (MLM), Huber-White robust SE estimator (MLR), and the bootstrap (BS). We found (a) ML point estimates can sometimes be quite biased at finite sample sizes if misfit and nonnormality are serious; (b) ML and MLM generally give egregiously biased SEs and CIs regardless of the degree of misfit and nonnormality; (c) MLR and BS provide trustworthy SEs and CIs given medium misfit and nonnormality, but BS is better; (d) given severe misfit and nonnormality, MLR tends to breaks down and BS begins to struggle.

**Structural Equation Modeling - Parallel Session: 10.6C: COMPARISON OF A NON-PARAMETRIC FACTOR SCORE ESTIMATOR WITH TRADITIONAL METHODS**
Tim Fabian Schaffland, Method Center, University of Tübingen; Stefano Noventa, Method Center, University of Tübingen; Augustin Kelava, University of Tübingen

Estimation of factor scores in latent variable models has repeatedly attracted the interest of researchers for decades. Already in 1935 Thurstone proposed the regression method, and in 1937 Bartlett suggested his well-known approach. Still today, factor score estimation and their properties, for example the bias of their moments, raise debate and interest (see, e.g., Grice, 2001; Hoshino & Bentler, 2013). In this talk we will compare the Bartlett estimator, the regression method, the least square estimation, and one new approach (Kelava, Kohler, Krzyzak, & Schaffland 2017) which makes no distributional assumptions on the latent variables. Factor scores are estimated by combining the empirical CDF and the independence assumption between the measurement errors and the latent factors. This results in factor score estimates that in theory could consistently replicate the true joint distribution of the latent variables and the measurement error. In a simulation study we vary the (multivariate) distribution of the underlying factors and examine the performance of the different approaches in recovering the first four moments of the joint distribution of the latent variables. Additionally, the influence of the factor loadings on the estimation is investigated. Two different ways of estimating the factor loadings are used as well as the true values of the loadings. This talk concludes with the implications and recommendations for factor score estimation in an applied context.

**Structural Equation Modeling - Parallel Session: 10.6D: ESTIMATION OF NONLINEAR STRUCTURAL EQUATION MODELS WITH ORDINAL DATA**
Johan Vegelius, Uppsala University

Structural equation models have been used extensively in social and behavioral sciences where relationships between latent variables are of interest. Although most established procedures assume linear relationships between the latent variables it is necessary to resort to nonlinear relationships motivated by psychological theories. As a result a vast variety of procedures have evolved in the last decades to estimate and test interaction and quadratic effects. However, most procedures and models assume continuous indicators. This study proposes a frequentist method to estimate model parameters by marginal maximum likelihood, in the presence of ordinal indicator variables. Inference on factor scores is also discussed. A simulation study is conducted in order to investigate the performance of the proposed method.

## Attendee Lunch: On Your Own: 12:30 PM – 2:00 PM

## Friday, July 13, 2018 PM

### Early Career Award: Jingchen Liu: 2:00 PM – 2:45 PM

Chair: Hua Hua Chang

**An Exploration of Latent Structure in Process Data**
Jingchen Liu, Columbia University

### Refreshment Break: 2:45 PM – 3:00 PM

### Awards Ceremony: 3:00 PM – 3:45 PM

Chair: Francis Tuerlinckx

### Presidential Address: 3:45 PM – 4:45 PM

**Psychometric Tools for Practical Problems in Educational Measurement**
Presidential Address: Cees Glas, University of Twente

### Closing Banquet Reception: 7:00 PM – 9:30 PM