University of Zurich^UZH

Psychometric Society

IMPS 2017

Abstracts: Talks

Monday, July 17, 2017

Workshop: 9:30 AM – 5:10 PM

Short Course 1: Causal Inference in Experiments and Observational Studies
Don Rubin

Short Course 2: Dynamic Structural Equation Modeling of Intensive Longitudinal Data Using Mplus Version 8
Bengt Muthén, Muthén & Muthén; Tihomir Asparouhov; Ellen Hamaker, Utrecht University

Short Course 3: IRT in R
Thomas Rusch, Vienna University of Economics and Business

Tuesday, July 18, 2017

Keynote: 9:15 AM – 10:15 AM

Heterogeneous Large-Scale Data: New Opportunities for Causal Inference and Prediction
Keynote Speaker: Peter Bühlmann
Chair: Carolin Strobl

Perhaps unexpectedly, heterogeneity in potentially large-scale or "big" data can be beneficially exploited for causal inference and more robust prediction. The key idea relies on invariance and stability across different heterogeneous regimes or groups. The resulting new procedures make use of regression analysis as a main tool and offer (possibly conservative) confidence guarantees. We will discuss the novel methodology as well as applications in biology and economics.

Symposium 1: 11:00 AM – 12:30 PM

Symposium 1: Psychometrics and Psychological Science
Chair: Paul De Boeck, Ohio State University & KU Leuven

Symposium 1a: Paul De Boeck
Paul De Boeck, Ohio State University & KU Leuven

One of the major contributions of psychometrics to psychological science is the development and evaluation of measurement methods. In his short introductory book Measurement David Hand (2016) states that measurement "equips us with some very powerful tools for teasing apart the secrets of the universe" (p.27) and subsequently he focuses on "measuring for operational purposes – to make a decision, improve a process, and so on" (p. 75). In discussions with colleagues from substantive areas in psychology, it is not uncommon to be questioned about our discipline. Here is a sample of questions: 1. What is the marginal value of further refining methods that are already so complex and basically work well? 2. Do you know of specific and rather recent examples showing how a psychometric model or approach has led us to revise a psychological theory or has led to a substantive breakthrough or a new insight in psychological phenomena? 3. Can psychometrics provide us with instrumentation that allows us to “see” phenomena and relationships that would otherwise remain out of sight and that are nevertheless important to uncover in order to make progress in psychological science? 4. Assuming that psychometric models can be formulated to reflect hypothesized psychological processes and dynamics, where is the largest need for such models? 5. Would not other quantitative methods be better suited to realize the goals of psychological science? These are all important questions for psychometricians who work in Departments of Psychology, and analogous questions may be relevant for psychometricians working in other departments. In response to this set of questions, four psychometricians will make an introductory statement to open the discussion with a broader audience during the symposium. We are also welcoming questions and considerations from Elsbeth Stern, professor of Research on Learning and Instruction at ETH Zürich. From each of the four psychometricians you will find here a few sentences to express possible ideas to discuss.

Symposium 1b: Han van der Maas
Han van der Maas, University of Amsterdam

Psychometrics supports psychological measurement and testing psychological theories. The relation between psychometrics and psychological science, however, should be bidirectional. Psychological science should be used to develop psychometric models.

Symposium 1c: Alberto Maydeu-Olivares
Alberto Maydeu-Olivares, University of South Carolina

Here are a few issues where psychometrics can help. The power to detect differences with a model decreases as the number of categories in rating ítems decreases. One can reach different substantive conclusions about the dimensionality of personality depending on the number of response categories the ítems have (because we do not check for power). The order of items in questionnaires may also matter. Our models assume that randomization of ítems is effective at eliminating order effects. I believe it fails. If so, we should model order effects and stop randomizing ítems, because it just makes our task of modeling order effects harder.

Symposium 1d: Carolin Strobl
Carolin Strobl, University of Zurich

I see our science as fundamental research in the sense that applicability might not be right around the corner for some or all of the ideas we develop - but that doesn't mean it won't be applied practically in the future. So I think we should not limit ourselves too much by looking for direct use. On the other hand, I also strongly promote the idea that methodological developments have associated free, well documented software, because then applied researchers will be able to actually use it when it fits their problem. I also think it is not a bad thing if we (at least in part of our research) ignore the current status quo in applied research, because applied researchers may be limited to the existing methods and think along the lines what is possible with them. If we communicate that we have new hammers, perhaps they will find "new kinds of nails". What we as a discipline need to improve is communication and documentation in a way that is comprehensible to applied researchers.

Symposium 1e: David Thissen
David Thissen, University of North Carolina at Chapel Hill

New psychometric linking methods have been developed to permit comparable inferences even from somewhat disparate measures, for example, new scales and legacy scales for anxiety and depression. These linking methods, along with more classic procedures, are an essential part of integrative data analysis, which combines data from multiple studies to obtain a single unified set of conclusions, something like what has been intended with meta-analysis or research synthesis, but with greater resolution. The new methods of test linking that involve careful examination of the dimensional structure of measures may also be extensible to examine the relation of, say, scales that measure clinical and non-clinical ranges of personality or affect. These facts offer great promise for the future of psychometric test theory, even while they raise critical questions.

Symposium 1f: Elsbeth Stern
Elsbeth Stern

Symposium 2: 11:00 AM – 12:30 PM

Symposium 2: How to Deal with DIF in Educational Surveys?
Chair: Robert Zwitser

Symposium 2a: A Comparison of Three Methods for Modeling DIF

Cees A. W. Glas, University of Twente ; Annemiek Punter, University of Twente; Martina R.M. Meelissen, University of Twente

Three methods for modeling country-specific DIF in large-scale educational assessments are compared. The first one adds country-specific item parameters to an IRT model as fixed effects (Glas & Jehangir, 2014). The second one adds country-specific parameters as random effects (De Jong, Steenkamp, & Fox, 2007). The third one is based on the bi-factor model (Punter, Glas, & Meelissen, 2016), and adds country-specific effects as an additional latent dimension per country. The three approaches are compared using the data of the parental involvement and reading literacy scales of the 2011 cycle of PIRLS (project funded by the IEA). The basic IRT model for the three approaches was the Generalized Partial Credit Model (GPCM, Muraki, 1992). The estimation procedure for all three approaches was Marginal Maximum Likelihood (MML). The results of the analyses showed that there was reasonable agreement regarding the items flagged as biased. Further, using plausible values generated by the three methods did not produce any marked differences when they were entered as predictors in a latent regression analysis. The overall conclusion was that the parental involvement and reading literacy scales of PIRLS are internationally valid.

Symposium 2b: On the Difference Between Modeling DIF and Correcting for DIF

Robert Zwitser; S. Sjoerd F. Glaser, University of Amsterdam; Gunter Maris, University of Amsterdam & Cito

DIF is likely to occur in international surveys. What is needed is a statistical approach that takes DIF into account, while at the same time allowing for meaningful comparisons between countries. Zwitser, Glaser, and Maris (2016) discuss some existing approaches and provide an alternative based on market basket scoring (Mislevy, 1998). The core of this approach is to define the construct as a large set of items, and to report in terms of summary statistics. Since the data are incomplete, measurement models are used to complete the incomplete data. For that purpose, different models can be used across countries. This talk will further discuss this approach. As part of that, the distinction between modeling DIF with, for instance, country-specific parameters, and correcting for DIF in the scoring rule will be emphasized. The method will be illustrated with PISA's reading literacy data.

Symposium 2c: Invariance vs. Fit or Stability vs. Variability in International Assessments

Matthias von Davier, National Board of Medical Examiners

One approach commonly seen in international assessments is ignoring item level misfit across groups. This type of misfit may be present even after careful translation procedures are applied, mainly due to differential population functioning that may be due to different curricular, or different types of instructional cultures across countries, or other reasons that are not directly observable. Other assessment programs allow for some level of adjustment of item parameter and link item sets based on the presence of misfit, and produce a model that is not fully comparable across all items. Another approach that completely does away with any type of invariance assumption was suggested, which appears to provide better fit if not examining the consequences closely, but can be shown to come at a cost. Essentially, estimation of different sets of item parameters for each of the participating populations is not a valid method, both from a perspective of model parsimony, but also from a perspective that looks at country level limitations of the data collection. It can be shown that an approach that estimates separate parameters for each country is inferior to approaches that aggregate information. The talk will summarize findings of a recent study that compared the fit of parameters between separate and combined estimation of models for international large-scale assessments.

Symposium 2d: A Network Perspective on Educational Surveys

Gunter Maris, University of Amsterdam & Cito; Timo Bechger, Cito

In a typical educational survey, a sample of students responds to items relating to one or more cognitive domains, and to a questionnaire. The cognitive domains are scaled using an item response theory model, and interest is in the conditional distribution of ability given student and school characteristics. We propose a network model in which all questions (cognitive and background) are nodes, connected by edges that encode the correlational structure. Interest is in modelling the joint distribution of all questions. The network structure consists of three components. First the edges between cognitive items encode the dimenstionality of the cognitive domains. Second the edges between questions from the background questionnaire encode its structure. Third, edges between cognitive items and questionnaire items provide information about Differential Item Functioning (DIF). We use this framework to put DIF in educational surveys in a novel perspective.

CDM 1: 11:00 AM – 12:30 PM

CDM 1: Attribute Hierarchies and Longitudinal

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

CDM 1a: Latent Attribute Space Identifiability and Q-Matrix Completeness for Cognitively Diagnostic Models with Attribute Hierarchies

Hans-Friedrich Koehn, University of Illinois at Urbana-Champaign

Educational researchers have argued that a realistic view of the role of attributes in cognitively diagnostic modeling should account for the possibility that attributes are not isolated entities, but interdependent in their effect on test performance. ("Attributes" is a collective term in cognitive diagnosis for skills, specific knowledge, aptitudes - any cognitive characteristic required to perform tasks.) Different approaches to modeling possible attribute interdependence have been discussed in the literature; among them the proposition to impose a hierarchical structure so that mastery of certain attributes is a prerequisite of mastering one or more other attributes. A hierarchical organization of attributes constrains the latent attribute space such that several proficiency classes, as they exist if attributes are not hierarchically organized, are no longer defined because the corresponding attribute combinations cannot occur with the given attribute hierarchy. Hence, the identification of the latent attribute space is often difficult - especially, if the number of attributes is large. As an additional complication, constructing a complete Q-matrix may not at all be straightforward if the attributes underlying the test items are supposed to have a hierarchical structure. (A Q-matrix is said to be complete if it allows for the identification of the attribute profiles of all realizable proficiency classes.) In this study, a framework based on lattice theory is proposed for examining the conditions of identifiability of the latent attribute space and of completeness of the Q-matrix if attributes are hierarchically organized.

CDM 1b: Bayesian Network for Modeling the Uncertainty of Attribute Hierarchy

Lihong Song, Jiangxi Normal University; Wenyi Wang, Jiangxi Normal University; Shuliang Ding, Jiangxi Normal University

In the Attribute Hierarchy Method (AHM), cognitive attributes are usually assumed to be organized hierarchically (Leighton, Gierl, & Hunka, 2004). This assumption requires content specialists to conduct a task analysis on a sample of items, in order to specify the latent cognitive attributes and order these attributes to create an attribute hierarchy. However, the problem-solving performances of experts and novices were almost certain to be different, because expert’s knowledge is highly organized in deeply integrated schemas, while novices view domain knowledge and problem-solving knowledge separately. This may bring uncertainty into the specification of attribute hierarchy and lead to generate different attribute hierarchies for a test (Wang & Gierl, 2011). Formally, a Bayesian network is a probabilistic graphical model that represents a set of random latent attributes or variables and their conditional dependencies via a directed acyclic graph. Here, a Bayesian network can be used to represent the probabilistic relationships between latent attributes of hierarchical structure. The purpose of this study is to apply Bayesian network for modeling attribute hierarchy under uncertainty. Created from the attribute hierarchy, the Bayesian network can be regarded as a flexible high-order model (de la Torre & Douglas, 2004) of the Reduced Reparameterized Unified Model (R-RUM; Hartz, 2002). This flexible high-order model incorporated into R-RUM has an advantage of taking account of both the subjectivity of attribute hierarchy from experts and the stochastic nature of item response data. Finally, fraction subtraction data is analyzed to evaluate the performance of the new model.

CDM 1c: Latent Transition Analysis with Log-Linear CDM

Qian Peng, Beijing Normal University

Latent variable models are extended to explore change in latent ability over time, which is recently beginning to receive attention in education and psychological measurement, e.g., mixture IRT with Latent Transition Analysis (LTA; Cho & Cohen et al., 2010), longitudinal DINA or DINO (Feiming & Cohen et al., 2016; Kaya & Leite, 2016). Besides, the Log-Linear Cognitive Diagnostic Model (LCDM; Henson &Templin et al., 2010) to discuss Cognitive Diagnostic Models (CDMs) more generally is gaining popular in diagnosis assessment. This study aims to explore the general longitudinal CDMs to study the change of attributes representing cognitive skills over time. The model combined the LTA with LCDM as measurement model was developed and then evaluated through a Monte Carlo simulation study. The simulation study is implemented with T=3 time points, I = 30 items, A = 3 attributes, the Q-matrix designed balanced, two class sizes at the first time (equal or unequal), three sample sizes (N=200,500, 1,000), and three guessing and slipping parameters (g=s=0~0.1, 0.1~0.2, 0.2~0.3). The study results indicate that the proposed model can measure the development of the cognitive skills over time well.

CDM 1d: Multilevel Hidden Markov Model for Learning Under Cognitive Diagnosis Framework

Susu Zhang, University of Illinois at Urbana-Champaign; Hua-Hua Chang, University of Illinois at Urbana-Champaign

There is a growing number of online courses and open-source educational repositories. Often there are various instructional videos, slides, and exercises available to students for the learning a specific skill, but the scarcity of time and mental energy limits the amount of materials a learner can study. Educators are interested in adaptively selecting the best content sequence for individual learners, specifically which skill s/he should learn next, and which material s/he should use to study, based on previous performance. Such adaptive selection would rely on preknowledge of how the learners' and the instructional materials' characteristics jointly affect the probability of acquiring a certain skill. Building upon previous research on latent transition analysis and learning trajectories, we propose a multilevel logistic hidden Markov model for learning under cognitive diagnosis: At each stage of learning, a learner receives an instructional material targeting a specific skill in the curriculum. The probability that the learner acquires the target skill depends not only at the general difficulty of acquiring the skill and his/her mastery of other skills in the curriculum, but also on the effectiveness of the particular learning material and the learning material's interaction with mastery on other skills, captured by random slopes and intercepts for each learning material. The latent trajectory of skill mastery changes is then observed through item responses in post-learning assessments after each learning stage, where the responses follow a cognitive diagnostic model. A Bayesian modeling framework and MCMC algorithm for parameter estimation are also proposed.

CDM 1e: Bayesian Modeling for Learning Trajectories in Cognitive Diagnosis Models

Yinghan Chen, University of Illinois at Urbana-Champaign; Steven Culpepper, University of Illinois at Urbana-Champaign; Shiyu Wang, University of Georgia; Jeff Douglas, University of Illinois at Urbana-Champaign

The increasing presence of electronic and online learning resources presents challenges and opportunities for psychometric techniques that can assist in the measurement of abilities and even hasten their mastery. Cognitive Diagnosis Models (CDMs) are ideal for tracking many fine-grained skills that comprise a domain, and can assist in carefully navigating through the training and assessment of these skills in e-learning applications. We propose a class of CDMs for modeling changes in attributes, which we refer to as learning trajectories. We focus on the development of Bayesian procedures for estimating parameters of a first-order hidden Markov model. We present an application of the developed model to a spatial rotation experimental intervention.

IRT 1: 11:00 AM – 12:30 PM

IRT: Rater Effects

Chair: Andreas Frey, Friedrich Schiller University Jena

IRTa: New Item Response Theory Models for Rater Errors

Xue-Lan Qiu, The Education University of Hong Kong; Wen-Chung Wang, The Education University of Hong Kong

When humans are enlisted to judge competency in many fields, rating errors are inevitable and can have serious consequences. Major rater errors include severity, inconsistency, centrality, and similarity (Myford & Wolfe, 2004). There are two IRT frameworks for rater effects. In the facets framework, raters are treated as independent judges. In the Hierarchical Rater Model (HRM) framework, the ratings given by raters are treated as indicators of the ideal "true" category for the work being judged (e.g., an essay). Both approaches focused on raters' severity and inconsistency; leaving centrality and similarity almost untouched. This study developed a class of IRT models that account for various rater errors in the HRM framework. Like the traditional HRM, the rating process in the new models contains two stages. In the first stage, the ideal rating that a rater with perfect reliability would assign to the item response follows a standard IRT model (e.g., the rating scale model). In the second stage, raters give ratings to an item response, which may be different to the ideal rating due to rater errors. By incorporating specific parameters for response criteria, the new models can handle raters' severity, inconsistency, and centrality simultaneously. A series of simulations were conducted to assess parameter recovery of the new models. Results found that the parameters can be well recovered with the freeware JAGS. An empirical example was provided to demonstrate the implications and applications of the new models. A discussion on extending the new models to capture raters' similarity was given.

IRTb: Assessing Raters' Reliability Without Multiple Ratings

Filip Kulon, Educational Research Institute

The rating process is extensively used in educational assessment (e.g. in scoring students' essays), although it is not limited to this field. The reliability of the process is usually controlled by using multiple ratings. Among other models, the hierarchical rater model with signal detection theory (DeCarlo, Kim & Johnson, 2011) was developed to utilize the information from multiple ratings for assessing the raters' performance. Kim (2009) presented an extension of this model which incorporates Multiple Choice (MC) items, but it still requires multiple ratings. I propose further modifications: to use MC items and only one rating of Constructed Response (CR) items to estimate both the students' ability and the raters' reliability. A Monte Carlo simulation study was developed to check usefulness and accuracy of the proposed model in situations that can take place in real-life assessment. Combinations of different number of MC (10, 20) and CR (1, 5) items with a range of scoring categories (3, 5, 10) and different number of raters and students were tested. Various rater effects were simulated by random generation of rater parameters. The model was estimated with a Bayesian approach using MCMC. The simulation results show that students' ability bias is similar to that of the Graded Response Model and that the item parameter recovery is good. The rater parameters should be treated with caution as their recovery is not as good as the item parameters recovery. Overall, the proposed model provides useful information on the raters' reliability in the absence of multiple ratings.

IRTc: Modeling Changes in Adolescents' Character with the Hierarchical Rater Model

Jodi M. Casabianca, Educational Testing Service

This research presents results from an application of the Longitudinal Hierarchical Rater Model (L-HRM; Casabianca & Junker, 2013) to data from the Character Development in Adolescence Project (CDAP; https://coa.stanford.edu/content/character-development-adolescence), a study that examines growth in adolescents' character traits such as grit and gratitude over a two-year period. The L-HRM is a longitudinal version of the HRM, a multilevel IRT model for ratings that controls for the dependency brought about by multiple raters assigning scores to the same work. Using the CDAP data, we describe the stages of the modeling process which compare several parameterizations of the L-HRM following McArdle, Petway, and Hishinuma (2015): (i) the L-HRM without a trend or autoregressive structure, (ii) the L-HRM with different trends (linear and logistic), (iii) the L-HRM with an autoregressive structure (no trend), and (iv) the L-HRM with a trend and an autoregressive structure. We explain the procedures for model selection as well as how to perform posterior predictive model checking (Sinharay, Johnson, & Stern, 2006) to evaluate the fit of the selected L-HRM parameterization. Importantly, the presentation will discuss how the L-HRM aligns with the structural equation model framework and also discuss the limitations of the L-HRM.

IRTd: Evaluating Rater Judgments with Cross-Classified Multilevel IRT Models
Jue Wang, University of Georgia; Zhenqiu (Laura) Lu, The University of Georgia; George Engelhard Jr., University of Georgia

The Brunswik's (1952) lens model can be used to explore rater judgments and identify rater effects within the context of rater-mediated assessments. Lens model studies have been limited in the past few decades with the dominant use of multiple regression techniques. Nestler and Back (2015) proposed to use cross-classified structural equation models (Asparouhov & Muthén, 2012) to examine the accuracy of personality judgments. This research can also lead to new methodologies based on IRT models for evaluating human judgements and cognitive processes represented by a lens model. Research studies on human judgments focus on individual raters with various purposes. IRT models accommodate item-level and person-level analyses, and this can be of great use for human judgment studies. Therefore, we propose a Cross-Classified Multilevel IRT model (CCM-IRT; Van den Noortgate et al., 2003) for evaluating rater accuracy and judgments. CCM-IRT treats rater and essay facets as two separate clusters that both predict rating scores. We can obtain individual rater and essay estimates, as well as evaluate rater judgments by including essay features and rater characteristics as covariates in each cluster. The CCM-IRT model is illustrated with a set of student essays from a Grade 7 writing assessment with 7 essay feature indices (i.e., word count, narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, & verb cohesion) obtained using Coh-Metrix text analyzer (McNamara et al., 2005) for each essay. Preliminary results show that narrativity and word count significantly affected essay measures so that these two factors influenced rater judgments in scoring.

IRTe: Cognitive and Psychometric Perspectives on Rater Accuracy in Writing Assessment
George Engelhard Jr., University of Georgia; Jue Wang, University of Georgia; Stefanie A. Wind, University of Alabama

The purpose of this study is to discuss rater accuracy based on recent advances in research on the quality of ratings in rater-mediated assessments. In order to obtain highly accurate ratings, it is important to consider both a theory of rater cognition and a measurement theory that are congruent. This study describes a model for rater judgment that is built on Brunswik's Lens model to represent rater cognitive processes. The psychometric model used to examine rater accuracy is based on a multiple-group Rasch model embedded within a structural equation modeling framework. Data from a large-scale writing assessment in the US are used to explore these issues. Specifically, we examine the ratings of 20 operational raters and 3 experts based on 100 essays written by Grade 7 students in two domains: (1) Idea, Development, Organization, and Cohesion (IDOC) and (2) Language Usage and Convention (LUC). We transformed the original ratings to accuracy ratings in order to define a latent variable representing accuracy (Engelhard, 2013). Preliminary results indicate that rater accuracy measures vary across the two domains, and that the raters are not invariant. Fit statistics reveal raters that have inconsistent judgments as compared to the experts. Explicit consideration of both cognitive and psychometric perspectives has important implications for rater training and maintaining the quality of ratings obtained from human raters.

CI 1: 11:00 AM – 12:30 PM

Causal Inference
Chair: David Kaplan, University of Wisconsin-Madison

CI 1a: Lewis, Rubin, Pearl: Compare and Contrast
Keith A. Markus, John Jay College of Criminal Justice of The City University of New York

Philosopher David Lewis developed a non-causal theory of counterfactual conditionals and drew from that theory to develop a reductive theory of causation based on counterfactual conditionals. Picking up on work by Jerzy Neyman, Donald Rubin and colleagues developed an account of the definition and estimation of causal effects based on an assumed primitive notion of causation consistent with manipulation theories of causation. Rubin adopted the term "counterfactual" without committing to a counterfactual theory of causation. Pearl and colleagues developed structural causal models as a method of expressing and reasoning about causal expressions including causal counterfactual conditionals. Pearl's framework adopts a broadly Humean approach to causation, a description that encompasses both manipulation theories and counterfactual theories, but the precise account of causation remains unsettled. Pearl has suggested that the frameworks are equivalent. Attention to the formal presentations of the suggested equivalences show them to rest on apparent misunderstandings. The differences and similarities between the frameworks shed light on issues of specifying causal hypotheses and making causal inferences. The comparison of Lewis' and Pearl's frameworks helps to clarify the scope and limits of Pearl's framework. The comparison of Rubin's and Pearl's frameworks helps to clarify two plausible but contrasting interpretations of Pearl's work.

CI 1b: A Latent Variable Modeling Approach to the Regression Discontinuity Design
Monica Morell, University of Maryland, College Park; Ji Seung Yang, University of Maryland, College Park

The utility of the conventional Regression Discontinuity (RD) design lies in the estimation of a local average treatment effect when the treatment eligibility and assignment is determined by an individual’s location on an observed continuous variable (Thistlethwaite & Campbell, 1960). In order to increase statistical power, applied researchers regularly include covariates in the RD model (Cattaneo & Escanciano, 2016). Often, these covariates, and the assignment variable itself, are measured by a set of observed categorical item responses (e.g., pre-test scores, social-economic status, academic motivation). Nevertheless, RD analysis has long been conducted with only summed scores for both assignment and covariate variables. However, Morell and Yang (2016) reported biased RD treatment effects when the measurement error in the assignment variable is not properly taken into account and proposed a latent variable (LV) modeling approach to the RD design. The purpose of this study is to extend the LV approach to not only assignment variable but also covariate to improve parameter recovery of the RD treatment effect. The impact of ignoring measurement error in both assignment variable and covariate is explored. Two one-stage estimation methods (Bayesian approach and Metropolis-Hastings Robbins-Monro (MHRM) algorithm to obtain full information maximum likelihood estimation) are developed and the performances are compared against two two-stage estimation methods (summed or Expected A Posteriori (EAP) scores from item response models). A Monte Carlo simulation study is conducted under different sample sizes, measurement conditions, and magnitudes of treatment to evaluate performance of LV approach to the RD design.

CI 1c: CART-Based Methods for Variable Selection in a Conditional Independence Framework
Bryan Keller, Columbia University; Tianyang Zhang, Columbia University

The ability of a conditioning strategy to consistently estimate an average treatment effect in an observational study depends on the assumption that a set of pretreatment variables has been observed that is sufficient to eliminate bias due to confounding. This assumption, referred to as ignorability (Rubin, 1978), is more likely to be satisfied if the conditioning set is comprehensive and large. Nevertheless, there are compelling reasons to not simply condition on all available variables. Inclusion of instrumental variables, collider variables, or other noise variables can increase the variability and/or inconsistency of causal effect estimators. De Luna, Waernbaum, and Richardson (2011) proposed an algorithm for variable selection based on conditional independence relationships. If ignorability holds with the complete set of variables, it will also hold with the reduced set after running the algorithm, with the added benefit that estimation will be more efficient. Available implementations of the algorithm use nonparametric tests of conditional independence to avoid strong parametric assumptions, however, the computational intensity limits the number of variables that can be accommodated. Regression trees (Breiman, Friedman, Olshen, and Stone) are a promising alternative because they handle complex functional forms algorithmically and provide a ranking of variable importance. Significance tests for variable importance were given by Breiman and Cutler (2008) and Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008). In this framework, we explore, through simulation, the performance of regression tree methods relative to nonparametric tests of conditional independence for variable selection as measured by ability to correctly identify the minimal subset and by the computational time required to run.

CI 1d: A General Framework of Multilevel Matching Strategies for Clustered Observational Data

Jee-Seon Kim, University of Wisconsin-Madison; Peter M. Steiner, University of Wisconsin-Madison; Youmi Suk, University of Wisconsin-Madison

This study develops and examines strategies for causal inference with observational multilevel data. Matching strategies for nonequivalent control group designs with clustered data are required because randomized experiments cannot always be conducted, and selection bias should be removed to draw proper causal conclusions from observational data. Matching techniques like propensity score analysis have been increasingly popular in the social sciences, but most methods are limited to single-level data. Matching strategies for multilevel data should be flexible and realistic, as selection or outcome processes may vary across clusters or individuals. Also, even in large-scale data, some or many clusters may have small sample sizes or lack sufficient overlap, resulting in serious problems in estimating treatment effects for each cluster. To handle various issues and challenges in multilevel matching, this study presents a general framework for multilevel matching strategies that include a series of previous matching techniques as special cases and provides extensive tools to examine and specify various forms of selection and outcome models. This framework can be viewed as a multilevel matching continuum for making causal inferences with clustered observational data. This talk shows that this continuum consists of the two most common multilevel matching strategies, within-cluster matching and across-cluster matching, as two opposite ends and contains diverse manifest and latent class modeling approaches in between. Properties of these matching strategies are discussed with practical guidelines and recommendations.

CI 1e: Causal Treatment Effect Analysis Using Two New SAS Procedures

Yiu-Fai Yung, SAS Institute Inc.

This presentation describes the general problem of causal treatment effect estimation in correlational studies and introduces two new SAS/STAT® procedures that offer solutions to the causal inference problem. The CAUSALTRT procedure estimates treatment effects by propensity score weighting, regression adjustment, or a combination of the two (the doubly-robust method). The PSMATCH procedure employs the propensity score model to create matched samples that behave like data from a randomized experiment. Alternatively, the PSMATCH procedure can create data sets that you can use to estimate causal effects by propensity score weighting or stratification. We use data examples to illustrate these procedures.

MF 1: 11:00 AM – 12:30 PM

Model Fit: Bayesian and Simulation-based

Chair: Jean Paul Fox

MF 1a: Testing Small Variance Priors Using Prior-Posterior Predictive p-Values

Herbert Hoijtink, University Utrecht; Rens van de Schoot, University Utrecht

Muthen and Asparouhov (2012) propose to evaluate model fit in structural equation models based on approximate (using small variance priors) instead of exact equality of (combinations of) parameters to zero. This is an important development that adequately addresses Cohen's (1994) "The earth is round (p<.05)", which stresses that point null-hypotheses are so precise that small and irrelevant differences from the null-hypothesis may lead to their rejection. It is tempting to evaluate small variance priors using readily available approaches like the posterior predictive p-value and the DIC. However, as will be shown, both are not suited for the evaluation of models based on small variance priors. In this paper a well behaving alternative, the prior-posterior predictive p-value, will be introduced. It will be shown that it is consistent, the distributions under the null and alternative hypotheses will be elaborated, and it will be applied to testing whether the difference between two means and the size of a correlation are relevantly different from zero.

MF 1b: The Impact of Moderately Informative Priors on Ability to Detect Model Misfit
Sierra Bainter, University of Miami

In psychological research, available sample sizes are often insufficient to test structural equation models of interest using traditional estimation methods, such as ML or modified WLS approaches. Bayesian estimation with common-sense, moderately informative priors can greatly improve efficiency of parameter estimates, power to detect meaningful effects, and stabilize estimation to avoid extreme estimates. Further, in the case of categorical indicators, appropriate methods for testing model fit are limited for ML estimation, and available statistics using WLS-based estimation are inappropriate in some realistic cases. In a Bayesian framework, there are a variety of additional tools available to describe model fit. However, it is not known how moderately informative priors influence ability to detect model misfit. Papers demonstrating Bayesian model fit assessment commonly assume flat priors, which have no utility over ML in limited data settings. Hoijtink and van de Schoot (in press) have shown that two commonly used indices, the DIC and posterior predictive checks, are inconsistent under small-variance priors, and recommend a prior-posterior predictive check to evaluate model fit. These results have not been extended to the case of moderately informative priors. In this presentation, I evaluate the impact of moderately informative priors on ability to detect model misfit for several standard indices.

MF 1c: Resampling-Based Approaches for Person fit Assessment in Cognitive Diagnosis Modeling
Kevin Carl Santos, University of the Philippines; Jimmy de la Torre, The University of Hong Kong; Matthias von Davier, National Board of Medical Examiners

Detection of aberrant response patterns is of primary importance in educational and psychological measurement because the discrepancy between examinee's ability and test performance could lead to inappropriate remediation measures wasting teacher's and student's efforts or invalid selection decisions leading to serious consequences. Person Fit (PF) analysis principally aims to identify such response patterns. However, in cognitive diagnosis modelling, there is a dearth of research on PF literature. This study attempts to bridge this gap by investigating the viability of resampling-based approaches (Sinharay, 2016), whether they can provide more accurate classification of aberrant examinees than the traditional approach of assuming normality. The first procedure for resampling utilizes the parametric bootstrap (Efron, 1979). It fixes the attribute pattern and uses it to simulate bootstrapped response patterns assuming the item parameters are known. The second procedure, on the other hand, employs the posterior distribution of each response pattern and uses it to generate response patterns, also, on the assumption that the item parameters are known. Likelihood- and residual-based PF statistics are considered in this study using the Generalized Deterministic Inputs, Noisy "And" gate (G-DINA; de la Torre, 2011) model. Preliminary simulation results showed that, compared to the traditional approach, the two resampling approaches have well-controlled Type I error rates for the likelihood-based statistics. It was found that there is a very small difference between the two resampling-based approaches and both of them work well with likelihood-based PF statistics. The residual-based statistics, on the other hand, perform very poorly in all three approaches.

MF 1d: Automatic Bayes Factors for Testing Equality and Inequality Constrained Hypotheses on Variances
Florian Böing-Messing, Tilburg University; Joris Mulder, Tilburg University

In comparing characteristics of independent populations, researchers frequently expect a certain structure of the population variances. These expectations can be formulated as hypotheses with equality and/or inequality constraints on the variances. In this talk we consider the Bayes factor for testing such (in)equality constrained hypotheses on variances. Application of Bayes factors requires specification of a prior under every hypothesis to be tested. However, specifying subjective priors for variances based on prior information is a difficult task. We therefore consider so-called automatic or default Bayes factors. These methods avoid the need for the user to specify priors by using information from the sample data. We discuss three automatic Bayes factors for testing variances. The first is a Bayes factor with equal priors on all variances, where the priors are specified automatically using a small share of the information in the sample data. The second is the fractional Bayes factor, where a fraction of the likelihood is used for automatic prior specification. The third is an adjustment of the fractional Bayes factor such that the parsimony of inequality constrained hypotheses is properly taken into account.

Results from a simulation study indicate that the adjusted fractional Bayes factor converges fastest to the true hypothesis.

MF 1d: WAIC as a Model Selection Method for Polytomous Items

Yong Luo, National Center for Assessment

The Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2010) is an emerging Bayesian model selection method that has the theoretical advantage of being fully Bayesian over other traditional information-criterion based model selection methods such as AIC, BIC, and DIC. WAIC is considered fully Bayesian because it uses the whole posterior distribution to compute both its deviance and penalty terms, whereas neither AIC nor BIC utilize the posterior distribution, and DIC only uses the posterior distribution partially to compute the penalty term, with its deviance term still based on point estimates. It has been shown (Luo & Al-Harbi, 2016) that in the context of dichotomous IRT model selection, WAIC performs better than AIC, BIC, DIC, and the likelihood ratio test due to its theoretical advantage of being fully Bayesian. However, it remains unknown whether the superior performance of WAIC is generalizable to scenarios of polytomous IRT model selection. In this study, the performances of AIC, BIC, DIC, and WAIC as model selection methods are compared. A simulation study with number of item categories, sample size, test length, and generating IRT model as manipulated factors is conducted to examine the statistical power of these model selection methods when selecting from a group of polytomous IRT models (i.e., the graded response model, the generalized partial credit model, the partial credit model, and the rating scale model). A real data set is also used to illustrate the use and relative performance of these methods in polytomous IRT model selection.

CFA 1: 11:00 AM – 12:30 PM

CFA and Generalizability Theory

Chair: Fan Wallentin, Uppsala University

CFA 1a: Recovery of Weak Factor Loadings When Adding the Mean Structure

Carmen Ximenez, Universidad Autónoma Madrid

The present study aims to examine the conditions that affect the recovery of weak factor loadings when the model includes the mean structure, compared to analyzing the covariance structure alone. The study extends previous research on the recovery of weak factor loadings in Confirmatory Factor Analysis (CFA) by exploring the effects of adding the mean structure. This issue has not been examined in previous research. A simulation study is presented in which recovery of weak factor loadings is studied under conditions of estimation method, sample size, constraints in the mean structure, and factor correlation. The study is based on the framework proposed by Yung and Bentler (1999), which proved that the reduction of asymptotic variance can be substantial for the estimation of factor loadings when the associated mean structure is added to the covariance structure model. Results show that adding the mean structure improves the recovery of weak factor loadings and reduces the asymptotic variances for the factor loadings, and that certain conditions are important for the design of the study. For instance, the recovery of weak factor loadings improves when adding the associated mean structure to the CFA model in the models with a smaller number of factors and a small sample size, and when the constraints imposed on the mean structure imply that all the items have the same units of measurement. Therefore, under certain circumstances, modeling the means should be seriously considered for covariance models containing weak factor loadings.

CFA 1b: Testing Measurement Bias in Restricted Factor Analysis Using Product Indicators

Laura Kolbe, University of Amsterdam; Terrence Jorgensen, University of Amsterdam

In the presence of measurement bias, observed differences in composite scores (e.g., scale means) might not represent true differences in the construct a scale was developed to measure. One common method to assess measurement bias is Restricted Factor Analysis (RFA), which can be extended with Latent Moderated Structures (LMS) to test for nonuniform bias. Although RFA is a powerful means to detect measurement bias, several simulation studies observed severely inflated Type I error rates. By means of a Monte Carlo simulation, this study will compare methods to estimate latent interactions in

RFA models to test measurement bias with respect to a categorical contextual variable. Data will be generated under several conditions that vary according to sample size and effect size (i.e., magnitude of the uniform and nonuniform bias). We will investigate whether the inflated Type I error rates observed when using LMS are better controlled by using product indicators to model latent interactions in RFA. Additionally, we will examine whether the use of product indicators have adequate power to detect measurement bias relative to LMS. The product-indicator approach has never been studied in light of testing measurement bias, but because LMS is only implemented in limited Structural Equation Modeling (SEM) software (e.g., Mplus), knowing whether product indicators work at least as well as LMS could provide more researchers the opportunity to test for nonuniform bias using any SEM software package.

CFA 1c: Examining the Cross-Cultural Applicability via Generalizability Theory

Sümeyra Soysal, Hacettepe University; Çiğdem Akın Arıkan, Hacettepe University

PISA, which has been designed to collect information about 15 year-old students in participating countries, is the most comprehensive education survey at the present time. So results provided by PISA are extremely valuable for researchers, policy makers, educators, parents and students. PISA could also provide internationally comparable information, which allows cross-cultural/national comparability of measures. All PISA reports have revealed that measures of tests and questionnaires vary across countries. One of the major challenges of an international study such as PISA is the cross-cultural validity and applicability of all instruments. So, this study will investigate what causes the variation of measures and examine cross-cultural applicability in the context of multidimensionality via generalizability theory. Ten subdomain scales of questionnaires will be used for research: mathematics interest, instrumental motivation for mathematics, subjective norms in mathematics, mathematics self-efficacy, mathematics anxiety, mathematics self-concept, attributions to failure in mathematics, mathematics work ethic, mathematics intentions, and mathematics behavior. Random samples from Turkey, Finland and USA will be used and the number of students per country will be balanced by using the smallest number of respondents. Our G study will have the four facets person (P), county (C), dimension (D) and item (I), in which the person facet is within the county facet and the item facet is nested within the dimension facet. Whereas county and person are specified as differentiation facets, item and dimension are specified as generalization facets.

CFA 1d: State Space Approach to Canonical Correlation Analysis

Zhenqui (Laura) Lu, University of Georgia; Fei Gu, McGill University

Canonical correlation analysis (CCA) is a statistical method for multivariate analysis that has been widely used in many research areas. However, CCA cannot provide the variance of each canonical variate. The two-set canonical variate (CV-2) model is mathematically equivalent to CCA in terms of the obtained Canonical correlation. It uses a different normalization method when compared with the normalization in CCA. It restricts all canonical weight vectors to have unit length so that it imposes invariance constraints on the canonical weight vectors between populations, and also provides the variance of each canonical variate, which makes it convenient for researchers to evaluate the representativeness of canonical variates. However, due to the normalization and restriction of the CV-2 model, the parameter estimation of this model is a major obstacle for applied researchers. Therefore, we propose an innovative state space approach to the CV-2 model to advocate the use of this model. The state space specifications for the CV-2 model are explicitly presented, and numeric examples are provided to illustrate the utility of the state space approach. Moreover, we integrate two additional concepts that are related to CCA in the state space approach. Examples related to the extended concepts are demonstrated. Finally, additional discussions on other modeling extensions are given and discussed.

EST 1: 11:00 AM – 12:30 PM

Estimation

Chair: Wicher Bergsma, The London School of Economics and Political Science

EST 1a: Gremlins in the Data: Identifying Information Content of Research Subjects

Peter Ebbes, HEC Paris; John R. Howell, The Pennsylvania State University; John Liechty, The Pennsylvania State University

In many research studies in the social sciences we fit choice models using survey-based methods. For instance, in marketing and economics empirical demand functions are often estimated using survey-based conjoint approaches. With any survey-based method, the quality of the estimated model is dependent on the quality of the information obtained from the respondents. We develop a model that classifies and reweights respondents based on the statistical information contained in the respondents' survey questionnaires. We illustrate our model using simulated data and several data from empirical choice-based conjoint studies. Our approach provides an automated way of determining which respondents are relevant for model estimation. Importantly, we show that reweighting the respondents based on the quality of their information can lead to substantial different insights than by treating each respondent equally in model estimation.

EST 1b: Bayesian Inference Based on Conditional Likelihood Functions

Clemens Draxler, UMIT – The Health & Life Sciences University

This talk is concerned with Bayesian inference in psychometric modeling. It deals with conditional likelihood functions obtained from discrete conditional probability distributions which are generalizations of the hypergeometric distribution. The influence of nuisance parameters is eliminated by conditioning on observed values of their sufficient statistics and Bayesian considerations are only referred to parameters of interest. Since such a combination of techniques to deal with both types of parameters is less common in psychometric research a wider scope in practice of data analysis may be gained. It is argued that this procedure is particularly beneficial in small sample scenarios. The focus is on the evaluation of the empirical appropriateness of assumptions of the Rasch model, thereby extending the frequentists' approach which is dominating in this context. A number of examples are discussed. Some are very straightforward to apply. Others are computationally intensive and may be unpractical. The suggested procedure is illustrated using real data from a study on vocational education.

EST 1c: Staying in the Loop: Specifying Prior Probabilities

Fayette Klaassen, Utrecht University; Herbert Hoijtink, University Utrecht

A Bayes factor describes the increase in relative evidence or belief for two hypotheses after observing the data. Prior odds, the subjective belief for the hypotheses considered, are multiplied with the Bayes factor into posterior odds, a ratio of posterior probabilities of the hypotheses given the data. Posterior odds can serve as prior odds as new data becomes available, and so relative belief can continuously be updated. In applied research, the specification of prior probabilities is often not done explicitly or implicitly equal prior probabilities are considered for all hypotheses. In fundamental research, the importance of prior probabilities is noted, but generally no guidelines or illustrations of this importance are provided. This research aims to provide concrete tools and examples to help the specification of prior probabilities for applied users. Interpreting and defining a probability based on subjective beliefs is not intuitive. This research provides several illustrations that try to make the concept of prior probability more tangible. Furthermore, different approaches to specify prior probabilities are discussed and illustrated: equal prior probabilities, the complexity of a hypothesis as prior probability, betting odds to express a priori surprise about hypotheses and transforming conclusions from similar earlier research into prior probabilities. The second goal is to illustrate the consequences of adding, adjusting or dropping hypotheses for the specification of prior probabilities. Illustrated by a multiple replicated study, the effects of prior probability specification and hypothesis adjustment are shown.

EST 1d: The Fused Graphical Lasso for Computing Psychological Networks

Giulio Costantini, University of Milano-Bicocca; Sacha Epskamp, University of Amsterdam

Networks have been recently proposed as plausible models of psychological phenomena in several domains, such as personality psychology and psychopathology. In these fields, nodes represent variables such as cognitions, behaviors, emotions, motivations, and symptoms, and edges represent their pairwise associations. Edge weights are typically estimated using regularized partial correlations, for instance via the graphical lasso. In several situations, it is necessary to compute networks on observations that belong to different classes (e.g., patients vs. controls). Previous studies estimated either a single network for all classes or several networks in each class independently. These strategies may be both suboptimal. The Fused Graphical Lasso (FGL) has been recently proposed for dealing with such situations (Danaher et al., 2014), but it has never been applied to psychology before. FGL allows simultaneously estimating multiple partial correlation networks from observations belonging to different classes. FGL does not assume that the networks are similar, but if similarities are present, they are exploited to improve parameter estimates. This method requires setting two tuning parameters: One is akin to the graphical lasso parameter and controls sparsity, the second one controls the amount of similarity among classes. We developed an R package that implements automatic tuning parameter selection according to information criteria (AIC, BIC, and extended BIC) or relying on k-fold cross-validation. We present FGL from a theoretical point of view, discuss its performance in simulation studies, and show examples of its applications to personality psychology and psychopathology.

EST 1e: Prediction or Production? A Bayesian Stochastic Frontier Structural Equation Approach

Rüdiger Mutz, ETH Zurich

In performance measurement (e.g., group performance, student achievement) the application of regression analysis is the method of choice to predict outcomes. Actually, one is not mainly interested in predicting the average performance of a unit (e.g., group, student), but in assessing the performance of a unit in relation to the maximal possible outcome (the frontier of production) given the input of the respective unit. In this case Stochastic Frontier Analysis (SFA), adopted from econometrics, should be preferred to ordinary regression analysis (Kumbhakar, Wang, & Horncastle, 2015). The application of SFA is, however, restricted to measurement-error free outcome and input variables, which is unlike econometrics not a realistic assumption in psychology. The main objective of this paper is, therefore, to formulate for the first time SFA as a structural equation model with a measurement component for both the outcome and the input variables as a further development of a univariate model (Mutz, Bornmann, & Daniel, 2017, under review). A Bayesian version is preferred to consider the complicated model structure, among others due to the residual part with a normally distributed error component and a half-normally distributed Technical InEfficiency (TIE) component (individual deviation from the frontier), and due to an explanation model for the TIE. Besides the economic foundation (production theory) and the statistical model formulation, a simulation study should demonstrate the behavior of the model under different sampling conditions (SAS, PROC MCMC). The Austrian Science Fund (FWF) provided performance data of N = 1,046 funded projects to illustrate the proposal.

Symposium 3: 1:30 PM – 3:00 PM

Symposium 3: Networks and Latent Variable Models: Equivalences, Distinctions and Combinations

Chair: Sacha Epskamp, University of Amsterdam

Symposium 3a: Three Representations of the Ising Model

Joost Kruis, University of Amsterdam

Examining the structure of observed associations between measured variables is an integral part of psychometrics. At face value, associations inform about a possible relation between two variables, yet contain no information about the nature and directions of these relations. Making causal inferences from these associations requires the specification of a mechanism that explains the emergence of the associations. With the arrival of the network perspective, as such a mechanism, in psychometrics we have a promising new contender in a field that has been historically dominated by latent variable modelling. However, with the ever-increasing popularity of applying network models to data, it is important to subject this approach to some scrutiny. In this talk we therefore add a footnote to the application of network models to (binary) data. Specifically, we discuss two topics that can have an effect on the substantive interpretation of an obtained network structure. First, we discuss a recent paper in which we describe the common cause (latent variable model), reciprocal affect (network model), and common effect (collider model) frameworks as theoretically very distinct mechanisms from which associations between variables can emerge. However, while theoretically distinct, we demonstrated in the paper that their associated statistical models for binary data are mathematically equivalent. Furthermore, we discuss a recently accepted paper that shows how the sparsity assumptions made by the lasso estimation method influence the network structure resulting from this procedure.

Symposium 3c: Generalized Network Psychometrics: Combining Network and Latent Variable Models
Sacha Epskamp, University of Amsterdam; Mijke Rhemtulla, University of California, Davis; Denny Borsboom, University of Amsterdam

The formalization of the Gaussian Graphical Model (GGM), a popular undirected network model of partial correlation coefficients, as a formal psychometric model allows for its combination with the general framework of Structural Equation Modeling (SEM; Epskamp, Rhemtulla & Borsboom, in press). The GGM conceptualizes the covariance between psychometric indicators as resulting from pairwise interactions between observable variables in a network structure. This contrasts with standard psychometric models, in which the covariance between test items arises from the influence of one or more common latent variables. Here, we present two generalizations of the network model that encompass latent variable structures. In the first generalization, we model the covariance structure of latent variables as a network. We term this framework Latent Network Modeling (LNM) and show that, with LNM, a unique structure of conditional independence relationships between latent variables can be obtained in an explorative manner. In the second generalization, the residual variance-covariance structure of indicators is modeled as a network. We term this generalization Residual Network Modeling (RNM) and show that, within this framework, identifiable models can be obtained in which local independence is structurally violated. These generalizations allow for a general modeling framework that can be used to fit, and compare, SEM models, network models, and the RNM and LNM generalizations. This methodology has been implemented in the software package lvnet, which contains confirmatory model testing and two exploratory search algorithms: stepwise search algorithms for low-dimensional datasets and penalized maximum likelihood estimation for larger datasets.

Symposium 3d: How to Think of Model Complexity?
Riet van Bork, University of Amsterdam; Mijke Rhemtulla, University of California, Davis; Denny Borsboom, University of Amsterdam

This talk explores how to think of model complexity of regularized partial correlation network models compared to latent variable models. Within the latent variable modeling approach the complexity of a model is usually expressed by the number of freely estimated parameters of a model. The more parameters are allowed to be freely estimated, the more adaptable the model is to the data. The effective degrees of freedom of a regularized partial correlation network model can be estimated as the number of zero-valued edges (Zou, Hastie, & Tibshirani, 2007). This number is typically much smaller than the degrees of freedom of a latent variable model on the same set of variables, suggesting that network models are vastly more complex. However, the number of freely estimated parameters does not always capture the complexity of a model. For example, models that have equivalent numbers of freely estimated parameters can differ in their flexibility to fit random data (Preacher, 2006). When comparing a network model and a latent variable model in how they fit to empirical data it is important to account for the model's flexibility to fit arbitrary data. We will consider several approaches to model complexity from philosophy of science. Our goal is not to provide a technical solution to the question of how to assess model complexity of network models and latent variable models, but to evaluate which possible approaches to model complexity are appropriate when dealing with network models and latent variable models that stem from such different modeling frameworks.

Symposium 3e: A Comparison of Latent Variable vs. Network Models Using Longitudinal Data
Abe Hofman, University of Amsterdam; Rogier Kievit, Univerity of Cambridge; Ingmar Visser, University of Amsterdam; Han van der Maas, University of Amsterdam

In psychology the correlational structure between items or subtests is often analyzed with latent variable models and more recently using different network modeling approaches. For example, in the study of intelligence a positive (cross-sectional) correlational structure - the positive manifold - is a well established empirical phenomenon which is often explained by introducing a general intelligence factor. Recent papers have shown that both latent variable and network models are in some cases mathematically equivalent, hence can result in the same observed correlational structures. However, these models differ greatly in their substantive explanations and imply that different mechanisms generated these correlations. This talk presents a comparison of a network model and a latent variable model using longitudinal data. We use a longitudinal structural equation modeling framework and different specifications of latent change score models that can capture the implied dynamics of different developmental theories. Using data from a large online learning platform for mathematics (Math Garden), we show that the development of learning to do mathematics is best described by a network approach that allows direct links between the development of different domains (e.g. counting and addition).

Symposium 4: 1:30 PM – 3:00 PM

Symposium 4: Computerized Adaptive and Multistage Testing: Developments, Challenges, and Solutions

Chair: David Magis, University of Liège

Symposium 4a: The Paradox of Adaptive Testing and Item Calibration

Peter van Rijn, ETS Global

The estimation of item parameters, or item calibration, using data from a Computerized Adaptive Test (CAT) can be problematic for certain Item Response Theory (IRT) models because the algorithms typically adjust the difficulty level of the test to the ability of the test takers. This can create at least two problems, which are not often addressed in research on online calibration: 1) Guessing behaviors become less likely, and 2) The observed ability range per item is restricted. It will be shown that certain approaches in which new items are placed in the test ignore these issues. Furthermore, several IRT models become hard to estimate in an adaptive context and their use becomes questionable. Several arguments will be made. One argument is that it does not seem to make sense to estimate parameters for test behavior that is hard to observe because of the CAT algorithm. Another argument is that it does not seem to make sense to create a calibration situation in which behaviors can be substantially different from the testing situation. A third argument is that it does not seem to make sense to keep item parameters fixed while plenty of new data becomes available. These issues will be illustrated by making use of straightforward simulations.

Symposium 4b: A New Probability-Based Classification Method for MST

Duanli Yan, Educational Testing Service

This presentation will introduce a probability-based approach for classification. It is based on recent studies on addressing the challenges of classification accuracy and fairness in MST. Up till now, classifying a student into one of several groups based on her test score remains a research problem in the areas of computerized classification testing for CAT and MST. The Sequential Probability Ratio Test (SPRT) is well studied, initially proposed by Wald (1947) and further developed by Spray (1993) for classification in computerized classification testing. However the SPRT methods always result in an indifference region in decision space in which no classification decision can be made and requires more test items to reach a decision. We attempt to develop a probability-based method that always provides a classification on the student score. In this presentation, several different probability functions are applied and tested. The method will improve classification accuracy and efficiency, especially for two-stage testing when stage 1 classification determines the groups and the type of stage 2 testing. This presentation will illustrate the probability-based classification method for a two-category example, design a classification rule that is both accurate and fair to students under tests, and discuss the results of the studies.

Symposium 4c: Routing and Scoring Issues in an MST Design

Alina von Davier, ACTNext

In this presentation, I will draw on two recent studies that address challenges in Multi-Stage Testing (MST). The first study of Ricarte, Curi and von Davier (in progress) investigates the issue of misrouting when smart test takers make mistakes on easy items in the routing module. We investigated a Rasch-version of the logistic positive exponent model (Samejima, 2000) with a fixed value (larger than 1) of the exponent parameter so that the model penalizes less dramatically than the traditional IRT models for mistakes on easy items. A simulation study was conducted, in which the sample ability is drawn from a mixture distribution: a part of the sample had the response patterns simulated according to the 3PL IRT model, and the other part had the response patterns simulated according to the 4PL IRT model, where a slip parameter leads to a lower probability of success for high-ability subjects. The second study investigates the issue of small samples for an MST administration and continues the research of non-IRT methods, such as regression tree, to routing and scoring in MST (Reshetnyak, von Davier, Lewis, & Yan, in progress). In particular, in this study, a comprehensive comparison of the IRT and CART methods is presented.

Symposium 4d: mstR: An R package to Generate Multistage Testing Designs

David Magis, University of Liège; Duanli Yan, Educational Testing Service; Alina von Davier, ACTNext

Multi-Stage Testing (MST; Yan, Von Davier, & Lewis, 2014) has become an important framework of tailored testing. Various operational testing programs are nowadays considering MST for practical administrations. However, despite its increasing use and assessment with respect to item-level adaptive testing, there is still very limited access to accurate and open-source software to generate MST scenarios for, for example, research purposes. This talk focuses on the presentation of a new package from the R software, called mstR (Magis, Yan, & Von Davier, 2017). This package was built in the same spirit of the package catR for CAT designs. However, mstR permits to generate repeated response patterns under a predefined MST scenario, by providing the set of modules and related item parameters, the number of stages and the connections between modules from successive stages. Several rules for optimal module selection and ability estimation (under IRT framework or based on test scores) are also available. Eventually, mstR was developed in such a way that CAT scenarios elaborated under catR use can easily be transposed to mstR by providing in addition the modules and MST structure as additional input arguments. In this talk we will briefly describe the package, its assets and functionalities, and we will illustrate its performance by describing an application using a real item bank.

IRT 2: 1:30 PM – 3:00 PM

IRT: Fit and Diagnostics

Chair: Johan Braeken, University of Oslo

IRT 2a: Assessing Item Fit of Unidimensional IRT Model Using Plausible Values

Bartosz Kondratek, Educational Research Institute

Most of the methods of item-level fit analysis of IRT models examine discrepancies between observed and expected frequencies over specific ranges of ability. Early attempts (Bock, 1972; Yen, 1981; McKinley & Mills, 1985) relied on grouping based on point estimates of θ. This drew a twofold criticism: the observed frequencies were model dependent and error of θ estimates was ignored. Orlando and Thissen (2000) addressed these issues by proposing grouping over number correct scores. Their approach, however, is not applicable in the areas where the IRT models flourish the most, like computerized adaptive testing or incomplete (booklet) designs used in large scale educational surveys. Recently, a comeback of residual analysis that is performed on the θ scale is observed, with Stone and Zhang (2003) describing a method for comparing posterior expectations with "pseudocounts" or Toribio and Albert (2011) providing evidence of the utility of the McKinley and Mills statistic when used in Bayesian posterior predictive checks. In this presentation I propose yet another approach to grouping on θ scale for item fit analysis that involves drawing plausible values from students' posterior distribution in order to determine a quantile group membership. Monte Carlo studies are reported that analyze the distribution of residuals in case of logistic models for dichotomous items. The issue of model dependency of observed proportions is discussed in order to pursue both graphical item fit analysis and a fit statistic. The simulations are performed with uirt software (Kondratek, 2016) that runs in the statistical package Stata.

IRT 2b: Goodness of Fit Assessment for Restrict Re-Calibration of IRT Models

Yang Liu, University of Connecticut; Ji Seung Yang, University of Maryland, College Park

In many Item Response Theory (IRT) applications, researchers borrow published item parameter estimates from previous studies to calculate scale scores or to build explanatory models in a new sample. Such a two-stage model-building scheme often calls for modifications to the original measurement model, which is termed Restricted Re-calibration (RR) in the current work. For instance, regression paths among latent variables need to be added/removed to describe the hypothesized association among latent variables. As another example, the mean and variance of the latent variable distribution may be re-estimated in order to accommodate the fact that the new sample and the original calibration sample were drawn from different populations. Consequently, it is necessary to account for two sources of sampling error in RR: One arises from the direct use of estimated parameters from a published calibration study, and the other is introduced by estimating additional parameters in the new sample. We study the Goodness Of Fit (GOF) assessment for measurement/structural models built in two stages, which extends the existing work of GOF testing in cross-validation samples (Joe & Maydeu-Olivares, 2006). We derive an adjusted asymptotic covariance matrix for the second-stage parameter estimates and a limited-information GOF test for the final model. The empirical Type I error and power performance of the test statistic is investigated via Monte Carlo simulations.

IRT 2c: Score Tests and Information Matrix Approximations for MIRT Under Misspecification

Carl F. Falk, McGill University; Scott Monroe, University of Massachusetts, Amherst

Score (or Lagrange Multiplier) tests have seen renewed interest in Item Response Theory (IRT). Score tests allow researchers to test whether a parameter differs from a fixed value, and can be used to diagnose model misspecification. Whereas score tests do not require re-fitting of the IRT model, they require an approximation to the information matrix that also includes the to-be-freed parameter. The observed cross-product approach is typically the fastest to compute, but may be the least accurate. The observed Hessian is often more accurate, but requires more computational time due to use of second-order derivatives. Although previous research usually employs one of these two information matrix approximations, a generalized score test designed specifically for use under misspecification has apparently not been studied in an IRT framework. This generalized score test is analogous to the sandwich approach in computing standard errors in that it requires both first- and second-order derivatives of model parameters. We present a Monte Carlo simulation study that examined the performance of these information matrix approximations in a multidimensional IRT framework under various levels of model size and misspecification. We argue that the generalized score test performed the best in these simulations, and we further discuss the utility of score tests under misspecification. When the initially fitted model is highly misspecified, we argue that score tests may not be useful in determining whether a parameter is correctly fixed to its data generating value.

IRT 2d: RMSEA2-Based Power for IRT Models: Theoretical Derivation and Empirical Validation

Daniel Serrano, Pharmerit; Charles Iaconangelo, Pharmerit

The field of psychometrics, as applied to the development of patient reported outcomes, is hampered by insufficient sample size. Sample size determinations are often derived from overly simplistic heuristics. As a result, sample size is often severely underestimated. Such sample sizes undermine the ability to detect important item- and model-dependent properties (e.g. dimensionality, parameterization, DIF, etc.). In this talk, we link two existing literatures to extend the full information RMSEA statistics of Maydeu-Olivares et al. to the RMSEA-based statistical power framework developed by MacCallum et al. (1996). The primary difference between the MacCallum et al. (1996) procedure for normal-theory EFA/CFA/SEM and the method developed here is that under normal theory, only the loadings contribute to the second-order moments and degree of freedom estimate, whereas with IRT models, both the slopes and thresholds contribute. A simulation study compared the theoretical estimates to the empirical power. Theoretical power was computed for a 10 binomial-item-set for exact fit (null RMSEA = 0) and not close fit (null RMSEA = 0.08) for sample sizes ranging from 100 to 700, by increments of 50. For each sample size and fit test (exact and not close fit), 1000 replications were simulated under a 2PL model fitting a 1PL model. Empirical power for the RMSEA was computed for each replication. Preliminary results indicate that theoretical and empirical power computations for exact fit and not close fit were correlated r = .99913 and r=.99973, respectively. Full details and results will be presented.

IRT 2e: On the Consistency of Latent Variable Models with the Observations

Rudy Ligtvoet, University of Cologne

A fundamental problem is psychometrics it is how to determine whether the response data is consistent with a latent variable model. One approach to address this problem is to look at the necessary observable properties for various latent variable models. Many such properties have been shown to be hierarchically related, imposing ever-increasingly tighter bounds on the latent variable models. In this talk, I will provide an overview of some relevant properties for the broad class of monotone latent variable models. I will do this not just in terms of the implications between these properties, but also in terms of the relative size of the outcome space they encompass. This relative size provides an expression of complexity of general classes of latent variable models, which is at least useful in weighting the evidence in favour of a latent variable model.

LDA 1: 1:30 PM – 3:00 PM

Longitudinal Data Analysis

Chair: Tom Loeys, Ghent University

LDA 1a: A Unified Framework of Cross-Lagged Models

Satoshi Usami, University of Tokyo; Kou Murayama, University of Reading; Ellen Hamaker, Utrecht University

Inferring mutual effects or causality between variables is a central aim of behavioral research and various longitudinal models that include cross-lagged relationships (e.g., Cross-Lagged Panel Model: CLPM, Random-Intercepts CLPM: RI-CLPM, Stable Trait Autoregressive Trait and State (STARTS) model, Latent Change Score (LCS) model, Autoregressive Latent Trajectory (ALT) model) have been proposed in different contexts. The present research provides a unified framework to understand the conceptual, mathematical and methodological differences among these models and to show potential alternative models to analyze longitudinal data through three different perspectives. Simulation and empirical examples are also provided to show that the different models can easily draw different conclusion about mutual effects and to show the greater risk of obtaining improper solutions in parameter estimates especially under the presence of model misspecifications. We will discuss how applied researchers should choose model(s) to evaluate mutual effects in future cross-lagged panel research.

LDA 1b: Evaluation of Quality of Estimate for Change in Two-Wave Data

Yasuo Miyazaki, Virginia Tech

Multilevel growth model or latent growth curve model by Structural Equation Modeling (SEM) framework provides a powerful tool for studying a change in behavioral and social sciences. These modeling frameworks, however, require at least three waves of information and in practice we often encounter the cases that have only two waves because of limitation of resources or nature of the study goals. In such cases, we can still obtain a reasonable estimate for the true change if information on multiple indicators such as test items/subscale scores are available, which is frequently the case. Of course, if there are at least three waves of data with multiple indicators, we can examine not only the change, but also the tenability of assumption of measurement invariance by fitting a second-order growth curve model (Sayer & Cumsille, 2001, Hancock, Kuo, & Lawrence, 2001). But, with two waves, we need to go with a first-order model and to estimate the true change by the difference in true scores in two waves. Intuitively, this estimate should be reasonably good, but there is no study about how good it is compared to the second-order model. Thus, the goal of the present study is to evaluate the amount of loss in terms of precision of the estimate by comparing the estimate from the first-order model that uses only the two-wave information by eliminating one wave to the one from the second-order model that fully utilizes the three-wave information via a Monte Carlo simulation and a derivation.

LDA 1c: Several Approaches to the Modeling of Ephemeral Effects

John Tisak, Bowling Green State University; Guido Alessandri, Sapienza University of Rome; Marie S. Tisak, Bowling Green State University

When studying longitudinal phenomena, the notions of traits and states can be a useful classification. Specifically, traits represent basic human characteristics that have a permanency or enduring property, while on the other hand, states are environmental or ephemeral that are more time specific (Tisak & Tisak, 2000). Admittedly, research often focusses on traits and the relationships of these traits to other important variables. In addition, one might be interested in the proportion of these properties for any psychological variable. Moving in a different direction, this presentation focusses on the more ephemeral aspects of longitudinal variables, that is, states. A very practical justification for this direction is model-fitting indices. Concretely, the modeling of states may improve the acceptability of one's statistical model or more precisely one's Structural Equations Model (SEM) without the inclusion of "nuisance" parameters. A probably more important justification for a more developed state model is a more accurate reflection of the situation under study. Given this framework, two possible approaches are suggested for the statistical model of ephemeral effects. The first is a one-factor model (Spearman, 1904), and the second will be ARMA models as developed by Browne and Nesselroade (2003).

LDA 1d: Parameter Estimation in Latent Growth Curve Item Response Models

Xiaying Zheng, University of Maryland, College Park; Ji Seung Yang, University of Maryland, College Park

When individual changes are measured by administering the same set of items multiple times, a second-order Latent Growth Curve model (LGC; Bollen & Curran, 2006) allows estimating the overall initial status and growth rate as well as variabilities. When the observed item responses are categorical, coupling item response models (IRT) as the measurement model with LGC is a natural extension. However, when testlet effects are introduced into the measurement model to address the local dependency attributed to repeated uses of the same items, the estimation of the LGC-IRT models becomes challenging due to involvement of high dimensional latent variables. This research compares three estimators of LGC-IRT model parameters in the presence of testlets, namely the Monte Carlo Expectation-Maximization algorithm (MCEM; Wei & Tanner, 1990), the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2010a) and the diagonally weighted least squares estimator (DWLS; Muthén, du Toit, & Spisic, 1997). MCEM and MH-RM were the computation algorithms to obtain full-information maximum-likelihood estimates in the presence of many latent variables, while DWLS is a limited-information estimator. A Monte Carlo simulation study is conducted to examine the three estimators under various data conditions, including test length, sample size, and existence of missing data (i.e., complete data, missing at random with respect to covariates only, and missing at random with respect to both covariates and observed outcomes). The performance of the estimators will be assessed with respect to convergence rate, estimation time, relative bias of item and structural parameters, adequacy of standard errors and confidence interval coverage.

LDA 1e: Stochastic Bivariate-Dual Change Score Model: Evaluation and Issues in Implementation

Linying Ji, The Pennsylvania State University; Sy-Miin Chow, The Pennsylvania State University

The ever increasing popularity of multivariate longitudinal data in the social and behavioral sciences in the past few decades has called for the development of more sophisticated models to understand how processes unfold over time. Among these models is the Bivariate-Dual Change Score (BDCS) model proposed by McArdle and Hamagami (2001), which uses a latent difference equation framework to represent the bivariate dynamics of two variables, their influence on each other, as well as selected interindividual differences in intraindividual changes. The original BDCS model assumes that the underlying latent dynamic processes are deterministic and do not show residuals at the latent level. In this study, we consider the properties of a stochastic BDCS model, wherein both process noises and measurement errors are included. We perform Monte Carlo simulations to exam the robustness of this model under different degrees of misspecifications in initial conditions and different sample sizes (both in terms of number of participants and number of measurement occasions). We also investigate how magnitudes of the self-feedback and coupling coefficients may affect the estimation results. We demonstrate the implementation of the stochastic BDCS model using the longitudinal data of reading and arithmetic performance scores from the Early Childhood Longitudinal Study, Kindergarten Class of

1998 (ECLS-K) study. We compare estimation results from different variations of the stochastic BDCS model and discuss the substantive implications.

IRT 3: 1:30 PM – 3:00 PM

IRT 3

Chair: George Engelhard Jr., University of Georgia

IRT 3a: Psychometric Models for Forecast Evaluation

Ed Merkle, University of Missouri

The aim of this talk is to describe and illustrate application of psychometric models to the domain of probabilistic forecasting. This domain is receiving increased attention, particularly in the context of voting (e.g., Brexit and the U.S. election), and methods are needed for evaluating the forecasters and items involved. Many of the concepts developed in psychometrics for standardized tests can be readily applied to such evaluations, yet these concepts have received little attention in the forecasting literature. The talk will include discussion of specific issues, solutions, and extensions, focusing on data from a recent forecasting tournament.

IRT 3b: Some Considerations on the Derivation of Logistic Item Response Theory Models

Stefano Noventa, Hector Research Institute of Education Sciences and Psychology; Luca Stefanutti, University of Padova; Giulio Vidotto , University of Padova

Logistic Item Response Theory (IRT) models are widely applied in psychological testing. A formal derivation of their families generally relies on functional equation methods and a combination of both specific assumptions on the item response function (e.g., continuity, monotonicity, and asymptotical behavior), and general assumptions or criteria (e.g., local stochastic independence, sufficiency of statistics, and specific objectivity). In particular, an important although often unnoticed requirement is the existence of a dense set of items. At the same time, in other fields of science like physics, engineering, and chemistry, the derivation of logistic models often relies on different methodologies. For instance, a classical approach is Boltzmann's Most Probable Distribution method (MPD), under which IRT logistic models become the distributions accounting for the maximum number of possible outcomes in a test, while introducing latent traits and item characteristics as constraints to the system. More in general, equilibrium statistical mechanics and the principle of maximum entropy describe logistic models as equilibrium solution accounting for the maximum lack of information, or maximum randomness in the system. In the present work, IRT logistic dichotomous and polytomous models were derived and their properties and assumptions discussed in lights of the MPD framework, with respect to both finite and infinite populations. In particular, a representational theory of measurement framework was considered to assess the measurement level of the latent scale. Considerations are provided on possible implications over model construction and connection with set theoretic approaches like knowledge space theory.

IRT 3c: A GRM Where Parameters are Functions of a Continuous Variable

Henk Kelderman, Leiden University; Thomas Bakker, Leiden University

We study a graded response model (Samejima, 1969) whose parameters are allowed to vary as smooth functions of other variables. To the extent that the parameters of the response model are not invariant, one or more observed indicators of latent variable(s) may suffer from response shift. If parameters of the distribution of the latent variable are not invariant, population properties may change, e.g. means or variances may change or over time. These functions are estimated using K-fold cross-classification of kernel-weighted local likelihoods over the relevant range of the collateral variable. We describe the algorithm and a method of post-processing for best interpretation of the parameter functions. The method was applied to a variety of simulated data sets in order to test its properties. We illustrate the method on a data set from a study of generational effects on the parameters of Big-Five personality items administered over a time period of 21 years (Smits, Dolan, Vorst, Wicherts, & Timmerman, 2011).

IRT 3d: Bias of Two-Level Scalability Coefficients and Their Standard Errors

Letty Koopman, University of Amsterdam, ; Mark de Rooij, Leiden University; B. J. H. Zijlstra, University of Amsterdam; Andries van der Ark, University of Amsterdam

Mokken's H-coefficients are popular indices to assess the scalability of item pairs, items, and sets of items. Recently, there has been renewed interest in Mokken's scalability coefficients for two-level data: item scores for subjects scored by multiple raters. Two-level data result in 9 types of H-coefficients: between-rater coefficients, within-rater coefficients, and their ratio. These coefficients are available for item pairs, items, and sets of items. Lately, estimates of standard errors (SEs) based on the delta method have been derived for all two-level H-coefficients, and both the point estimates and SEs have been implemented in R. In this study, we also considered SE estimates based on the bootstrap and cluster bootstrap. We investigated the bias and computation time for the point estimates and the three SEs estimates under several conditions. Bias of point estimates of the scalability coefficients was negligible regardless of condition. The SEs were slightly underestimated when item discrimination was high, there was little within subject variance, or the number of raters was low. In all other conditions the bias of all SE estimates was negligible. The ratio coefficient and its SE appeared only unbiased when the scalability coefficients themselves were above the desired threshold, regardless of condition. Difference in bias favored the individual level bootstrap as SE estimation method, although the difference was in most conditions close to zero. Since the difference was very small and computation time was much higher for the bootstrap methods, the delta method is preferred in practice.

IRT 3e: Person Scoring in 2PL Model between CB-IRT and FB-IRT

Karen M. Schmidt, University of Virginia; Kwanghee Jung, ACTNext; Heungsun Hwang, McGill University; J. Patrick Meyer, University of Virginia; Ji Hoon Ryoo, University of Virginia

Component-Based IRT (CB-IRT) utilizing Generalized Structured Component Analysis (GSCA) in IRT has been considered as an alternative to Factor-Based IRT (FB-IRT). Based on the advantages in GSCA such as no distributional assumptions and providing proper solutions avoiding factor score indeterminacy, the efficiency and applicability of GSCA have been shown in the many other areas including human brain research, genetic epidemiological research, commerce, etc. IRT would also benefit from these advantages, especially for either small or large data. Since we demonstrated the applicability of CB-IRT, we have explored its stability in terms of estimates and compatibility with FB-IRT. In this paper, we examined the associations of person parameter estimates from CB-IRT and FB-IRT with true ability in 2-parameter logistic model (2PLM) applying a simulation study. The simulation design included various conditions on sample sizes (50, 100, 250, 500, 1000, and 2000), number of items (5 items and 9 items), discriminant parameters ranging from 0.5 to 4.0, and difficulty parameters ranging from -2.0 to 2.0. The results show that their associations of CB-IRT and FB-IRT with true ability via correlations of scorings are almost identical between CB-IRT and FB-IRT, but CB-IRT indicated more efficient (small variances) over different sample sizes than FB-IRT. Even if CB-IRT ran with 100 Bootstraps, running times of CB-IRT in the gesca package in R is shorter than those of FB-IRT in the ltm package in R.

MLSA 1: 1:30 PM – 3:00 PM

Measurement and Large Scale Assessment

Chair: Rianne Janssen, KU Leuven

MLSA 1a: Producing a Reliable Collaborative Problem Solving Scale in PISA 2015

Qiwei He, Educational Testing Service; Haiwen Chen, Educational Testing Service; Matthias von Davier, National Board of Medical Examiners; Mary Louise Lennon, Educational Testing Service; Hyo Jeong Shin, Educational Testing Service

Collaborative Problem Solving (CPS) is a critical and necessary skill across educational settings and in the workforce. The Programme for International Student Assessment (PISA) 2015 first introduced the testing of CPS in an international large-scale assessment, which not only highlights the importance of CPS among the students' competence in the 21st century but also proposes new research questions such as how to develop a reliable scale for examining the students' CPS skills. The purpose of this study is to present how the CPS skills were reliably estimated in PISA 2015 by addressing two major challenges in

CPS assessment from item development and IRT estimation. Specifically, we will illustrate the item development with bonus and rescue items design, estimate the factors that impact the CPS item difficulty levels, and present a residual analysis using empirical data to introduce an approach to solving the potential problems in item local dependence.

MLSA 1b: Can Score Jump Threaten the Stability of Linking Scales?
Rongchun Zhu, ACT, Inc.

This study aims to explore potential threat of test-taker score jump to the stability of linking scales using PISA 2003 and 2012 data. The stability of linking scales is fundamental for large-scale testing programs to monitor educational status over time. Scores across testing cycles are linked to the same scale using linking items, so the linking item set offers a window for observing the linking process. As a small subset of items, linking items represent separate testing item pools used in different testing cycles on test content and statistics under similar test context. The consistency of group performance and item parameter estimates over time is desired for linking items. However, individual countries or regions can have sizable score jump, possibly benefited from effective educational reform, which may become a threat to the accuracy of score reporting for these groups (not likely for PISA full sample). To look into this issue, this study selects PISA data from two countries: Portugal, whose average math score increases 37 points from 2003 to 2012 (as major subject), and one country that has nearly no score change but shares common characteristics to Portugal. The data from several math clusters will be analyzed that contain at least four linking items per cluster. For the linking items, item parameter estimates will be compared using different item response theory models, including Rasch, 2-PL, and 3-PL models. Implications for selecting and using linking items appropriately will be discussed.

MLSA 1c: A Comparison of Methods of Subscoring with Unbalanced Item Structures
Chong Min Kim, Gyeongin National University of Education; Moonsoo Lee, Korea Institute for Curriculum and Evaluation

Subscores are influential in that they provide meaningful information about students' strengths and weaknesses in the assessment of students’ growth and development. Previous studies showed that psychometric qualities with focus on the reliability, validity, and value of subscores might vary per different subscoring methods which were based on CTT and IRT. In the actual test development situation, however, the perception and interest of subscoring methods are not so high. Specifically, the number of test items per domain would be decided according to importance of each domain without considering the reliability of the subscores, which lead to unbalanced data with uncertain psychometric qualities. Thus, the purpose of this study is to compare the methods of subscoring in tests with unbalanced number of items across domains. Specifically, we will compare the reliability of three subscoring methods (Wainer's augmentation method, Haberman's three subscoring methods, and Yen's objective performance index) based on multidimensional item response theory with simple structure and complex structure in tests with unbalanced items. First, we will use simulated test data which consist of six domains with 3, 4, or 5 items across dimensions (total 30 items). Second, we will analyze actual ICT literacy test data (six domains and total 30 items) for elementary and secondary students in Korea usingthe R package 'subscore'. Based on our results, we will discuss the usefulness of MIRT subscoring methods and suggest implications in growth referenced evaluation with standard setting.

MLSA 1d: Reporting Reliable Aggregate-Level Subscores
Usama Ali, Educational Testing Service; Joseph Rios, Educational Testing Service

The demand for aggregate-level subscore reporting has increasingly grown as institutions, states, and countries have become more interested in the relative strengths and weaknesses of their students when compared to other institutions, states, and countries (Haberman, Sinharay, & Puhan, 2009). Although there has been extensive research investigating the necessary test design considerations (e.g., number of items) for reporting reliable subscores for identifying student learning needs at the individual-level (e.g., Feinberg, 2012; Sinharay, 2010), no such investigations have been conducted for aggregate-level subscore reporting. In fact, research on aggregate-level subscore reporting has been limited to designing methodologies to evaluate the utility of providing aggregate-level subscores from operational data (Haberman et al., 2009; Longford, 1990; Wainer, Sheehan, & Wang, 2000); however, understanding the minimal design considerations for reporting reliable subscores will have implications for planning data collections for aggregate-level assessments, which presents its own unique challenges. This study investigates three variables that may impact the reliability of such scores: (a) the number of within-form items, (b) the number of total test forms, and c) the sample size within a group. Data from an operational large-scale aggregate-level assessment are used. The results of this research will help provide psychometric guidelines for development of future testing programs that report subscores at the aggregate-level given the testing challenges associated with group-based assessments (e.g., a short test to fit within a short testing timing).

MLSA 1e: Comparing Different Standard Setting Methods in a Higher Education Context
Iris Yocarini, Erasmus University Rotterdam; Samantha Bouwmeester, Erasmus University Rotterdam; Joran Jongerling, Erasmus University Rotterdam

Ideally, standards for pass/fail cut-off scores are set using expert panels. In a higher educational context however, using mostly in-house designed tests, such panels are often too time and cost intensive. Consequently, a pre-fixed percentage of test items to answer correctly is often used to determine cut-off scores. Problem with such absolute standards is that test difficulty is not taken into account. Incorporating item difficulty is especially difficult in higher education tests as item characteristics for these non-standardized tests are generally unavailable. Alternatively, relative standards may be problematic as strategic behavior from students may lead to low cut-off scores. Therefore, a compromise method such as the Hofstee method or Cohen method might be applied. In this paper the validity of these methods is evaluated. Do competent students pass and incompetent students fail when using these compromise standard methods? Specifically, the pre-fixed percentage method, the Hofstee method, and the Cohen method are compared in this simulation study. Here, the specific percentages used in each methods are varied. Additionally, to evaluate these methods within realistic higher education settings, several variables are varied such as the sample size, the average true ability of students, test difficulty, test length, and test discriminability. The expected results include the mean square error and correlation between the simulated and estimated true knowledge scores.

STAT 1: 1:30 PM – 3:00 PM

Statistics
Chair: Matthew Johnson

STAT 1a: Covariance Model Simulation Using Regular Vines
Steffen Grønneberg, BI Norwegian Business School; Njål Foldnes, BI Norwegian Business School

We propose a new and flexible simulation method for non-normal data with user-specified marginal distributions, covariance matrix and certain bivariate dependencies. The VITA (VIne To Anything) method is based on regular vines and generalises the NORTA (NORmal To Anything) method. Fundamental theoretical properties of the VITA method are deduced. Two illustrations demonstrate the flexibility and usefulness of VITA in the context of structural equation models.

STAT 1b: Regression Modelling with I-Priors
Wicher Bergsma, The London School of Economics and Political Science; Haziq Jamil, The London School of Economics and Political Science

This is an overview of a unified methodology for fitting parametric and nonparametric regression models, including additive models, multilevel models, and models with one or more functional covariates. We also discuss an associated R-package called iprior. An I-prior is an objective prior for the regression function, and is based on its Fisher information. The regression function is estimated by its posterior mean under the I-prior, and scale parameters are estimated via maximum marginal likelihood using an Expectation-Maximization (EM) algorithm. Regression modelling using I-priors has several attractive features: it requires no assumptions other than those pertaining to the model of interest; estimation and inference is relatively straightforward; and small and large sample performance can be better than Tikhonov regularization. We illustrate the use of the iprior package by analysing three well-known data sets, in particular, a multilevel data set, a longitudinal data set, and a dataset involving a functional covariate.

STAT 1c: Small sample Asymptotics for Multinomial Goodness of Fit Tests

Simone Giannerini, University of Bologna; Greta Goracci, University of Bologna

In this work we propose a (small sample) asymptotic approximation for multinomial goodness-of-fit tests. We focus on the power divergence family described in Read and Cressie (1988) [Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer]. The family includes as special instances Pearson's X^2, the likelihood ratio and the Hellinger distance statistics and, under the null hypothesis, all its members have a chi-square asymptotic distribution. We derive the (approximated) asymptotic distribution for the whole family under a general framework that includes both the null and the alternative hypothesis. Moreover we prove the asymptotic normality for the whole family. The results allow to compute analytically the power function for non local alternatives so that theoretical comparisons of the performance of the different statistics are possible.

STAT 1d: Informative Hypothesis Testing

Leonard Vanbrabant, Ghent University

In Informative Hypothesis Testing (IHT) prior knowledge about the population of interest is included in the hypotheses in terms of order constraints. To give some typical examples, in an ANOVA setting we might expect that the group means follow a certain order (e.g., $H0 : \mu 1 < \mu 2 < \mu 3 = \mu 4$), or that the variances are monotone increasing (e.g., $H0 : \sigma 1 < \sigma 2 < \sigma 3 < \sigma 4$). In a linear model, we might expect that the regression coefficients are subject to multiple one-sided constraints (e.g., $H0 : \beta 1 > 0; \beta 2 > 0; \beta 3 > 0$). In a confirmatory factor analysis, the factor loadings might be fully ordered (e.g., $H0 : \lambda 1 < \lambda 2 < \lambda 3 < \lambda 4$). In this presentation, we will discuss several examples of such informative hypotheses. Moreover, we will show how they can be tested using the available software tools. In addition, we will present simulation results to show the substantial gain in power in small samples.

STAT 1e: Correlation Computation for Ratings Using the Idea of Pearson (1913)

Kenpei Shiina, Waseda University; Saori Kubo, Waseda University; Takashi Ueda, Waseda University

It is widely known that the correlation coefficient, r, calculated from ordered categories is distorted, and thus the use of polychoric correlation is recommended. Karl Pearson (1913) recognized the distortion and proposed correction formulae, which have been forgotten. We revisited his formulae and arrived at a new formula with a classical flavor. If variable X has p categories and variable Y has q categories, the data are represented as a p by q contingency table. Pearson’s corrected r is computed as follows: 1) Assume a hidden bivariate normal distribution with zero means, unit variances, and rho; 2) estimate the category thresholds for both variables from the marginal distributions of the table; 3) compute each category score by estimating the mean of the truncated normal distribution defined between adjacent thresholds; 4) compute r by using the category scores; and finally, 5) compute the correlation between X and its category scores and the correlation between Y and its category scores, and then divide the r obtained in Step 4 by the product of the correlations in Step 5 to obtain the corrected r. Whereas Steps 1–4 are often used presently, Step 5 is truly unique. However, because Step 5 may include a problematic assumption, our new formula does not include Step 5. Instead, in our formula, different scores are assigned to each cell of the contingency table in Step 3, while in Pearson’s formula, category scores are assigned to rows and columns. We report that the computational results obtained were promising.

Symposium 5: 3:20 PM – 4:50 PM

Symposium 5: Response Bias in International Large-Scale Assessments

Chair: Matthias von Davier, National Board of Medical Examiners; Lale Khorramdel, Educational Testing Service

STAT 1a: Mixture and Multi-Process IRTree Models for Measuring Response Styles

Lale Khorramdel, Educational Testing Service; Matthias von Davier, National Board of Medical Examiners; Artur Pokropek, European Commission

Personality constructs, attitudes and other noncognitive variables are often measured using rating or Likert-type scales which does not come without problems. Respondents may not be motivated, or may experience fatigue effects, or have problems understanding the questions, and therefore give invalid responses in form of Response Styles (RS) that reduce validity and comparability of the measurement. RS can produce artefacts and spurious differences and are especially a problem in low-stakes assessments. A multi-process IRTree approach to detect RS in rating data (introduced by 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). The findings of these studies show that the approach can be a successful tool to control for RS bias but they also show 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 the IRTree approach with multigroup and mixture IRT models to differentiate between groups of respondents who give valid responses versus RS. Furthermore, we are examining the relation between RS and other variables (e.g., timing information) that might be helpful in identifying meaningful groups in regard to RS. The goal is to provide tools that will improve the comparability of data, meaningful scores, and unbiased measurement of group differences by correcting test scores for RS.

STAT 1b: Modeling Response Styles with Ordinal and Multidimensional IRTree Models
Thorsten Meiser, University of Mannheim; Hansjörg Plieninger, University of Mannheim; Mirka Henninger, University of Mannheim

Different IRT approaches have been proposed to account for response styles in rating data, like the general tendency towards midscale or extreme responses irrespective of item content. The families of IRT models include mixture-distribution and multidimensional IRT models, random threshold models and IRT decision trees. While mixture-distribution, multidimensional and random-threshold IRT models are based on an ordinal response process, where rating categories are conceptualized as reflecting gradual degrees of agreement with the item content, IRTree models postulate a sequence of binary decisions, where each decision reflects a qualitatively distinct aspect of the judgment process. In this talk, we focus on extended IRT decision trees that allow researchers to exploit ordinal information and to combine the strengths of theory-driven tree models with ordinal IRT models. We first show that existing IRTree models resemble ordinal models of response intensity, and we present ordinal IRTrees with multidimensional node models to capture specific response styles. Multidimensional models of node probabilities can be reparameterized as random threshold models that accommodate interindividual differences in the use of the rating response format. The family of IRTrees with ordinal and multidimensional node models is used to test the symmetry of extreme and moderate response styles across rating categories of agreement versus disagreement, to analyze the dynamics of response styles across a sequence of items, and to investigate the role of response styles as a function of item complexity.

STAT 1c: What Anchoring Vignettes Measure in Children Self-Reports?
Ricardo Primi, Universidade São Francisco

The IRT models for likert-type items assumes a set of thresholds – where in the latent scale given likert scale response is more likely than the other responses – is a property of the item and is assumed to be equal for all persons. If persons vary in responses styles this assumption may not hold. This is called person differential functioning (PDIF). Anchoring vignettes is way to learn how persons translate the latent trait into likert responses and a way to assess individual differences related to item thresholds (interaction effect of persons X items thresholds). It presents hypothetical persons differing on the attribute of interest (usually low, medium and high) and asks persons to rate those persons in the same likert scale used in self-assessment. This can then be used to resolve PDIF potentially producing measures that are more comparable. We present a study in the context of a large scale educational assessment of 31,715 children and adolescents from 500 schools of State of São Paulo Brasil with ages from 11 to 18 attending grades 6 to 12. They answered a 162-item inventory measuring 18 facet within five broad domains: E: Engaging with others, A: Amity, ER: Emotional resilience, WO Work orientation, and O: Open-mindedness and five vignettes sets on the same five domains. We investigated if the patters of responses to vignettes have a developmental trend, if they vary by domains and if they are related to cognitive capacity. We then investigated if anchor-adjusted scores produce more reliable and valid measures.

STAT 1d: Using Response Times to Deal With Missing Responses Due to Time Limits

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

In data on competence measurement, due to time limits not all test takers reach the end of the test, resulting in item nonresponse. These missing values are usually nonignorable and if not appropriately accounted for, they can result in biased parameter estimation. There are models within the missing data framework that deal with these missing values (Pimentel & Glas, 2008; Rose, von Davier, & Xu, 2010). These rely on data one can obtain from paper-and-pencil testing and simultaneously model a latent ability and a latent missing propensity. With the change to computer-based assessment (CBA), further information on the test taking behavior and, thus, on the missing process is available. In our talk, we aim at making use of response times for dealing with nonignorable missing data due to time limits. We draw on the response time model of van der Linden (2007), in which item responses and response times are simultaneously modeled. As such, a latent ability as well as a latent speed variable are estimated. We argue that the response time model of van der Linden, which has not been designed for treating missing values, does account for nonignorable missing values. Theoretically and via a simulation study, we show that the missing model has many similarities to the response time model, with the response time model depicting the missing process in much more detail. We also discuss the limitations of the response time model for treating missing values due to time limits and outline further extensions of the model.

STAT 1e: A Flexible Mixture Modeling Approach to Correct Guessing Bias

Artur Pokropek, European Commission; Lale Khorramdel, Educational Testing Service

Within low-stakes assessments like PISA (Programme for International Student Assessment), PIAAC (Programme for the International Assessment of Adult Competencies) or other educational studies, concern about test-taking motivation is extremely crucial. The results of these studies are important for decision makers, though not necessarily for examinees – as they bear no personal consequences. Low test-taking motivation can be reflected through fast responses, omitted responses, or guessing in both cognitive and non-cognitive assessments. It is assumed that during the solution behavior of motivated examinees, they take certain time by trying actively to solve an item. Low motivated examinees on the other hand might show rapid guessing taking too little time to analyze the items fully. Computer-based testing and response filtering based on timing information have attracted the attention of researchers because these methods promise notable increases in the accuracy of estimates and the ease of application (Wise & Kong 2005). While simple item filtering brings practical and methodological problems recent extensions of IRT mixture models that allow for inclusion of collateral information promise to bring more accurate solution. In this presentation we will further develop a flexible modeling approach based on the Grade of Membership (GoM) model (Erosheva, 2005) proposed by Pokropek (2015). Using PIAAC data we will show how information about different characteristics of items and respondents together with response time might help for detecting guessing behaviors, correcting the ability estimates and identify items that are particular prone to guessing.

CDM 2: 3:20 PM – 4:50 PM

CDM: CAT and Nonparametric

Chair: Gongjun Xu, University of Michigan

CDM 2a: Computerized Adaptive Testing for Cognitive Diagnosis in Classroom: A Nonparametric Approach

Yuan-Pei Chang, National Taiwan Normal University; Chia-Yi Chiu, Rutgers, The State University of New Jersey; Rung-Ching Tsai, National Taiwan Normal University

Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) has been suggested by researchers as a diagnostic tool for assessment and evaluation. While model-based CD-CAT is relatively well-researched in the context of large-scale assessments, this type of system has not received the same degree of development in small-scale settings, where it would be most useful. The main challenge is that the statistical estimation techniques successfully applied to the parametric CD-CAT require large samples to guarantee the reliable calibration of item parameters and accurate assignments of examinees. In response to the challenge, a nonparametric approach that does not require any parameter calibration, and thus can be used in small educational programs, is proposed. The proposed nonparametric CD-CAT relies on the same principle as the regular CAT algorithm, but uses a nonparametric classification method (Chiu & Douglas, 2013) to assess and update the student's ability state while the test proceeds. Based on a student's initial responses, a neighborhood of candidate proficiency classes is identified, and items not characteristic of the chosen proficiency classes are precluded from being chosen next. The response to the next item then allows for an update of the skill profile, and the set of possible proficiency classes is further narrowed. In this manner, the nonparametric CD-CAT cycles through item administration and update stages until the most likely proficiency class has been pinpointed. The simulation results show that the proposed method outperformed the compared parametric CD-CAT algorithms and the differences were more significant when the item parameter calibration was not optimal.

CDM 2b: Implications of Model Comparison on Cognitive Diagnosis Computerized Adaptive Testing
Miguel A. Sorrel, Universidad Autónoma de Madrid; Francisco J. Abad, Universidad Autónoma de Madrid; Julio Olea, Universidad Autónoma de Madrid

Paper and pencil tests based on Cognitive Diagnosis Models (CDMs) have been found to be a useful tool to provide diagnostic information about examinees' strengths and weakness on a variety of fields. A new area of application is the inclusion of adaptive testing methodologies that are based on these models. Typically, the same CDM is applied to all the items in the item bank. This is inconsistent with results of real test data indicating that no one model can be deemed appropriate for all the test items. Several indices have been proposed for the purposes of model comparison, including the Wald test. Recent item selection methods rely on the estimation of the posterior distribution of the attribute profiles. In this sense, the prior at the beginning of the test administration is set to be the posterior distribution computed in the calibration sample. A better estimation of the posterior distribution can be expected if the item bank is calibrated using a combination of different reduced models, which would lead to a better performance of the item selection methods available. Additionally, this combination of models will result in a better generalization performance of the item parameter estimates. In the present study, a simulation study is conducted to investigate these two topics within the generalized deterministic inputs, noisy, "and" gate model framework. Several factors are varied, including the generating model, the item quality, test length, and the item bank size. Implications for practical settings include an improvement in terms of ability estimation accuracy.

CDM 2c: Investigating the Constrained-Weighted Item Selection Methods for CD-CAT
Ya-Hui Su, National Chung Cheng University; Hua-Hua Chang, University of Illinois at 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 Modified Cognitive Diagnostic Discrimination Index (MCDI) and Modified Attribute-level Discrimination Index (MADI) have been proposed by Kuo, Pai, and de la Torre to assemble such tests; however, it was found that item usage from different attributes was still unbalanced for nonhierarchical structure. In addition, Zheng and Chang proposed the Posterior-Weighted Cognitive Diagnostic Discrimination Index (PWCDI) and Posterior-Weighted Attribute-level Discrimination Index (PWADI) and suggested that these indices can be integrated with constraint-weighted methods for item selection. Therefore, this study has two fold (a) integrate MCDI, MACI, PWCDI, and PWADI indices with the constraint-weighted methods for item selection, and (b) investigate the efficiency of these item selection methods in CD-CAT.

CDM 2e: A Cognitive Diagnosis Method Based on the Mahalanobis Distance
Jianhua Xiong, Jiangxi Normal University; Fen Luo, Jiangxi Normal University; Shuliang Ding, Jiangxi Normal University

Cognitive diagnosis is a popular issue in psychological and educational measurement. Cognitive Diagnostic Models (CDMs) are important for cognitive diagnosis and the primary purpose for CDM is to classify examinees into mutually exclusive categories. Researchers have proposed many new CDMs and, among them, the Generalized Distance Discrimination (GDD) and the Hamming Distance Discrimination

(HDD) have some advantages, and thus receive increasing attention. The purpose of this study is to extract the essence of GDD and HDD, put forward a generalized CDM, and to evaluate it in various conditions. The generalized CDM is called Mahalanobis Distance Discrimination (MDD). MDD uses the Mahalanobis Distance (MD) to measure the distance between an examinee's observed response pattern and an ideal response pattern, and specifies the Shannon entropy as covariance. A Monte Carlo simulation was used to compare the classification accuracy for respondents using MDD with that of HDD and GDD. The pattern match ratio and average attribute match ratio were used as criteria to evaluate the classification accuracy of different methods. Five attribute hierarchical structures are discussed (linear, convergent, divergent, unstructured and independent). The results show that the MDD performs better than HDD and GDD in terms of classification accuracy. For all mehtods, the classification accuracy is lower under unstructured and independent hierarchies. Further research should focus on the design of the Q-matrix in order to increase the classification accruacy.

IRT 4: 3:20 PM – 4:50 PM

IRT: Local Dependence
Chair: Herbert Hoijtink, University Utrecht

IRT 4a: Bayesian Elastic Constraints to Screen and Model Local Item Dependencies
Johan Braeken, University of Oslo

Detecting so-called Local Item Dependencies (LID) has diagnostic value for assessing whether the dimensionality structure of the assessment instrument is correctly specified and whether the measurement model can be safely used in measurement practice or for research purposes. The usual approaches to deal with LID are somewhat cumbersome and impractical. In this talk I will discuss and evaluate a conceptually simple simultaneous screening and modelling procedure for LID in IRT. The approach builds on recent developments in the use of Bayesian informative priors for regularization of overparameterized models.

IRT 4b: Considering Local Dependencies: Person Parameter Estimation for IRT Models of Forced-Choice Data
Safir Yousfi, Federal Employment Agency

The Thurstonian IRT model of Brown and Maydeu-Olivares was a breakthrough in estimating the structural parameters of IRT models for forced-choice data of arbitrary block size. However, local dependencies of pairwise comparisons within blocks of more than two items are only considered for item parameter estimates, but are explicitly ignored by the proposed methods of person parameter estimation. This present talk aims at supplementing the Thurstonian IRT model for forced choice data by introducing methods of person parameter estimation that adequately deal with local dependencies. The likelihood function of the response pattern is derived. This allows for consistent maximum-likelihood and Bayesian modal estimation. Numerical methods allow to determine the observed Fisher Information which results in accurate estimates of the covariance matrix of person parameter estimates. Consequences of disregarding local dependencies will be addressed analytically and by means of simulation studies.

IRT 4c: A Non-Compensatory Rasch Testlet for Items Embedded in Multiple Contexts
Hong Jiao, University of Maryland, College Park; Peida Zhan, Beijing Normal University; Yong Luo, National Center for Assessment

Recent years have seen the surge of development of innovative assessments with innovative items. Items embedded in multiple contexts are such an innovation. Answering such items often requires reading two passages in large-scale educational testing programs or requires reading of one passage and listening to an audio prompt in language tests. The use of such an item format usually intends to assess high-order cognitive skills in extracting and synthesizing information from multiple contexts. This type of items used in large-scale testing programs could pose psychometric issues and challenges as items embedded in multiple contexts are likely to be locally dependent due to the doubly contexts from each prompt. This study addresses complex local item dependence in items embedded in multiple contexts, sources or media. A new non-compensatory doubly Rasch testlet model is proposed in this study for testlet-based assessments with paired passages. A real data set from a large-scale educational testing program is analyzed with multiple models which deal with local item dependence in different ways. A simulation study is conducted to demonstrate model parameter recovery for the proposed model using the Bayesian Markov chain Monte Carlo estimation method. A simulation study will be conducted with the condition reflecting the real test structure. The recovery of the model parameters will be examined in terms of estimation errors. The impact of ignoring such complex local item dependence in multiple contexts on model parameter recovery in Rasch modeling will be investigated when compared with other competing models.

IRT 4d: Is a Test Sufficiently Unidimensional and Why 1D-IRT Will Not Work

Edward Ip, Wake Forest School of Medicine; Tyler Strachan, The University of North Carolina at Greensboro; Terry Ackerman, ACT, Inc.; Yanyan Fu, The University of North Carolina at Greensboro; John Willse, University of North Carolina at Greensboro

A traditional method to examine the unidimensionality of a test is to fit both 2D (or even higher-dimensional) and unidimensional IRT (1D-IRT) models to the data and examine goodness-of-fit indexes as well as Local Dependency (LD) patterns of the items. Chen and Thissen (1997) identified two types of LDs: surface LD, which arises from nuisance factors such as items that are similar in wording, and underlying LD, which arises because there exists a separate trait that is common to the set of LD items. For identifying underlying LD, we argue that fitting a 1D-IRT directly to the data and then examining the LD patterns is not an ideal approach. The fitted 1D-IRT would represent a composite dimension that maximizes the total variance in the multidimensional space in cases where multiple underlying dimensions besides the major dimension do exist. As a result, the magnitude of LD may be subdued as the composite dimension includes some components of the "underlying" dimension. Our solution is to apply projective IRT (PIRT, Ip & Chen, 2012) such that the projected unidimensional LD model can be used as a basis for studying true LD patterns in the data. In other words, instead of 1D-IRT, a "purified" unidimensional model would be fitted for LD assessment. We conducted simulation experiments to compare the LD patterns obtained from 1D-IRT and PIRT. We showed that directly fitting 1D-IRT to multidimensional response data could lead to obscured LD patterns. A real data set was analyzed to illustrate the method.

CMLD 1: 3:20 PM – 4:50 PM

Complex Models for Longitudinal Data

Chair: Sy-Miin Chow, The Pennsylvania State University

CMLD 1a: Gaussian Process Panel Modeling – A Unification of Longitudinal Modeling Approaches

Julian Karch, Max Planck Institute for Human Development; Andreas Brandmaier, Max Planck Institute for Human Development; Manuel Völkle, Humboldt-Universität zu Berlin

In machine learning, Bayesian Gaussian process models are known as a flexible modeling technique for time-series data. In psychology, Structural Equation Modeling (SEM) is a popular modeling technique for the analysis of longitudinal panel data. We integrate ideas from both approaches to create a novel longitudinal panel data analysis method that we call Gaussian Process Panel Modeling (GPPM). The main advantage of GPPM is its flexibility. Most popular modeling approaches for time-series and longitudinal panel data can be considered a special case of GPPM. This does not only include SEM and hierarchical linear modeling, but also state-space modeling in its time-discrete and its time-continuous variant as well as generalized additive models. GPPM allows specifying all aforementioned model families in a consistent modeling language. In addition, the generality of GPPM enables specifying novel models by either combining approaches from different traditions or by relying on the wealth of models used in Gaussian process time series modeling. As an example, we present the exponential squared model, which implements the generic assumption of smooth process trajectories. We show that it is related to the continuous-time autoregressive model as well as the generalized additive model, which has recently been proposed for the analysis of psychological time series data. In summary, GPPM is a unification as well as an extension of existing panel data modeling techniques.

CMLD 1b: Subjects and Time Points Needed for Multilevel Time Series Analysis in Mplus Version8
Mårten Schultzberg, Uppsala University; Bengt Muthén, Muthén & Muthén

Dynamic Structural Equation Modeling (DSEM) provides new methods for analyzing intensive longitudinal data such as that obtained with ecological momentary assessments, experience sampling methods and ambulatory assessments. DSEM uses two-level modeling with time on level 1 and individuals on level 2. It models intra-individual changes over time and allows the parameters of these processes to vary across individuals using random effects. This is made possible using Bayesian estimation. As an extension of conventional multilevel modeling, DSEM allows random effects to not only be dependent variables regressed on level 2 covariates but also allows them to be predictors of various outcomes. There are three key random effects of interest in psychological research of longitudinal data, random mean, random autocorrelation and random residual variance. A series of increasingly more complex models of this kind are used in a Monte Carlo study that varies several factors: Number of subjects, number of time points, variance explained in dependent variables, intra-class correlation and model misspecification. Preliminary results indicate that it is more difficult to obtain good estimates for models using the random effects as predictors. Practical considerations of Monte Carlo simulations using Bayes are also discussed.

CMLD 1c: What Do We Gain and Lose by Using a Person-Specific Modeling Approach
Siwei Liu, University of California, Davis

Person-specific time series autoregressive models and multilevel autoregressive models are two common approaches for analyzing intensive longitudinal data. However, their relative performance under different conditions is largely unknown. This study compares the two approaches in their ability to capture individual-level dynamic processes, and examines the influences of sample heterogeneity, time series length, sample size, and estimation method on their relative performance. It is found that multilevel modeling generally outperforms the person-specific approach when there is pattern homogeneity in the data (i.e., all individuals are characterized by autoregressive processes with the same order), and when the multilevel model is general enough to include the most complicated autoregressive pattern. In contrast, the person-specific approach is more desirable when there is weak homogeneity (i.e., individual vary in the order of their autoregressive processes) and when there are sufficient measurement occasions. Estimation method also has an impact on the accuracy of autoregressive modeling. These findings have important implications on model selection and research design.

CMLD 1d: Topic Modeling for Longitudinal Text Data
Seohyun Kim, The University of Georgia; Zhenqiu (Laura) Lu, The University of Georgia; Allan S. Cohen, The University of Georgia

We propose an approach to analyze texts of answers to constructed response (i.e., open-ended) items that were collected under a longitudinal study design. Analyzing texts associated with multiple time points has been studied using time series models. Blei and Lafferty (2006) developed a dynamic topic model for analyzing documents that evolve over time. Glynn (2016) proposed a dynamic linear topic model to investigate how latent profiles underlying texts change over time by incorporating one of the time series models into the dynamic topic model. The objectives of time series models are different from the objectives of models for analyzing longitudinal data. Therefore, we investigate a model for analyzing text data within the context of a longitudinal data analysis framework. We focus particularly on text data consisting of students' answers to constructed response data that were collected from two forms of an assessment across four time points. This study is based on Glynn's model, but modifies that approach to reflect the longitudinal data structure and by incorporating a latent growth curve model instead of a time series model. The focus of this model will be to describe how students' use of latent profiles changes over time. In addition, the latent growth curve model will be modified to handle the situation in which students' written answers came from two different forms of the test.

MIS 1: 3:20 PM – 4:50 PM

Missing Data
Chair: Victoria Savalei, University of British Columbia

MIS 1a: Congenial Imputation for Designed Missing Data Using Plausible Values

David Kaplan, University of Wisconsin-Madison

A recent paper by Kaplan and Su (2016) investigated the problem of matrix sampling of context questionnaire items with respect to the estimation of the plausible values of cognitive outcomes in large-scale educational assessments. Drawing on earlier work by Adams, Lietz, and Berezner (2013) and motivated by the design of PISA 2012, Kaplan and Su found that matrix sampling of context questionnaire material followed by predictive mean matching imputation could recover the known marginal distributions of the plausible values quite well; however, bias was found in the estimation of correlations between background questionnaire variables and plausible values. It is speculated that this bias is due to the fact that the plausible values were not part of the missing data model used for the imputation and hence not "congenial" in the sense of Meng (1994). This paper examines the consequences of adding plausible values to the imputation model for the context questionnaire through a detailed simulation study motivated by the questionnaire design of PISA 2012. Consistent with the notion of congeniality, our results show that including plausible values in the imputation of the context questionnaire substantially reduces bias in correlations. Furthermore, we found that congenial imputation under a partially-balanced incomplete design for the context questionnaire leads to even greater reduction in bias compared to the design of Adams et al. (2013).

MIS 1b: F-tests and Estimates for R2 for Multiple Regression in Multiply Imputed Datasets: A Cautionary Note on Earlier Findings

Joost van Ginkel, Leiden University

Multiple imputation (Rubin, 1987) has become a widely accepted method for handling missing data. The procedure estimates multiple plausible values for the missing data, resulting in several complete versions of the incomplete data set. Next, each of the imputed data sets are analyzed using standard statistical analyses, and the results are combined into one overall analysis, using specific combination rules. For F-tests in regression models testing either R2 or the change in R2, combination rules have been proposed by Rubin (1987). In the early literature on multiple imputation it was never explicitly mentioned how estimates of R2 themselves were supposed to be pooled. However, more recently Harel (2009) and Chaurasia and Harel (2014) proposed combination rules for R2, along with alternative combination rules for significance tests for R2 and the change in R2. In this presentation it is argued and shown by means of simulations that their proposed methods lead to incorrect type-I error rates whereas the originally proposed combination rules give better type-I error rates. It is therefore recommended to stick to the originally proposed rules by Rubin.

MIS 1c: Comparing Parametric and Non-Parametric Data Integration Methods

Justin M. Luningham, University of Notre Dame; Gitta H. Lubke, University of Notre Dame & Vrije Universiteit Amsterdam

Integrative Data Analysis (IDA) refers to the retrospective combination of item-level datasets from existing studies (Curran & Hussong, 2009). Analysis of the pooled item-level data can offer advantages over the meta-analysis of summary statistics, especially when the summary statistics overlook potential psychometric structure in the items to be integrated. Recently, researchers have distinguished between measurement model integration (MMI), which estimates a latent trait score from all available items, and less structured approaches such as multiple imputation of missing items with decision trees (Carrig et al., 2015). This project compares a parametric MMI approach with a proposed non-parametric IDA approach using gradient boosting machines for the imputation model (Friedman, 2001). MMI is performed using the moderated non-linear factor analysis model, which directly accounts for measurement non-invariance across studies in the factor score estimates (Bauer & Hussong, 2009; Curran et al., 2014). Boosting the item responses in the Multivariate Imputation with Chained Equations (MICE) framework extends recent data integration applications that sequentially fitted single decision trees (Burgette & Reiter, 2010; Carrig et al., 2015; van Buuren & Groothuis-Oudshoorn, 2000). The power to detect covariate effects on an integrated outcome is compared under a series of simulation conditions with one- and two-dimensional latent traits. Results indicate IDA improves power to detect effects over standard meta-analysis approaches, but the better integration approach is not uniform across conditions.

MIS 1d: Latent Variable Modelling with Non-Ignorable Item Nonresponse: Cross-National Analysis
Jouni Kuha, The London School of Economics and Political Science; Myrsini Katsikatsou, The London School of Economics and Political Science; Irini Moustaki, The London School of Economics and Political Science

When missing data are produced by a non-ignorable nonresponse mechanism, analysis of the observed data should include a model for the probabilities of responding. We consider such models for item nonresponse in survey questions which are treated as multiple-item attitude scales and analysed using latent variable models. The nonresponse models include additional latent variables (latent response propensities) which determine the response probabilities and which may be associated with the latent attitudes. Such models have previously been used most often in applications in educational research, and the response propensity has been specified as a continuous variable. We propose flexible models where the response propensity is a categorical (latent class) variable, applied to the analysis of data from general social surveys. We consider in particular cross-national surveys, where the nonresponse model may also vary across the countries. The models are applied to analyse data on welfare attitudes in 29 countries in the European Social Survey.

MIS 1e: Planned Missing Data Designs in Large-Scale Assessments
Dan Su, University of Wisconsin-Madison

Planned missing data designs in large-scale assessments can efficiently reduce respondents' burden and lower the cost associated with data collection, without cutting down on the questionnaire items. If the missing data are not appropriately planned, the parameter estimates will be biased. For a fixed sample size, the extent of bias depends on three major characteristics of design and data: the missing percentage, the overlap percentage (i.e., the portion of cases where two items are observed jointly), and the distribution of variables. The simulation studies investigate how the bias in marginal means, correlations and regression coefficients depends on the chosen planned missing data designs and the related characteristics. In large-scale assessments, the available planned missing data designs have not yet been systematically investigated with respect to the missing percentage, overlap percentage, distribution of variables, and sample sizes. The findings will guide researchers in choosing a planned missing data design with optimally arranged items in forms (i.e., with optimal missing and overlap percentages).

FMM 1: 3:20 PM – 4:50 PM

Finite Mixture Models
Chair: Chun Wang, University of Minnesota

FMM 1a: Profiles of Students on Account of Complex Problem Solving Strategies
Michela Gnaldi, Università degli Studi di Perugia; Silvia Bacci, Università degli Studi di Perugia; Samuel Greiff, University of Luxembourg; Thiemo Kunze, University of Luxembourg

In this contribution we aim at identifying profiles of testees that are homogeneous as regard to their ability in the Complex Problem Solving (CPS), which can be conceptualized in terms of a multi-dimensional latent variable made of three components, that is, rule identification, knowledge acquisition, and knowledge application. The analysis is carried out on a Finland data set related to the MicroDYN test, which is made of nine independent tasks. In each task, the testee has to dynamically interact with a computer, discovering the relations between input and output variables (which can be manipulated), draw causal models on the relationships between them, and reach target goals given the acquired knowledge in the previous steps. The ability in CPS is investigated through the estimation of a discrete two-tier item response theory model. The model at issue is characterized by two independent vectors of latent variables: the first vector denotes the three components of CPS above mentioned, whereas the second vector is introduced to account for the correlation among responses on items related to the same task. Each latent variable is assumed to have a discrete distribution with a finite number of support points, identifying homogeneous latent classes of testees, and related weights, denoting the class membership probabilities. Preliminary results show that testees may be allocated on seven latent classes and confirm the three-dimensional structure for CPS. Besides, the most frequent profile is the one made of students who score low in rule identification and knowledge acquisition, but average on knowledge application.

FMM 1b: Three-Step Bias-Correction Methods for Auxiliary Variables in Longitudinal Mixture Modeling

Ai Ye, University of North Carolina at Chapel Hill; Jeffrey Harring, University of Maryland, College Park

An important issue in mixture modeling is the nature of relations between Latent Class (LC) variables and observed auxiliary variables (Clark & Muthén, 2009). Building on early attempts to correct parameter bias of auxiliary effects in procedures that uncoupled the process of enumerating LC from covariate estimation (see, Bolck et al., 2004), new bias-corrected "three-step" approaches have recently been proposed (Vermunt, 2010, 2016) and built in modern statistical software such as Mplus (Asparouhov & Muthén, 2014). In contrast with intensive studies focused on evaluating the accuracy of covariate inclusion procedures within the contexts of Latent Class Analysis (LCA) or Growth Mixture Modeling (GMM), where one latent categorical variable is estimated, investigations under a longitudinal framework where multiple latent categorical variables are observed, are scant. Although such challenge is largely due to a software restriction - that the automation function for the "three-step" procedure in Mplus can only accommodate one latent variable. Using Monte Carlo simulation, the primary objective of this study is to extend the evaluation of the newer three-step approaches for estimating auxiliary effects to longitudinal mixture modeling with multiple waves of measurement models under various conditions, including class separation, covariate effect size, type of parameters estimated, and sample size. We first address the issue associated with the manual approach by using R statistical programming; for the evaluation of the method, we focus on parameter and standard error (SE) bias under realistic data analytic conditions found in practice.

FMM 1c: Latent Markov Modeling with Covariates in the Presence of Direct Effects

Roberto Di Mari, University of Catania; Zsuzsa Bakk, Leiden University

We evaluate the performance of the one-step and the bias-adjusted three-step (Di Mari, Oberski, & Vermunt, 2016) approaches for Latent Markov (LM) modeling with covariates in the presence of unmodeled direct effects of the covariates on the indicators of the LM model. Unmodeled direct effects violate the assumption of local independence: in the general context of latent variable modeling it is well known that unmodeled direct effects can severely bias estimates. However, the effect of unmodeled direct effect was thus far not evaluated in the context of LM modeling. At the same time, we propose two alternative approaches that might do better than the currently used methods in cases where direct effects of covariates on the indicators of the LM model are present. One is based on the three-step method, in which the residual association due to direct effects is modeled by using a second latent variable in step one. Alternatively, one can use a newly introduced two-step approach, which estimates the LM model without covariates in step one. In step two the covariates are added to the model, while keeping the measurement model fixed at the estimates obtained in step one. Using available methods for detecting any residual association between the indicators, with the two-step approach it is possible to free the fixed value restrictions on some of the step one parameters, and freely re-estimate them, including the direct effects. We evaluate the methods through an extensive simulation study, and illustrate them in a real data application.

FMM 1d: Identifying Latent Structures in Restricted Latent Class Models

Zhuoran Shang, University of Minnesota

This study focuses on a family of restricted latent structure models, where the model parameters are restricted via a latent structure matrix to reflect pre-specified assumptions on the latent attributes. In application such a latent structure matrix is often provided by experts and assumed to be correct upon construction, yet it may be subjective and mis-specified. Recognizing this problem, researchers have been developing methods to estimate the structure matrix from data. The first goal is to establish identifiability conditions that ensure the estimability of the structure matrix. The results provide theoretical justification for the existing estimation methods as well as a guideline for the related experimental designs. The second part proposes an information-based model selection method to estimate the latent structure. We consider two important cases in practice: (1) misspecification detection in confirmatory analysis where a pre-specified matrix is provided by experts; and (2) estimation of the whole latent matrix in exploratory analysis. Simulation and case studies show that the proposed method outperforms the existing approaches.

FMM 1e: A Graphical Latent Class Model for Multivariate Binary Data

Hok Kan Ling, Columbia University; Guanhua Fang, Columbia University; Jingchen Liu, Columbia University; Zhiliang Ying, Columbia University

In this paper, we propose a graphical latent class model that combines a latent class model and a graphical structure. In a latent class model, the observed variables are conditionally independent given the latent class membership. However, this could be an over-simplified model as some variables are likely to be dependent even within the same latent class. For example, in educational testing, we could have different groups of students but questions that test for the same set of skills are likely to be highly correlated within each group. As a result, we propose to include a graphical structure to the conditional probability of the observed variables given the latent class membership to capture the remaining dependence between the variables. The resulting model can be viewed as a finite mixture of Ising models with the same potential matrix. We believe that the latent class model is largely correct and so the graph should be sparse. Hence, during estimation, we achieve a sparsity structure on the graph using L1 regularization. To allow the data to inform the appropriate number of classes, we further impose a Dirichlet process prior on the model, resulting an infinite mixture of Ising models. We propose a stochastic approximation algorithm with Markov chain Monte Carlo method for the estimation of this extension model. The model is applied to several educational testing and psychiatric assessment data sets.

Symposium 6: 3:20 PM – 4:50 PM

Symposium 6: Adaptive Testing with Echo-Adapt

Chair: Wim van der Linden

Symposium 6a: Shadow-Test Approach to Adaptive Testing and Its Generalization to Other Testing Formats

Seung W. Choi, Pacific Metrics Corporation; Wim van der Linden

The shadow-test approach to adaptive testing provides a flexible mechanism to render different test formats with the same test specifications while controlling their level of adaptation. Although maximum adaptation is realized in a fully adaptive format whereby the shadow test is reassembled upon administering each item, its freezing/thawing mechanism (Van der Linden & Diao, 2014) allows for the real-time assembly of any conceivable testing format as a special case of the approach. A few common test formats with reduced levels of adaptation are: (1) Fixed-form testing - a single shadow test assembled to target a specific ability level or a distribution administered to all examinees; (2) Linear-on-the-fly (LOFT) testing - an individualized shadow test assembled to target the examinee's anticipated specific location on the ability continuum administered to the examinee in its entirety; (3) on-the-fly multistage (MST) testing - a shadow test re-assembled at only a few predetermined points during testing to be optimal at the examinee's ability update; and (4) any hybrids of the above. The mechanism is further generalized to administering test batteries or tests with multiple components based on unidimensional or higher-order IRT models, which will be discussed in this session.

Symposium 6b: Adaptive Testing with Bayesian Parameter Updating and Optimal Design

Wim van der Linden; Hao Ren, Pacific Metrics Corporation; Bingnan Jiang, Pacific Metrics Corporation

The posterior uncertainty about all model parameters in adaptive testing lends itself beautifully to an application of a combination of Bayesian parameter updating and statistical optimal design. In the application, at each step the new response would be used to update both the posterior distribution of the ability and the item parameters, whereupon the next item would be selected to guarantee maximum expected reduction of the remaining uncertainty. In a setup with embedded field-testing of new items, the application would enable us to optimize both the selection of the operational and the field-test items for each examinee. Although attractive, computational complexity has prevented its real-time application so far. However, a simple real-time implementation of an MCMC algorithm along with a straightforward shadow-test approach can do the job. The same combination offers us a natural framework for importing prior empirical information about the examinees to initialize their adaptive test or optimize the sequence of subtests in a test battery using a second-level of adaptation.

Symposium 6c: Adaptive Testing Simulation with Echo-Adapt

Michelle Barrett, Pacific Metrics Corporation

Following more theoretical considerations for adaptive testing, the session will turn to simulation of adaptive testing. Simulation is necessary for ensuring the settings selected for a specific adaptive test will result in desired characteristics for the test's purpose, such as degree of adaptation, error of measurement, bias, item and stimulus exposure. Simulation with Echo-Adapt, a cloud-hosted shadow test adaptive testing engine, will be demonstrated along with an approach to analyzing the results. Attendees are encouraged to bring personal data sets for assistance with running simulation over the course of the conference; research accounts for the simulator will be made available for personal research purposes.

MULTI 1: 3:20 PM – 4:50 PM

Multiway FA, PC, and Rotation

Chair: Marieke Timmerman

MULTI 1a: Cattell's Parallel Proportional Profiles: The Triumph of a Prodigal Rotation

Pieter M. Kroonenberg, Leiden University & The Three-Mode Company

In a primarily informal and conceptual way we trace the history of Cattell's parallel proportional principle for factor rotation, also known as confactor rotation. Both its original idea in connection with standard and confirmatory factor analysis of sets of covariance matrices from random samples is discussed as its use in the non-stochastic framework of three-mode analysis. It will be shown that, unfortunately the principle as an alternative for simple structure rotation has not led to a wide-spread use in the social and behavioural sciences, but that it has celebrated triumphs in sciences like chemistry, signal processing, and several other disciplines mostly under the flag of the Parafac/Candecomp model, tensor decomposition, and canonical polyadic decomposition.

MULTI 1b: Multiway Factor Analysis with Functional and Structural Constraints

Nathaniel E. Helwig, University of Minnesota

In many longitudinal studies, multiple variables are collected to measure some latent construct(s) of interest. In such cases, the goal is to understand temporal trends in the latent variables, as well as individual differences in the trends. Multiway factor analysis models provide a powerful framework for discovering latent trends in longitudinal data. However, classic implementations of multiway models do not take into consideration functional information (i.e., the temporal sequence of the collected data) or structural information (i.e., which variables load onto which latent factors) about the study design. In this paper, we reveal how functional and structural constraints can be imposed in multiway models (Parafac and Parafac2) in order to elucidate trends in longitudinal data. As a motivating example, we consider a longitudinal study on per capita alcohol consumption trends conducted from 1970-2013 by the U.S. National Institute on Alcohol Abuse and Alcoholism. We demonstrate how functional and structural information about the study design can be incorporated into the Parafac and Parafac2 alternating least squares algorithms to understand temporal and regional trends in three latent constructs: beer consumption, spirits consumption, and wine consumption.

MULTI 1c: Factor Rotation to a Simple Structure With Prescribed Inter-Factor Correlation

Shin-ichi Mayekawa, Tokyo Institute of Technology; Yoshinori Oki; Yume Yamamoto, Tokyo Institute of Technology; Naoto Yamashita, Osaka University

Factor analysis is one of the multivariate analysis techniques for exploring latent factors behind observed variables. In exploratory factor analysis, resarchers usually rotate factors in order to facilitate the interpretation of the factors. Especially, if an oblique factor rotation method is used, not only the factor pattern matrix, but also the factor correlation matrix will be presented as the result of roation. In this research, we developed a method that rotates the factors to a simple structure while keeping the factor correlation matrix exactly the same as the prescribed one. The proposed method can be used in the following situations: Firstly, when an existing oblique factor rotation method resulted in a simple structure with very high factor correlations, the new method can be used to reduce the amount of factor correlation while keeping the simple structure obtained. Second, researchers can prescribe inter-factor correlation matrix as prior information by the proposed method. In order to evaluate the effectiveness of the proposed method, we rotated several factor loading matrices published in academic journals by the new method by changing the factor correlations and compared the similarities between the original factor loading matrices and the new rotated factor loading matrices. Since the proposed method guarantees that the resulting factor correlation matrix is exactly the same as the prescribed one, we could successfully reduce those high factor correlations associated with the original factor rotation method while, in most of the cases, keeping the factor loading matrices as simple as the original ones.

MULTI 1d: Rotation to Optimal Bi-Factor and Multi-Factor Structures

Henk A.L. Kiers, University of Groningen; Marieke Timmerman; Eva Ceulemans, KU Leuven

Bi-factor analysis can be used to estimate loadings for an a priori defined bi-factor structure, using a confirmatory approach. However, often for a given set of items the structure is at least partly uncertain. In such cases an exploratory approach is needed. For this purpose, Jennrich and Bentler (2011, 2012) proposed to use exploratory factor analysis followed by optimizing an orthogonal or oblique bi-factor rotation criterion. Such rotations yield high loadings on one “general” factor and encourage a perfect cluster structure for the loadings on the "specific" factors. In the present paper, we propose to use constrained versions of orthogonal and oblique Simplimax (Kiers, 1994), to directly aim for low and high loadings at particular positions in the loading matrix. This yields a flexible exploratory bi- and multi-factor rotation approach. The first variant has the same aim as Jennrich and Bentler’s approach (high loadings on the general factor, and exactly one high loading per item on specific factors). Other variants will be introduced where a) cross-loadings are allowed, i.e., some items have high loadings on more than a single specific factor, yielding a tri-factor (or generally multi-factor) rotation criterion; b) more than a single "general" factor is specified; c) items within the same subscale are treated equally, i.e., aiming for high and low loadings on the same specific factors. The approach is tested by means of a simulation study and its usefulness is illustrated by means of empirical examples.

MULTI 1e: Orthonormal Polynomial Principal Component Analysis

Takashi Murakami, Chukyo University

An exploratory procedure for analyzing responses to Likert-type items was developed. Classical principal component analysis is applied to a set of quantified variables defined on each category of items by using orthonormal polynomials modified for adjusting differences of response rates. Quartimax rotation is applied to characteristic vectors with unit length. Although the procedure is completely metric, the obtained component scores coincide with individual scores yielded by standard multiple correspondence analysis except for orthogonal rotation. The entire indeterminacy of quantities given to categories is shown, and the property justifies the use of orthonormal polynomials in multiple correspondence analysis. The pattern matrix also coincides with that from independent cluster rotation by Harris & Kaiser (1964), which makes it possible to interpret high dimensional solutions of multiple correspondence analysis in the same manner as the output from principal component analysis with rotation. The procedure is viewed as an extended canonical correlation analysis as well. Quadratic and cubic components obtained by analyses of real data sets demonstrate the existence of respondents who make consistent responses using modest categories, and it suggests that sum scores of Likert-type items are not necessarily appropriate for assessment of individual differences. On the other hand, the use of sum scores for correlational studies may be acceptable since the linear composites tend to explain a large amount of variances of scores given by multiple correspondence analysis.

Invited : 5:00 PM – 5:45 PM

Statistical Quality Control in Psychometrics

Invited Speaker: Ying (Alison) Cheng
Chair: Terry Ackerman

Statistical process or quality control (SQC) in its early days means the application of statistical methods to monitor and control an individual industrial process. In modern psychometrics research, there has been abundant application of SQC, particularly in the area of educational measurement. For example, researchers have developed various CUSUM (cumulative sums) control chart procedures to detect outliers and/or person misfit in adaptive or non-adaptive testing since 1998. Very recently, a handful of studies have investigated the utility of change point analysis in psychometrics in detecting speeded responses, person misfit, or unusual fluctuations in the mean score of an assessment. In this talk I will first review briefly the studies using CUSUM procedures, then discuss how change point analysis procedures differ from the CUSUM procedures, and focus on using change point analysis to detect aberrant responses with item responses, or response time data, or both.

State: 5:00 PM – 5:45 PM

Psychometrics and Genetics

State-of-the-Art Speaker: Dorret Boomsma
Chair: Irini Moustaki

Quantitative genetic methodology, in particular genetic structural equation modeling, can assist in assessing and understanding the dimensionality of psychometric instruments as often used in psychology and psychiatry. The covariance structures that are observed among sets of items in such instruments are a function of the underlying genetic and environmental covariance structures that may be estimated from studies that include genetically informative designs. The relationship between the observed covariance structure and the underlying genetic and environmental covariance structures may be such that it hampers obtaining a clear estimate of dimensionality using standard tools for dimensionality assessment alone. One situation in which dimensionality assessment may be impeded is when genetic and environmental influences differ from each other in structure and dimensionality. In such situations settling dimensionality issues may be problematic and employing quantitative genetic modeling to uncover the (possibly different) dimensionalities of the underlying genetic and environmental structures may resolve some problems. The approach is illustrated in empirical data on childhood problems and personality in adults, where the use of twin data ensures the identification of the genetic and environmental covariance structures. A second illustration involves estimating the genetic (co)variance between personality items based on genotyped markers in samples of unrelated subjects.

Keynote: 5:50 PM – 6:20 PM

Making the Most out of Response Times: Moving Beyond Traditional Assumptions

Dissertation Prize Winner: Maria Bolsinova, Utrecht University & Cito
Chair: David Thissen

With the increasing popularity of computerised testing, in many applications of educational and cognitive testing not only the accuracy of the response is recorded, but response time as well. This additional information provides a more complex picture of the response processes. The two most important reasons to consider response times are: 1) to increase precision of the estimation of ability by using response times as collateral information; 2) to gain further insight in the underlying response processes. The hierarchical modelling framework for response times and accuracy (van der Linden,

2007) which arguably has become the standard approach for jointly modeling response time and accuracy in educational measurement provides a clear structure for studying both response times and accuracy, but is based on a set of assumptions which may not match the complex picture that arises when realistic response processes are considered. In this presentation, I will consider models that move beyond the simple structure assumption and the assumption of conditional independence between response times and accuracy and consider their added value from a statistical and substantive point of view.

Wednesday, July 19, 2017

Symposium 7: 8:30 AM – 10:00 AM

Symposium 7: iRtool: A New Set of Tools for Psychometricians
Chair: Timo Bechger, Cito

Symposium 7a: Dexter: Classical Test and Item Analysis Using the Interaction Model
Ivailo Partchev, Cito

Dexter is the entry point for analyses with iRtools as it helps users to organize their data in a data-base that can then be used by all modules and indeed for all other purposes. It provides the usual statistics to assess the quality of items and of a test as a whole. A new feature is that Haberman's interaction model is used to produce estimates of item-total regressions. The item-total regressions allow one to see the appropriateness of the items for different groups of respondents and evaluate the fit of a Rasch model. Dexter also provides a way to investigate differential item functioning based on plots of respondent profiles.

Symposium 7b: Enorm: Bayesian Analysis With the Extended Nominal Response Model
Frans Kamphuis, Cito

Enorm does Bayesian inference for extended Rasch models and can be considered a Bayesian successor of the OPLM software. We discuss the Gibbs sampler that works under the hood. This algorithm is new and an improvement of the one proposed by Maris, Bechger and San-Martin (Psychometrika 2015) in the sense that it handles complete and incomplete data in the same way and with great efficiency.

Symposium 7c: Roger: Enter Student Abilities
Timo Bechger, Cito

The extended nominal response model contains no abilities. After calibration, Roger provides a number of ability estimators as well as plausible values; i.e., draws from the posterior of ability. The talk is focused on the algorithms that produce plausible values. We discuss how they work and why they work so well. In fact they become more efficient when data sets become larger.

Symposium 7d: 3DC: Using iRtools in Standard Setting
Jesse Koops, Cito

The data driven consensus method procedure (3DC) is a proven method for standard setting that is based on the direct consensus method with added psychometric information based on data. While stand-alone software is available, iRtools can be used to support the standard-setting procedure and add some new features.

Symposium 8: 8:30 AM – 10:00 AM

Symposium 8: Psychometric Research at Educational Testing Service
Chair: Kadriye Ercikan, Educational Testing Service

Symposium 8a: Overview of Psychometric Research at ETS
Kadriye Ercikan, Educational Testing Service

As the Vice President of Statistical Analysis, Data Analysis and Psychometrics Research at Educational Testing Service, I will give an overview of research in this area. In particular, I will describe the role of psychometric, statistical and data science research in supporting operational assessment programs and in developing tools and systems for next-generation assessments.

Symposium 8b: Using Constrained Optimization to Promote Academic Excellence and Student Diversity in University Admissions
Rebecca Zwick, Educational Testing Service

In making admissions decisions, American universities typically consider not only the academic performance of the entering class, but also its diversity on a number of dimensions. What is the best way to simultaneously consider all these factors? In this presentation, I will discuss the use of constrained optimization (more specifically, zero-one programming), an operations research technique, to incorporate both academic requirements and diversity goals in college admissions. For example, the incoming class's average admissions test score could be maximized while requiring the percentage of low-income students to exceed 20%. I will include illustrations based on analyses of data from the Education Longitudinal Study (ELS:2002) and data from a professional school. I will end by giving a short overview of a group of technical studies of university admissions that are being conducted under an ETS umbrella project called, "Towards a Better Understanding of Educational Admissions Decisions."

Symposium 8c: Using Item Response Theory Models to Detect and Correct for Rater Effects in Test Scores
Jodi M. Casabianca, Educational Testing Service

Testing programs typically take an observed score approach and use an aggregate of raw ratings (human ratings, sometimes combined with a machine rating) to produce a score for constructed-response items. However, raters are subject to various performance issues stemming from rater bias, centrality, or inaccuracy, which negatively affect score distributions, correlations, and reliability. IRT models can be used in test scoring to mitigate these rater effects. In this presentation I first present a systematic review of a wide assortment of explanatory IRT rater models, which differ in terms of the types, parameterizations, and effectiveness of rater effects indices they provide. I then provide analyses of simulated datasets that demonstrate the advantages of scoring with these models relative to more traditional IRT models that do not explicitly accommodate these effects. I close my presentation with a few recommendations for operational practice and a brief overview of related research on integrated scoring approaches at ETS.

Symposium 8d: A Generalized Speed-Accuracy Response Model for Dichotomous Items
Peter van Rijn, ETS Global; Usama Ali, Educational Testing Service

The availability of process data information including response time in computer-based assessments provides an opportunity for researchers to investigate the use of such information for improving measurement. We propose a generalization of the Speed-Accuracy Response Model (SARM) introduced by Maris and van der Maas (2012). In these models, the scores that result from a scoring rule that incorporates both the speed and accuracy of item responses are modelled. Our generalization is similar to that of the one-parameter logistic (or Rasch) model to the two-parameter logistic (or Birnbaum) model in item response theory. A mixed expectation-maximization Newton-Raphson algorithm for estimating model parameters and standard errors was developed. Furthermore, methods to assess model fit are provided in the form of generalized residuals for item score functions and saddlepoint approximations to the density of the sum score. The presented methods were evaluated in a small simulation study, the results of which indicated good parameter recovery and reasonable Type I error rates for the residuals. Finally, the methods were applied to two real data sets. It was found that the two-parameter SARM showed improved fit, in both relative and absolute senses, compared to the one-parameter SARM in both data sets.

Symposium 8e: Detecting Test Fraud: Erasure Analysis, Detection of Item Preknowledge, Person-Fit Analysis and More

Sandip Sinharay, Educational Testing Service

Producers and consumers of test scores are increasingly concerned about fraudulent behavior before and during the test. Naturally, statistical procedures for detection of test fraud, or data forensics analyses, are employed by virtually all testing organizations. Data forensics analyses usually focus on one of five areas: analysis of gain scores, detection of answer-copying, detection of fraudulent erasures, detection of item pre-knowledge, and detection of person misfit. After a brief description of each of these research areas, I will discuss my recent research on three of these areas. A signed likelihood ratio test statistic will be introduced. Then I will discuss how the statistic can be applied to erasure analysis and detection of item pre-knowledge. The statistic will be shown to provide useful information from two unique data sets for which test fraud—fraudulent erasures in one and use of item pre-knowledge in the other—actually occurred. Finally, I will discuss an extension of the popular l*z statistic (Snijders, 2001) for assessment of person fit to polytomous items. I will end by briefly mentioning additional research by my colleagues on detection of test fraud.

Symposium 9: 8:30 AM – 10:00 AM

Symposium 9: ANACONDA: Analysis of Conditional and Average Causal Effects

Chair: Rolf Steyer, Friedrich Schiller University Jena

Symposium 9a: ANACONDA – Analysis of Conditional and Average Causal Effects: Introduction

Rolf Steyer, Friedrich Schiller University Jena

I summarize the core of the theory of causal total treatment effects and the data analysis associated with this theory referred to as ANACONDA. This sets the stage for the subsequent presentations on topics of methodological research that evolve for this kind of data analysis. Core concepts of the theory are true outcomes variables and their differences, the atomic causal total treatment effect variables. The expectation of the latter is the average causal total treatment effect. The crucial link between these theoretical concepts and empirically estimable conditional expectations is the concept of unbiasedness of a conditional expectation E(Y|X, Z), where Y denotes the outcome variable, X a discrete treatment variable with values 0, 1, …,i,..., n, and Z a (qualitative or quantitative, uni- or multidimensional, manifest or latent) covariate. I present some causality conditions, that is, sufficient conditions implying unbiasedness of E(Y|X, Z). Under unbiasedness of E(Y|X, Z), the differences E(Y|X = i, Z) - E(Y|X = 0, Z) are the Z-conditional total treatment effect functions, the expectations E[E(Y|X = i, Z)] are the Z-adjusted expectations of the outcome variable Y in treatment i, and the difference E[E(Y|X = i, Z)] - E[E(Y|X = 0, Z)] is the average total treatment effect.

Symposium 9b: Standard Errors of Adjusted Probabilities of Success in Logistic Regression

Andreas Neudecker, Friedrich Schiller University Jena; Gregor Kappler, Friedrich Schiller University Jena; Rolf Steyer, Friedrich Schiller University Jena

In a study with a dichotomous outcome (success/failure, sick/cured) we are interested in the relative frequency of patients sick and cured under different treatment conditions. In terms of theory, this directly translates to probabilities conditioning on the treatment variable and possibly other covariates Z. Moreover, Z-adjusted probabilities of success for a given treatment condition (APS) are defined as expectation of these probabilities. In practice we estimate logistic regression coefficients. Using these regression coefficients it is easy to estimate the conditional probabilities mentioned above - and this is what logistic regression is good for. What we do not get are parameters quantifying the sampling error of the Z-adjusted probabilities in the treatment conditions. This is why we focus in our presentation on standard errors on the level of probabilities for APS. We illustrate three methods to estimate these standard errors: For fixed (non-stochastic) regressors we can apply the delta method to the variance-covariance matrix of the estimated logistic regression coefficients. To incorporate stochastic covariates we use a Taylor expansion of the inverse logit. Finally, we show how to use bootstrapping in this setting. In a simulation study we compare the performance of these methods. We also discuss the limitations of standard errors for bounded parameters such as probabilities.

Symposium 9c: ANACONDA with a Latent Outcome Variable Measured by Binary Variables

Jan Plötner, Friedrich Schiller University Jena; Rolf Steyer, Friedrich Schiller University Jena

Evaluating treatments is an important goal in applied social sciences. Often we are not only interested in the Average Treatment Effect (ATE), but also in Conditional Treatment Effects (CTEs) as well, which inform us about the expected treatment effects of a person with observed values on one or more (qualitative or quantitative) covariates. In many cases, the dependent variable is a latent variable that can be measured only by binary variables. In such a case, it is possible to estimate ATEs and CTEs with multi-group Structural Equation Modeling (SEM), which also allows for interactions between treatment and (latent) covariates. We present such a model for evaluating a training program for inductive reasoning. In this empirical study, 279 first-grade pupils either participated in a training to improve inductive reasoning or continued normal classroom activities. In the study, the children's fluid intelligence was measured by Raven's Colored Progressive Matrices before the training and after the training in a 6-month follow-up. We discuss model specification, assumptions and the results of this empirical study.

Symposium 9d: Latent Covariates for Identifying the Average Treatment Effect

Marie-Ann Sengewald, Friedrich Schiller University Jena; Steffi Pohl, Freie Universität Berlin; Peter M. Steiner, University of Wisconsin-Madison; Rolf Steyer, Friedrich Schiller University Jena

Covariate adjusted treatment effects are commonly estimated in non-randomized studies either with generalized ANCOVA or propensity score methods. Based on the stochastic theory of causal effects, we present conditions under which the Average Treatment Effect (ATE) can be identified by adjusting for manifest covariates and conditions that require adjusting for latent covariates. We use empirical data of a within-study comparison to evaluate the bias in ATE estimates when we adjust for a manifest or the respective latent covariate (a) in the single covariate case, and (b) with additional covariates. In the empirical study, 202 students were randomly assigned to a randomized and a non-randomized condition. These conditions vary in the used assignment strategy to establish participation in an English or a mathematics training (i.e., random assignment vs. self-selection). In the non-randomized condition, we use a generalized ANCOVA to estimate the ATE of the English training. We model the English pretest ability either as a manifest or as a latent covariate and use additional covariates that are relevant or non-relevant for the bias in ATE estimates. We compare the ATE estimates to the respective ATE estimate in the randomized condition. Based on this comparison, we show that adjusting for the latent covariate reduces more bias than adjusting for the manifest covariate. Adjusting for additional relevant covariates partly compensates the bias of adjusting for the manifest covariate. Limitations of the study as well as implications for applications are discussed.

EST 2: 8:30 AM – 10:00 AM

Estimation: Categorical-Response LV Models

Chair: Myrsini Katsikatsou, The London School of Economics and Political Science

EST 2a: On the Use of Pairwise Likelihood Estimation for Clustered Data

Mariska Barendse, Ghent University; Yves Rosseel, Ghent University

Social and behavioural research frequently involves multilevel data, with individuals and groups defined at separate levels. Multilevel analysis then often leads to the use of models with a large number of (latent) variables (i.e., random slopes, random intercepts, and hypothetical constructs) at different levels. The analysis of categorical multilevel data often requires the evaluation of high-dimensional integrals. Current full-information approaches typically involve computationally intensive numerical methods (e.g., adaptive Gauss-Hermite quadrature). Alternatively, in the Pairwise Likelihood (PL) approach, the full likelihood is replaced by a sum of (bivariate) pairwise likelihoods, which are easier to handle. PL estimation has already been proven to be quite successful in single level datasets with a small number of categorical variables. In this presentation, we will explore various possibilities of PL estimation for clustered data. We will use PL both as a general (i.e., a single algorithm) and a local solution (e.g., only for the latent part of the model) to avoid computationally intensive numerical methods. We will present several applications and show how the freely available R package "lavaan" can be used for estimation.

Finally, a small simulation study with a focus on bias, efficiency, and computational time will be presented.

EST 2b: Pairwise Likelihood Estimation Based on a Sample of Pairs

Irini Moustaki, The London School of Economics and Political Science; Ioulia Papageorgiou, Athens University of Economics and Business

Pairwise likelihood estimation has been recently developed for estimating the parameters of latent variable and structural equation models. Pairwise likelihood is a special case of composite likelihood methods that use lower order conditional or marginal log likelihoods. The composite likelihood to be maximised is a weighted sum of marginal or conditional loglikelihoods. Weights can be chosen to be equal or unequal for increasing efficiency. In this paper, we approach the problem of weights from a sampling perspective. More specifically, we propose a sampling method for selecting pairs that is based on those pairs that contribute more to the total variance from all pairs. We demonstrate the performance of our methodology using simulated examples.

EST 2c: Approximate Maximum Likelihood Estimation for IRT Models with Dichotomous Items

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

Parameter estimation in Item Response Theory (IRT) by Marginal Maximum Likelihood (MML) integrating out the latent variables, tends to be computational intensive for larger latent dimensions. Under certain conditions, item response models for dichotomous items can be represented as association models (Holland, 1990; Anderson & Vermunt, 2000). Such a representation reduces the complexity in the latent variables, but for large number of items, the Maximum Likelihood Estimation (MLE) of the associated model parameters is also not feasible in practice. Anderson, Li, and Vermunt (2007) and Paek and Anderson (2016) proposed a pseudo likelihood approach for association models, leading to fast calculated estimates, which are close to standard MLEs. Here, we propose an alternative approach that applies directly on the IRT model. This new algorithm is evaluated in terms of computational expenditure and estimation quality via a simulation study.

EST 2d: Estimation of Latent Regression IRT Models Using the Laplace Approximation

Björn Andersson, Beijing Normal University

The estimation of high-dimensional latent regression IRT models in large scale assessment programs is made difficult by the need to approximate integrals required to calculate the likelihood function. Proposed solutions in the literature include using Monte-Carlo approximations and a two-step procedure using a Laplace approximation. We propose using a second-order Laplace approximation of the likelihood to estimate IRT models with ordinal observed variables and fixed covariates where all parameters are estimated simultaneously. The method applies when the IRT model has a simple structure, meaning that each observed variable loads on only one latent variable. By estimating all parameters simultaneously, it is straight-forward to incorporate the variability associated with the parameter estimation when generating the plausible values which are used in many large scale assessment programs. We compare the performance of the proposed method to a two-step method currently used and illustrate the utility of the method using data from a large scale assessment program in China.

EST 2e: Latent Variable Models for the Analysis of Cognitive Functioning over Time

Silvia Cagnone, University of Bologna; Silvia Bianconcini, University of Bologna

Dimensions of cognitive functioning are potentially important but often neglected determinants of the central economic outcomes that shape overall well-being over the life course. In this regard, the Health and Retirement Study and the Asset and Health Dynamic study (HRS/AHEAD) aims to examine the impact of cognitive performance and decline on key domains of interest (e.g., health and daily functioning, retirement, economic and health decision making, use of economic and social resources). In this work, the analysis of the HRS/AHEAD cognitive data is performed using latent variable models, that easily allow to determine common factors of the cognitive items. The estimation of these models is cumbersome when the observed cognitive items are of different nature, either continuous or discrete as in the HRS/AHEAD study. Indeed, problems related to the integration of the likelihood function arise since analytical solutions do not exist. This problem is more evident in presence of longitudinal data when the number of latent variables increases proportionally to the number of observed items. In this study, we analyse the performance of a new integration method, known as Dimension Reduction Method (DRM), in the estimation of the latent individual cognitive status over time of the HRS/AHEAD data. It provides parameter estimates as accurate as techniques commonly applied in the literature, but without sharing the same computational complexity of the latter. We show that it can be applied in common situations in which standard techniques are unfeasible.

IRT 5: 8:30 AM – 10:00 AM

IRT: Forced Choice and Faking

Chair: Safir Yousfi, Federal Employment Agency

IRT 5a: A Forced-Choice Model for Enhancing Interpersonal Comparisons

Jyun-Hong Chen, National Sun Yat-sen University; Hsiu-Yi Chao, National Chung Cheng University; Chi-Chen Chen, National Sun Yat-sen University; Ching-Lin Shih, National Sun Yat-sen University

To deal with the response bias (e.g., faking) caused by Likert-type format, the forced-choice format is proposed as one of the alternatives. The forced-choice format has been widely applied to reduce the effect of item responses involving faking and intentional response distortion. However, tests with forced-choice items encounter a critical problem: Test scores from forced-choice questionnaire are ipsative and are problematic for interpersonal comparisons. The main reason for this problem might be similar to the one that solves the simultaneous equations with a system of n equations in n+1 unknowns. Since the system is underdetermined, there is an infinitude of solutions instead of one solution. To overcome the problem, the study proposed a Within-item Multidimensional Model (WMM) for forced-choice tests. In contrast to the existing models (e.g., Thurstonian IRT model; Brown & Maydeu-Olivares, 2011), WMM employs statements with within-item multidimensionality in addition to unidimensional statements for constructing forced-choice items. Items that include at least one statement with within-item multidimensionality can efficiently solve the underdetermined problem suffered by the forced-choice format. According to the simulation study, results indicated that WMM can produce validly interpersonal comparisons, no matter what condition is. For test practitioners, WMM can extend the usage of forced-choice format in more advanced ways, such as to serve as the criteria for personnel selection.

IRT 5b: An Optimization Procedure for Assembling Multidimensional Forced-Choice Blocks

Rodrigo Schames Kreitchmann, Universidad Autónoma de Madrid; Daniel Morillo, Universidad Autónoma de Madrid; Vicente Ponsoda, Universidad Autónoma de Madrid; Iwin Leenen, Instituto Nacional para la Evaluación de la Educación

Although Forced-Choice Questionnaires (FCQ) are well-known to have ipsativity issues under classical test theory, the use of recent Item Response Theory methods can offer non-ipsative, normative scores. However, some features of the blocks may affect identification and compromise the accuracy of the latent trait estimates. Yousfi and Brown (2014) and Morillo (2015) define some optimization criteria for block assembly that can reduce trait indeterminacy and enhance the quality of trait estimates. The former authors propose an optimization method based on linear binary programming. However, their approach (1) assumes a linear relation between these criteria and the decision variables, and (2) implies a decision space that makes the algorithm computationally complex as the item pool size increases. This study presents a new approach for block assembly using a Non-dominant Sorting Genetic Algorithm (NSGA-II) variant for permutations, which specifies the constraints through the codification of the decision variables and allows for the optimization of nonlinear criteria. A simulation study illustrates the benefits of this approach on the recovery of person parameters in three dimensions. The MUPP-2PL model (Morillo et al., 2016) was assumed and four optimization criteria were compared. Random block assignment was included as a control condition. The following factors were also manipulated: (1) percentage of the item pool included in the FCQ (50% and 100%), (2) questionnaire length (18 and 36 blocks), (3) correlation between the latent dimensions (0, .25, and .50), and (4) proportion of opposite-polarity blocks (0, 1/3 and 2/3).

IRT 5c: Modelling Faking Behaviour in High-Stakes Personality Assessments
Anna Brown, University of Kent

Impression management (also known as faking) on self-report questionnaires is a threat to validity and fairness of psychological assessments, which has been driving the quest for "faking-resistant" designs, for example forcing choice between equally desirable questionnaire items. Truly effective methods, however, cannot be developed unless response behaviours in high stakes are understood. This research aimed to propose a model incorporating cognitions typical to high-stakes assessments into responses collected with single-stimulus and force-choice questionnaires. Mixed methods were used, beginning with a qualitative study comprising 10 one-to-one interviews, in which participants reflected on their responses to single-stimulus and forced-choice questions in a job application setting. Having identified common decision points and actions from the interviews and published research (e.g. Robie, Brown & Beaty, 2007), response models incorporating faking behaviour into single-stimulus and forced-choice questions were proposed. These involve the "retrieve" and "edit" stages as in Böckenholt's (2014) model for sensitive survey questions, but in selecting the final response, job requirements are the new defining factor. To validate the proposed models, a series of empirical studies tested five samples of archival data collected in low, medium and high stakes. Simulation studies are underway to confirm the conditions in which the main effects can be identified reliably. The initial results are encouraging in that the proposed models seem to capture the prominent features of faking behaviour; however, my discussion will discourage continuing the quest for methods of detection and correction of biases, but investing effort in the development of better assessments methods.

IRT 5d: Application of Thurstonian IRT Model in the Development of Fake-Resistant Forced-Choice Questionnaires
Luning Sun, University of Cambridge; Fang Luo, Beijing Normal University; Hongyun Liu, Beijing Normal University

Self-report questionnaires are widely used in occupational and educational settings. In order to effectively reduce the response bias and faking behaviour, forced-choice questions have been frequently adopted. However, this format results in ipsative data, which is incompatible with traditional scoring methods. Recently, a Thurstonian Item Response Theory (IRT) model has been proposed as a solution to this problem. The current research innovated the development of fake-resistant forced-choice questionnaires by incorporating the social desirability ratings of the items and applied the Thurstonian IRT model to the ipsative data. A simulation study and an empirical study were conducted, in order to investigate the effects of a range of factors on the estimation accuracy of the latent attributes. The results showed that the Thurstonian IRT model exhibited considerable strengths in analysing response data of fake-resistant forced-choice questionnaires, particularly when a large number of items were available to construct multi-dimensional item pairs. Additionally, a newly designed forced-choice questionnaire of neuroticism was administered in a high-stake situation. The attribute scores estimated by the Thurstonian IRT model showed no correlation with the social desirability scores, indicating remarkable resistance to faking behaviour. In the discussion, further suggestions were provided to guide the application of the Thurstonian IRT model in the development of fake-resistant forced-choice questionnaires in future research and practice.

REL 1: 8:30 AM – 10:00 AM

Reliability I
Chair: Klaas Sijtsma, Tilburg University

REL 1a: Information Gain and Informational Correlation as Measures of Reliability
Matthew Johnson

In the classical true score model there are a number of equivalent ways to define the reliability of a test including: (1) the correlation between parallel forms of a test; (2) the squared correlations between the observed and true scores; and (3) the ratio of the true score variance to the observed score variance. Once we move beyond the classic true score model to more latent variables models (e.g., factor analysis, IRT) the equivalencies no longer hold. The classical definitions of reliability are especially difficult to apply to multivariate models, or to models with discrete latent variables (e.g. latent class models or diagnostic classification models). In this presentation I will review the work from the field of information theory by Linfoot (1957) and Kent (1983) who develop measures of association based on the idea of information gain, and discuss how these informational correlation coefficients can be used to report the reliability of an estimator of a latent variable in a wide-variety of psychometric models. The properties of these new informational reliability measures will be examined with a Monte Carlo experiment, and the utility of the measures will be demonstrated with a real data example.

REL 1b: Planning Inter-Rater Reliability Studies

Andries van der Ark, University of Amsterdam; Terrence Jorgensen, University of Amsterdam; Hannah Rós Sigurðardóttir, University of Amsterdam

We developed a stepwise data-collection procedure to obtain precise Inter-Rater Reliability (IRR) estimates when a single rating (i.e., one rater providing a score for one object) is expensive. The study was motivated by a request to estimate the IRR of various outcomes of an expensive assessment procedure where an officer from Child Protection Services assesses the recidivism risks of a juvenile delinquent. Given a predetermined minimum level of precision of the IRR estimate, the procedure seeks to minimize the number of ratings. It is assumed that the number of raters, the number of objects, team size (i.e., number of raters per object), and workload (i.e., number of objects per rater) are the only relevant factors. In a preliminary study, we showed that the number of ratings equals both the product of team size and number of objects and the product of workload and number of raters. We also showed that given a fixed number of ratings, the ratio "team size–number of objects" has a positive effect on precision, whereas other factors showed no discernable effect. The results of the preliminary study were used in the stepwise data-collection procedure: In each step, a relatively large team of raters rate a relatively small number of objects. With each step, more ratings become available, and the precision of the IRR estimate increases. The procedure stops if all IRR estimates have the required precision. The procedure seems rather robust to missing ratings, and therefore attractive for practical purposes.

REL 1c: Measurement versus Prediction in Test Construction

Niels Smits, University of Amsterdam; Judith M. Conijn, University of Amsterdam; Andries van der Ark, University of Amsterdam

Two important test goals are (i) measurement: questionnaires are constructed to assign numerical values that accurately represent the test taker's attribute, and (ii) prediction: the questionnaire is constructed to give an accurate forecast of an external criterion. Construction methods aimed at measurement prescribe that items should be reliable, leading to questionnaires with high inter-item correlations. By contrast, construction methods aimed at prediction prescribe that, in addition to a high correlation with the criterion, items should have low inter-item correlations. The latter approach has often been said to produce a paradox concerning the relation between reliability and validity (Gulliksen, 1950; Lord & Novick 1969, McDonald, 1999), because a well-known result from classical test theory states that the validity of a test with respect to any criterion is bounded above by its reliability, by which it may appear more appropriate to use items with high rather than low inter-item correlations. The first goal of the present paper is to provide insight into this seemingly paradoxical situation and the second goal is to show that there is a trade-off between the measurement and prediction qualities of a questionnaire. To that end, both a theoretical exposition and an empirical illustration are presented. We claim that a trade-off will exist for any pair of test goals, and the implications for the construction of questionnaires are discussed.

REL 1d: Reliabilities of Intraindividual Variability Indicators with Autocorrelated Longitudinal Data

Han Du, University of Notre Dame; Lijuan Wang, University of Notre Dame

Researchers have found that individual differences in intraindividual variability exist and can be related to important individual outcomes. To quantify intraindividual variability, different indicators have been used. Some indicators measure the amplitude of intraindividual fluctuations, such as the Intraindividual Standard Deviation (ISD) and the intraindividual variance (ISD^2), some estimate the temporal dependency (i.e., how the current and future states are correlated with the prior ones), such as the estimated hth-order autocorrelation coefficient (ρ □(h)), whereas some others measure a combination of both, such as the Mean Square Successive Difference (MSSD). Studying reliabilities of intraindividual variability indicators is helpful for accurately evaluating effects of intraindividual variability and designing longitudinal studies. We consider a multilevel time series model, where each individual's true scores have an AR(1) process. We obtained the reliability estimates of the indicators through simulations. The simulation results showed that all the indicators are generally more reliable with more reliable measurement scales, more assessments, and lower average autocorrelation. The reliabilities of ρ □(1) were generally lower than those of ISD^2 and ISD, reliabilities of MSSD were usually between the reliability of ISD^2 or ISD and that of ρ □(1), whereas reliabilities of y¯were generally the highest. This study is the first that has evaluated reliabilities of ρ □(1) and MSSD. The results showed that we need to be particularly cautious when using ρ □(1), ISD^2, ISD, and MSSD as predictors in regressions, especially when the number of assessments is not large enough.

REL 1e: On the Fallacies About and the Usefulness of Change Scores

Zhengguo Gu, Tilburg University; Wilco H.M. Emons, Tilburg University; Klaas Sijtsma, Tilburg University

Change scores obtained in pretest-posttest designs in psychotherapy are important for evaluating treatment effectiveness in clinical research and for assessing change of separate individuals in clinical practice. However, over the years the use of change scores has raised much controversy. In this presentation, from a multilevel perspective, we provide a structured treatise on several persistent fallacies about change scores in the literature and show that these fallacies were promoted by the confounding of the effects of within-person change on change-score reliability and between-person change differences. We argue that psychometric properties, such as reliability and measurement precision, of change scores should be placed at suitable levels within a multilevel framework. We show that, if examined at proper levels with such a framework, fallacies about change scores can be broken down convincingly. Once the fallacies have been clarified, we discuss an alternative, relatively unknown method for estimating change score reliability and compare it to the conventional method often seen in literature. We show that the estimated change score reliability by means of this alternative method is very close to the true reliability (hence, low bias) and is robust against correlated measurement errors in pretest and posttest, whereas the conventional method greatly underestimates true reliability especially when correlated measurement errors exist.

EFA 1: 8:30 AM – 10:00 AM

EFA & Classification and Regression Trees

Chair: Henk Kiers

EFA 1a: Matrix Results for Elucidating the Essence of Factor Analysis

Kohei Adachi, Osaka University; Nickolay Trendafilov, The Open University

Factor Analysis (FA) can be formulated as a matrix factorization problem, in which all FA model parameters (common and unique factors, loadings, and unique variances) are treated as fixed unknown matrices. For obtaining them, the discrepancy between a data matrix and the FA model is minimized with specific matrix factorization. The properties of the resulting parameter matrices are investigated in this study. Major results are summarized as follows: [1] the model part and residuals can be identified, though the matrix of common and unique factors is undetermined since it has higher rank than the data matrix; [2] common factors are not correlated with residuals, while the unique factors vary with the residuals; [3] the covariance determines the amount of residuals; [4] for a certain data set, FA always fits better than Principal Component Analysis (PCA), while it tends to give larger squared loadings than FA. Those theoretical results are illustrated using a real data example.

EFA 1b: Assessing Dimensionality of Bi-factor and Second-Order Models With Parallel Analysis

María Dolores Nieto, Universidad Autónoma de Madrid; Francisco J. Abad, Universidad Autónoma de Madrid; Luis E. Garrido, Universidad Iberoamericana en Santo Domingo; Vicente Ponsoda, Universidad Autónoma de Madrid

Horn's Parallel Analysis (PA) has consistently shown to be a superior method for the assessment of latent dimensionality, a critical phase in the validation of measurements and the development of theory. Even though PA has been extensively evaluated across first-order structures, there is no previous evidence about its performance with second-order and bifactor models. Due to the recent interest in distinguishing between these two models in applied settings, it becomes relevant to understand the performance of PA in such scenarios. Thereby, the current study assessed the accuracy of PA with second-order models and a range of increasingly complex bifactor structures (simple, with cross-loadings in the group factors, with pure indicators of the general factor, and mixed structures with cross-loadings and pure indicators). Six variables were manipulated using Monte Carlo methods: Type of structure, sample size, number of group factors, number of variables per group factor, loading size on the group factors, and loading size on the general/second-order factor. The results indicated that PA generally tends to underfactor with hierarchical and bifactor latent structures. In this regard, with strong group factors, PA generally yielded solutions with as many factors as group factors present in the hierarchical/bifactor structure. Paradoxically, with mixed bifactor structures and a weak general factor (i.e., with low loadings), PA often suggested the extraction of one more dimension. The performance of PA is discussed in light of the complexity of the different structures and practical guidelines are offered.

EFA 1c: Modified Parallel Analysis for Data with Planned Missingness

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

Parallel analysis is often suggested as an effective and accurate heuristic for determining the number of factors to retain in exploratory factor analysis. However, only recently have researchers begun considering the limitations of the procedure in the presence of missing data. A commonly cited paper implementing parallel analysis in statistical software (O'Connor, 2000) suggests handling missing data by listwise deletion prior to utilizing parallel analysis. Liu and Rijmen (2008) offer a modification to parallel analysis that is able to handle small amounts of missing data and provides an alternative to listwise deletion. However, in practice this method is limited in handling only a small proportion of missing data – utilizing the method with large proportions of missing data results in the procedure suggesting a model with an unreasonably large number of factors. This is caused by neglecting to account for additional uncertainty in the randomly generated correlation matrices due to the missing data. We propose a modification to parallel analysis that accounts for large proportions of missing data by generating random data matrices with the same missing pattern as the original data set. The modified procedure applies to both continuous and ordinal data. The current study compares and contrasts the performance of existing methods with the proposed procedure via an empirical simulation study, offers a motivating example for the use of the new method in the context of planned missingness, and discusses the practical implication of the results.

EFA 1d: Variable Importance Measures from Random Forests and Bayesian Model Averaging

Chansoon Lee, University of Wisconsin-Madison

Classification and regression problems where the number of variables is relatively large to the sample size are becoming increasingly common in behavioral sciences. Variable Importance Measures (VIMs) from random forests (Liaw & Wiener, 2002) are used to efficiently reduce the predictor space while preserving predictive accuracy in genome-wide association studies. The VIMs, however, do not offer a threshold, which requires users to figure out how large the VIMs need to be for the variables to be important. Moreover, tree models have not been compared to Bayesian Model Averaging (BMA; Hoeting et al., 1999), which is known to outperform any single model in prediction. Existing BMA algorithms, however, are not applicable to data where the number of variables exceeds the sample size. The purpose of this research is to propose a Cross-Validated permutation (CV-permutation) threshold for random forest VIMs and a hybrid approach of random forests and BMA (RF-BMA), in order to investigate variable selection performance of VIMs from random forests and BMA with predictive accuracy. The CV-permutation threshold approach finds an appropriate threshold for variables using permutation of the outcome variable and k-fold cross validation. VIMs based on the proposed approaches will be compared to other tree VIMs using simulation studies and real data. VIMs from random forests and RF-BMA can be attractive alternatives to variable selection to traditional approach, when the number of variables is large relative to the sample size.

EFA 1e: Assessing the Stability of Trees and Other Statistical Learning Results

Michel Philipp, University of Zurich; Thomas Rusch, Vienna University of Economics and Business; Kurt Hornik, Vienna University of Economics and Business; Achim Zeileis, University of Innsbruck; Carolin Strobl, University of Zurich

Classification and regression trees as well as model-based recursive partitioning methods, such as Rasch trees and mixed model trees, have become quite popular in many scientific fields, including psychometrics. Their popularity is largely due to the straightforward interpretability of their graphical representation. However, trees are also known to be instable and provide no means of statistical inference to judge whether the results generalize to other samples. Ensemble methods, like random forests, are more stable but lack the interpretability of a single tree. Therefore, from a user's perspective, the question is: When is it OK to interpret a single tree and when should it be considered with caution? In a first attempt to address this question, we introduce and illustrate an easy-to-use toolbox of summary statistics and plots for assessing the stability of trees. Furthermore, we will outline a framework for measuring the stability of supervised statistical learning results in general. All presented methods will be freely available in the R package stablelearner.

Keynote: 10:20 AM – 11:20 AM

Keynote Speaker- Disentangling Active Treatment Effects from Placebo Effects in Randomized Double-blind Trials

Keynote Speaker: Don Rubin
Chair: Sophia Rabe-Hesketh

When approving pharmaceutical drugs (for example by the US FDA or the EU EMA) for general sale, it is common to rely on double-blind randomized trials, where the active drug is compared to an inactive placebo. The logic is that the seller of the drug should not get the drug approved and make money selling it if the drug has no effect beyond that of an inactive placebo. This position is predicated on an ethical argument, suggesting that even if the originator of the drug is the first to market some clever idea that capitalizes on a placebo effect, the originator should not profit from that idea. Despite this position, once a drug is approved and patients get their doctors’ prescriptions filled for this new drug (which patients have been told has been shown to work in patients with similar conditions), some patients taking the drug will experience the actual treatment effect of the active drug as well as the placebo effect.   Disentangling these effects, the medical treatment effect from the psychological effect of being told the drug should help them, is a difficult but important practical problem, similar in some ways to the problem of estimating the causal effect of an active drug when there is non-compliance with assignment to take or not to take the drug. There is now a substantial literature on the problem of non-compliance, a substantial thread following Angrist, Imbens & Rubin (1996, J American Statistical Association), which showed how the econometric idea of “instrumental variables (IV)” could be applied to address the non-compliance problem using the “Rubin Causal Model” to formulate causal inference problems in terms of “potential outcomes”. A generalization of that approach, “Principal Stratification" (Frangakis & Rubin, 2002, Biometrics), can be used to disentangle treatment effects from placebo effects under explicitly stated assumptions. Some basic statistical theory will be presented and then applied to some real data where placebo effects are expected to be substantial, at least for some subset of patients.

Thursday, July 20, 2017

Symposium 10 : 8:30 AM – 10:00 AM

Symposium 10: Intensive Longitudinal Data: Past, Present, and Future

Chair: Ellen Hamaker, Utrecht University; Ted Walls

Symposium 10a: A Brief History of Dynamic Modeling in Psychology
Ellen Hamaker, Utrecht University

In this presentation I provide a brief overview of the dynamic modeling developments since the first publication of Cattell's P-technique in 1947. I will discuss the contributions of pioneers in this area, including Peter Molenaar, John Nesselroade, Jack McArdle, and Michael Browne. Since then, the invention of technological devices like smartphones has proven a demarcation point, as it has led to a new class of easy to use data collection techniques to obtain intensive longitudinal data. I will discuss some promising modeling approaches for intensive longitudinal data, along with the kind of research questions that they can help tackle. I will end by discussing how these developments have accumulated in the construction of Dynamic Structural Equation Modeling (DSEM), which is now implemented in Mplus version 8. This extensive modeling toolkit allows for single-subject time series analysis, as well as multiple-subject, multilevel extensions thereof.

Symposium 10b: Time Stops for No One. A Continuous Time Perspective on Dynamic Modeling
Manuel Völkle, Humboldt-Universität zu Berlin

The goal of this presentation is to introduce continuous time dynamic modeling and to review recent developments in theory, applications, and software. Dynamic models for the analysis of longitudinal data become increasingly popular in modern psychometrics. Most modeling approaches, however, treat time as a discrete variable, with discrete time vector autoregressive cross lagged models being a prototypical example. Especially in the case of panel data with many individuals being observed at many different time points, the use of discrete time models can be problematic. Based on stochastic differential equations, continuous time models resolve this problem, provide optimal parameter estimates even for highly complex patterns of measurement occasions, and offer additional information not provided by discrete time models. In addition, they may permit insights into dynamic processes at different time scales that are difficult to obtain using conventional discrete time approaches. After a basic introduction to the idea of continuous time modeling, I will reconsider the design and analysis of longitudinal studies from a continuous time perspective. Special emphasis will be put on the handling of time, missing values, and the analysis of panel data in the presence of unobserved heterogeneity (N large, T small), as well as time series analyses (T large, N small). I will briefly introduce ctsem, an R package for continuous time modeling and will point to recent developments in hierarchical modeling as well as higher order continuous time models. I will end with discussing the advantages and limitations of continuous time modeling for applied psychological research.

Symposium 10c: Dynamical Models for Intensive Longitudinal Data
Francis Tuerlinckx, KU Leuven

In this talk, I want to give a general overview of dynamical modeling approaches for (intensive) longitudinal data. It will be discussed how these models relate to more traditional psychometric models. In addition, I will give a personal account of what I believe are the most important future prospects and challenges for the domain of dynamical modeling.

Symposium 10d: New Directions in Intensive Longitudinal Data Analysis
Theodore A. Walls, University of Rhode Island

When Binet set out to classify children by intelligence through testing, he was responding to the demands of French society to determine how to place schoolchildren into classrooms. Hence, early psychometric origins were driven by nomothetic demands, to match people's abilities and put them by groupings into settings. What if Binet had had access to volumes of psychosocial, contextual and health data on every child and took the view that cognitive or other outcomes were latent, malleable outcomes arising from dynamic multivariate trajectories? The tension between the need to match resources to affect optimal outcomes remains as strong today, however our data and inferential tools are growing much more robust. The future of modeling intensive longitudinal data, even as these data become available in real-time with individuals interacting within them, will lead us to revisit fundamental questions of research design, privacy, ethics of educational and economic opportunity, and the ubiquity of numerical and stochastic models as social tools. Models currently emerging in the arena of intensive longitudinal modeling, including the dynamical models presented in this symposium, time-scale dependent longitudinal design, functional models, control systems models, non-linear models of self-organization, machine learning techniques, and "assemblies" of models in which data are passed from one framework to another, increasingly drive the innovation of collaborative team science in the ILD community. This talk will characterize trends towards future intensive longitudinal data forms and modeling frameworks.

CDM 3: 8:30 AM – 10:00 AM

CDM: Q-Matrix Inference

Chair: Hok Kan Ling, Columbia University

CDM 3a: Bayesian Estimation of the Reduced Reparameterized Unified Model Q Matrix

Steven Culpepper, University of Illinois at Urbana-Champaign; Yinghan Chen, University of Illinois at Urbana-Champaign

The reduced Reparameterized Unified Model (rRUM) is a well known Cognitive Diagnosis Model (CDM) and has received significant attention in the psychometrics literature. The popularity of the rRUM is attributed to the fact that it is a more flexible conjunctive CDM than the Noisy Inputs, Deterministic, "And" gate (NIDA) and the Deterministic-Input, Noisy-And-gate (DINA) models. A Bayesian formulation for estimating the rRUM Q matrix is presented. We show that estimating the rRUM Q matrix is complicated by the existence of identifiability issues. A Bayesian framework is presented that enforces identifiability constraints and uses a spike-slab prior for item parameters to select the necessary attributes for each item. The formulation uses an efficient data augmentation strategy and estimates the Q matrix with Metropolis-within-Gibbs sampling. The developed algorithm demonstrated accurate recovery of the rRUM Q matrix in Monte Carlo simulation studies. We present applications to several classic cognitive diagnosis datasets.

CDM 3b: An Exploratory Discretized Factor Loading Method for Q-Matrix Specification

Wenyi Wang, Jiangxi Normal University; Lihong Song, Jiangxi Normal University; Shuliang Ding, Jiangxi Normal University

The Q-matrix plays an important role in cognitive diagnostic assessment. However, the Q-matrix is usually unknown for many existing tests. If a provisional Q-matrix from subject matter experts contains a large amount of misspecification, the recovery of a high-quality Q-matrix through validation methods will be difficult because the performance of validation methods rely on the classification precision of attribute patterns (de la Torre, 2008; Rupp & Templin, 2008). Under these two situations above, an exploratory technique is necessary. There has been a study about the adoption of principal components analysis as an exploratory technique for finding the Q-matrix by assuming that items measuring the same skill set will load on the same component (Close, 2012). However, since a large number of attribute sets is expected to yield a large number of components, it is hard to determine the meaning of components (Close, 2012). The purpose of this study is to explore a simple method for Q-matrix specification, called an Exploratory Discretized Factor Loading (EDFL) method. This method includes two critical steps. First, the factor loadings matrix is estimated through an exploratory factor analysis regarding latent attributes as latent factors; then a discretization process is employed on the factor loadings matrix to obtain a binary Q-matrix. Simulation studies are conducted to investigate the performance of the EDFL method under various conditions. The simulation results showed that the EDFL method can provide high-quality provisional Q-matrix for any validation Q-matrix method.

CDM 3c: On the Identifiability of Diagnostic Classification Models

Guanhua Fang, Columbia University

This paper establishes fundamental results for statistical inference of Diagnostic Classification Models (DCM). The results are developed at a high level of generality, applicable to essentially all diagnostic classification models. In particular, we establish identifiability results of various modeling parameters, notably item response probabilities, attribute distribution, and Q-matrix-induced partial information structure. Consistent estimators are constructed. Simulation results show that these estimators perform well under various modeling settings. We also use a real example to illustrate the new method. The results are stated under the setting of general latent class models. For DCM with a specific parameterization, the conditions may be adapted accordingly.

CDM 3d: Alternative Implementations of the GDI Q-Matrix Validation Procedure

Yu Bai, Columbia University; Wenchao Ma, Rutgers, The State University of New Jersey; Jimmy de la Torre, The University of Hong Kong

Misspecification of the Q-matrix under Cognitive Diagnostic Models (CDM) can affect correct classification of examinees. Previous researchers have developed several methods to validate a Q-matrix based on special CDMs (Templin & Henson, 2006; DeCarlo, 2012; Chiu & Douglas, 2013). De la Torre and Chiu (2016) proposed a discrimination index (sigma squared) that can be used to identify and replace misspecified Q-matrix entries using a general CDM (i.e., the Generalized Deterministic Input, Noisy, "And" gate [G-DINA] model) and applicable to the reduced models it subsumes. However, a cutoff ε for the Proportion of Variance Accounted For (PVAF) by a particular q-vector relative to the maximum sigma squared needs to be predetermined. It was set to be 0.95 in the study (de la Torre & Chiu, 2016), without further justification. Despite promising results, choosing the best cutoff value can in practice be difficult. This study proposes two methods to validate a Q-matrix based on the PVAF using the G-DINA model, but without the need to specify ε a priori. Method 1 selects the best q-vector based on the "mesa" plot, which shows the PVAF values against the number of attributes specified; Method 2 selects the best candidate q-vector based on their AIC and BIC indices. The proposed methods, together with the current implementation, will be compared in terms of their q-vector recovery rates. The factors considered in the simulation study includes the number of attributes (K), the number of items (J), sample size (N), and the pattern of Q-matrix misspecifications.

CDM 3e: Different Expressions of a Knowledge State and Their Applications

Shuliang Ding, Jiangxi Normal University; Fen Luo, Jiangxi Normal University; Wenyi Wang, Jiangxi Normal University; Jianhua Xiong, Jiangxi Normal University

For Boolean matrices, any non-zero knowledge state is a column of a potential Q matrix Qp (Ding, Luo, Wang, & Xiong, 2016) and it can be expressed as a Boolean union of the columns of the reachability matrix R based on the Augment algorithm (Ding, Luo, Cai, Lin, & Wang, 2008; Ding et al., 2016). If x is a column of Qr, let Sx={r|(r is a column of R) and (r≤x)}, then the Boolean union of all elements in Sx is called as the redundant expression (RE) of x. Let S'x be the subset of Sx and any two elements in S'x have no prerequisite relationship, i.e., they are not comparable, then the Boolean union of all elements in S'x is called the concise expression (CE) of x. The definitions of RE and CE may be applied to prove the fact that under some conditions R or its equivalent class plays an important role in the design of cognitive diagnostic testing (CDT). Some other interesting applications of these definitions are discussed and some analogous results for a polytomous Q matrix are given under some modifications. For example, the sum of all elements in a knowledge state, say x, equals to the number of the vectors in RE of x. And if the scoring rubric formatis changed, the importance of the quasi-reachability matrix (Ding et al., 2016) in a design of CDT using a polytomous Q matrix is proved.

CAT 1: 8:30 AM – 10:00 AM

CAT

Chair: Ying 'Alison' Cheng, University of Notre Dame

CAT 1a: Some Promising Advancements Concerning CAT Foundations and Implementations

Hua-Hua Chang, University of Illinois at Urbana-Champaign; Shiyu Wang, University of Georgia; Susu Zhang, University of Illinois at Urbana-Champaign

This presentation introduces several promising new advancements concerning Computerized Adaptive Testing (CAT) foundations and implementations. The first one is the establishment of a mathematical foundation to demonstrate that an examinee should be allowed to revise the answers to previously administered items during the course of testing. Currently almost all operational CAT programs forbid item revision, which is allowed by the traditional paper-and-pencil tests. This has become such a main concern for both examinees and testing companies that some testing programs have decided to switch from CAT to other modes of testing. Recently Wang, Felloris, and Chang (2015) demonstrated that, under the Nominal Response Model, allowing item revision will not compromise test efficiency and security. Most recently, such result has been generalized to most commonly used IRT models, including the 2PL and 3PL models. Then, we address a number of issues emerging from large scale implementation and show how theoretical works can solve practical problems. Our new focus will be on Cognitive Diagnostic CAT (CD-CAT), which has become a powerful tool for schools to assess students' mastery of various skills. In particular, we will present our research results on the use of CD-CAT to improve STEM learning and retention. We will also discuss the possibilities to combine CAT and learning, i.e., to adaptively select the most appropriate contents for the individual learners. Lastly, we will ruminate on and discuss some possible future directions of research on CAT.

CAT 1b: Indicators of the Amount of Adaptation by Computerized Adaptive Tests

Mark D. Reckase, Michigan State University; Unhee Ju, Michigan State University; Sewon Kim, Michigan State University

Computerized Adaptive Testing (CAT) is gaining wide acceptance with the ready availability of computer technology. The general intent is to adapt the difficulty of the test to the capabilities of the examinee so that measurement accuracy is improved over fixed tests, and the entire testing process is more efficient. However, many computer administration designs, such as two-stage tests, stratified adaptive tests, and those with content balancing and exposure control, are called adaptive, but the amount of adaptation greatly varies. In this paper, several measures of the amount of adaptation for a CAT are presented along with information about their sensitivity to item pool size, distribution of item difficulty in the item pool, and exposure control. A real data application will also be presented to show the level of adaptation of a mature, operational CAT. Some guidelines will be provided for how much adaptation should take place merit the label of an "adaptive test."

CAT 1c: A Dynamic Stratification Method for Improving Trait Estimation in CAT

Hsiu-Yi Chao, National Chung Cheng University; Jyun-Hong Chen, National Sun Yat-sen University

In Computerized Adaptive Testing (CAT), a fundamental problem known as CAT's attenuation paradox has not yet been solved and is a potential threat to test efficiency. The attenuation paradox means that a high-quality item can only provide little information for trait estimates that are far away from the item's difficulty. In the early stages of a CAT where the trait estimate is less precise, it stands a good chance to produce an item administration involving the attenuation paradox. Consequently, CAT's performances will unavoidably decrease in reliability. To solve the attenuation paradox, this study introduces a new concept of "global adaptiveness" in contrast to the ordinary "local adaptiveness." CAT with local adaptiveness only considers how to maximize information within single item administration, whereas CAT with global adaptiveness additionally considers how to maximize the test information for individuals and all examinees. Based on the definition, this study develops a Stratification method based on Dominance Curve (SDC) to solve the attenuation paradox. SDC stratifies an item pool into strata according to the dominance curve derived by Fisher information and utilizes a dynamic process for item-stratum adjustment to optimize high-quality items' usage. According to the simulation study, SDC outperforms the other methods in terms of lower RMSE and higher test information. For test practitioners, SDC that enhances global adaptiveness is expected to improve CAT's efficiency in most test scenarios, especially for conditions with stringent exposure control (e.g., achievement tests).

CAT 1d: Measuring Non-Compensatory Multidimensional Structures with Computerized Adaptive Testing

Andreas Frey, Friedrich Schiller University Jena; Anna Mikolajetz, Friedrich Schiller University Jena

Multidimensional Computerized Adaptive Testing (MCAT) is an efficient and flexible way to measure the ability of an individual simultaneously on several dimensions. Until now, MCAT has been limited to compensatory Multidimensional Item Response Theory (MIRT) models in which a low ability on one dimension is compensated for by higher abilities on other dimensions. However, for some constructs non-compensatory MIRT models are sometimes more appropriate. To overcome this shortcoming, we introduce how the non-compensatory Multicomponent Latent Trait Model (MLTM; Whitely, 1980) can be used for fixed-length MCAT. We examined the performance of MCAT with the MLTM in a simulation study based on a fully crossed factorial design, comprising the independent variables item selection (random, adaptive), MIRT model (compensatory, non-compensatory), correlation between dimensions (.00, .25, .50, .75), and test length (1 to 80 items). In each condition, two dimensions were measured with an item pool of 400 items. The simulation was carried out with SAS, while calling the R package mirt (Chalmers, 2017) for ability estimation. MCAT exhibited a higher precision of the estimated abilities (lower MSE) compared to random item selection in all conditions. The advantages of MCAT became smaller with increasing correlations between the dimensions. The level of precision obtained with the compensatory and the non-compensatory models was similar, with a slightly higher precision for the non-compensatory model, especially for no (.00) and low correlations (.25) between the dimensions. The present study expands the applicability of MCAT to cases in which a non-compensatory connection between the measured dimensions is appropriate.

CAT 1e: Item selection in Multidimensional Adaptive Testing for Noncompensatory IRT Models

Chia-Ling Hsu, The Education University of Hong Kong; Wen-Chung Wang, The Education University of Hong Kong

In the literature, computerized adaptive testing is used in conjunction with compensatory Multidimensional Item Response Theory (MIRT) models (denoted as MCAT-C). Since items may measure multiple latent traits that cannot compensate each other, noncompensatory MIRT models have been developed to account for such items. This study aimed to develop computerized adaptive testing algorithms that were based on noncompensatory MIRT models (denoted as MCAT-N). Four item selection methods, namely, the Fisher information method, the Kullback-Leibler information method, the Shannon entropy method, and the mutual information method, were investigated under two major factors, namely, the correlation between latent traits and the level of termination criterion. Results of a series of simulations showed that all the four item selection algorithms were successfully implemented; a high correlation between latent traits or/and a strict level of termination improved measurement precision and test reliability; test reliabilities for all dimensions were similar regardless of test administration stage because of the noncompensatory nature of MCAT-N. Moreover, among the four methods, the Fisher information method was the superior and the Kullback-Leibler information method the inferior.

Symposium 11: 8:30 AM – 10:00 AM

Symposium 11: Missing Data in Large-Scale Educational Assessments

Chair: Tyler Matta, University of Oslo

Symposium 11a: Design and Treatment of Missing Auxiliary Data in Large-Scale Assessments

Leslie Rutkowski, University of Oslo; Tyler Matta, University of Oslo

With a 20-year history of modern International Large-Scale Assessments (ILSAs) and a growing interest in non-achievement outcomes (i.e., socioemotional or affective domains), pressure to increase the length of the ancillary context questionnaire is a natural consequence. In response, PISA 2012 developed and administered questionnaires with planned missingness that follows a so-called three-form design (Graham, Taylor, Olchowski, & Cumsille, 2006). In essence, the three-form design is a type of matrix sampling (Shoemaker, 1973) that features the administration of different items to different respondents and allows correlations to be estimated for all pairs of variables. Although this strategy was abandoned in PISA, planned missingness remains of interest in other studies such as the U.S. National Assessment for Educational Progress. As such, we propose to investigate modifications of the three-form design in terms of recovering fully observed moments (e.g., means and covariances) of context questionnaire variables. In particular, we consider whether design aspects that deal directly with correlations within and between constructs are better suited to the task. We also evaluate discrepancies between state-of-the-art missing data methods, including a straightforward implementation of an EM algorithm for means and covariances and multiple imputation. Given the primacy of achievement measures, we consider the impact of several planned missing designs and associated treatments on achievement distributions, as currently estimated with operational methods. In each case, we balance respondent burden against ease of missing data treatment implementation, stability of achievement estimates, and the usability of resultant public-use data for secondary analysts.

Symposium 11b: An Item Response Model for Omitted Responses in Performance Tests

Alexander Robitzsch, Leibniz Institute for Science and Mathematics Education; Oliver Lüdtke, Leibniz Institute for Science and Mathematics Education

Dealing with omitted item responses in performance tests is challenging because it is often unclear whether the omitted response is an indicator of low ability or low test motivation. In the psychometric literature, different approaches for modelling omitted item responses have been proposed. In some articles, it has been argued that treating missing item responses as incorrect lead to biased item and person parameter estimates (Pohl et al., 2013). Moreover, multidimensional item response models with an additional missingness indicator for each item have been introduced to model non-ignorable omitted item responses (Rose et al., 2016). However, these models assume an underlying (unidimensional) latent variable for all missingness indicators, which seems to be a very restrictive assumption. In this talk, we propose an item response model in which the omission of an item is modelled as a function of ability and the unobserved (true) item response (see Mislevy & Wu, 1996). This model contains different treatments of omitted item responses as special cases: the treatment as incorrect, the treatment as correct, modelling as ignorable data and the non-ignorable multidimensional IRT model of Rose et al. (in press). We show that item parameters can be statistically identified and can be consistently estimated and model selection based on information criteria provides valid statistical inference. Finally, we illustrate the modelling approach using PIRLS data. The results suggest that omitted responses for constructed item responses should be treated as incorrect while the omission of item responses for multiple-choice items can be practically modelled as ignorable responses.

Symposium 11c: Imputation of Missing Data at Level 2 Using Plausible Values

Simon Grund, Leibniz Institute for Science and Mathematics Education; Oliver Lüdtke, Leibniz Institute for Science and Mathematics Education; Alexander Robitzsch, Leibniz Institute for Science and Mathematics Education

Educational assessment often involves the analysis of multilevel data, in which missing values can occur at Level 1 (e.g., students) and Level 2 (e.g., teachers). Multiple Imputation (MI) can be used to address missing data at Level 2 while taking into account auxiliary information that is available from variables at Level 1. Several authors have considered multilevel MI for missing data at Level 1 (e.g., Enders, Mistler, & Keller, 2016). By contrast, the treatment of missing data at Level 2 has received relatively little attention, and open questions remain about how to include auxiliary information available from variables at Level 1 (Resche-Rigon & White, in press). Using theoretical arguments and computer simulations, we compare Joint Modeling (JM) and the Fully Conditional Specification (FCS) of MI as well as different strategies for including auxiliary variables at Level 1 using either manifest or latent cluster means. The plausible values technique (Mislevy, 1991) is used to generate latent cluster means within the FCS approach. We show that (a) an FCS approach that uses latent cluster means is comparable to JM, and (b) the use of manifest cluster means, while not fully equivalent with JM, provides similar results unless in relatively extreme cases with strongly unbalanced data. We outline a computational procedure for including latent cluster means in an FCS approach using plausible values, and we provide an example using data from the PISA 2012 study.

Symposium 11d: Assessing Missing Data Assumptions Using Posterior Predictive Checks

Tyler Matta, University of Oslo

One of the most critical aspects of analyzing incomplete data are the assumptions pertaining to the missing data mechanism. Because the ignorability property rests on the data that went uncollected, one cannot empirically validate their missing data assumptions from the observed data alone (Molenberghs, Beunckens, Sotto, & Kenward, 2008). This has led to the development of methods that assess the sensitivity of model inferences to departures from the ignorability assumption. This paper will introduce a novel approach for assessing missing data assumptions using Posterior Predictive Checks (PPC). Traditional PPCs compare simulated values from a posterior predictive distribution to the observed data as a means of understanding the probability that the observed data came from such a model. When the missing data mechanism is assumed to be a function of the data the went uncollected, parts of the model may appear to fail when compared to the observed data alone. This is because the posterior predictive distribution was designed to produce the unrealized full data. Thus, this paper will demonstrate how the PPC can be used to better understand the connection between our assumptions about a missing data mechanism and the potentially observable, yet unrealized, full data. A series of simulated examples will illustrate the utility of this approach.

Symposium 12: 8:30 AM – 10:00 AM

Symposium 12: The Role of Testing Organizations in Promoting Psychometric Research

Chair: Alina von Davier & Terry Ackerman

Symposium 12a: The Role of Testing Organizations in Promoting Psychometric Research. ACT's Perspective

Alina von Davier, ACTNext

In this panel, we will discuss the strategies that testing companies may implement to support innovation in assessments and psychometric research through regulation, processes, and companies' initiatives. Partnerships with universities and the role of writing grants in advancing a psychometric agenda will be discussed. We will also have a conversation about international employment, flexible employment, and students' internships. Moreover, we will address the foray made by machine learning, data mining and technology in the testing industry and their impact on psychometrics. In this panel, the focus is on ACT 's perspective and examples of processes developed at ACT will be provided.

GI 1: 8:30 AM – 10:00 AM

Guessing and Indecision

Chair: Allison Ames, James Madison University

GI 1a: Simulating Response Patterns in Multiple-Choice Questions

Qian Wu, KU Leuven; Tinne De Laet, KU Leuven; Charlotte Lewyllie, KU Leuven; Rianne Janssen, KU Leuven

Correct for guessing is a commonly used scoring method in multiple-choice questions. A penalty is used for incorrect responses to discourage guessing. However, combining Item Response Theory (IRT) and prospect theory of decision making under uncertainty, Budescu and Bo (2015) showed that a penalty has detrimental effects for examinees, especially for those who are risk averse. Another drawback of correct for guessing is its insensitivity to the differences between various knowledge levels. This can be dealt with by a scoring method that credits partial knowledge, such as elimination scoring (Coombs, Milholland, & Womer, 1956). It requires examinees to eliminate alternatives that they consider to be incorrect. This study investigates the combined effect of ability and risk aversion on expected scores on multiple-choice items, and compares these effects under two testing instructions, namely, correction for guessing and elimination scoring. A model is proposed to simulate expected answering patterns on multiple-choice items, combining the IRT and prospect theory. It consists of two steps: (1) probabilities of a correct response to each of the alternatives in a multiple-choice question are modeled using Rasch model based on ability; (2) the decision making of giving a particular answering pattern is modeled using the prospect theory taking into account risk aversion. The results from the simulation revealed that overall ability had a predominant effect on expected scores, while risk aversion had a decisive impact on expected answering patterns especially for examinees with lower abilities. Examinees with medium ability levels benefited from using elimination scoring.

GI 1b: Responding to Rating Questions with Inaccuracy: A Multivariate Mixture Model

Sabrina Giordano, University of Calabria; Roberto Colombi, University of Bergamo

This work proposes a multivariate model for ordinal rating responses, allowing for inaccuracy in answering. In responding to rating questions, an individual may give answers either according to his/her knowledge (feeling) or to his/her level of indecision (inaccuracy). Since ignoring this inaccuracy may lead to misleading results, we define the joint distribution of the ordinal responses via a mixture of components, characterized by inaccuracy in answering to a subset of variables. Uncertain people can be assumed to give an answer at random, by assigning equal probability to every category in the scale

(Uniform distribution), but in most cases, wavering respondents tend to use only a small number of the available rating scale options: someone may skip extreme values, optimists may overvalue their feelings and pessimists may underrate them, someone else can take shelter in the middle category. The proposed approach aims to adequately model the distribution of responses, given with inaccuracy according to these different behaviors, by U-shaped, bell shaped, unimodal, symmetric and skewed distributions. Moreover, our proposal enables us to separate association from inaccuracy and describe how concordant or discordant the responses are that aware people give to a set of questions. Since the proposed models are based on some restrictive assumptions which assure identifiability, the true distribution is not necessarily in the model we are using. For this reason, testing procedures for comparing competitive models will be dealt with by avoiding the assumption of correct specification of the models.

GI 1c: Parameters Interpretation and a Selection Test for the 1PL-G Model.

Paula Fariña, Universidad Diego Portales

The contribution of this research is twofold: first it clarifies how to interpret the usually called the guessing and difficulty parameters in the 1 Parameter Logistic with Guessing model (1PL-G); second it offers a specification test to decide between the semiparametric random effect Rasch and 1PL-G models. Regarding our first contribution we found that both parameters are not able to isolate the concepts of difficulty and guessing of an item. In fact, these concepts are interwoven with each other. Our approach for the interpretation of the parameters is based on a formal methodology which consists on linking the parameter with the sampling distribution, taking advantage of identification results presented by (San Martín et al., 2013). Regarding our second contribution we propose a test statistic. The test is free of any distributional assumptions of the random ability, since it is based only on sampling frequencies. This property is an advantage with respect to other typically used hypothesis tests, as the likelihood ratio test, that depends on an unknown distribution of abilities. We present simulations to study the performance of two versions of the hypothesis test (an asymptotic version and a version using bootstrap). Simulations suggest that the test has a good performance, in terms of both confidence and power. In the case of exams of 30 and 50 items a bigger sample size of at least 10000 is required.

GI 1d: The Effect of Probing "Don't Know" Responses on Measurement Quality

Myrsini Katsikatsou, The London School of Economics and Political Science; Jouni Kuha, The London School of Economics and Political Science; Sarah Butt, City University of London; Chris Skinner, The London School of Economics and Political Science

In survey interviews, "Don't know" (DK) responses are commonly treated as missing data. One way to reduce the rate of such responses is to probe initial DK answers with a follow-up question designed to encourage respondents to give substantive, non-DK responses. However, such probing can also reduce data quality by introducing additional or differential measurement error. We propose a latent variable model for analysing the effects of probing on responses to survey questions. The model makes it possible to separate measurement effects of probing from true differences between respondents who do and do not require probing. We analyse new data from an experiment which compared responses to two multi-item batteries of questions with and without probing. In this study, probing reduced the rate of DK responses by around a half. However, it also had substantial measurement effects, in that probed answers were found to be weaker measures of constructs of interest than were unprobed answers. These effects were larger for questions on attitudes than for pseudo-knowledge questions on perceptions of external facts.

SEM 1: 8:30 AM – 10:00 AM

SEM: H-likelihood and Ridge Estimation

Chair: Silvia Cagnone, University of Bologna

SEM 1a: Factor Analysis with Ordinal Data: An H-Likelihood Approach

Youngjo Lee, Seoul National University

Marginal likelihood-based methods are commonly used in factor analysis with ordinal data. In order to obtain the marginal likelihood, Full Information Maximum Likelihood (FIML) uses the Gauss-Hermite quadrature or adaptive quadrature. However, the computational burden increases rapidly as the number of factors increases, which renders FIML impractical for large factor models. Another limitation of the marginal likelihood-based approach is that it only provides inference for fixed parameter but nothing has been said about the factors. In this study, we propose a Hierarchical Likelihood (HL) approach for factor models with ordinal data, following the extended likelihood principle: given the data, all information regarding the fixed parameters and random effects are contained in the extended likelihood function. HL remains computationally efficient as the number of the latent variables increases. It simultaneously estimates fixed unknown parameters and predicts the latent variables. We also assess the joint maximization of latent variables and mean parameters. A simulation study is conducted to evaluate the estimation accuracy for the proposed approach.

SEM 1b: Robust Nonlinear Structural Equation Modelling Using Double Hierarchical Likelihood

Shaobo Jin, Uppsala University

Various analytical distributional approaches have been proposed for nonlinear structural equation models with latent interactions. Two well-known analytical distributional approaches, latent moderated structural equations and quasi-maximum likelihood, are based on the normality assumption. When the normal assumption fails, they are subject to bias. The Hierarchical Likelihood (HL) approach has been proposed recently. However, it also relies on the normality assumption. In this study, we propose a robust HL approach, namely the Double Hierarchical Likelihood (DHL) approach. Instead of assuming normally distributed exogenous latent variables, DHL assumes that the exogenous latent variables are conditionally normal by introducing an extra latent variable to account for non-normality. Thus, it is more robust against distributional misspecification. A simulation study shows that the DHL approach is as accurate as the existing distributional approach when the exogenous latent variables are normal and is more accurate when the exogenous latent variables are fat tailed or skewed.

SEM 1c: Nonlinear Structural Equation Modelling with Hierarchical Likelihood Estimation

Fan Wallentin, Uppsala University

Nonlinear Structural Equation Modelling (SEM) has become widely used in practice. Due to its nature of non-normally distributed indicators, various approaches have been proposed, including the product indicator approach, distribution analytic approach, and method of moments. Two well-known distributional approaches are latent moderated structural equations and quasi-maximum likelihood. The former is Gauss quadrature based and the latter is based on an approximation to the conditional distribution. In this study, an alternative approach, the Hierarchical-Likelihood (HL) approach, is proposed. The mean parameters are estimated by a Laplace approximation of the likelihood and the dispersion parameters are estimated from an approximated restricted likelihood. The proposed approach is computationally appealing and predicts the value of latent variables alongside with estimating the fix parameters. A simulation study shows that the proposed approach produces accurate parameter estimates if the distribution is correctly specified.

SEM 1d: Optimizing Tuning Parameter in Ridge Generalized Least Squares for Structural Equation Modeling

Miao Yang, University of Notre Dame; Ke-Hai Yuan, University of Notre Dame

The ridge generalized least squares (GLS) procedure is a combination of least squares (LS) and GLS. Empirical studies have indicated that the ridge GLS procedure for structural equation modeling can yield more accurate and more efficient parameter estimates than either GLS or the maximum likelihood method when the population distribution is unknown. In the formulation of ridge GLS, there is a tuning parameter whose value determines the relative contribution of LS and GLS. The tuning parameter might be selected according to the empirical efficiency of the parameter estimates. However, for real data analysis, how to choose a value for the tuning parameter has not been formally studied. In the current study, a formula has been developed to predict the optimum ridge tuning parameter. For the formula to have a wide scope of applicability, it is calibrated via many conditions on population distribution, sample size, number of variables and model structure.

SEM 1e: Improving Parameter Estimates for Factor Analysis of Ordinal Data
Ge Jiang, University of Notre Dame; Ke-Hai Yuan, University of Notre Dame

Data in psychology are often collected using Likert-type scales to measure a smaller number of latent dimensions. In theory, factor analysis of Likert-type data is best conducted on the polychoric correlation matrix using Generalized Least Squares with an Asymptotically correct weight matrix (AGLS). However, AGLS performs poorly due to finite sample sizes and/or a large number of indicators (e.g., items in a test). Consequently, estimation methods such as Least Squares (LS) and Diagonally Weighted Least Squares (DWLS) that do not require a full weight matrix are routinely used in practice. The issue with LS and DWLS is that they ignore the association among the polychoric correlations and thus are not optimal. Recently, two ridge methods have been proposed to yield more efficient parameter estimates for ordinal data, and they can be regarded as a weighted combination of AGLS and LS/DWLS. A tuning parameter, a, is involved and needs to be determined in each of the two ridge methods. The research to be presented aims to develop a formula of the ridge tuning parameter a to yield most efficient estimates of model parameters in practice. Empirical modeling is used in which a is predicted by sample size, number of indicators, number of categories, skewness/kurtosis of the observed frequency, and the conditions of asymptotic covariance matrix. An example is used to illustrate the application of the obtained formula and its effect in real data analysis.

Invited: 10:05 AM – 10:50 AM

Optimal Scoring as an Alternative to IRT and Sum Scoring
Invited Speaker: Marie Wiberg, Umeå University
Chair: Wim van der Linden, Umeå University

Test constructors often use item response theory (IRT) in the design and evaluation of items and tests. When delivering test results from an academic test or an industrial constructive test, it is however still common for a test taker to receive a sum score, i.e. a sum of number of correct answers of the items in the test. Optimal scoring is an alternative to IRT and sum scoring. Optimal scoring is built on the ideas of functional data analysis, and in particular, uses parameter cascading in the estimation step. By applying maximum likelihood estimation, a weighted score is obtained– where the weights are specific to the test takers performance level. Simulations show that optimal scoring perform better than sum scores in terms of lower root mean squared error, especially for test takers at performance extremes. A real data example illustrates how optimal scoring can be used in practice. The advantages of using optimal scoring in academic tests and industrial constructive tests as an alternative to IRT and sum scoring are discussed.

State: 10:05 AM – 10:50 AM

Network Analysis: Goodbye to Independence Assumptions
State-of-the-Art Speaker: Tom Snijders, University of Groningen & University of Oxford
Chair: Carolyn J. Anderson, University of Illinois at Urbana-Champaign;

Network modeling has developed in the new millennium from a niche topic to the scientific mainstream, but in a large diversity of approaches. This presentation will tell something about the current state of statistical modelling of data sets representing social networks, as distinct from other network approaches such as probabilistic modelling, algorithmic developments, and the use of networks to represent dependencies in "regular" multivariate data. The basic mathematical structure for representing network data is a graph or digraph (directed graph): a set of nodes some of which are tied by edges or directed edges. For a social network, the nodes usually represent social actors and the ties some kind of social relation. Data sets often also contain nodal variables, representing behaviour and other characteristics of the actors. Some conceptual issues are fundamental in social network analysis, and these imply a basic contrast between network data and traditional "rectangular" multivariate data.

The first is the dependence between ties, mirroring social phenomena such as reciprocity, transitivity ("friends of my friends are my friends"), and differential popularity of actors. The second issue is that ties between actors will imply social influence or others kinds of dependence between the tied actors. The third issue is the importance of indirect connections: what matters for my well-being and the information available to me are not only those to whom I am connected, but also their further connections. All this means that for statistical modelling we are in a mess, because we cannot base models on independence assumptions any more. There is some room for permutation-based procedures, but their use is limited. Instead, parametric statistical models have been proposed that are based on conditional independence assumptions. Examples are Stochastic Blockmodels and Latent Space Models, assuming conditional independence given an assumed latent structure; and Exponential Random Graph Models ("ERGMs"), making certain conditional independence assumptions directly about the ties. For social science research, especially important are questions about the importance of networks for changing actor variables (i.e., characteristics of the nodes) such as individual performance, attitudes, and behavioural tendencies. Given the mutual dependence between networks and actor variables, it is clear that such questions are best studied using longitudinal data. Here a crucial conditional independence assumption is the regular Markov assumption for stochastic processes, but this has to be supplemented by further parametric assumptions to obtain workable models. The Stochastic Actor-oriented Model is used a lot for modelling dynamics of networks as well as modelling the interdependent dynamics of networks and actor variables. The presentation will explain some things about this bestiary of models, although only the surface can be scratched. In addition, some attention shall be paid to the question how much it matters whether or not the usual independence assumptions are made for analysing actor variables; and to current developments, including multilevel network analysis.

Invited: 11:10 AM – 11:55 AM

Psychometrics for Complex Systems

Invited Speaker: Han van der Maas, University of Amsterdam
Chair: Francis Tuerlinckx, KU Leuven

Humans, like eco-systems, the weather, and the stock market, are non-linear complex systems. With this in mind, it is possible to develop suitable formal measurement models of human psychological functioning. Complex systems can be understood as networks that, depending on connection strength, behave linearly or nonlinearly, or even discretely. There are many interesting technical links and equivalences between network models and the dominant latent variable approach in psychometrics, but conceptually they are very different. This will be discussed in the context of the measurement of cognitive functions and modeling of general intelligence. First, I will present a network model that combines mutualism, central processes and sampling within one integrated (non-g) model of general intelligence. Second, I will present a new approach to educational measurement, motivated by the requirement of high frequent measurements in complex systems research. As an example I will present the results of a web-based computerized adaptive training and monitor systems used by thousands of schools in the Netherlands, yielding over 1 billion item responses.

State: 11:10 AM – 11:55 AM

SEM, Robust Corrections, and Missing Data

State-of-the-Art Speaker: Victoria Savalei, University of British Columbia
Chair: Alberto Maydeu-Olivares, University of South Carolina

In this talk I will review the logic of robust standard errors as used in SEM by first showing how these standard errors are derived in the case of linear regression under the violation of assumptions, and then viewing SEM as a special type of a nonlinear regression model. I will then review the most common applications of robust standard errors in SEM. While these are often known as "corrections for nonnormality", I will illustrate that they should be more accurately described as "corrections for inefficiency", and have many applications that have nothing to do with the specification of the distribution of the data. Lastly, I will zero in on what is arguably the most popular application of these corrections, and review the different options for robust standard errors, test statistics, and fit indices available to accompany the normal-theory ML estimator for nonnormal and/or incomplete data.

Keynote: 1:30 PM – 2:30 PM

Item-Response Models for the Analysis of Self-Reports
Ulf Böckenholt, Northwestern University

Many policy decisions rely on the information extracted from self-reports obtained in surveys or non-cognitive assessment tests. The importance of the self-report tool in the social sciences explains why, ever since it emerged, researchers and practitioners have advanced theories about psychological processes involved in responding to questionnaire items. However, there is a substantial gap between cognitive theories explaining the response process and current psychometric models utilized to analyze self-report data. This talk highlights this gap by focusing on the modeling of such well-established phenomena as response styles, item-reversal effects, halo effects, and self-enhancing/-protective responding. I present item-response models that build on cognitive theories and inform about the adaptive and goal-directed nature of the item-response process. Both for estimation and validation purposes, I also discuss the use of supplemental information in the form of eye-tracking and response-time data. From a modeling perspective, the presented item-response models join a growing literature on psychometric approaches that, by explicitly accounting for different types of response processes, go beyond the classic view of random sources of measurement error.

Symposium 13: 2:50 PM – 4:20 PM

Symposium 13: Recent Extensions to Autoregressive Models in Psychology
Chair: Casper Albers, University of Groningen

Symposium 13a: Modelling Smooth and Sudden Changes in Temporal Dynamics of (V)AR-Models
Casper Albers, University of Groningen; Laura Bringmann, University of Groningen

Interest in studying temporal dynamics has seen a massive increase over the past years. Standard autoregressive models assume time-invariance of the parameters: whilst the data will change over time, the generating process can not. Two extensions to the AR-model are in vogue: regime change models and Time-Varying (TV-AR) models. Regime change models are useful whenever sudden, abrupt changes in the generating mechanisms can be expected (i.e. when a life event occurs, or a new phase in treatment begins). TV-AR models, on the other hand, allow for the modelling of smooth changes in the generating process, thus modelling gradual changes in e.g. someone's emotion dynamics. So far, one could either model regime changes or smooth changes, but not both. In this presentation, I will outline a novel methodology that combines both approaches. This model can be applied both confirmatory and exploratory. Confirmatory analyses are useful when studying whether the dynamics have changed after a certain speficied time-point. Exploratory analyses are useful to find out where unexpected sudden changes occured, whilst allowing for gradual changes at the same time.

Symposium 13b: Measurement Error and Person-Specific Reliability in Multilevel Autoregressive Models
Noémi Schuurman, Utrecht University

Many researchers in the social sciences are discovering intensive longitudinal data, using ambulatory assessment techniques such as experience sampling. An increasingly popular way to analyze these data is autoregressive time series modeling; either by modeling the repeated measures for a single individual using classic $n = 1$ autoregressive models, or by using multilevel extensions of these models, with the dynamics for each individual modeled at level 1 and interindividual differences in these dynamics modeled at level 2. However, while it is widely accepted that psychological measurements usually contain measurement error, the issue of measurement error is largely neglected in applications using these models. In this talk we discuss extensions of the (multilevel) autoregressive model that take measurement error into account, for models with either multiple or single indicators for each construct.

Further, we discuss the consequences of disregarding measurement error in the autoregressive model, and how to obtain person-specific and between-person reliability estimates based on these models.

Symposium 13c: Bayesian Sensitivity Analysis of Dynamic Factor Analysis Models with Nonignorable Missingness
Sy-Miin Chow, The Pennsylvania State University; Niansheng Tang, Yunnan University; Joseph G. Ibrahim, University of North Carolina at Chapel Hill; Hongtu Zhu, The University of Texas MD Anderson Cancer Center

Many psychological concepts are unobserved and usually represented as latent factors apprehended through multiple observed indicators. When single- or multiple-subject multivariate time series data are available, dynamic factor analysis models offer one way of modeling patterns of within-person changes in these data by combining factor analysis and time series (e.g., vector autoregressive) analysis at the factor level. Unfortunately, evaluating the optimal lag structures of unobserved factors and the tenability of common missing data assumptions in these models are important but computationally prohibitive tasks. We propose and illustrate the use of a Bayesian local influence method to carry out sensitivity analysis involving the simultaneous perturbations of multiple elements in a dynamic factor analysis model, including perturbations to detect potential misspecifications in the dynamic functional forms and lag structures of the factors, and the assumption of ignorable missingness.

Symposium 13d: Interpretation and Identification of Path-Specific Effects in CT-VAR(1) Models
Oisín Ryan, Utrecht University; Ellen Hamaker, Utrecht University

Repeated measurement data in psychology is typically analysed by regressing a set of variables on their lagged values at a previous measurement occasion. The most widely used model of this type is the Discrete-Time (DT) first-order Vector Auto-Regressive (VAR(1)) model, also known as the Cross-Lagged Panel Model (CLPM). In recent years there has been a surge of interest in applying these models to learn about the relationships between sets of three or more variables. Typically researchers assume that the regression coefficients in these models may be interpreted as direct effects. Continuous-Time (CT) VAR(1) models, based on first-order stochastic differential equations, have been proposed as an alternative to the more widely used DT VAR(1): in practical terms CT models deal well with measurements taken at unequal intervals, a typical characteristic of many momentary-assessment designs. However, authors such as Aalen et al. (2014, 2016) and De Boeck and Preacher (2016) have argued that treating variables as representing CT processes leads to important conceptual differences regarding the interpretation of VAR(1) model parameters; namely that regression co-efficients may be better interpreted as representing total effects rather than direct effects. In light of this, we present work examining the use of CT models to investigate path-specific effects between three or more processes. Specifically we approach this topic from the interventionist perspective on mediation (VanderWeele, 2015) and examine how these opposing conceptualisations of path-specific effects relate to different types of hypothetical interventions.

Symposium 13e: Studying Time-Lagged Effects: To Be Continued or to Be Considered
Rebecca M. Kuiper, Utrecht University

The emergence of devices, such as smartphones, led to an exponential increase in real-time self-report data studies in the social and medical sciences. In these data, referred to as Ambulatory-Assessment data or Experience-Sampling-Method-(ESM-)data, participants registered their, for example, feelings or symptoms multiple times a day for several consecutive days. ESM-data offer the unique opportunity to model everyday processes as they unfold over time and to investigate cross-lagged relationships, that is, the effects variables have on each other. The latter are of interest when researchers want to examine hypotheses such as "Stress causally dominates anxiety". Unfortunately, the existing techniques for analyzing cross-lagged relationships in ESM-data fall short. In this presentation, I will discuss what type of models can be estimated (discrete-time models and multilevel continuous-time models) and what their pros and cons are; and what type of statistical techniques for evaluating hypotheses exists for these models.

Symposium 14: 2:50 PM – 4:20 PM

Symposium 14: The Replicability Crisis: Diagnosis, Etiology, and Treatments

Chair: Jelte Wicherts

Symposium 14a: Replicability of Structural Equation Modeling and the Availability of Syntaxes

Elise A. V. Crompvoets, Tilburg University; Jelte M. Wicherts, Tilburg University

A straightforward way to replicate a Structural Equation Model (SEM) described in the literature is to re-run the model with either the original data or novel data using the original syntax or code. Such SEM syntaxes convey valuable information on model specifications and the manner in which SEMs were estimated. In this study, we requested SEM syntaxes from 229 articles (published in 1998-2013) that ran SEMs using LISREL, AMOS, or Mplus. After exchanging over 500 emails, we ended up obtaining a meagre 57 syntaxes used in these articles (24.9% of syntaxes we requested). Results considering the 129 (corresponding) authors who replied to our request showed that the odds of the syntax being lost increased by 22% per year passed since publication of the article, while the odds of actually obtaining a syntax dropped by 13% per year. So SEM syntaxes that are crucial for reproducibility and for correcting errors in the running and reporting of SEMs are often unavailable and get lost rapidly. The preferred solution to heighten replicability of SEM results is mandatory sharing of SEM syntaxes alongside articles or in data repositories.

Symposium 14b: Bayes and Nonsignificant Findings: A Solution for the Replication Crisis

Rink Hoekstra, University of Groningen; Rei Monden, University Medical Center Groningen; Don van Ravenzwaaij, University of Groningen; Eric-Jan Wagenmakers, University of Amsterdam

An important issue in the replication crisis is the clear bias for publishing findings that show evidence for the existence, rather than for the absence of an effect. A possible explanation might be the fact that classical statistics are unable to quantify the degree to which the data actually support the null hypothesis. Nevertheless, there are many situations for which showing the absence of an effect is highly interesting from a substantive point of view. To underscore the importance of a solution for this discrepancy, we reanalyzed data from papers claiming the absence of an effect by conducting a Bayesian hypothesis test. This reanalysis revealed that the degree of evidence was highly variable between papers and could not be reliably predicted from the p-value. Instead, sample size was a better (albeit imperfect) predictor for the strength of evidence in favor of the null hypothesis. Our findings do not only underscore the problems with currently used methods, but also suggest that the Bayesian hypothesis test might well be a promising tool to select papers in which claims are made that are in need of additional evidence.

Symposium 14c: Why Replications Fail

Jelte M. Wicherts, Tilburg University

Several recent large-scale collaborative projects have attempted to replicate – as closely as possible – previous results from highly cited psychology papers or papers published in top psychology journals. Despite considerations of power and rigorous pre-registration of analysis plans, the majority of these replications failed when judged on the basis of significance or when the size of the effect was compared to that in the original studies. Here I discuss substantive, statistical, and methodological reasons for these (apparently) discrepant results. I discuss the so-called hidden moderator account of failures to replicate, which states that replications differ from original studies in a priori unknown yet crucial ways. I also discuss the widespread failure to publish non-significant results and considerable potential bias caused by researcher's tendency to exploit the flexibility of designs and analyses of data in original studies. I conclude that although it is wise to always consider moderators even in fairly close replications, publication bias and common exploitation of researcher degrees of freedom in designing, running, analysing, and reporting of psychological studies together likely explain the vast majority of failed replications. I discuss some preferred solutions.

Symposium 14d: Multilevel Multivariate Meta-Analysis with Application to Choice Overload
Ulf Böckenholt, Northwestern University; Blakely B. McShane, Northwestern University

We introduce multilevel multivariate meta-analysis methodology designed to account for the complexity of contemporary psychological research data. Our methodology directly models the observations from a set of studies in a manner that accounts for the variation and covariation induced by the facts that observations differ in their dependent measures and moderators and are nested within, for example, papers, studies, groups of subjects, and study conditions. Our methodology is motivated by data from papers and studies of the choice overload hypothesis. It more fully accounts for the complexity of choice overload data relative to two prior meta-analyses and thus provides richer insight. In particular, it shows that choice overload varies substantially as a function of the six dependent measures and four moderators examined in the domain and that there are potentially interesting and theoretically important interactions among them. It also shows that the various dependent measures have differing levels of variation and that levels up to and including the highest (i.e., the fifth, or paper, level) are necessary to capture the variation and covariation induced by the nesting structure. Our results have substantial implications for future studies of choice overload.

Symposium 15: 2:50 PM – 4:20 PM

Symposium 15: Psychometrics for NextGen Educational Assessments
Chair: Alina von Davier, ACTNext

Symposium 15a: Network Time: Network Models for Response Time and Accuracy
Gunter Maris, University of Amsterdam & Cito

We demonstrate how the hierarchical models of Van der Linden (and collaborators), the diffusion type process models of Tuerlinckx (and collaborators) and the scoring rule based models of Maris (and collaborators) can be cast as network models. In the framework of network psychometrics, accuracy and time are conceived of as separate nodes, and statistical dependencies are represented by edges between them. Even though the three classes of models for response time and accuracy are conceptually distinct, they share a common network representation in which the differences between the models is coded in the values (function of response time) the response time nodes can take. The network perspective offers a unified framework, which can easily be extended to capture more intricate statistical dependencies (beyond one dimensional constructs) and more diverse sources of information (beyond accuracy and time).

Symposium 15b: A Random-Effects Extension of the Hawkes Process
Peter Halpin

Past research has shown that the Hawkes process can provide a useful framework for modeling process data obtained from technology-enhanced assessments. Two challenges that arise in this setting are (a) to apply the model to many examinees simultaneously, and (b) to accommodate situations in which each examinee provides a relatively short time series (~ 100 events). To this end, a modification of the Hawkes process is proposed, in which the response intensity parameter (branching ratio) is treated as a time-invariant random effect over examinees. The resulting model is similar in nature to "frailty models" from survival analysis, although the proposed estimation procedure is based on the likelihood for a general marked point process. A number of approaches to modeling the response kernel of the Hawkes process are also considered, as well as appropriate distributions for random effects in the multivariate case. The utility of the new model is illustrated with an application in which pairs of examinees collaborate via online chat.

Symposium 15d: The Estimation of the Nonlinear Mixed-Effects Continuous-Time Models
Lu Ou, The Pennsylvania State University; Sy-Miin Chow, The Pennsylvania State University

The increased popularity of irregularly spaced intensive longitudinal measurements presents a need for continuous-time models (i.e., differential equation models) that may be nonlinear in form and include mixed effects in key dynamic parameters of interest. For fitting such models, innovative estimation approaches have been proposed but have not been examined in a systematic way. To examine the existing and potential methods for estimating mixed-effects nonlinear continuous-time models, the current study implements and improves the accuracy, robustness, and computational efficiency of the Continuous-Discrete Extended Kalman Filter (CDEKF) approach to fitting a nonlinear stochastic differential equation model of emotions to data from the Affective Dynamics and Individual Differences study; and performs a Monte Carlo simulation study to evaluate the strengths and limitations of the CDEKF approach. Results from the simulation study and empirical application are used to offer practical suggestions on ways to use the CDEKF approach and the associated continuous-time models to answer substantive questions that are otherwise cumbersome to test within a discrete-time framework.

Symposium 15e: NLP and Data Mining Methods for Performance Assessments

Pravin Chopade, ACT Inc.; Jiangang Hao, Educational Testing Service; Mengxiao Zhu, Educational Testing Service; Steve Polyak, ACTNext

Performance tasks in virtual environments result in rich data about the test takers' behavior. The more realistic the task, the more difficult it is to identify meaningful signals in these data from the noise and the artifacts. Several approaches that have been considered in educational assessments have been inspired by the work conducted in the data mining and Natural Language Processing (NLP) communities. In this presentation several approaches are described: a data model for creating a structure for the log files; a visualization method using networks, and a scoring method using NLP and clustering methods. The methods will be illustrated with data from a collaborative task from ETS and a collaborative game from ACT.

CDM 4: 2:50 PM – 4:20 PM

CDM: Classification Accuracy and Misspecification

Chair: Chanho Park, Keimyung University

CDM 4a: Attribute Classification Accuracy Improvement: Monotonicity Constraints on the G-DINA Model

Jimmy de la Torre, The University of Hong Kong; Miguel A. Sorrel, Universidad Autónoma de Madrid

Cognitive Diagnosis Models (CDMs) are restricted latent class models developed to identify students' mastery and nonmastery of multiple attributes. A common indicator of reliability in CDM is Attribute Classification Accuracy (ACA). In this work, we explore the consequences of assuming an inappropriate model, and propose a new version of a general CDM, G-DINA, where a monotonic constraint is included. A simulation study is conducted to investigate how the ACA of monotonic G-DINA compares with those of G-DINA and other reduced CDMs. The comparison involves both calibration and validation samples. We also introduce the use of the Likelihood Ratio (LR) test to evaluate the appropriateness of imposing this nonlinear constraint. LR Type I error and power in this context is evaluated. For comparison purposes, the performance of AIC and BIC is also documented. Results show that the ACA of the monotonic G-DINA model is always better than that of the G-DINA model, and approaches that of the generating reduced CDMs. These differences were more pronounced in the validation sample indicating that the lack of parsimony of the G-DINA model affects the generalizability and suitability of the item parameter estimates across samples. The results also show that the LR test can be used to determine whether or not monotonicity can be assumed. Overall, this study finds that the appropriateness of the constrained version of the G-DINA model can be tested empirically, and its proper use (i.e., in situations where the true CDMs cannot be assumed) leads to improved ACA.

CDM 4b: Generalizing Indices of Classification Accuracy for Cognitively Diagnostic Assessments

Charles Iaconangelo, Pharmerit

An index of classification accuracy was recently proposed that, in addition to being relatively straightforward and fast to compute, estimates accuracy conditional on the latent class rather than marginalized to the test level. This can inform the practitioner of the effectiveness of the assessment in correctly classifying specific latent classes of interest. Additionally, weighting and summing over the latent classes returns an estimate of the test-level classification accuracy for any attribute distribution of interest. This is of particular significance because a key component of the validity argument is understanding how the classification accuracy generalizes across other examinee populations. A simulation study was designed to evaluate how well the index predicted test-level accuracy for samples drawn from different attribute distributions. First, the CDM was fitted to item responses from examinees drawn from a uniform attribute distribution. The index was computed, and then weighted according to the latent-class proportions of the higher-order attribute distribution, and summed. This value was compared to the empirical classification accuracy of the CDM under the higher-order distribution. The reverse scenario was also considered. Additionally, factors manipulated include sample size, test length, and item quality. Results suggested that using the proposed index to predict classification accuracy for a different attribute distribution led to estimates close to the empirical values under all but the least favorable test conditions. In 38 out of 48 total conditions, the proposed index predicted the classification accuracy within 0.03 of the empirical value. Detailed findings will be presented.

CDM 4c: The Effect of Minor Q-Matrix Misspecification in Cognitive Diagnosis Models
Jordan Prendez, University of Maryland, College Park

Cognitive Diagnosis Models (CDM) are an increasingly important class of models and techniques that are designed to provide richer information when compared to many other psychometric techniques. Previously, many different methods for understanding important human constructs have focused on providing a single score to describe an unobservable (i.e. latent) trait. Because these scores are often seen by stakeholders as providing insufficient information, there has been a strong push that tests provide more actionable and detailed information rather than information that is simply normative in nature. CDMs are meant to provide more granular information by demonstrating evidence of mastery or non-mastery of attributes (e.g., ability to perform inverse operations, or add three digit numbers etc.). Within CDMs, the Q-matrix is the mathematical link between item responses and those attributes measured within the assessment, and its correct specification is extremely important. Incorrect specification can lead to biased parameter estimates and, importantly, poor classification accuracy (Rupp & Templin, 2008; Kunina-Habenicht, Rupp, & Wilhelm, 2012). Previous research, however, has not examined the effect of ignoring (underspecifying) weak relationships between items and attributes. Because these relationships are weak and less apparent, the subject matter expert is more likely to ignore them. A simulation study is conducted to evaluate classification accuracy under different misspecification scenarios. This project will address two primary questions from a log-linear CDM framework: 1. At what strength does a loading lead to a decrease in classification accuracy when ignored (misspecified)? 2. How do the number of minor-misspecifications influence the classification accuracy rate?

CDM 4d: Identifying Poor-Fitting Items Using Limited-Information Statistics for CDM
Yan Sun, Rutgers, The State University of New Jersey; Kevin Carl Santos, University of the Philippines; Miguel A. Sorrel, Universidad Autónoma de Madrid; Jimmy de la Torre, The University of Hong Kong

The Pearson χ2 and G2 statistics are the most often used full information statistics for testing goodness-of-fit for latent variable models. However, they are effective only when all the expected frequencies are greater than 5. This requirement becomes difficult to achieve when the number of examinees is relatively smaller than the number of possible response patterns resulting in a sparse response contingency table. Recently, various alternatives using information from the lower margins have been developed (see, e.g., Christofferson, 1975; Bartholomew & Leung, 2002; Reiser, 1996; Maydeu-Olivares & Joe, 2005; Hansen, Cai, Monroe, & Li, 2016). By adopting the limited information model fit indices, this study aims to develop an algorithm that can eliminate poor-fitting items not due to Q-matrix or model misspecification but other reasons e.g. omission of latent attributes within the cognitive diagnosis modeling framework. Two simulations studies are conducted. First, the performance of various limited-information statistics such as log odds ratio of item pairs (Chen, de la Torre & Zhang, 2013), Christofferson's test, Bartholomew and Leung's test and M2 statistics will be compared with each other and with full information statistics in terms of type I error and power. Next, an algorithm is developed using previously selected methods to eliminate poor-fitting items. Various types of misfit are covered including 1) misspecification of Q-matrix, 2) omission/addition of latent attributes, 3) misspecification of CDMs, and 4) dependency of residuals.

CDM 4e: Investigating Differential Item Functioning under DINA and G-DINA Models

Yuan-Ling Liaw, University of Oslo; Phonraphee Thummaphan, University of Washington

The purpose of this paper is to explore how Cognitive Diagnosis Models (CDMs) with higher dimensions can be used to account for Differential Item Functioning (DIF) detection. CDM is different from traditional unidimensional Item Response Theory (IRT) models in that the latent traits (i.e., attributes) in CDM are multidimensional and binary. The outcome of a CDM analysis is a mastery profile for each examinee that indicates whether the examinee possesses each measured attribute. In the CDM context, hence, DIF occurs when examinees with the same mastery profile but from different groups (e.g., race/ethnicity or gender) have different probabilities of answering an item correctly. Few studies have investigated DIF within CDM and only focused on limited number of attributes. In practice, test lengths are different and the numbers of attributes vary. A simulation study is conducted to evaluate recovery of the CDM-based DIF. Data are generated from two different CDM models: DINA (Deterministic Inputs, Noisy "And"-gate) model G-DINA (Generalized Deterministic Inputs, Noisy "And"-gate) model. DINA is more restrictive but G-DINA allows different probabilities of success for various mastery profiles. We investigate how different test lengths and number of attributes, and the complexity of the Q-matrix impact DIF detection. Additional factors are manipulated including sample size, percentage of DIF items, and the size of DIF. Type I error rate, power as well as bias and Root Mean Square Error (RMSE) in the latent trait estimates are computed and reported.

IRT 6: 2:50 PM – 4:20 PM

IRT: Response Style and Fatigue

Chair: Kensuke Okada, Senshu University

IRT 6a: Use of the Mixed-Effects Location-Scale Model to Detect Inconsistent Responding

Deborah L. Bandalos, James Madison University; Donald Hedeker, University of Illinois at Chicago

In this study we examined the utility of the mixed-effects location-scale model (Hedeker, Mermelstein, Demirtas, & Berbarum, 2016) for instruments containing both positively and negatively keyed items. In addition to the IRT location, difficulty, and discrimination parameters, the location-scale model includes: (a) a scale parameter allowing for differences in items' levels of dispersion across the response categories, (b) a random scale effect allowing for differences in scale across respondents, and (c) an item-level scale discrimination parameter that indicates items' ability to differentiate respondents with different amounts of dispersion. When scales include both positively and negatively keyed items, some research participants respond in opposite ways to positively and negatively keyed items, as expected. However, others do not notice or are confused by the negative keying and provide similar responses to both types of keying. Because positively keyed items do not engender such confusion in respondents, we hypothesized that those items would yield lower scale (less variability in responses). Results supported this hypothesis: the average scale parameter value for positively and negatively keyed items were -.37 and +.63, respectively. In addition, negatively keyed items had higher scale discrimination (average of .94) than positively keyed items (average of .85), indicating that negatively keyed items are better at distinguishing consistent and inconsistent respondents. These results suggest that the mixed-effects location-scale model can be useful for detecting both respondents with inconsistent response patterns and items that can differentiate such respondents, as well as for instrument development and detection of respondents exhibiting various response styles.

IRT 6b: Detecting Rater Fatigue Using Change Point Analysis

Allison Ames, James Madison University; Nick Curtis, James Madison University; Madison Holzman, James Madison University

To make accurate inference about students' gains or proficiency, assessment instruments must be well-aligned with the student learning outcome and scores must be accurately assigned. For performance assessments, scoring is often accomplished via raters using a rubric. However, even well-developed rubrics have weaknesses. Limiting characteristics include inconsistent application of rubric scoring attributable to rater effects such as drift, leniency/harshness, and rater fatigue. Such rater effects can lead to low inter-rater reliability, decreasing the trustworthiness of scores. A novel method of scoring ethical reasoning essays has been introduced and piloted at our university. Using traditional rubrics, raters ideally use a series of implicit questions to arrive at a score. Making the process of asking questions more explicit requires raters to use the same thought process to arrive at their scores, but may be less cognitively demanding. Using an online system to facilitate skip/display logic and score assignment should also reduce burden and fatigue. To investigate whether the new, logic-question approach to rating essays reduces the burden and fatigue on raters, two methods will be employed. The same set of ethical reasoning essays were rated both by traditional scoring methods and also by the new method. Change point analysis will be used to indicate the point at which raters begin to experience fatigue in both methods (Sinharay, 2016), and then compared across methods. A burdensome scale, the NASA-TLX (Hart & Staveland, 1988), and focus group feedback are used to augment the findings from the change point analysis.

IRT 6c: Performance Decline in Educational Assessments: Different Mixture Modeling Approaches

Marit K. List, Leibniz Institute for Science and Mathematics Education; Alexander Robitzsch, Leibniz Institute for Science and Mathematics Education; Oliver Lüdtke, Leibniz Institute for Science and Mathematics Education; Olaf Köller, Leibniz Institute fo

In low-stakes educational assessments, test takers might show a Performance Decline (PD) on end-of-test items. In order to assess PD, mixture models have been proposed by Bolt, Cohen, and Wollack (2002), Yamamoto (1995), and Jin and Wang (2014). We show how these models can be extended to allow for multigroup comparisons. Using data from a German low-stakes mathematics assessment in Grade 5, we investigate school track differences in PD and examine how accounting for PD affects parameter estimates and the estimation of school track differences in proficiency. In addition, we assess the differences between the three mixture models. Our results show that accounting for PD has a small impact on parameter estimates and school track differences in proficiency. All three mixture models show similar results for item parameter and proficiency estimation. We discuss the differences between the mixture models with regard to their assumptions about the relationship between PD and item responses.

IRT 6d: An Acquiescence Model at the Interface of Psychometrics and Cognitive Psychology

Hansjörg Plieninger, University of Mannheim; Daniel W. Heck, University of Mannheim

Responses to questionnaire items are influenced not only by the target trait, but also by so-called response styles (i.e., individual preferences for specific response categories). Recently, Böckenholt (2012) proposed an item response model that allows to disentangle the target trait and two such response styles, namely, extreme and midpoint responding. The model is a multidimensional IRT model and, additionally, has the appeal that it can be represented with a psychologically meaningful tree-structure. We extend Böckenholt's idea to acquiescence, which is the tendency to agree with both regular and reverse-coded items. The novel mixture model builds on item response theory, multinomial processing tree models (from cognitive psychology), and Bayesian hierarchical modeling. Specifically, the new model assumes a mixture distribution of affirmative responses, which are either determined by the underlying target trait or by acquiescence. A simulation study was carried out to study properties of the Bayesian implementation. Furthermore, the model is illustrated using an empirical data set from personality psychology. The proposed approach is a viable alternative to existing acquiescence models. Furthermore, it is only one example of the framework that emerges at the interface of item response theory and multinomial processing tree models that is an interesting route for further developments and applications.

IRT 6e: A New Rasch Facets Model for Rater's Centrality/Extremity Response Style

Kuan-Yu Jin, The Education University of Hong Kong; Wen-Chung Wang, The Education University of Hong Kong

The standard Rasch facets model was developed to account for facets data, such as student essays graded by raters. The standard facets model accounts for only rater severity. In practice, raters may exhibit different levels of tendency of using middle or extreme scores in their ratings, referring to centrality/extremity response style. To achieve better measurement quality in facets data, it is desirable to consider rater's severity and centrality/extremity jointly. A new facets model is thus developed to represent rater's centrality/extremity by adding a weight parameter on thresholds for each rater to the standard facets model. Simulation results show that the parameters of the new facets model could be well-recovered; failing to consider centrality/extremity would deflate the ability difference between ratees. Two empirical examples are provided for illustration, along with the implication of the new model.

EN 1: 2:50 PM – 4:20 PM

Equating and Norming
Chair: Jorge Gonzalez, Pontificia Universidad Catolica de Chile

EN 1a: CEFI AdultTM: Generalized Additive Model vs. Generalized Linear Model Scoring

Vivian W. Chan, University of Waterloo; Gregory Gunn, Multi-Health Systems, Inc.; Gill Sitarenios, Multi-Health Systems, Inc.

Newer methods for producing test norms include accounting for non-parametric distributions. One of these methods is the generalized additive models for location, scale, and shape (GAMLSS; Ribgy & Stasinopoulos, 2005), and it could be used to normalize, standardize and smooth assessment scores by age. However, very few studies have applied the new method in psychometric assessments and examined its implication with modeling the relationship between raw scores and derived norm scores after scores are normalized by age. In this study, we compare a traditional method of scoring (Zachary & Gorsuch, 1998), which is based on the generalized linear model, with the newer GAMLSS method, which is based on the generalized additive models. Analyses were based on 1,660 adults who completed a self-report executive functioning assessment (CEFI AdultTM). As expected, the two methods yielded slightly different solutions for normalizing assessment scores by age. Moreover, both solutions required similar effort in hand-smoothing raw scores to derived norm scores to minimize discontinuities or "jumps" in scores across ages. Thus, for this CEFI AdultTM assessment, neither scoring approach appears superior to the other, such that the quality of the psychometric instrument is not compromised when either scoring approach is applied.

EN 1b: Uncertainty in Normed Test Scores Due to Sampling Variability

Lieke Voncken, University of Groningen; Casper Albers, University of Groningen; Marieke Timmerman

Test publishers usually provide Confidence Intervals (CIs) for normed test scores. These CIs reflect the uncertainty due to the unreliability of the tests. This implies that another source of uncertainty, namely due to sampling variability in the norming phase, is ignored in practice. To enable a fair positioning of the person under study relative to the norm population, it is important to account for both sources of uncertainty. Some methods to do so were proposed recently, but they are not applicable in the context of continuous norming and they are restricted to a certain distribution of test scores. We propose a method that is both applicable in continuous norming and very flexible in terms of the score distribution, using the Generalized Additive Models for Location, Scale, and Shape (GAMLSS; Rigby & Stasinopoulos, 2005) framework. We assessed the performance of this method in a simulation study, by examining the quality of the resulting CIs. We varied the procedure of estimating the CI, CI size, sample size, value of the predictor, extremity of the test score, and the procedure of estimating the variance-covariance matrix. The results showed that good quality of the CIs could be achieved in most conditions. We recommend test publishers to use this approach to arrive at CIs, and thus properly express the uncertainty due to both test unreliability and norm sampling fluctuations, in the context of continuous norming. Adopting this approach will help (e.g., clinical) practitioners to obtain a fair picture of the person assessed.

EN 1c: Can Artificial Intelligence Learn to Equate?

Tom Benton, Cambridge Assessment

Equating involves discovering a transformation of the score scale on one test form such that the transformed scores can be interpreted interchangeably with those on another test form. This research attempts to generate new approaches to classical (non-IRT) equating in the Non-Equivalent Anchor Test (NEAT) design using techniques from the Artificial Intelligence (AI) and machine learning community. Software enabling complex machine learning has become widely available in recent years and has been successfully employed to tackle a variety of difficult computational problems. For example, AI can learn to play the game of Go by playing against itself millions of times and learning which moves are likely to lead to victory in different situations. This research examined whether we can train AI to equate in a similar way. Specifically, we simulated a wide variety of equating scenarios with the NEAT design. For each simulation, we recorded a "true" equating function as a vector of values denoting the form Y score that should be associated with each possible form X score. We also saved input data in the form two cross-tabulations of the simulated raw scores on each form against the anchor test. Using data from many such simulations, we then trained a machine learning algorithm to reproduce the true equating functions as accurately as possible from the available cross-tabulations. Once trained, the resulting algorithm was then applied to perform equating against real (as opposed to purely simulated) data and displayed a superior performance to existing classical equating techniques.

EN 1d: Optimal Bandwidth Selection in Kernel Equating for Different Test Types

Gabriel Wallin, Umeå University; Jenny Häggström, Umeå University; Marie Wiberg, Umeå University

A common equating procedure is to match the scores from the separate test forms by their percentiles using the equipercentile equating transformation. All equipercentile equating methods need to address the issue of continuizing the discrete score distributions. Kernel equating uses kernel smoothing techniques for this purpose, and it involves selecting a smoothing parameter, the bandwidth, which determines the smoothness of continuized score distributions. This choice is critical for the equating transformation and hence the possibility to make fair comparisons between test-takers. There have been several suggestions on the optimal choice of bandwidth, both in kernel equating and in the general field of kernel density estimation. The aim of this study was to compare the existing bandwidth selection methods in kernel equating, and also to suggest a new way of selecting it. The comparisons were made using both real test data and a simulation study. The length of the tests and the distribution of the test scores were altered, and the methods were compared using both the equivalent groups design and the non-equivalent groups with anchor test design. The preliminary results suggest that the different bandwidth selection methods lead to different equated scores, especially in the end-points of the score scale. The new method shows potential in comparison with the existing alternatives, especially for skewed score distributions and in the upper tail of the score distribution.

REL 2: 2:50 PM – 4:20 PM

Reliability II

Chair: Niels Smits, University of Amsterdam

REL 2a: Investigating Psychometric Properties of Composite Scores with Multivariate Generalizability Theory

Qing Xie, The University of Iowa & ACT, Inc.; Yi-Fang Wu, ACT, Inc.; Xiaohong Gao, ACT, Inc.

Scores from test batteries are usually used for making high-stakes decisions for placement, admission, and certification. It is thus vital to develop appropriate composite score and make adequate classification decisions. Psychometric properties of composite scores via different weight combinations have been addressed with multivariate generalizability theory (G-theory) analyses (e.g., He, 2009; Jarjoura, Early, & Androulakakis, 2004; Moses & Kim, 2014; Powers & Brennan, 2009). Various sources of errors from populations, tests, and/or their interaction may affect the accuracy of variance components estimation, and further, have an impact on the estimates of composite score reliability as well as classification consistency and accuracy. So far few studies have focused on the invariance of weighting schemes in forming composite scores. To better understand how stable weighting schemes are against different examinee populations and test form characteristics warrants further exploration. This study investigates the psychometric properties of composite scores, with a focus on the invariance of four weighting schemes in forming composite scores across different examinee populations and test form characteristics. We use a test battery which contains three multiple-choice tests. Multivariate G-theory will be applied in estimation of composite score reliability, classification consistency, and classification accuracy. Factors of interest include weighting schemes, population distributions, and pairwise correlations between individual tests of a battery. Both empirical analyses and simulations from the multivariate G-theory model will provide useful information for practitioners to understand potential weighting issues for composite scores under various testing situations.

REL 2b: Composite Score Reliability Indices for Mixed-Format Tests that Contain Testlets

Won-Chan Lee, University of Iowa; Kuo-Feng Chang, University of Iowa; Mingqin Zhang, University of Iowa; Jaime Malatesta, University of Iowa

This study discusses and illustrates the complexities involved in computing composite score reliabilities for mixed-format tests that contain testlets. The caution concerning the use of potentially inflated item-based reliability estimates for testlet-based tests is not a new phenomenon (Anastasi, 1988; Guilford, 1936; Thorndike, 1951). However, research related to this topic has generally been confined to tests comprised of a single item type (i.e., either multiple-choice or free response) (DeMars, 2006; Hendrickson, 2001; Lee & Frisbie, 1999; Lee, 2000; Lee & Park, 2012; Wainer & Thissen, 1996). This study expands upon the existing literature by providing a thorough framework to address this issue using three prominent psychometric theories: classical test theory, generalizability theory, and Item Response Theory (IRT). The first part of the study explicitly outlines and compares the assumptions, flexibilities, and limitations of each theory, as they relate to reliability. Specific attention is given to how each theory defines true score and error score variance. In the second part of the study, a real data analysis is performed using two mixed-format exams that differ with respect to testlet composition and content area. First, the IRT assumption of local independence is examined by comparing within-testlet interitem correlations with between testlet interitem correlations as well as by using Chen and Thissen's (1997) χ2 index. Then, using each psychometric theory, several composite score reliability indices are computed and compared. A discussion follows regarding the appropriateness of each reliability index under certain situations and suggestions for future research are offered.

REL 2c: Reliability of Multistage Tests Using Generalizability Theory

Hyung Jin Kim, University of Iowa

As Multi-Stage Tests (MSTs) have become increasingly popular in recent years, test reliability concerns for MST are essentially identical to such concerns for tests in any other format. Based on the current literature, Livingston and Kim (2014) present an approach for estimating the MST reliability from the perspectives of Classical Test Theory (CTT), and van Rijin (2014) presents another method using Item Response Theory (IRT). This study proposes a new approach to measuring reliability for MSTs using Generalizability Theory (GT). Note that, for most operational tests, multiple tests are administered to examinees to cover different subjects (contents); those contents are fixed, not random. Nested within each content, there are multiple panels; in doing so, all examinees do not necessarily take the same routing module. Within each module, there are different paths that examinees can be routed based on their scores on a previous module. Therefore, from the GT perspectives, contents (c) are fixed, panels nested within contents (a) are random, and paths within panels (m) are also random. Therefore, the data collection design becomes p•×(m•:a•:c°). Note that, each examinee is associated with a single panel and a single path. However, when one or more facets have a single condition, the issue of confounded effects arises as Brennan (2017) addresses. Brennan (2017), then, presents an approach to reporting an interval for reliability-like coefficients and error variances. Using his approach, this study will find intervals for reliability-like coefficients and compare them to reliability estimated using CTT and IRT.

REL 2d: Rater Agreement Indices for Large-Scale Writing Assessments

Dongmei Li, ACT, Inc.; Qing Yi, ACT, Inc.; Benjamin Andrews, ACT, Inc.

Rater agreement is an important factor affecting the reliability and validity of test scores in performance testing, such as in writing assessments. However, as has been shown in numerous studies (e.g., Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Warrens, 2012; Zhao, Liu, & Deng, 2013), rater agreement indices can be hard to interpret. Interpreting rater agreement indices reported from large-scale writing assessments is even more challenging because agreement indices are often based on scores from different rater pairs from a large rater pool, or with one rater being artificial intelligence. In addition, the reliability of the final scores is also affected by other factors such as the number of prompts and the number of raters that contribute to the final scores. The purpose of this study is to compare a few commonly used rater agreement indices (percent of agreement, Kappa, quadratic weighted Kappa, and various types of rater correlations), together with a couple of more recently proposed indices (AC1 and

AC2; Gwet, 2014) under the framework of generalizability theory, by taking into consideration both variations among raters and variations among prompts. Assuming that rater scores are categorizations of underlying continuous variables, data were generated under various conditions (informed by real data) so the impact of factors such as true score distributions, rater and prompt pool characteristics, and the number of score categories can also be evaluated. This research provides guidance for the interpretation of both overall and conditional rater agreement indices from large-scale writing assessments.

REL 2e: Comparison of Scoring Methods for Multiple-Multiple-Choice Items

Sayaka Arai, National Center for University Entrance Examinations; Hisao Miyano, National Center for University Entrance Examinations

The Multiple-Choice (MC) format is the most widely used format in objective testing. The "select all the choices that are true" item, which also called as Multiple-Multiple-Choice (MMC) item, is a variation of the MC format, which has no instruction to indicate the number of correct choices. Although many studies have developed and compared scoring methods for the MMC item, the results have often been inconsistent. Whereas most scoring methods that have been developed are based on the number of choices correctly selected, Arai & Miyano (in press) have proposed new scoring methods based on the similarity or the degree of association between response patterns and key patterns. In this study, we consider some other scoring methods based on similarity measures, and clarify their scoring features. Conducting numerical simulation, we also examine the relationships between examinees' abilities (true score) and scores for each scoring methods, and illustrate that the response patterns with high scores are basically identical among the methods.

Symposium 16: 4:30 PM – 6:00 PM

Symposium 16: Networks in Psychology: Recent Developments

Chair: Laura Bringmann, University of Groningen

Symposium 16a: The Process Factor Model as a Framework for Network Building

Marieke Timmerman; Laura Bringmann, University of Groningen

A network built from intensive longitudinal data aims at expressing the dynamic relations between the observed variables. The network can be inferred with different regression-based models. The reduced-form vector autoregressive model of lag 1 (VAR(1)) is currently popular, but suffers from several limitations. To remedy the problem that white noise (e.g., measurement error) is not properly separated from innovation, a VAR(1) + White Noise (+WN) model has been proposed (Schuurman, 2016). To address the shortcoming that in a VAR(1) based network only unique direct effects of variables are represented, Bulteel et al. (2016) proposed to express shared effects via relative importance metrics derived from the VAR(1) model. We propose to use the process factor analysis (PFA) model (see Browne & Zhang, 2007) to integrate both improvements, while simultaneously enhancing the interpretability of shared effects. It will be shown that the PFA model can be seen as a general framework, covering the VAR(1) and VAR(1)+WN models for a single subject as special cases. It will be discussed how to use the PFA model to succinctly express shared and unique effects, and how this relates to using relative importance metrics. The approach is tested by means of a simulation study. The use of the PFA model as a basis for building a network, including model selection and interpretation, will be illustrated using an empirical example.

Symposium 16b: Inferring Changing Networks with Time Varying Vector Autoregressive Models

Laura Bringmann, University of Groningen; Emilio Ferrer, University of California, Davis; Ellen Hamaker, Utrecht University; Denny Borsboom, University of Amsterdam; Francis Tuerlinckx, KU Leuven

Humans are complex dynamic systems, whose emotions, cognitions, and behaviors constantly fluctuate and interact over time. Several methods have been used to study the interaction or dynamics of, for example, emotions over time. The models used typically assume stationarity, meaning that the dynamics, for example time-lagged relations, are invariant across time periods. However, this is generally an unrealistic assumption. Whether caused by an external (e.g., divorce) or an internal (e.g., rumination) event, dynamics are prone to change. When this is the case, there should not be a single network of the dynamics, but a movie representing the evolution of the network over time. We have developed a new data-driven model that can explicitly model the change in temporal dependency within an individual without pre-existing knowledge of the nature of the change: the semi-parametric time-varying vector autoregressive method (TV-VAR). The TV-VAR proposed here is based on the easy applicable and well-studied Generalized Additive Modeling techniques (GAM), available in the software R. Using the semi-parametric TV-VAR one can detect and model changing dynamics and thus changing networks for a single individual or system. The TV-VAR model is applied here to empirical data on daily feelings of positive affect (PA) from a single heterosexual couple.

Symposium 16c: Time, Dynamics and Psychology Using the Gaussian Graphical Model

Tiago Cabaço, Humboldt-Universität zu Berlin & International Max Planck Research School on the Life Course (LIFE); Sacha Epskamp, University of Amsterdam; Manuel Völkle, Humboldt-Universität zu Berlin; Florian Schmiedek, Max Planck Institute for Human Dev

The use of network models to study the dynamics of psychological variables over time has received an increasing attention over the past years. Alongside other reasons, the focus on the temporal dynamics of psychological variables is considered to be essential for understanding the mechanisms that underlie certain psychological phenomena. In this presentation I will introduce a recently developed method that allows for exploratory discovery of the relationships between psychological variables, derived from the well-known multi-level Vector-AutoRegression model (VAR). However, instead of interpreting the temporal coefficients as a directed network, this extension of the VAR model estimates a network of partial correlation coefficients - a Gaussian Graphical Model (GGM) – that represents the lagged relationships between variables. In addition to the temporal networks, this methodology also allows the estimation of contemporaneous and between-subjects networks – both in form of a GGM. During the course of this presentation I will discuss the assumptions that underlie the VAR model and the potential interpretations of the resulting network structures: temporal, contemporaneous and between-subjects. Moreover, using empirical data, I will showcase how the VAR model can be estimated through the R package mlVAR, and highlight how it can be used to study psychological mechanisms and their individual differences.

Symposium 16d: Transforming Mutations into Models: Inferring Causal Networks from Experimental Data

Jolanda J. Kossakowski, University of Amsterdam; Lourens Waldorp, University of Amsterdam; Han van der Maas, University of Amsterdam

In psychology, researchers are often interested in the interaction between certain variables, and more specifically, how variables influence eachother. Observational data may show possibly interesting interactions between variables, while experimental data, where specific variables are manipulated, can be used to discover possible cause-effect relations between variables. Current network methodology typically focuses on estimation interactions between observed variables; it remains difficult to estimate the unique influence that one variable may have on another. Also, there has not been done much research on how to combine data from manipulations with observational data when estimating a network. In this talk, we will present a methodology with which we can determine the unique influence of a manipulated variable onto other variables in a network setting. We will first review methodologies, which were introduced in the field of genetics. Then, we will demonstrate how independent, univariate manipulations are used to create a perturbation graph that shows unique, causal paths between variables. The accuracy of our metholodology is compared to those existing methodologies by means of a simulation study. Results show that one can accurately estimate the unique influence of one variable on another, by using the data with manipulations.

Symposium 16e: Estimating Cross-Source Relationships from Big Data Using Component- and Networks-Analysis

Pia Tio, University of Amsterdam & Tilburg University; Katrijn van Deun, Tilburg University; Lourens Waldorp, University of Amsterdam

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 (possible) 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 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. Here we present the results of a simulation study demonstrating the benefits of SNAC in estimating cross-source relationships from big data.

Lecture: 4:30 PM – 5:30 PM

Public Lecture: After the PISA 2000 Debacle in Switzerland: Lessons Learned and Measures Taken

Martin Tomasik, University of Zurich; Stéphanie Berger, University of Zurich
Chair: Carolin Strobl

Switzerland – a small country in Europe with eight million inhabitants and four different official languages – has always been proud of its federalistic educational system with 26 local school administrations and 26 different school curricula. This pride was ruffled after the OECD had published the results of the Programme for International Student Assessment (PISA) where Switzerland was ranked just above the average of all participating countries, especially for reading competencies. In the first part of this talk, we will introduce the peculiarities of the Swiss educational system, highlight the most important results of the PISA assessments, and introduce the measures taken as a consequence of the country's mediocre performance. One related measure was the launch of a regional initiative for the assessment of school performance in Northwestern Switzerland, which we will describe in more detail in the second part of our talk. Here, we will introduce the idea behind a large-scale computer-assisted system for summative and formative assessment that we are developing as part of this initiative. We will point to the methodological and practical challenges we have encountered, and introduce item response theory as the mathematical measurement model that stands behind the system. We will argue why it is not only useful but even necessary to have such a model, if one wants to provide sound feedback on the students' performance. Furthermore, we will also discuss possible developments of the system in the future. The entire talk will take an applied perspective and is meant to demonstrate the challenges and benefits that the application of psychometric theory creates for students, teachers, local educational authorities, item developers, programmers, and educational scientists.

Symposium 17: 4:30 PM – 6:00 PM

Symposium 17: Response Times for Measurement and Understanding

Chair: Paul De Boeck, Ohio State University & KU Leuven

Symposium 17a: Hidden Markov Mixture Modeling of Responses and Categorized Response Times

Dylan Molenaar, University of Amsterdam; Maria Bolsinova, Utrecht University & Cito; Jeroen Vermunt, Tilburg University

In item response theory, modeling the item response times in addition to the item responses may improve the detection of possible between- and within-subject differences in the process that resulted in the responses. For instance, if respondents rely on rapid guessing on some items but not on all, the joint distribution of the responses and response times will be a multivariate within-subject mixture distribution. Suitable parametric methods to detect these within-subject differences have been proposed. In these approaches, a distribution needs to be assumed for the within-class response times.

In this paper, it is demonstrated that these parametric within-subject approaches may produce false positives and biased parameter estimates if the assumption concerning the response time distribution is violated. In the present talk, an approach is presented based on the categorized response times which hardly produces false positives and parameter bias. In addition, the new approach has approximately the same power to detect within-subject differences in responses and response times as compared to the parametric approaches.

Symposium 17b: Detecting Aberrant Behavior and Item Pre-knowledge - Mixture Modeling vs. Residuals

Chun Wang, University of Minnesota; Gongjun Xu, University of Michigan; Zhuoran Shang, University of Minnesota

The modern web-based technology greatly popularizes computer administered testing, also known as online testing. When these online tests are administered continuously within a certain "testing window", many items are likely to be exposed and compromised, posing a type of test security concern. Besides, if the testing time is limited, another widely recognized aberrant behavior is rapid guessing, which refers to quickly answering an item without processing its meaning. Both cheating behavior and rapid guessing result in extremely short response time. This paper introduces a mixture hierarchical item response theory model, using both response accuracy and response time information, to help differentiate aberrant behavior from normal behavior. The model-based approach is compared to the Bayesian residual-based fit statistic in both simulation studies and two real data examples. Results show that the mixture model approach consistently outperforms the residual method in terms of correct detection rate and false positive error rate, in particular when the proportion of aberrance is high. In the presence of cheating behavior, the model-based approach is also able to correctly identify compromised items as a by-product.

Symposium 17c: Mixed Effects Modelling of Differences in the Speed-Accuracy Tradeoff Function

Frank Goldhammer, German Institute for International Educational Research; Ulf Kroehne, German Institute for International Educational Research; Merle Steinwascher, University of Mannheim

Response times provide valuable information about test-taking behavior and lend themselves to explain inter- and intra-individual differences in task success. The relation of response time to response accuracy is traditionally described by two conceptually different functions: The Speed-Accuracy Tradeoff Function (SATF) across multiple speed conditions relating the condition-average of response time to the condition-average of accuracy, and the Conditional Accuracy Function (CAF) within a test-taking condition describing the proportion correct conditional on response time.

This presentation focusses the SATF and applies an item response modeling framework for item response and response time data from multiple experimental speed conditions. The proposed SATF model is based on the Generalized Linear Mixed Modeling (GLMM) approach and is suitable for measures with a strong speed component. Among others, it can accommodate response time effects between conditions (i.e., SATF slope) as well as related person and item differences. The model provides new insights into test-taker characteristics; particularly, the SATF slope informs about how persons make a compromise between accuracy and time demands, and how fast information can be accumulated. The model is applied to data from a visual lexical decision task administered under conditions with item-level time limits ranging from slow to fast and no time limit at all. We will present results on the average SATF, its shape (linear and non-linear components), related individual and item differences, determinants of these differences, and the relation of differences in the person intercept and the SATF slope components to reading ability.

Symposium 17d: Interwovenness of Response Time and Response Accuracy in Cognitive Tests

Paul De Boeck, Ohio State University & KU Leuven

The relationship between response time and accuracy in cognitive tests is a complicated relationship. One popular and elegant approach is to focus on the relationship simultaneously across persons and across items based on a model with two correlated latent variables: ability and speed and with correlated item parameters for speed and accuracy. However, the model leaves within-person and with-item dependencies between response time and response accuracy out of consideration, while there is more and more evidence for such dependencies. The dependencies are potentially important for measurement reasons and they may throw light on theoretical issues. In the presence of dependencies ability and speed are confounded and cannot be easily separated for measurement purposes, but, because the dependencies appear to show individual differences, they may offer an opportunity to measure an extra dimension. A clue for the interpretation of the extra dimension is that shorter response times tend to be associated with higher accuracy. An evident explanation is that the cognitive capacity varies during the test (e.g., based on attention fluctuations) so that larger variations would lead to more negative dependencies. Interestingly, the dependencies cannot be explained by fast guessing or fluctuations of the speed-accuracy balance. A further exploration shows that the relationship is even more complicated. Based on a study with double-centered response times, the dependencies seem to be curvilinear: a rather short increase of accuracy with response time is followed by a longer decrease.

DIF 1: 4:30 PM – 6:00 PM

DIF: Country-Level and Bayesian

Chair: Edward Ip, Wake Forest School of Medicine

DIF 1a: Investigating DIF Using the IRT Projective Model

Terry Ackerman, ACT, Inc.; Edward Ip, Wake Forest School of Medicine; Shyh-Huei Chen, Wake Forest School of Medicine; Yanyan Fu, The University of North Carolina at Greensboro; Tyler Strachan, The University of North Carolina at Greensboro

Ackerman (1992) and Shealy and Stout (1993) demonstrated that one of the main causes of DIF is the multidimensional nature of tests interacting with groups of examinees who differ in their ability distributions on the invalid traits being measured. If the test were strictly unidimensional, which is probably unrealistic, or the nuisance dimensions could be marginalized out by integration, there would be no DIF. This conceptualization provides a natural link to projective IRT (Ip, 2010; Ip & Chen, 2012). In the projective IRT formulation the Multidimensional IRT (MIRT) model containing both valid and nuisance dimensions is projected onto the dominant dimension of primary interest, extracting the contamination of nuisance dimensions which could result in DIF. Part of this study will be based on simulated data in which two-dimensional data will be generated for a reference and a focal group that differ in their ability distributions on the second (nuisance). Both a two-dimensional calibration and a unidimensional calibration of the generated data will be done using FlexMIRT. Projective unidimensional IRT parameters will be computed from the estimated two-dimensional item parameters. The R package DifR (Magis, Beland, & Raiche, 2016) will be used to examine DIF using Raju's area method. DIF results for the two groups using the FlexMIRT's unidimensional 2PL estimated parameters and the estimated projective IRT parameters will be compared. Factors that will be examined include: percentage of DIF items, sample size, and correlation between traits. The study will be repeated using PISA data. Implications concerning validity and DIF will be discussed.

DIF 1b: A Multidimensional Testlet Response Model for DIF and DTLF Detecting

Dan Wei, Beijing Normal University; Danhui Zhang, Beijing Normal University; Hongyun Liu, Beijing Normal University

Testlet Response Models (TRM) have been applied to detect Differential Item Functioning (DIF) and Differential Testlet Functioning (DTLF) in much research (e.g. Beretvas & Walker, 2012; I. Peak, 2014; Ravand, 2015). However, previous research has been conducted under the condition that only a one-dimensional ability trait was modeled. The detection of both DIF and DTLF under the condition that multiple latent traits exist has not been addressed. This study aimed to serve two primary objectives: First, explore a multidimensional testlet response model for detecting both DIF and DTLF based on an extension of the MRCMLM (Multidimensional Random Coefficients Multinomial Logistic Model); Second, compare the new developed model with two previously established methods, DFIT and SIBTEST. A Monto Carlo methodology was used to compare different methods in terms of the power and type I error rate, with manipulation of the following factors: number of items, number of examinees, magnitudes of tesltets effect, proportion of DIF among items, difference between the mean group abilities, and level of DIF contamination of anchor items. It was discovered that the new proposed method outperformed the other two methods in reducing type I error rates, regardless of the number of items. In addition, the proposed method was much more efficient than DFIT and SIBTEST, when the difference between group abilities enlarged and the magnitudes of testlets effect increased.

DIF 1c: Penalized Conditional Likelihood: Improving Recent DIF-Detection Methods

Can Gürer, UMIT – The Health & Life Sciences University ; Clemens Draxler, UMIT – The Health & Life Sciences University

Recent developments in detection of Differential Item Functioning (DIF) include approaches like Rasch Trees (Strobl, Kopf & Zeileis, 2015), DIF Lasso (Tutz & Schauberger, 2015) and Item-focussed Trees (Tutz & Berger, 2016) that are able to handle metric covariates inducing DIF in comparison to well established methods, which usually have been restricted to categorical variables. Still each of these methods have downsides which shall be addressed with a new estimation method that mainly aims to combine three central virtues of the three approaches: the use of conditional likelihood for estimation, the incorporation of genuinely linear influence of covariates on difficulty of items and the possibility to detect different DIF types: Either certain items showing DIF, certain covariates inducing DIF, or certain covariates inducing DIF in certain items. Each of the recent methods mentioned lacks in two of these aspects. In this talk we will introduce a method for DIF detection, which firstly uses the conditional likelihood for estimation to tackle the problem of nuisance parameters (Andersen, 1970) combinded with an L1-penalization for variable selection, secondly is based on the DIF model used in DIF Lasso to include genuinely linear effects instead of approximation through step functions, and thirdly leaves the user the option to decide, which of the three DIF types they would like to investigate. The method will be described theoretically, the challenges in implementation will be discussed and performance of the approach illustrated presenting first results.

DIF 1d: Detection of DIF Based on Non-IRT Generalized Regression Approaches

Patricia Martinkova, The Czech Academy of Sciences; Adela Drabinova, Charles University & The Czech Academy of Sciences

Logistic Regression (LR) models have been widely used for detection of Differential Item Functioning (DIF) since work of Swaminathan and Rogers (1980). In this work we consider various extensions of LR models for DIF detection while accounting for other aspects of the model. These include guessing or non-attention parameters in dichotomous and nominal data, which can bring better precision as well deeper understanding of the variables examined. Some of these models have been accounted for within Item Response Theory (IRT) framework. We argue that presented non-IRT models fill a logical gap in methodology and as such are important for educational purposes. Simulation studies suggest good properties as well as advantages in case of smaller samples. Models are implemented within difNLR library in R and offered to wider audience through ShinyItemAnalysis web application.

DIF 1e: Permutation Tests of Measurement Equivalence with Binary and Ordinal Indicators

Terrence Jorgensen, University of Amsterdam; Benjamin A. Kite, University of Kansas; Po-Yi Chen, University of Kansas

Tests of measurement equivalence in confirmatory factor analysis are typically conducted by calculating the difference in chi-squared fit statistics between nested models with and without the invariance constraints. When using diagonally weighted least squares estimation for binary or ordered-categorical data, test statistics do not follow a chi-squared distribution. A popular solution is to apply a robust transformation to the chi-squared-difference statistic, e.g., using a scaling factor and shift parameter. Sass, Schmitt, and Marsh (2014) showed the Type I error rates under this robust transformation are still inflated. Inflated error rates can be even more severe for an individual model's fit statistic, particularly with asymmetric thresholds and small-to-moderate sample sizes (Bandalos, 2014), leading to frequent incorrect rejections of configural models. Permutation randomization is an alternative method for controlling Type I error rates by repeatedly shuffling the grouping variable and saving the test statistic. The resulting empirical sampling distribution is derived under the conditions of the observed data, p-values calculated from the permutation distribution can be robust under conditions in which the scaled-and-shifted chi-squared statistic is not. This study was conducted to evaluate tests of measurement equivalence with binary and ordinal indicators. We compare Type I error rates and power between permutation and robust transformation, and we reveal why some reports of inflated Type I rates may have resulted from overly restrictive models rather than the specific Monte Carlo conditions. We also recommend attention to certain software defaults that can greatly affect error rates and power.

IRT 7: 4:30 PM – 6:00 PM

IRT: Multidimensional and Equating

Chair: Dries Debeer, University of Zurich

IRT 7a: Issues in Applying Multidimensional Item Response Models

Karen Draney, University of California, Berkeley; Leah Feuerstahler, University of California, Berkeley; Kerry Kriener-Althen, WestEd; Diah Wihardini, University of California, Berkeley; Tian Xia, University of California, Berkeley

The Desired Results Developmental Profile (DRDP) has been used to assess children in infant-toddler, preschool, and kindergarten programs in California for over a decade. The system consists of 8 domains, including language, math/science, social/emotional and physical development. Each domain is represented by at least four items, using a teacher observational rating system with between 5 and 11 levels per item. The IRT model used in the current system includes several unidimensional domain-level scores and three multidimensional score sets, based on requirements of the Office of Special Education Programs. This calibration was performed on statewide data gathered in the 2014-15 school year. A new multidimensional model was desired for the five domains of school readiness as set out in the Race to the Top - Early Learning Challenge. In addition, performance criteria for kindergarten readiness and a clear path to achieve these were desired. This paper describes the issues that we dealt with as we developed a justifiable method to set the performance criteria, as well as methods for the multidimensional analysis, and for equating to earlier scoring systems. The new model was fit using the Multidimensional Random Coefficients Multinomial Logit (MRCML) model (Adams, Wilson, & Wang, 1991) and the delta dimension alignment technique (Schwartz & Ayers, 2010) to allow comparability between the metrics of the dimensions. Performance criteria cut points were chosen using a modification of the construct mapping technique (Draney & Wilson, 2000). Finally, equipercentile equating was used to equate old and new cut points.

IRT 7b: Small-Sample MIRT Calibration and Model Selection Using the Grit Data

Ji Seung Yang, University of Maryland, College Park; Monica Morell, University of Maryland, College Park; Hyojin Im, Seoul National University of Education ; Allan Wigfield, University of Maryland, College Park; Katherine Muenks, Indiana University

Recent developments in Multidimensional Item Response Theory (MIRT; e.g., Reckase, 2009) provide more options in choosing flexible item response models that reflect complex structures of psychological constructs. Accordingly, MIRT applications have become more common not only for large-scale assessment data (e.g., Programme for International Student Assessment or Patient Reported Outcomes Measurement Information System) but also for relatively limited data collected in research settings (e.g., Reise & Waller, 2009). However, the small sample size issue in MIRT can easily puzzle researchers with respect to convergence of the model solutions and model fit diagnostics because the typical asymptotic properties are not guaranteed. The purpose of this study is to illustrate the phenomena and demonstrate utility of a targeted simulation study (known as parametric bootstrap) in examining the structure of grit, a newly emerging construct that has gained substantial attention recently in education and psychology (Duckworth & Quinn, 2009). With the empirical data, using different model fit indices leads to divergent conclusions. A parametric bootstrap simulation helps explain the behaviors of likelihood-based model fit indices, and provides guidance to choose the best statistic when different statistics disagree. Preliminary results found that the power of Bayesian Information Criterion (BIC) to detect a true model is very low with a small sample size when the data generated from a bifactor model, while the Akaike Information Criterion (AIC) is more efficient with small sample size to detect a true data generating model.

IRT 7c: Misspecification of the MIRT Model in Higher-Dimensional Projective IRT

Tyler Strachan, The University of North Carolina at Greensboro; Edward Ip, Wake Forest School of Medicine; Yanyan Fu, The University of North Carolina at Greensboro; Shyh-Huei Chen, Wake Forest School of Medicine; Terry Ackerman, ACT, Inc.

Projective Item Response Theory (PIRT) is designed to allow comparison of a latent trait of interest across tests that may contain different mixes of multiple latent traits in their items. The PIRT approach requires first fitting a Multidimensional Item Response Theory (MIRT) model to the response data before projecting the MIRT onto a unidimensional IRT model. This study aims to explore how robust the results are in misspecification of the fitted MIRT model. This proposed study plans to show the importance of this underlying issue because if the PIRT model shows to be sensitive to the misspecified MIRT, the whole projective approach would require exact methods to get the MIRT model right. However, if this method shows to be robust to the misspecification of the MIRT, the approach is justifiable even when an approximate and/or misspecified MIRT model is used in the first step. In this study, data from the 3D-MIRT model will be fitted using a misspecified MIRT model (i.e. 2D-MIRT, 4D-MIRT), and then projected onto a specific dimension (i.e. θ_1). The PIRT model associated with the 3D-MIRT will be used as a comparison model. Various simulation conditions will be manipulated including: sample size, test length, number of latent dimensions, and association between latent dimensions. The RMSE will be used to assess recovery of both the estimated θ of the PIRT model and estimated θ_1 of the MIRT model to the true θ_1 of the 3D-MIRT model. Various model fit indices will be examined as well.

IRT 7d: Several MIRT Models for Equating of Testlet-Based Test Scores

Guemin Lee, Yonsei University; In-Yong Park, Korea Institute for Curriculum and Evaluation; Moonsoo Lee, Korea Institute for Curriculum and Evaluation; Euijin Lee, Korea Institute for Curriculum and Evaluation; Hyejin Kang, Yonsei University

Testlets, as the name implies, have been defined as smaller subsets of a larger test (Wainer & Kiely, 1987). A one-factor solution for the item responses of a testlet-based test may not entirely reflect the underlying structure of the data. One way to take testlet effects into account is to incorporate more dimensions or factors in addition to a general dimension. The residual dependence within testlets after controlling for the influence of the primary factor could be modeled by introducing additional latent dimensions or factors. Lee, Kolen, Frisbie, and Ankenmann (2001) were probably the first that investigated the effects of testlets in the context of IRT equating. Li, Bolt, and Fu (2005) provided a test characteristic curve linking method under the testlet response model. Lee and Brossman (2012) proposed observed-score equating procedures under a simple-structure MIRT framework for mixed-format tests without mentioning testlets. Recently, Lee et al. (2015) provided a bi-factor MIRT true-score equating for testlet-based tests and Lee and Lee (2016) developed a bi-factor MIRT observed-score equating for mixed-format tests. The main purposes of this study are to specify several plausible MIRT models, (a) bifactor model, (b) simple structure model, and (c) second-order model, for testlet-based tests and to evaluate relative appropriateness of those models in the context of equating. Equating methods using dichotomous IRT, polytomous IRT, and testlet response models investigated by previous studies are also implemented for the purpose of comparison with specified models in this study.

IRT 7e: Equity Property in MIRT Equating

Won-Chan Lee, University of Iowa; Stella Kim, University of Iowa; Jaime Malatesta, University of Iowa

Equity is often considered as one of the desirable properties in equating. The notion of equity as first introduced by Lord (1980) states that for a given true score, the observed score distributions on two forms should be the same after equating is performed. If the equity property is met, it is a matter of indifference to any examinee which form of a test is taken. Since Lord's notion of equity cannot be achieved perfectly in practice, it has been customary to consider the first two moments of the score distributions, which are called First-Order Equity (FOE) and Second-Order Equity (SOE), respectively. Evaluation of the equity property typically requires a psychometric model that involves the notions of true score or latent trait such as a strong true score model or an Item Response Theory (IRT) model. It has been found that using different psychometric models as a framework for evaluating equity could lead to substantially different conclusions about the extent to which equity holds for particular equating results. In the present study, FOE and SOE are considered in equating with various Multidimensional IRT (MIRT) models such as simple structure, bifactor, and full MIRT models. Some approaches are considered to deal with the complexities and issues associated with the multidimensional nature of data.

SEM 2: 4:30 PM – 6:00 PM

SEM: Estimation and Inference

Chair: Shaobo Jin, Uppsala University

SEM 2a: A Penalized Likelihood Method for Multi-Group Structural Equation Modeling

Po-Hsien Huang, National Cheng Kung University

In the past two decades, statistical modeling with sparsity became an active research topic in the field of statistics and machine learning. Recently, Huang (2014) and Jacobucci, Grimm, and McArdle (2016) propose sparse estimation methods for Structural Equation Modeling (SEM). Under both methods, users can flexibly specify which model parameters should be penalized and then obtain a final sparse estimate by choosing the penalty level. However, their methods are restricted to the case of single group analysis. The aim of the present work is to establish a Penalized Likelihood (PL) method for Multi-Group SEM (MGSEM). The proposed method decomposes the group model parameter into a common reference component and a group-specific increment component. By penalizing the increment components, the heterogeneity of parameter values across the populations can be efficiently explored since the null group-specific effects are expected to be shrunk to zero. An Expectation-Conditional Maximization (ECM) algorithm is developed to optimize the PL criterion. A real data example is illustrated to show how to use the proposed method to explore the pattern of partial invariance.

SEM 2b: How to Handle Outlying Observations in SEM: Deletion or Robust Methods?

Xin Tong, University of Virginia

Structural Equation Modeling (SEM) is a popular multivariate modeling technique in social and behavioral sciences because it allows for testing complex theories by modeling latent variables and measurement errors simultaneously. Among procedures developed for SEM estimation, Normal-distribution-based Maximum Likelihood (NML) has been widely used because it generates consistent and efficient parameter estimates. However, these good properties may not hold when data are not normally distributed. Unfortunately, empirical data are rarely normal and often contain outlying observations. In practice, researchers usually remove those observations prior to fitting a model to their datasets to avoid biased parameter estimates. The performance of this technique depends on the accuracy of outlying observation diagnostic methods. Previous literature has shown that no method can find outlying observations with complete assurance and methods may suffer masking or swamping problems. As an alternative, different robust methods have been developed and shown to outperform NML. However, are robust methods always better than outlying observation deletion? The purpose of this study is to systematically investigate the impact of removing multivariate outlying observations in SEM. The tradeoff between unbiased parameter estimates and statistical power are evaluated through a Monte Carlo simulation study under different conditions of sample size, proportion of outlying observations, geometry of outlying observations, and alpha level. Two robust approaches for nonnormal data are also applied as a comparison. At the end, we provide recommendations on under what circumstances we should delete multivariate outlying observations or use robust methods when nonnormality is suspected.

SEM 2c: Robust Confidence Intervals for Structural Equation Modelling with Improved Robustness

Paul Dudgeon, University of Melbourne

Large sample robust confidence intervals (CIs) in Structural Equation Modelling (SEM) programs such as Mplus, EQS, and lavaan are equivalent to classical Huber-White heteroscedastic-consistent CIs in linear regression (Arminger & Schoenberg, 1989; White, 1980, 1982). The latter is often denoted using the shorthand label HC0, because improved likeminded "sandwich" interval estimators were subsequently developed and labelled HC1, HC2, HC3, HC4m, and HC5 (MacKinnon & White 1985; Cribari-Neto & Da Silva, 2011). The enhanced robustness in these latter estimators was achieved by using leverage values for each case in the sample. ML estimation in structural equation modelling is typically conditioned on any exogenous predictors, and therefore the potential benefits from incorporating leverage into robust CIs are lost. This talk presents a new method for calculating CIs in SEM, when exogenous observed variables are present, that incorporates sample leverage values, thereby opening up the possibility of improving the performance of existing robust CIs being reported by SEM programs. The proposal is tested on a model containing three exogenous predictors and three latent variables in a Monte Carlo simulation design in which non-normality, model misspecification, sample size, and parameter strength are manipulated. The results demonstrate that the HC3-type and HC5-type CIs are notably more robust than the existing HC0-type CIs reported in current SEM programs, especially at smaller sample size and for greater non-normality. Additional robustness for CIs of variance parameters is achieved by using an interval transformation proposed by Browne (1982).

SEM 2d: Nonparametric Estimation of a Latent Variable Model

Tim Fabian Schaffland, Hector Research Institute of Education Sciences and Psychology; Augustin Kelava, Hector Research Institute of Education Sciences and Psychology; Michael Kohler, Technical University Darmstadt; Adam Krzyzak, Concordia University

In this talk we present a new nonparametric latent variable approach (Kelava, Kohler, Krzyzak, & Schaffland). In this approach the model is estimated without specifying the underlying distributions of the latent variables. In a first step, we fit a common factor analysis model to the observed variables. The main trick in the estimation of the common factor analysis model is to estimate the values of the latent variables in such a way that the corresponding empirical distribution asymptotically satisfies the conditions that uniquely characterize the distribution of the latent variables. The main condition is the independence of the latent variables from the error terms of the measurement models. In a second step, we apply suitable nonparametric regression techniques to analyze the relation between the latent variables in this model. Theoretical results (e.g., concerning consistency of the estimates) are briefly presented. Furthermore, the finite sample size performance of the proposed approach is illustrated by applying it to simulated data in three simulation studies.

SEM 2e: Two-stage Estimator for Item-Level Missing Data in Linear Regression

Lihan Chen, University of British Columbia; Victoria Savalei, University of British Columbia

In psychology, researchers often use scales composed of multiple items to measure underlying constructs. Missing data often occur at the item level. A typical approach for dealing with item-level missing data, called Available-Case Maximum Likelihood (ACML), is to use the mean of available items for each subject as the personal score. Another intuitive approach is to treat an entire score as missing if one or more item is missing, and applies Full Information Maximum Likelihood (FIML) at the scale level. Neither approach can always produce consistent estimates. A new maximum-likelihood based analytical approach, called the Two-Stage approach (TS), was recently developed as an alternative (Savalei & Rhemtulla, 2016). Stage 1 produces saturated estimates of means and the covariance matrix using FIML at the item level. Stage 2 fits the model and applies a correction to the test statistics. The original work showed this approach outperforming ACML and scale-level FIML in the context of structure equation models with parcels. Current work expands the findings to linear regression. A simulation study was conducted under various missing data mechanisms: missing completely at random, strong and weak linear Missing At Random (MAR), as well as nonlinear MAR. Informed by earlier findings (Mazza, Ender, & Ruehlman, 2015), we also contrasted the performance of each approach based on whether items in the composite scales have equal or unequal means and loadings. Data was analyzed under ACML, scale-level FIML, and TS. Recommendations to psychological researchers are made based on their performances.

CLU 1: 4:30 PM – 6:00 PM

Clustering

Chair: Hans-Friedrich Köhn

CLU 1a: Graph Theory Approach to Detect Test Collusion

Dmitry Belov, Law School Admission Council; James Wollack, University of Wisconsin-Madison

Test Collusion (TC) is a large-scale sharing of test materials or answers to test questions. There are many potential sources of shared information, including teachers, test preparation entities, the Internet, or even examinees collaborating during the exam. Because of the potentially large number of examinees involved, TC poses a serious threat to the validity of score interpretations; hence accurately identifying individuals involved in collusion is important. TC is expected to produce response similarity on common items for a group of examinees. Recently, cluster analysis and factor analysis have been applied to answer similarity data for purposes of detecting TC. The proposed approach operates similarly but applies graph theory methodology for purpose of identifying groups. A graph is built as a set of vertices connected by edges, with each vertex representing an examinee. For each statistically significant response similarity index, an edge is created between the vertices corresponding to those two examinees. This research will study cliques - subsets of vertices where each vertex is connected to all other vertices. Our primary focus will be on the utility of clique size as an index for detecting TC. For each measure of response similarity, clique size critical values will be computed via simulations. Varying amounts and magnitudes of TC will be simulated, and graphs will be built. All cliques with sizes exceeding the critical value will be identified and removed from each graph. Performance of the new approach will be analyzed for different similarity measures. The methods will also be applied to a real dataset.

CLU 1b: Cross Data Biclustering for Multiple Matrices of Different Sizes and Sources

JiYao Li, Osaka University; Kohei Adachi, Osaka University

We propose a novel clustering procedure named Cross Data Biclustering (CDBC) for exhibiting the latent linear structure across entirely different groups. It is performed for multiple data matrices of different sizes and sources, such as three matrices of 15 Chinese by 6 traits, 10 Japanese by 9 attributes, and 12 Swiss by 8 records respectively. In CDBC, different entities are clustered simultaneously in a cross-data manner, regardless of their original arrangements. The goal of CDBC is to modify each data set as a product of two unique membership matrices representing differences and one common center matrix identifying relationships. These parameters are estimated by an alternating least squares algorithm. Simulation studies show that our proposed method has an excellent recovery accuracy. Some real data examples are discussed in order to illustrate CDBC.

CLU 1c: New Features in Clustering Objects on Subsets of Attributes (COSA)

Maarten Kampert, Leiden University; Jacqueline Meulman, Leiden University; Jerome H. Friedman, Stanford University

Friedman and Meulman (2004) proposed "Clustering Objects on Subsets of Attributes" (COSA) to extract signal from the overwhelming noise in high-dimensional data settings. COSA is an unsupervised algorithm that outputs a "cluster-happy" dissimilarity matrix that one can use for subsequent analysis by a variety of proximity methods. In this talk we will address two topics. First, there are recent state of the art methods that claim to outperform COSA on finding clustering structures. It is strange, however, that these new methods have not applied COSA to their specific simulated and real data examples to obtain a complete comparison. Second, COSA has been improved lately. There is an optimimzation strategy for the tuning parameters, and an upgraded feature that concerns the clustering on targeted attribute distances. These recent developments have made COSA more powerful and versatile. We will compare the (old) default and the updated COSA with the other state of the art methods. The comparison is based on similated and real omics data sets. The COSA results on these data sets were obtained with the software package rCOSA in R (Kampert, Meulman, & Friedman, 2017), available for free at https://github.com/mkampert/rCOSA.

CLU 1d: SPARK: A New Clustering Algorithm for Obtaining Sparse and Interpretable Centroids

Naoto Yamashita, Osaka University; Kohei Adachi, Osaka University

Recent advances in data collection technology result in larger scale datasets and, at the same time, growing demand of efficient data-mining technique in order to correctly and sufficiently capture homogeneity that lies in such large datasets. K-means clustering is considered to serve this purpose;

nevertheless the resulting centroid matrix often is difficult to interpret: it cannot allow us to easily capture what each cluster stands for. In order to consider the difficulty, we propose a new K-means algorithm that produces sparse and thus interpretable centroid matrix called SPARK (SPARse K-means clustering). The sparse centroid matrix, which contains a number of zero elements, provides well-emphasized clusters that facilitate easy and coherent interpretation. Performances of the proposed algorithm are assessed by simulation studies and illustrated with real data examples.

UNF 1: 4:30 PM – 6:00 PM

Unfolding

Chair: Andries van der Ark, University of Amsterdam

UNF 1a: Dimensionality Misspecification in the Multidimensional Generalized Graded Unfolding Model (MGGUM)

James S. Roberts, Georgia Institute of Technology; Riesling Meyer, Georgia Institute of Technology; David R. King, Pacific Metrics Corporation

The MGGUM is a multidimensional generalization of a popular unidimensional item response theory model for unfolding responses to stimuli. The effects of underfitting or overfitting the dimensionality of proximity-based item responses on MGGUM parameter estimates has yet to be systematically studied, but limited evidence suggest that such effects depend on whether the items have simple or complex structure. Furthermore, anecdotal evidence suggests that overestimating the number of dimensions may lead to arch effects like those seen in correspondence analysis. Results of a simulation study which varies the dimensionality of the item responses, the dimensional structure of the items (simple versus complex structure), the number of MGGUM dimensions fit to the data, and the estimation method (MCMC for all parameters, MMAP estimates for item parameters followed by EAP estimates of person parameters, or MH-RM estimates of item parameters followed by EAP estimates of person parameters) will be examined. One and two-dimensional data will be generated and fit with either a one or two-dimensional MGGUM. For the two-dimensional data condition, the item responses will have either simple or complex structure. Parameter estimation for each data-model combination will be repeated with each of the three estimation methods. The accuracy and general characteristics of the model parameter estimates will be discussed.

UNF 1b: DIF and Violations of Local Independence in Unfolding Models

Giulio Flore, Leiden University; Willem Heiser, Leiden University; Mark de Rooij, Leiden University

In psychometric unfolding models the acceptance or rejection of an item depends only on the relative position of an individual with respect to an item on the latent trait dimension. These models are not as widely used as the dominance models prevalent in the Item Response Theory (IRT) literature. To date the only psychometric unfolding parametric models supported by software are GHCM, PARELLA and GGUM. This paper presents the results of a comparative analysis of simulated data assessing the sensitivity of each model to Differential Item Functioning and violations in Local Independence. The Marginal Maximum Likelihood/Non-Parametric Maximum Likelihood (MML/NPML) approach used by PARELLA appears to be fairly sensitive to violations of item functioning and local independence, insofar that the effect of the violations is not necessarily localized to specific items. In addition, violations lead to clear distortions in the distribution of the recovered person parameters. Conversely the MML / Expected A Posteriori (EAP) approach of GGUM proves to be more robust against these violations, and the Joint Maximum Likelihood Estimate (JMLE) approach of GHCM is in some circumstances superior to PARELLA's approach. These results suggests that MML approaches benefit from the use of parametric constraints. The flexibility of non-parametric approaches and their relatively good performance when there is no DIF or conditional dependence may be offset by excessive sensitivity to violations of these assumptions.

UNF 1c: Modified MH-RM for Efficient Estimation of the Multidimensional GGUM

David R. King, Pacific Metrics Corporation; James S. Roberts, Georgia Institute of Technology

The Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2010a, 2010b, 2010c) is an efficient method for estimating high-dimensional item response theory models. The stochastic imputation of person parameter estimates allows runtime to be a linear function of the number of dimensions in the model, in contrast to the exponential function observed with rectangular quadrature. The current study examined the performance of the MH-RM in the estimation of the Multidimensional Generalized Graded Unfolding Model (MGGUM; Roberts & Shim, 2010), a distance-based, unfolding multidimensional item response theory model for measuring person and item characteristics from graded or binary disagree-agree responses. Initial attempts to estimate the MGGUM with the MH-RM resulted in severe misestimation of item parameters, although estimation accuracy was markedly improved through modifications to the MH-RM. Namely, the Newton-Raphson step for updating item parameters was replaced with the L-BFGS-B method for constrained optimization (Byrd, Lu, Nocedal, & Zhu, 1995). Runtime and estimation accuracy of the modified MH-RM were examined through a parameter recovery study that varied test length (10, 20, or 30 items), sample size (1000, 1500, or 2000 persons), number of response categories (2, 4, or 6), dimensional structure of items (simple or complex), and dimensionality (2 or 3 dimensions). Furthermore, the practical utility of the method was explored through a real data analysis of facial affect responses. Results indicate that the modified MH-RM is an efficient method for estimating high-dimensional MGGUMs and that estimation accuracy is comparable to other commonly used methods.

UNF 1d: Unfolding IRT Models: Do They Always Fit When Expected?

Jorge N. Tendeiro, University of Groningen; Rob R. Meijer, University of Groningen

Item Response Theory (IRT) models are nowadays popular statistical tools in educational, psychological, and clinical assessment. However, the quality of the results based on IRT models crucially depends on attaining a minimum fit quality. Hence, it is important to assess which model best fits the data at hand. Two important classes of IRT models exist, namely the so-called cumulative and unfolding models. Unfolding models in particular are commonly suggested when measuring attitudes and preferences (as it is typically the case in clinical settings). However, little exploratory research has actually illustrated how well unfolding models outperform cumulative models in this type of context. Based on empirical data, we propose to see how well does the most popular unfolding model in use (the generalized graded unfolding model) fare against the more classical polytomous IRT cumulative models. We fit both cumulative and unfolding models to clinical questionnaires data. Results indicate that not always the most obvious model choice was the one that fit the data best. The most important implication is that practitioners need to be very careful when assessing model fit, in spite of what the apparent data structure at hand may suggest.

UNF 1e: Information Functions for the GGUM-RANK Multidimensional-Forced-Choice Model

Seang-Hwane Joo, University of South Florida; Phil Seok Lee, South Dakota State University; Stephen Stark, University of South Florida

Multidimensional Forced Choice (MFC) measures have been proposed to address concerns about response biases (e.g., socially desirable responding, central tendency, acquiescence) associated with Likert-type measures (Stark et al., 2012). To date, one of the more widely used Item Response Theory (IRT) models for MFC measures is the Multi-Unidimensional Pairwise Preference model (MUPP; Stark et al., 2005) due to its suitability for adaptive testing and validity evidence in field contexts. Stark et al. (2005) chose the dichotomous-type Generalized Graded Unfolding Model (GGUM; Roberts et al., 2000) for computing MUPP statement endorsement probabilities. Hontangas et al. (2015) later extended the MUPP to handle more complex MFC formats, such as triplets and tetrads. One extension, which we henceforth refer to as the GGUM-RANK, applies when examinees must rank the statements in an MFC item from most preferred to least preferred. Although Hontangas et al. (2015) conducted a simulation to examine GGUM-RANK trait recovery in some representative conditions, they did not develop information functions, which are helpful in constructing test forms that meet reliability targets. Thus, we derived GGUM-RANK item and test information functions for MFC triplets and tetrads then conducted an experiment to see how the statement parameters composing MFC pair, triplet, and tetrad measures jointly influence GGUM-RANK item and test information and scoring. In our presentation, we will review the GGUM-RANK model, information indices, scoring methods, and complete experimental findings. We will also discuss the implications of these results and make recommendations for constructing MFC measures to use in psychological contexts.

Friday, July 21, 2017

Symposium 18: 8:30 AM – 10:00 AM

Symposium 18: Big Data Analysis in Cognitive Assessment and Latent Variable Models

Chair: Jingchen Liu, Columbia University

Symposium 18a: Exploratory Item Classification via Spectral Graph Clustering

Yunxiao Chen, Emory University; Xiaoou Li, University of Minnesota; Jingchen Liu, Columbia University; Gongjun Xu, University of Michigan; Zhiliang Ying, Columbia University

Large-scale assessments are supported by a large item pool. An important task in test development is to assign items into scales that measure different characteristics of individuals and a popular approach is cluster analysis of items. Classical methods in cluster analysis, such as hierarchical clustering, the K-means method, and latent class analysis, often induce a high computational overhead and have difficulty handling missing data, especially in the presence of high-dimensional responses. In this talk, we propose a spectral clustering algorithm for exploratory item cluster analysis. The method is computationally efficient, effective for data with missing or incomplete responses, easy to implement, and often outperforms traditional clustering algorithms in the context of high dimensionality. The spectral clustering algorithm foots on graph theory, a branch of mathematics that studies the properties of graphs. The algorithm first constructs a graph of items, characterizing the similarity structure among items. It then extracts item clusters based on the graphical structure, grouping similar items together. The proposed method is evaluated through simulations and an application to the revised Eysenck Personality Questionnaire.

Symposium 18b: Using MMCAR to Explore the Structure of Personality and Ability

William Revelle, Northwestern University

Personality and ability item pools can be very large (> 5,000) but no one person likes to answer more than 50-150 items. Using web based "Synthetic Aperture Personality Assessment" at sapa-project.org we collect data using a Massively Missing Completely at Random (MMCAR) design. Items are taken from the open source International Personality Item Pool (http://ipip.ori.org) as well as the International Cognitive Ability Resource (icar-project.org). As an incentive for subjects to participate, we offer personality feedback based upon 5 and 27 factor models. Approximately 50,000 subjects participate per year. 1,000 x 1,000 covariance matrices based upon 250,000 subjects using pairwise covariances have roughly 1 - 2,000 observations/pair. Using conventional covariance algebra and R functions in the psych package, we can examine the joint structure of personality, ability, and interests. Although standard errors of individual correlations reflect the pairwise sampling based N, because the pairs are given with a MMCAR design, standard errors of composite scores reflect much larger effective sample sizes. The advantages of a SAPA/MMCAR design compared to the more conventional use of short forms will be discussed. In particular, by using large samples with many items, the structure of personality can be examined at many levels of resolution, from the conventional 3-5 factors to our preferred 27 homogeneous scales, down to the unique (but stable) correlates of individual items.

Symposium 18c: Optimal Item Selection and Stopping for Computerized Adaptive Testing

Xiaoou Li, University of Minnesota

Computerized Adaptive Testing (CAT) is a form of test that is tailored to an examinee's latent trait. When designing an efficient CAT, one usually faces three challenges: 1) how to select the next item based the examinee's previous responses? 2) for a variable length test, when to stop the test to balance the accuracy of inference and sample size? 3) how to evaluate the performance of a CAT? In this talk, I will formulate the problem of assessing the performance of a CAT through the theory of Markov Decision Process (MDP). I will also present some new results regarding the optimal item selection and stopping.

Symposium 18d: Analysis of Local Dependence for Latent Variable Models

Jingchen Liu, Columbia University

Latent variable models and latent class models take advantage of the fact that the dependence of a high-dimensional random vector is often induced by just a few latent (unobserved) factors. When the number of dimensions grows higher and the dependence structure becomes more complicated, it is hardly possible to find a low-dimensional parametric latent variable model that fits well. In this talk, I present several real data examples in education, political science, psychology, and finance, for which we include a graphical structure to capture the local dependence that cannot be explained by the low-dimensional latent structure.

DIF 2: 8:30 AM – 10:00 AM

DIF: Country-Level & Bayesian

Chair: Steven Culpepper, University of Illinois at Urbana-Champaign

DIF 2a: Marginal Measurement Invariance Testing Using Bayes Factor

Jean Paul Fox

A new Bayesian method to detect measurement invariance violations is proposed. The method is based on a marginal multilevel IRT model, where random item effects are integrated out. This marginal IRT model accounts for mean differences, where a violation of measurement invariance reveals itself as a positive correlation between group-specific item responses. A fractional Bayes factor is derived to test this additional positive correlation using (noninformative) improper priors. Although the test is designed for cross-national survey studies, group-specific item parameters are no longer estimated. Multiple measurement invariance hypotheses can be tested simultaneously without the need for anchor items. Furthermore, additional identification restrictions, as for the random item effects model, are not required. The test can be applied to randomly selected groups as well as a fixed number of groups. The proposed tests can quantify evidence in favor of a null hypothesis, representing, for instance, full measurement invariance. The test method is consistent and will select the true covariance structure with probability one, when the sample size goes to infinity. Furthermore, the method is based on observed data and does not depend on large sample theory. The quantification of the relative data evidence between two possible IRT models is accurate for any sample size. It is shown that the method can also be applied to the two-parameter IRT model. For the two-group situation, a comparison is made with the Mantel-Haenzsel test. The method is illustrated using data from the European Social Survey for a two-group and multi-group situation.

DIF 2b: Using Multilevel Logistic Regression to Detect Item Difficulty Variance Between Random Groups

Johannes Hartig, German Institute for International Educational Research; Carmen Köhler, German Institute for International Educational Research; Alexander Naumann, German Institute for International Educational Research

In cross-national educational assessments, item difficulties are typically assumed to be invariant across countries. This assumption will often be violated, meaning that item difficulties differ between groups. We will refer to this violation of measurement invariance as Random Group Differential Item Functioning (RG-DIF). Variances and covariances of item difficulties on group level can be estimated with Generalized Linear Mixed Models (GLMMs) with responses nested in individuals nested in groups. This is computationally intensive, as the models include a large number of correlated random effects. The present study examines the suitability of Two-Level Logistic Regression (TL-LR) as a screening method for the presence of RG-DIF. In a simulation study, item responses of individuals nested in groups were generated. The magnitude of MG-DIF was varied between items, and the covariance of the item difficulties on group level was varied between conditions. In TL-LR, MG-DIF was estimated as the between-group variance in item responses after controlling for the estimated trait score. Additionally, item difficulty variances and covariances were estimated by means of a three-level GLMMs. When group level effects are independent, TL-LR recovers variances on group level very well. For items with correlated item effects, variances on group level are underestimated, since common variance on the group level is partialled out by the estimated trait score. TL-LR can serve as an easily accessible method to screen for MG-DIF. However, the more comprehensive analysis with a three-level model seems advisable to obtain more detailed information about the structure of item variation on group level.

DIF 2c: Model Comparison in Bayesian Item Response Models for Anchoring Vignettes

Kensuke Okada, Senshu University; Daiki Hojo, Senshu University

In this study, we perform a comparison of Bayesian item response models for anchoring vignette data. In the dataset we used - the Survey of Health, Ageing and Retirement in Europe (SHARE) - respondents in eight countries answered both self-rated and vignette items in seven health domains. We had two main objectives in modeling this dataset: First, we aimed to determine whether the respondents' response style was consistent irrespective of health domain. Second, we aimed to determine whether there were country-specific differences in response style. For these purposes, we used a fully Bayesian approach. Specifically, we built four models with differing complexities and then evaluated their fit. The four models could be summarized using a 2 by 2 table: constrained vs. unconstrained, and hierarchical vs. non-hierarchical. The constrained models assumed that the response style of a respondent was consistent across the health domains, whereas the unconstrained models assumed that it varied. By contrast, the hierarchical models assumed that there are two levels of variability (respondents and countries) in response style, while the non-hierarchical models assumed no country specific differences. We compared the fit of the models using both posterior predictive checking and various other relative measures of model fit. The psychometric implications of the results are discussed.

DIF 2d: A Hierarchical Bayesian Framework for Trait Measurement When Items Might Have DIF

Matthew Zeigenfuse, Fordham University; Carolin Strobl, University of Zurich

In many applications, measuring a latent trait from test responses typically comprises two steps. First, we check for variability in the properties of the test items across one or more covariates, i.e., Differential Item Functioning (DIF). We then estimate individuals' trait values using the DIF-free items. This approach has two difficulties. The first is the selection of anchor items. In many cases, we have no a priori reason to assume that a given item is invariant, and selecting the wrong items as anchors can lead to poor performance in the subsequent item-wise DIF tests. The second is that, once item-wise DIF tests have been performed, the items whose tests are not significant are simply assumed to be invariant and items whose tests are significant are tossed out. This ignores any residual uncertainty we have about whether a given item has DIF when estimating the test takers' scores on the latent trait. In the case of uniform DIF, we can employ hierarchical Bayesian modeling techniques to estimate the latent trait without having to select anchor items or condition on a set of DIF-free items. Specifically, we employ an explanatory IRT framework wherein related covariates, such as gender or nationality, are modeled hierarchically. Preliminary results indicate that this approach may allow accurate estimation of the person parameter as well as easily interpretable effect size estimates for DIF.

DIF 2e: A Monte Carlo Study on Bayesian Measurement Invariance and Non-Invariance

Deana Desa, IEA Hamburg

This Monte Carlo study aims at identifying the potential practical and operational benefits of using a Bayesian approach in Measurement Invariance and Non-Invariance evaluation (MINI). The focus of this simulation study is to fully account uncertainties related to MINI models and parameter values within the context of large-scale assessment programs. To facilitate the designs implemented in many large-scale assessment programs, where samples are largely heterogeneous, we examine several conditions include the number of groups (i.e., 10, 20, 40), group sizes (i.e., 100, 500, 1000), and different constraints on the prior means and variances of the MINI model parameters. We will use the information from the existing data sets (i.e., the previous two test cycles) to determine the sensitivity to the choices of our priors. Each simulation condition will be repeated 50 times. Parameter recovery of MINI models will be evaluated by the correlation between the true and estimated values, bias and root-mean-square error. In addition, the study will monitor the computing capacity that is needed for implementing Bayesian approach. Results of the study will give us the potential conditions for accurate parameter estimates in MINI modelings and its practicality for large-scale assessment settings.

SEM 3: 8:30 AM – 10:00 AM

SEM: Fit and Model Selection
Chair: Jouni Kuha, The London School of Economics and Political Science

SEM 3a: Assessing Fit in Structural Equation Models: A Monte-Carlo Evaluation of RMSEA vs. SRMR Confidence Intervals and Tests of Close Fit
Alberto Maydeu-Olivares, University of South Carolina; Dexin Shi, University of South Carolina; Yves Rosseel, Ghent University

We compare the accuracy of confidence intervals (CIs) and tests of close fit based on the RMSEA with those recently proposed based on the SRMR. Investigations used normal and non-normal data with models ranging from p = 10 to 60 observed variables. CIs and tests of close fit based on the SRMR are generally accurate across all conditions (even at p = 60 with non-normal data). In contrast, CIs and tests of close fit based on the RMSEA can only be used with the smallest models considered (p = 10). CIs and tests for the RMSEA outperformed the SRMR only in small models (p = 10) where the degree of misspecification is small (SRMR ≤ 0.02).

SEM 3b: Confidence Intervals of Fit Indices by Inverting a Bootstrap Test
Chuchu Cheng, Boston College; Hao Wu, Boston College

Fit indices are an important tool in the evaluation of model fit in Structural Equation Modeling (SEM). Currently, the newest proposed confidence interval (CI) for fit indices proposed by Zhang and Savalei (2016) is based on the quantiles of a bootstrap sampling distribution at a single level of misspecification. This method, despite a great improvement over naive and model based bootstrap methods, still suffers from unsatisfactory coverage. In this work, we propose a new method of constructing bootstrap Cis for various fit indices. This method directly inverts a bootstrap test and produces a CI that involves levels of misspecification that would not be rejected in a bootstrap test. Similar in rationale to a parametric CI of RMSEA based on a non-central chi-square distribution and a profile-likelihood CI of model parameters, this approach is shown to have better performance than the approach of Zhang and Savalei, with more accurate coverage and more efficient widths.

SEM 3c: Corrected Goodness-of-Fit Test in Covariance Structure Analysis
Kazuhiko Hayakawa, Hiroshima University

Many previous studies report simulation evidence that the goodness-of-fit test in covariance structure analysis or structural equation modeling suffers from the over-rejection problem when the number of manifest variables is not small compared to the sample size. In this study, we first investigate the reason and show that the test statistic involves a bias term that gets larger as the number of manifest variables grows. Then, we propose a test statistic by subtracting that bias term. Incidentally, it is demonstrated that the corrected test coincides with one of the tests considered by Amemiya and Anderson (1990). We also propose a simple modification of Satorra and Bentler's mean and variance adjusted test for non-normal data. A Monte Carlo simulation is carried out to investigate the performance of the corrected test in the context of a confirmatory factor model and a cross-lagged panel (panel vector autoregressive) model. The simulation results reveal that the corrected test overcomes the over-rejection problem and outperforms existing tests in most cases.

SEM 3d: New Testing Procedures for Structural Equation Modeling
Njål Foldnes, BI Norwegian Business School; Steffen Grønneberg, BI Norwegian Business School

We introduce and evaluate a new class of hypothesis testing procedures for moment structures. The methods are valid under weak assumptions and includes the well-known Satorra-Bentler adjustment as a special case. The proposed procedures applies also to difference testing among nested models. Also, we introduce a bootstrap selection mechanism to optimally choose a p-value approximation for a given sample. Bootstrap procedures for assessing the Asymptotic Robustness (AR) of the normal-theory maximum likelihood test, and for the key assumption underlying the Satorra-Bentler adjustment (Satorra-Bentler consistency) are presented. Simulation studies indicate that our new p-value approximations performs well even under severe nonnormality and realistic sample sizes, but that our tests for AR and Satorra-Bentler consistency require very large sample sizes to work well.

IRT 8: 8:30 AM – 10:00 AM

IRT: Equating

Chair: Marie Wiberg, Umeå University

IRT 8a: Incorporating Information Functions in IRT Scaling

Alexander Weissman, Law School Admission Council

Item Response Theory (IRT) scaling via a set of items common to two test forms assumes that those items' parameters are invariant to within a linear transformation. Characteristic curve methods rely on this assumption; scale transformations are conducted by minimizing a loss function between test characteristic curves, as in the case of Stocking-Lord (1983), or item characteristic curves, as in the case of Haebara (1980). In practice, however, minimizing the loss function between characteristic curves does not guarantee that the same will hold for information functions. This study examines differences in information functions that can arise after characteristic curve scaling, and proposes new methodologies for incorporating information functions into IRT scaling.

IRT 8b: Reducing Conditional Error Variance Differences in IRT Scaling

Tammy Trierweiler, Law School Admission Council; Charles Lewis, Fordham University; Robert L. Smith, Smith Consulting

In Item Response Theory (IRT), when item parameters are estimated for different forms using data based on two different groups of test takers they need to be placed on a common scale. A standard wayto put new form estimates onto a reference form scale using a NEAT (Non-Equivalent groups Anchor Test) design is to use the item parameter estimates for the common items to estimate A and B transformation constants. These constants are used to transform the new form parameter estimates to place them onto the reference scale. Although there are a number of methods that can be used to estimate the A and B transformation constants used in the linear scaling process, one of the most popular methods is the Stocking-Lord Test Characteristic Curve (TCC) method. In this study, we propose a hybrid scaling procedure based on the Stocking-Lord (1983) method that takes into account common item conditional error variances in addition to TCC differences to estimate transformation. Monte Carlo simulations are used to evaluate the performance of this hybrid scaling method across several different sample sizes, distributional conditions and form conditions.

IRT 8c: Bayesian Ability Estimation in IRT Equating with NEC Design

Valentina Sansivieri, University of Bologna; Mariagiulia Matteucci, University of Bologna

We use test score equating to compare different test scores from different test forms. When the Equivalent Groups (EG) design cannot be used, although the optimal choice would be to use the Non-Equivalent groups with Anchor Test (NEAT) design, it is often impossible to administer an anchor test due to several possible reasons (for example, test security). A possibility, then, is to use Non-Equivalent groups with Covariates (NEC) design. The overall aim of this work is to propose the use of Item Response Theory (IRT) equating with a NEC design (Sansivieri & Wiberg, 2017) by estimating abilities through Bayesian Markov chain Monte Carlo (MCMC) with and without covariates. In particular, we want to show that the standard error of equating is lower under the Bayesian approach than under a classical approach where the abilities are estimated through marginal maximum likelihood. The proposed approach is examined with both simulations and real empirical data. The results are compared with IRT observed-score equating methods using the EG and NEAT designs. The results from the simulations show that the standard errors of the equating are lower when we use Bayesian MCMC to estimate abilities. One real test dataset illustrates how the proposed approach to estimate abilities can be used in practice.

IRT 8d: An Alternative View on the NEAT Design in Test Equating

Jorge Gonzalez, Pontificia Universidad Catolica de Chile; Ernesto San Martin, Pontificia Universidad Catolica de Chile

The Non-Equivalent groups with Anchor Test design (NEAT) is widely used in test equating. Under this design, two groups of examinees are administered separate test forms with each test form containing a common subset of items. Because test takers from different populations are assigned only one test form, missing score data emerge by design rendering some of the score distributions unavailable. The equating literature has treated this problem from different perspectives all of them making different assumptions in order to estimate the missing score distributions. In this paper, we offer an alternative view for the estimation of the equating transformation under a NEAT design that is free of these types of assumptions. Comparisons with current methods are made and illustrated.

IRT 8e: Simultaneous Equating of Multiple Forms

Michela Battauz, University of Udine

When test forms are calibrated separately, Item Response Theory (IRT) parameters are not comparable because they are expressed on different measurement scales. The equating process converts the item parameter estimates to a common scale and provides comparable test scores. Various statistical methods have been proposed to perform equating between two test forms. However, many testing programs use several forms of a test and require the comparability of the scores of every form. To this end, Haberman (2009) developed a regression procedure that generalizes the mean-geometric mean method to the case of multiple test forms. A generalization to multiple test forms of the mean-mean, the Haebara, and the Stocking–Lord methods was proposed in Battauz (2016). In this paper, the asymptotic standard errors of the equating coefficients were derived for all the methods. These methods simultaneously estimate the equating coefficients that permit the scale conversion of the parameters of all forms to the scale of the base form. After the estimation of the equating coefficients, it is possible to determine comparable test scores using the methods available in the literature, as the true score equating method or the observed score equating method. In this talk, the new methods for the estimation of the equating coefficients will be illustrated and their performance will be explored by means of a simulation study. Results show that all the methods give nearly unbiased estimates of the coefficients and their standard errors.

CDM 5: 8:30 AM – 10:00 AM

Complex Dynamic Models

Chair: Peter Halpin

CDM 5a: Hierarchical Bayesian Continuous Time Dynamic Modelling

Charles Driver, Max Planck Institute for Human Development; Manuel Völkle, Humboldt-Universität zu Berlin

Continuous time dynamic models are similar to popular discrete time models such as autoregressive cross-lagged models, but through use of stochastic differential equations they can accurately account for differences in time intervals between measurements and specify complex dynamics more parsimoniously. As such they offer powerful and flexible approaches to understand ongoing psychological processes and interventions, and allow for measurements to be taken a variable number of times, and at irregular intervals. However, limited developments have taken place regarding the use of continuous time models in a fully hierarchical context, in which all model parameters are allowed to vary over individuals. This has meant that questions regarding individual differences in parameters have had to rely on single-subject time series approaches, which require far more measurement occasions per individual. We present a hierarchical Bayesian approach to estimating continuous time dynamic models, allowing for individual variation in all model parameters. We also describe an extension to the ctsem package for R, which interfaces to the Stan software and allows simple specification and fitting of such models. To demonstrate the approach, we use a subsample from the German socio-economic panel and relate overall life satisfaction and satisfaction with health.

CDM 5b: Bayesian Nonparametric Dynamic Item Response Model with Shape Constraints

Yang Liu, University of Connecticut

Parametric methods, such as autoregressive models or latent growth modeling are usually inflexible to model the dependence and nonlinear effects among the changes of latent traits whenever the time gap is irregular and the recorded time points are individually varying. Often in practice, the growth trend of latent traits are subject to certain monotone and smooth conditions. To incorporate such conditions and to alleviate the strong parametric assumption on regressing latent trajectories, a flexible nonparametric prior has been introduced to model the dynamic changes of latent traits for item response models over the study period. Suitable Bayesian computation schemes are developed for such analysis of the longitudinal and dichotomous item responses. Simulation studies and a real data example from educational testing have been used to illustrate our proposed methods.

CDM 5c: Phase Problem: An Example from Caffeine Consumption and Sleep Duration

Yueqin Hu, Texas State University; Katelyn Stephenson, Texas State University

In dynamical systems, multiple processes are often not phase aligned, that is the time to respond to an initial stimulus may vary across individuals, which may result in a non-significant correlation between two actually related processes. Moreover, abrupt phase changes are frequently seen in human behavioral data, which also place difficulties in portraying the overall trend when examining the association between multiple processes. This study recognized the challenges of the phase problem, and suggested to estimate local first derivatives to identify and extract phase jump information. The example data came from a national longitudinal study MIDUS. Participants were measured on sleep behavior and daily activities for seven consecutive days. Neither a zero-order correlation nor a linear mixed-effects model detected a significant association between caffeine consumption and sleep duration. In contrast, dynamical system analysis with an additional measure on phase jump revealed that sudden phase changes in caffeine consumption significantly predicted a shorter average sleep time, even with the variability of caffeine consumption controlled. The challenges and opportunities that phase problem brings into dynamical system analysis are also discussed.

CDM 5d: Continuous-Time Models with Time-Varying Parameters

Meng Chen, The Pennsylvania State University; Sy-Miin Chow, The Pennsylvania State University

In the field of dynamical systems modeling, self-organization occurs when a system shows distinct shifts in dynamics due to variations in the parameters that govern the system. The variations may unfold as a function of time, space, other exogenous variables, or possibly as triggered by unknown sources of influence either in a deterministic or a stochastic (probabilistic) fashion. Myriad work exists in the state-space, economical, statistical and psychological literature for capturing how parameters vary in discrete-time dynamic models. Related work for representing time-varying parameters in continuous-time models (e.g., differential equation models) remains scarce. We propose a Stochastic Differential Equation (SDE) modeling framework with time-varying parameters as a way to capture self-organization in continuous time. Incorporating time-varying parameters into SDE models provides one way of modeling self-organization, as well as multi-time scale processes (for example, simultaneous representation of gradual developmental changes and nuanced intraindividual variability from moment to moment) in continuous time. We present several examples of SDEs with time-varying parameters, including a stochastic damped oscillator model with stagewise and other hypothesized functional shifts in set-points and damping parameters, and discuss plausible approximation functions that may be used to approximate changes in the time-varying parameters. The extent to which information criterion measures can be used to select the best-fitting approximation functions will be discussed.

Symposium 19: 8:30 AM – 10:00 AM

Symposium 19: IRT is Just Another Network Model

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

Symposium 19a: Relating Ising Network Models to Item Response Theory

Maarten Marsman, University of Amsterdam; Denny Borsboom, University of Amsterdam; Joost Kruis, University of Amsterdam; Sacha Epskamp, University of Amsterdam; Lourens Waldorp, University of Amsterdam; Gunter Maris, University of Amsterdam & Cito

In recent years, network models have been proposed as an alternative representation of psychometric constructs such as depression. In such models, the covariance between observables (e.g., symptoms like depressed mood, feelings of worthlessness, and guilt) is explained in terms of a pattern of causal interactions between these observables, which contrasts with classical interpretations in which the observables are conceptualized as the effects of a reflective latent variable. However, few investigations have been directed at the question how these different models relate to each other. To shed light on this issue, we explore the relation between one of the most important network models -the Ising model from physics- and one of the most important latent variable models -the Item Response Theory (IRT) model from psychometrics. The Ising model describes the interaction between states of particles that are connected in a network, whereas the IRT model describes the probability distribution associated with item responses in a psychometric test as a function of a latent variable. Despite the divergent backgrounds of the models, we show a broad equivalence between them.

Symposium 19b: Network Multidimensional Item Response Models: Beyond Simple Structure

Carolyn J. Anderson, University of Illinois at Urbana-Champaign

Network item response models are statistical graphical models for discrete and continuous variables where the discrete variables are items and the continuous variables are latent constructs. Graphs can be used to represent both formative and reflective multidimensional item response models. In the formative case, the graphical models yield Log-Multiplicative Association (LMA) models for observed data (i.e., responses to items). Although the reflective and formative measurement models are philosophically and theoretically different, they behave empirically in similar (but not identical) ways. Anderson and Yu (2007, 2017) studied log-multiplicative association models for uni-dimensional models for binary items and multidimensional models with simple structure for polytomous items. They investigated the properties of LMA models as response models. We generalize the work by Anderson and Yu to graphical models where items are related to multiple latent variables, including bi-factor models and exploratory models with correlated latent variables.

Symposium 19c: A Fused Latent and Graphical Model for Item Response Analysis

Yunxiao Chen, Emory University

In this talk, we propose a new model that consists of a low dimensional latent variable component and a sparse graphical component for analyzing item response data in psychological measurement. Standard approaches to item response data in psychological measurement adopt multidimensional nonlinear factor models, also known as the multidimensional Item Response Theory (IRT) models. However, human mental process is typically complicated and may not be adequately described by just a few factors. Consequently, a low-dimensional latent factor model is often insufficient to capture the structure of the data. The proposed model adds a sparse graphical component that captures the remaining ad hoc dependence. It reduces to a multidimensional IRT model when the graphical component becomes degenerate. The model is proposed under an exponential family model framework that is applicable to data with categorical responses or a combination of both continuous and categorical responses. Model selection and parameter estimation are carried out simultaneously through construction of a pseudo-likelihood function and properly chosen penalty terms. Efficient algorithms are developed for the computation. Desirable theoretical properties are established under suitable regularity conditions. Under the new modeling framework, the possible inflation of test reliability under the misspecification of local independence assumption is studied. The method is applied to the revised Eysenck's personality questionnaire, revealing its usefulness in item analysis. Simulation results are reported that show the new method works well in practical situations.

Symposium 19d: On the Symmetric Quadratic Exponential Distribution

Nanny Wermuth, Johannes Gutenberg-University Mainz

The quadratic exponential distribution, also known as the Ising model, is a joint distribution for binary variables which has been described as being closest to a joint Gaussian distribution since in its log-linear representation, all higher than two-factor interactions are missing. We discuss the corresponding model for symmetric binary variables and identify a subclass of graphical Markov models in which sequences of logit regressions, by which the joint distribution can be generated, are indeed identical to linear regressions. We discuss further the relations to some item response models and the relevance for general Ising models. Especially for large data sets, the possible local fitting and testing is essential for comparing independence and dependence structures with available knowledge about the variables under study.

META 1: 8:30 AM – 10:00 AM

Meta-Analysis and Replicability

Chair: Melanie Wall

META 1a: Estimating Replicability of Science by Taking Statistical Significance into Account

Robbie C. M. van Aert, Tilburg University; Marcel A. L. M. van Assen, Tilburg University

Consider the following common situation in science nowadays: A researcher reads about a (statistically significant) effect in the literature and replicates the original study. As a result, the researcher has two effect size estimates, and his/her key objective is to evaluate effect size based on these two study outcomes. This objective is particularly challenging if, opposed to the original effect size, the replication's effect size is small and not statistically significant. These challenging situations are omnipresent in science. For instance, 63.9% and 31.3% of the replicated studies in the Reproducibility Project Psychology (RPP) and the Experimental Economics Replication Project (EE-RP) were characterized by a statistically significant effect size in the orignal study and nonsignificant replication effect size. The RPP and also EE-RP assessed effect size using fixed-effect meta-analysis. However, fixed-effect meta-analysis yields overestimated effect sizes because it does not take into account the statistical significance of the original study. We developed the snapshot Bayesian hybrid meta-analysis method (snapshot hybrid for short) that does take into account the statistical significance of the original study. This method quantifies the amount of evidence in favor of a zero, small, medium, or large underlying true effect size by computing posterior model probabilities. We will explain snapshot hybrid, and show the results of analytically approximating its statistical properties. We will also present results of applying the method to data of the RPP and EE-RP.

META 1b: The Effect of Publication Bias on the Assessment of Heterogeneity

Hilde E. M. Augusteijn

One of the main goals of meta-analysis is to test and estimate the heterogeneity of effect size. We examined the effect of publication bias on the Q-test and assessments of heterogeneity, as a function of true heterogeneity (absent, small, medium, large), publication bias, true effect size, number of studies, and variation of sample sizes. The expected values of heterogeneity measures H^2 and I^2 were analytically derived, and the power and the type I error rate of the Q-test were examined in a Monte-Carlo simulation study. Our results show that the effect of publication bias on the Q-test and assessment of heterogeneity is large, complex, and non-linear. Publication bias can both dramatically decrease and increase heterogeneity, depending on the true effect size and the other factors. When sample sizes are equal, extreme homogeneity can be expected even when the population heterogeneity is large. Particularly if the number of studies is large and population effect size is small, publication bias can cause extreme type I error rates close to 0 or 1, and power may change from .80 to close to 0 or 1 as well. We therefore conclude that the Q-test of heterogeneity homogeneity and heterogeneity measures H^2 and I^2 are generally not valid in assessing and testing heterogeneity when publication bias is present, especially when the true effect size is small and the number of studies is large. More research should be conducted to enhance meta-analytic methods that are able to deal with publication bias in their assessment and tests of heterogeneity.

META 1c: A New Approach to Moderation in Meta-Analysis: Meta-CART

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

In many meta-analytic studies on intervention effects, moderation analysis is performed nowadays to establish which intervention components (i.e., moderators) influence the pooled estimate of effectiveness. When many moderators are available, traditional meta-analysis methods often lack sufficient power to perform such moderation analysis. Furthermore, when no a priori hypotheses are available, the traditional methods are not appropriate for interaction detection. To solve this problem, meta-CART was proposed by integrating Classification And Regression Trees (CART; Breiman et al., 1984) into meta-analysis (Dusseldorp et al., 2014). Meta-CART appeared to be successful in detecting combinations of intervention components that result in a higher average treatment outcome (Li et al., 2017). In the present paper, the meta-CART method was improved upon two aspects: 1) the stepwise approach was changed into an integrated approach, and 2) the fixed- or random-effects assumption was taken into account in the tree growing procedure. The performance of the improved version of meta-CART was investigated by means of an extensive simulation study including all types of moderators (i.e., nominal, ordinal, and numerical). The results suggest that the method achieves satisfactory performance (power > 0.80 and Type I error < 0.05) if the number of included studies is large enough. The recommended minimum number of studies ranges between 40 and 120 depending on the complexity and strength of the interaction effects, and the residual heterogeneity. Using an example data set, we will illustrate that the improved version of meta-CART applies the fixed- or random-effects assumption consistently.

META 1d: Meta-Analytic Comparison of Latent Means via Logit Percent Correct Scores

Philipp Doebler, TU Dortmund University; Anna Doebler, University of Mannheim

Standardized mean differences like Cohen's d and Hedges' g are popular effect sizes for the comparison of groups. Poor reliability attenuates d and g relative to the standardized mean difference of the underlying latent variables, so correction factors have been recommended. However, if groupwise means of percent correct scores are compared, two problems limit the applicability of existing methods: (i) Reliability estimates are often not available and (ii) ceiling effects in one or both groups restrict the range of latent variables. We show that both problems are addressed by assuming a Beta-Binomial Model (BBM) for the observed number correct scores (Keats & Lord, 1962). The BBM entails beta-distributed latent variables and a KR21-reliability estimate results. Standardized mean differences of the log-odds of the latent variables are an alternative to the use of d or g. It is shown that moment estimation of BBM parameters is feasible, so that the method is applicable without access to the complete test data, e.g. in meta-analyses. The procedure is illustrated with data from a systematic review of (non-)symbolic magnitude processing in mathematically low-achieving children.

META 1e: Bootstrapping – Enhancing Successful Replication of Effect Size Estimates

Yongtian Cheng, University of Manitoba; Johnson C. Li, University of Manitoba; Rory Waisman, University of Manitoba

The purpose of this study is to evaluate the capture rate based on the conventional algorithm-based and bootstrap-bias-corrected-and-accelerated interval (BCaI) for four common Effect Sizes (ES) estimated in an independent-sample research scenario. Replication, conceptualized as the likelihood of whether or not a published ES estimate can be obtained by an independent researcher using the same methods, has received increasing attention in psychological research (Linda, 2015). Open Science Collaboration (OSC; 2015) suggested that to examine whether a study effect can be successfully replicated researchers should evaluate capture rates based on the ES's confidence interval (CI). Yet little research has empirically evaluated the accuracy of CI as a criterion for assessing the capture rate. Specifically, OSC recommended that researchers should use 83% as a reasonable capture rate, but there is no empirical evidence to justify this criterion. We simulated 145 conditions varying sample sizes, SDs, and standardized mean differences. Two CIs (algorithm-based and BCaI) for four common ES estimates were evaluated: A (Ruscio, 2008), dR (Algina, Keselman, & Penfield, 2005), Cohen’s d, Rpb, and Fisher r-to-z transformation. Although the algorithm-based CI yielded a satisfactory mean capture rate of 83%, with a SD of .10, there is still a problem that around 17% of ES from simulated samples have a poor capture rate (< 80%). In some extreme conditions, the success rate dropped to 60%. On the contrary, the

BCaI produced a more consistent capture of 84.5% with a SD of .025. We suggest that researchers could use and report the BCaI which provides more desireable capture rates for ES replications.

Invited: 10:20 AM – 11:05 AM

Profile Analysis: A Generalization of DIF Analysis

Invited Speaker: Norman Verhelst, Eurometrics
Chair: Paul De Boeck, Ohio State University & KU Leuven

To investigate if some groups of testees are (dis)advantaged by a certain category of items, sometimes Differential Item Functioning (DIF) analysis is applied to all items of this category. It was hypothesized, for example, that younger testees might be disadvantaged by items belonging to the occupational domain in the Michigan English Test (MET). DIF analysis, however, showed unclear and non convincing results. It will be argued that DIF analysis in this context is not a very good method: it lacks clarity in the interpretation and statistical power in the applications. A new method, called profile analysis, will be presented. In this method the focus is not on single items but on categories of items and the focus is shifted from exploratory to confirmatory analysis. For each category the sequence of observed scores (the observed profile) from a testee are compared to their expected value under the measurement model used (the expected profile). The difference between observed and expected profile is called the deviation profile. These deviation profiles are aggregated in each of two or more groups, and give rise to some statistical tests which show if different groups react differently to categories of items. Profile analysis turns out to be statistically powerful and very flexible in its use: it is not restricted to two categories and the number of groups to be compared is unlimited. It is also shown that DIF analysis is a special case of profile analysis. It is argued that profile analysis is a more constructive approach to deviations from the measurement model used than the usual approach by DIF analysis which is aimed mainly at detecting ‘bad’ items. An example using the data from TIMSS 2011 will sustain our point of view.

State: 10:20 AM – 11:05 AM

Value-Added Modeling

State-of-the-Art Speaker: J.R. Lockwood, Educational Testing Service
Chair: Jee-Seon Kim, University of Wisconsin-Madison

The use of standardized test scores to measure the performance of individual teachers has been one of the most controversial topics in U.S. educational research and policy over the past two decades. There have been substantial disagreements among policy makers, educators, researchers and other stakeholders about whether such teacher "value-added (VA)" measures can provide valid, fair and reliable inferences about the effects of individual teachers on student achievement progress. The essential challenge to estimating teacher VA is that groups of students taught by different teachers may have vastly different background characteristics, including different levels of prior achievement. VA modeling generally refers to the use of statistical models to adjust for these differences using longitudinal data on individual students that is now routinely archived by schools, districts and U.S. states. The pros and cons of different model specifications have been the subject of intense debate by econometricians, statisticians and psychometricians, with some of the essential issues lying at the nexus of these fields. This presentation will summarize the basic evaluation problem of estimating teacher VA, and will discuss recent empirical evidence for and against the validity of inferences about teachers made from VA models. It will discuss key areas of future research, including ways in which the psychometric community may be able to contribute to enhanced model specifications

Keynote: 11:10 AM – 12:10 PM

Keynote Speaker- Early Psychometric Contributions to Gaussian Graphical Modelling: A Tribute to Louis Guttman

Career Award For Lifetime Achievement: Willem Heiser, Leiden University

Chair: Ulf Böckenholt, Northwestern University

Graphical models and network analysis are increasingly being used in several areas of psychology and cognitive science, such as neuroimaging, psychopathology, and cognitive development. These models are also brought to the attention of psychometricians: a recent Psychometrika paper speaks of network psychometrics. Present-day authors using Gaussian graphical modelling refer to authors citing other authors who in turn credit Arthur Dempster's 1972 Biometrics paper. Closer scrutiny shows that the key ingredient leading to Gaussian graphical models is the inverse of the correlation matrix including its statistical interpretation as the partial association between two variables given all others. It was exactly this concept that was central in Louis Guttman's 1953 Psychometrika paper about image theory, once quite influential in psychometrics. Several elements of this story might still be relevant for us today.

Symposium 20: 1:10 PM – 2:40 PM

Symposium 20: Never Waste a Good Crisis: Towards Responsible Data Reproducibility, Transparency and Management

Chair: Klaas Sijtsma, Tilburg University

Symposium 20a: How Has the New Policy of Data Reproducibility Affected Research Published in Psychometrika?

Irini Moustaki, The London School of Economics and Political Science

It is almost two years since Psychometrika decided to encourage authors to produce data and code together with their journal submissions. The new policy is part of the general movement to make submitted and published research more transparent to users and reviewers. Authors have positively reacted to the new policy and when possible submit code and data that verifies their results. It is probably about time to start thinking of ways to measure the effectiveness of the new policy to research and to the society. I will present some statistics to show the type and volume of information that has been submitted as part of the new policy and how much it has been used by our readership. Issues that arise with the additional material submitted will be discussed. It is however clear that to be able to make the policy more effective to users and to the general research community, more resources need to be allocated to those efforts.

Symposium 20b: Improving Reproducibility by Archiving and Sharing Our Research Data

Jelte M. Wicherts, Tilburg University

Despite professional guidelines concerning data accountability and the clear importance of keeping and sharing data for reproducibility of empirical results, data from published studies are often insufficiently archived, documented, and shared for verification purposes. In this talk, I will relay my own experiences as data re-analyst and present results from empirical studies into the poor availability of research data. I discuss the cultural, educational, and institutional impediments to creating a truly reproducible science and highlight the benefits for both scientific progress and the individual researcher of proper data archiving and of sharing data (and codes). I will discuss what journals, funders, individual researchers, and academic institutions can do to improve data handling and data sharing. I focus on how the journal Psychological Science changed open data practices almost overnight by handing out badges for open data, and how –after the infamous Diederik Stapel fraud case- the Tilburg School of Social and Behavioral Sciences implemented a successful data policy in which a science committee checks on a random basis whether authors archive their data rigorously.

Symposium 20c: Increasing Transparency Through a Multiverse Analysis

Francis Tuerlinckx, KU Leuven; Sara Steegen, KU Leuven; Andrew Gelman, Columbia University; Wolf Vanpaemel, KU Leuven

Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using an example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction and gives pointers as to which choices are most consequential in the fragility of the result.

Symposium 20d: Statistics Playing with Researchers—Policy to Prevent Questionable Research Practices

Klaas Sijtsma, Tilburg University

Questionable Research Practices (QRPs) and worse, Fabrication, Falsification and Plagiarism (FFP), have probably been around for as long as scientific research exists. Recent incidents involving FFP in The Netherlands in a wide variety of research areas have had the effect of a wake-up call to critically reflect upon how we do our research. I contend that while the relatively rare FFP are obviously intentional they are less interesting from an etiological perspective (the perpetrator meant to deceive for personal profit—resources, status, admiration), and that the real interest should reside with the more widespread QRPs, which may be caused by methodological and statistical misunderstandings researchers are unware of as well as strategies used consciously but without much awareness of their potentially devastating consequences for the validity of the results. I will discuss three topics. First, I discuss the most likely causes of errors researchers make when designing research, analyzing data, and interpreting results from data analysis. Second, I explain why relief from QRPs should come from research policy rather than improved methodological and statistical methods. Third, I will say a few words about the way Dutch schools of social and behavioral sciences implement policy to improve data management in an effort to professionalize data handling and preventing QRPs.

Symposium 21: 1:10 PM – 2:40 PM

Symposium 21: Recent Developments of Item and Person Fit Measures in IRT

Chairs: Carmen Köhler, German Institute for International Educational Research; Janine Buchholz, German Institute for International Educational Research

Symposium 21a: Group-Level Item-Fit Statistics for the Analysis of Invariance

Janine Buchholz, German Institute for International Educational Research; Johannes Hartig, German Institute for International Educational Research

In its most recent cycle, the Programme for International Student Assessment (PISA) implemented an innovative approach to testing the invariance of IRT-scaled constructs: On the basis of a concurrent calibration with equal item parameters across all groups, a group-specific item-fit statistic (Root Mean Square Deviance; RMSD) was calculated and used as a criterion for the invariance of item parameters for individual groups. RMSD quantifies the magnitude of the shift of observed data from the estimated ICC. The present simulation study explores the statistic's distribution under different kinds of non-invariance in polytomous items: invariance, non-invariance related to shifts in item location, and non-invariance related to differences in item discrimination. Responses to five four-point Likert items were generated under the GPCM for 1000 simulees in 50 groups each. For one of the five items each, either location or discrimination parameters were manipulated between simulation conditions. The respective parameters for the 50 groups were drawn from a normal distribution with different levels of variance, thus manipulating the magnitude of invariance. Preliminary findings show that the RMSD's ability to detect invariance is better for invariance related to shifts in item location than it is to detect differences in item discrimination. The study's findings may be used as a starting point to sensitivity analysis aiming to define cut-off values for determining non-invariance.

Symposium 21b: Bias Correction of the RMSD Item Fit Statistic Using Bootstrap
Carmen Köhler, German Institute for International Educational Research; Alexander Robitzsch, Leibniz Institute for Science and Mathematics Education; Johannes Hartig, German Institute for International Educational Research

Testing model fit is considered an important step in Item Response Theory (IRT) modeling in order to reliably interpret model parameters (Wainer & Thissen, 1987). In the literature, numerous item fit statistics exist. Many of them show severe limitations such as (1) lack of theoretical proof about the distribution of the statistic or (2) finite-sample bias. Parametric bootstrapping provides the sampling distribution of a statistic under the null hypothesis, and has thus far been successfully applied to several item fit statistics (e.g., Douglas & Cohen, 2001; Habing, 2001; Sueiro & Abad, 2011). The nonparametric bootstrap can be used to construct the sampling distribution of a statistic in the underlying population of the given sample to correct an estimator's finite-sample bias (Efron, 1979). The current presentation employs parametric and nonparametric bootstrap methods to the Root Mean Squared Difference (RMSD) from the software mdltm (von Davier, 2005), which is used in the scaling of PISA. In simulation studies and an empirical example, we compare the performance of several fit statistics—Infit and Outfit (Wu, 1997), S – X2 (Orlando & Thissen, 2000), and RMSD—with regard to their Type I error rates and empirical power to detect misfit. Preliminary results show that the Infit/Outfit and S – X2 statistics give inconsistent Type I error rates. The RMSD outperforms the other approaches in detecting item misfit. The parametric bootstrap provides adequate confidence intervals for the RMSD statistic, and the bias corrected RMSD eliminates the positive bias of the RMSD due to finite-sampling error.

Symposium 21c: Bayes Factors for Testing Latent Monotonicity in Polytomous IRT Models
Jesper Tijmstra, Tilburg University; Maria Bolsinova, Utrecht University & Cito

For the modeling of ordinal polytomous item response data a large variety of parametric and nonparametric Item Response Theory (IRT) models have been proposed. These models differ in how they model the item step response functions that define the category probabilities, with some models focusing on an adjacent-category formulation, some on a continuation-ratio formulation, and some on a cumulative-probability formulation. Although the different polytomous IRT models differ in their choice of building blocks (as well as the parametric form that is considered for these building blocks), they share the assumption of latent monotonicity. This means that the functions that they consider for each of the item categories are assumed to behave monotonically: as the latent variable increases, so does for example also the probability of at least "partially agreeing" to a Likert item increase, or of obtaining the maximum score on an achievement test item. As this assumption of latent monotonicity is crucial for inferences based on the IRT model to be valid and fair, it is important to determine whether each item on a test adheres to this property. We propose the use of Bayes factors to assess whether the data provide evidence that latent monotonicity holds for the item that is considered, or whether it is likely to be violated. Bayes factors are proposed for the adjacent-category, continuation-ratio, and cumulative-probability formulation of polytomous IRT models. Their performance is evaluated through simulation study, and their application to empirical data is illustrated.

Symposium 21d: Person Fit Analysis and Robust Estimation Methods
Christian Spoden, Friedrich Schiller University Jena

While psychometricians have proposed the usage of either robust estimation methods or person fit analysis to deal with outliers in test data (e.g., Smith, 1985), this talk gives a summary on three studies investigating the performance of robust estimation methods in person fit analysis for item response models. In study 1 a logistic regression approach by Emons, Sijtsma and Meijer (2004) was compared to a robust (robit) regression model in terms of type I error rate and power to detect model violations in simulated data. Results indicate comparable detection rates of both methods and less frequent convergence issues in the robit model. In study 2 a bootstrap method for parametric person fit statistics based on conventional or robust ability estimates were contrasted. Type I error and power in simulated data depended on the specifications of the robust estimates. Bootstrap person fit tests based on Warm's (1989) estimates seemed more beneficial than tests based on robust estimates. In study 3 parametric person fit analysis based on 2PLM parameters were compared to person fit based on parameters from a robust IRT model (Bafumi, Gelman, Park & Kaplan, 2005) in a Bayesian framework (e.g., Glas & Meijer, 2003). Substantial differences between person fit tests based on these two methods in terms of type I error and power did not stand out. In summary, results of the three studies do not indicate that person fit tests based on robust estimation methods offer substantial advantages compared to person fit analysis with conventional estimation methods.

CDM 6: 1:10 PM – 2:40 PM

CDM: Extensions and Alternatives
Chair: Guanhua Fang, Columbia University

CDM 6a: A Cognitive Diagnosis Mixture Model for Learners Applying Different Test-Taking Strategies
Chanho Park, Keimyung University; Sookyung Cho, Hankuk University of Foreign Studies

The purpose of this study is to present a cognitive diagnosis mixture model for learners who apply different strategies when solving test items. Cognitive Diagnosis Models (CDMs) are gaining popularity as a means to identify remedial strategies for learners. In most CDMs, classification of examinees is based on a Q matrix, a binary incidence matrix representing the relationship between the test items and the skills or attributes necessary to complete the items. When learners employ different strategies in taking a test, multiple Q matrices may be possible. In this study, a modified version of the Deterministic Inputs, Noisy, "And" gate (DINA) model will be presented as a cognitive diagnosis mixture model with an application to an English proficiency test for learners employing multiple strategies for reading comprehension. Monte Carlo simulation analyses will be conducted to examine the model. Comparative advantages and necessary cautions regarding this modeling approach are also discussed.

CDM 6b: A Cross-Classified Diagnostic Classification Model for Dual Local Item Dependence
Manqian Liao, University of Maryland, College Park; Hong Jiao, University of Maryland, College Park

Traditional Diagnostic Classification Models (DCM) often assume local item independence. The violation of local item independence (LID) assumption or the presence of local item dependence can jeopardize model parameter estimation and classification accuracy (e.g., Chen & Thissen, 1997; Sireci et al., 1991; Wainer, 1995; Tuerlinckx & De Boeck, 2001; Yen, 1984). DCMs accounting for LID due to the testlet effects were proposed to deal with the violation of this assumption (Hansen, 2013). However, in some testing programs, LID could be caused by two clustering factors simultaneously. For example, in a scenario-based science assessment, items could cluster due to the use of scenario or testlets and content clustering. This is known as dual local item dependence (DLID). Cross-classified models were developed in the Item Response Theory (IRT) framework to deal with the DLID issue (e.g., Jiao, Wang, Wan, & Lu, 2009; Xie, 2014), but such issue has not been investigated for assessments with diagnostic purposes. Therefore, this study proposes a cross-classified DCM that can address the DLID issue in Cognitive Diagnostic Assessments (CDA). Specifically, a cross-classified structure for DLID is incorporated into the Deterministic Inputs, Noisy “And” Gate (DINA) model which is one of the most parsimonious and widely-used DCMs. This study evaluates model parameter recovery and classification accuracy using the Bayesian Markov chain Monte Carlo (MCMC) estimation method under different simulation conditions. Further, it examines the effect of ignoring one or both sources of LID on model parameter estimation and classification outcomes in the DINA model.

CDM 6c: Hierarchical Rater Model in Cognitive Diagnosis
Youn Seon Lim, Hofstra University

The Hierarchical Rater Model (HRM, e.g., Patz, Junker, & Johnson, 2000) for constructed or open-ended item response data scored by multiple raters is considered in the cognitive diagnosis framework. The HRM recognizes that such item responses have a hierarchical structure: at the first level is measuring the ideal rating of an item based on multiple raters’ observed ratings whereas at the second level is relating the ideal rating to the examinee ability. In this study, a cognitive diagnosis model is proposed for the second level of the HRM, especially to provide multiple fine-grained diagnostic information regarding examinees skill mastery level. Simulation and real data analysis were conducted with the proposed models as well as with the HRM-IRT model, and results are compared.

CDM 6d: Partial Mastery Models for Cognitive Diagnosis

Gongjun Xu, University of Michigan

Cognitive Diagnosis Models (CDMs) are popular statistical tools in cognitive diagnosis. Standard CDMs usually assume that latent attributes - specific skills - are either fully mastered or not mastered by a subject when responding to all test items. Each attribute then only classifies the subjects into two respective groups (mastery and non-mastery of a given skill). As a consequence, when partial mastery is possible, heterogeneity in response data may not be well accounted for by standard CDMs. To overcome this issue, we propose a new modeling approach for cognitive diagnosis which allows a subject to have partial mastery of each attribute. We show that this mixed membership approach generalizes CDMs and can also be interpreted as a restricted latent class model. Such property is then employed for the construction of a Bayesian estimation algorithm. Simulation studies and a data analysis are presented to examine the performance of the proposed method.

CDM 6e: Bi-factor MIRT as a Basis for Diagnostic Score Reporting

Daniel Bolt, University of Wisconsin-Madison

Psychometric models for diagnostic classification such as the DINA model assume discrete skill attributes. Among other concerns related to possible model misspecification is the potential for continuity in both the underlying skill attributes and the skill requirements of test items. Using both simulation and real data applications we explore the consequences of such forms of misspecification on the invariance of skill attribute mastery metrics across groups of differing skill distributions, as might be applicable in studies of skill acquisition over time. In this context, the bifactor MIRT model is presented as an appealing alternative. The usefulness of the bifactor approach follows from the tendency for items measuring multiple conjunctively interacting skill attributes to in actuality primarily distinguish only with respect to the most difficult of the required skill attributes, especially in the presence of a higher order factor underlying the skill attributes. We illustrate these results by simulation and in application to frequently analyzed fraction subtraction datasets. Issues in the potential use of bifactor models for diagnostic assessment, including several advantages of this alternative approach, are considered.

Symposium 22: 1:10 PM – 2:40 PM

Symposium 22: Recent Developments of Predictive Modeling in Psychology and Related Areas

Chair: Heungsun Hwang, McGill University

Symposium 22a: An Initial Test of CART-Based Missing Data Methods

Timothy Hayes, Florida International University; John J. McArdle, University of Southern California

Data mining algorithms such as Classification And Regression Trees (CART) and random forests provide a promising means of exploring potentially complex nonlinear and interactive relationships between auxiliary covariates and missing data. Recently, two CART-based missing data methods have been proposed. The first uses CART to create predicted probabilities of response and form data weights. The second uses CART to multiply impute the data. Although both methods have shown promise in prior research, these procedures have not been systematically compared to each other under small sample sizes. The present simulation compares these methods under a wide variety of sample sizes (N = 125, 250, 500, 100), MAR missing data mechanisms (linear, quadratic, cubic, interactive), and degrees of nonnormality (normal, severe nonnormal). Overall, CART-based weights outperformed CART-based multiple imputations under low sample sizes, but were much less efficient than CART-based multiple imputations under large Ns.

Symposium 22b: Recursive Partitioning of Extended Redundancy Analysis

Sunmee Kim, McGill University; Heungsun Hwang, McGill University

Extended Redundancy Analysis (ERA) is a multivariate technique for exploring the directional relationships between multiple sets of predictor and response variables. In ERA, a component is extracted from each set of predictor variables in such a way that it accounts for the maximum variation of response variables. We propose to combine ERA with model-based recursive partitioning to investigate if there exist differences in the effects of predictor components on response variables across several subgroups derived from interactions of additional covariates such as gender, ethnicity, socioeconomic status, etc. This proposed method, called ERA-Tree, utilizes covariates to create a decision tree that splits a population into homogeneous subgroups, each of which involves different sets of ERA parameters. We present two examples to illustrate the usefulness of the proposed method.

Symposium 22c: Propensity Score Analysis Using Super Learner for Longitudinal Data

Ji Hoon Ryoo, University of Virginia; Catherine Bradshaw, University of Virginia; Joseph Kush, University of Virginia

In practice, the Propensity Score Method (PSM) has been applied under the assumption that the true PS model is unknown. This means that PSM admits some degree of model misspecification and allows bias after propensity score matching and/or weighting. As shown in Lee, Lessler, and Stuart (2009) and Pirracchio, Petersen, and van der Laan (2015), a nonparametric approach in PSM, Super Learner (SL) by van der Laan, Polley, and Hubbard (2007), performed better for highly unbalanced variables and the SL-based estimators were associated with the smallest bias. Colson et al., (2016) further extend the discussion of bias reduction and efficiency in the previous two studies by applying SL ensemble machine learning algorithm in longitudinal data. They found that double robust analysis performed best in both the simulations and the applied example when combined the machine learning method. In this study, we extend the Colson et al.'s discussion by considering (1) fine balance matching that can be applicable to longitudinal study (Rosebaum, Ross, & Silber, 2007; Yang, Small, Silber, & Rosenbaum, 2012) and (2) marginal structural model that its weighting can appropriately adjust for measured time-vary confounders affected by prior exposure (Cole & Hernán, 2008; Daniel et al., 2012). In the study, we use empirical data of School-Wide Positive Behavioral Interventions and Supports (SWPBIS) across the state of Maryland and also conduct a simulation study to imitate the empirical data so that the results address issues in observational settings.

Symposium 22d: Variable and Individual Selection by Constraining the Singular Value Decomposition

Vincent Guillemot, Pasteur Institute; Derek Beaton, The Rotman Institute; Tommy Löfstedt, Umeå University; Arnaud Gloaguen, Supelec; Hervé Abdi, The University of Texas at Dallas

The Singular Value Decomposition (SVD) constitutes the core of most popular multivariate methods such as principal component analysis, canonical correlation analysis, multiple correspondence analysis, among many others. To analyze a data table, these methods use the SVD to generate components or factor scores that extract the important information in the original data tables. Loadings are used to interpret the corresponding components and this interpretation is greatly facilitated when only few variables have large loadings (particularly when the number of variables is large). If this pattern does not naturally hold, several procedures can be used to select the variables important for a component. The early psychometric school, for example, used rotation, but recent approaches favor computationally based methods such as bootstrap ratios, or select important variables with an explicit optimization problem such as the LASSO. With the advent of big data, we need to find new ways to select important variables or individuals for components. Several techniques exist, but sparsification is obtained at the cost of orthogonality. Here we propose a new approach for the SVD that includes sparsity constraints on the columns and rows of a rectangular matrix while keeping the singular vectors orthogonal: To do so, this framework integrates general constraints with projection operators. We show how our algorithm add desirable properties to many existing multivariate methods. We illustrate our method with real and simulated data and compare it to state of the art methods such as penalized matrix decomposition (Witten et al, 2009).

Symposium 22e: A Drug Response Prediction Model Using a Component-Based Structural Equation Modeling Method

Taesung Park, Seoul National University; Sungtae Kim, Seoul National University; Sungkyoung Choi, Seoul National University; Youngsoo Kim, Seoul National University; Jung-Hwan Yoon, Seoul National University

Component-based structural equation modeling has been used in sciences, business, education and other fields. This method approximates unobservable latent variables by components of observed variables, and investigates the structural relationships between observed and latent variables. We have extended the method to the analysis of various biologically structured data. In this paper, we propose a component-based drug response prediction model for peptide level data with Multiple Reaction Monitoring (MRM) mass spectrometry for liver cancer patients. MRM is a highly sensitive and selective method for targeted quantitation of peptide abundances in complex biological samples. The advantage of the proposed prediction model is that it first merges peptide level data into protein level information, which helps better biological interpretations later. It can also handle correlations between variables effectively, avoiding a multiple testing problem. We select significant protein biomarkers with permutation tests and construct a Sorafenib drug response prediction model. We demonstrate that the proposed model successfully predicts a drug response for liver cancer patients with the high Area Under the Curve (AUC) score.

CAT 2: 1:10 PM – 2:40 PM

CAT & Response Times
Chair: Dylan Molenaar, University of Amsterdam

CAT 2a: Investigating Item-Exposure Control in Shadow-Test Approach to Computerized Adaptive Testing

Yi-Fang Wu, ACT, Inc.

The purpose of this simulation study is to investigate item-exposure control in the Shadow-Test Approach (STA; van der Linden, 2000) for constrained item pools. Recently, STA has gained increasing attention in the literature and gradual popularity in practice for the reasons that shadow tests are full-length tests that meet all the content specifications, and in terms of precision measurement, they are optimal at the current trait estimate for the examinee. However, because the assembly of shadow tests is more a simultaneous than sequential item selection algorithm based on optimization of the test information as well as content constraints, severe item overexposure and underexposure could happen in a Computerized Adaptive Testing (CAT) administration. How various item-exposure control and item pool design methods work with STA and to what degree these methods work for constrained item pools needs further investigation. The current study addresses these issues. Four methods are included: an a-stratified method (Chang & Ying, 1999), an a-stratified method with b-blocking (Chang, Qian, & Ying, 2001), the Sympson-Hetter (1985) method, and a variation of the randomesque strategy (Kingsbury & Zara, 1989; Shin & Chien, 2009). Three pool sizes (200, 400, and 800 items) with some content constraints are used. The R package lp_SolveAPI (Diao & van der Linden, 2011) is used to execute the algorithm of STA. The results of this study are expected to inform researchers and testing organizations of the performance of STA with exposure control for CAT subject to certain item pool characteristics and content constraints.

CAT 2b: Automated Test Assembly of Multidimensional Parallel Test Forms

Dries Debeer, University of Zurich; Usama Ali, Educational Testing Service; Peter van Rijn, ETS Global

Many educational testing programs require different test forms with minimal or not overlap. At the same time, the test forms should be equivalent in terms of their statistical and content-related properties. That is, the test forms should be parallel. A well-established method to assemble parallel test forms is to apply combinatorial optimization using Mixed Integer Linear Programming (MILP). Using this approach, in the unidimensional case, the Fisher Information is commonly used as the statistical target to obtain parallelism. In the multidimensional case, however, the Fisher Information is a multidimensional matrix, which complicates its use as a statistical target. Previous research addressing this problem focused on item selection criteria for multidimensional adaptive tests. Yet, these criteria are not directly transferable to the linear assembly of parallel test forms. To fill this gap, different statistical targets are derived, based on either the Fisher information or on the Kullback-Leibler divergence, that can be applied in MILP models to assemble parallel test forms in the uni- as well as in the multidimensional case. Using simulated as well as operational item pools, the targets are compared and evaluated. Promising results with respect to the Kullback-Leibler based targets are presented and discussed.

CAT 2c: Facilitating Screening for Psychopathology in Primary Health Care Settings: CATja

Jan van Bebber, University Medical Center Groningen

I will discuss two aspects of IRT modeling. First, I will present an example of a practical application of IRT modeling: A demonstration of CATja, the adaptive test battery that we developed to screen clients in general practices in the Netherlands for psychopathology. Second, I will discuss a more technical aspect of IRT modeling and argue that model misfit is not problematic, as long as adaptive test scores (based on non-fitting item parameters) "behave" the way they are supposed to. That is, slight deviations from theoretical models do no hinder practical applications as long as adaptive test scores show good construct- and predictive validity. I will illustrate this with our recently published results where we report on the calibration of positive and negative symptoms of psychosis. Although model fit is weak for both symptom domains, results of Real Data Simulations are encouraging in terms of the relative sizes of validity coefficients. More specifically, the CATs for positive and negative symptom experiences (respectively 10 and nine items on average) outperform the existing short form (PQ-16) of the original instrument (PQ-92). The CATs even function as well as the original subscales of the instrument that utilize 45 positive and 19 negative symptoms.

CAT 2d: A Dynamic Response Time Model for Speeded Tests

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

In time-limited tests, the missing data process leading to not-reached items at the end of the test can be understood by studying examinees' working speed. Thus, modeling responses and response times jointly when estimating ability has been recognized as a promising approach by several researchers. Response times allow estimating working speed and, as a consequence, to account for potential mechanisms underlying the missing data process at the end of the test. However, recently developed test models for joint modeling of speed and accuracy assume a constant working speed across the test. This assumption is violated whenever examinees encounter tight time restrictions and are either running out of time, or perceive to do so. Under time limits, examinees might try to compensate by adjusting their pace in order to reach the end of the test. Building on previous research on dynamic processes in Item Response Theory (IRT) models, a dynamic response time model is presented. This model allows for varying speed as a function of the examinees' progress in terms of both time passed and the number of attempted items. The model allows assessing test speededness and describes the mechanisms leading to not-reached items. Parameter recovery of the proposed model as well as its ability to account for missing data due to not-reached items is investigated within a simulation study. An illustration of the model by means of an application to real data is provided.

CAT 2e: Cognitive Diagnosis Modeling Incorporating Item Response Times

Peida Zhan, Beijing Normal University; Hong Jiao, University of Maryland, College Park; Dandan Liao, University of Maryland, College Park

In order to provide more refined diagnostic feedback with collateral information in item Response Times (RTs), this study proposed joint cognitive diagnosis modeling of attributes and response speed using item responses and RTs simultaneously. To model the relationship between attributes and latent speed, a general ability was introduced to link the correlated attributes first, then a bivariate normal distribution was assumed between the general ability and latent speed in a similar manner as traditional lognormal RT models do. Further, the Deterministic-Inputs, Noisy "And" gate model (DINA) was used as a diagnostic model for illustration. An extended DINA model was proposed for joint modeling of responses and RTs. Model parameter estimation was explored using the Bayesian MCMC method. The PISA 2012 computer-based mathematics data were analyzed firstly to explore the direction and the degree of correlations among model parameters. Then real data estimates were treated as true values in a subsequent simulation study. The results indicated that model parameters could be well recovered using the MCMC approach. Further, incorporating RTs in the DINA model would improve individual correct classification rates and result in more accurate and precise estimation of the general ability when the test length was short.

MULTI 2: 1:10 PM – 2:40 PM

Multilevel Modeling
Chair: Ed Merkle, University of Missouri

MULTI 2a: Caution When Centering Lower Level Interactions in Multilevel Models
Haeike Josephy, Ghent University; Tom Loeys, Ghent University

In hierarchical designs, the effect of a lower level predictor on an outcome may often be confounded by an (un)measured upper level variable. When such confounding is left unaddressed, the effect of the lower level predictor will be estimated with bias. To remove any such bias in a linear random intercept model, researchers often separate the lower level effect into a within- and between-component (under a specific set of confounding-assumptions). When the effect of the lower level predictor is additionally moderated by another lower level predictor, an interaction between both predictors needs to be included into the model. To again address any possible unmeasured upper level confounding, this interaction term also requires partitioning into a within- and between- cluster component. This can be achieved by first multiplying both predictors and by consequently centering that product term, or vice versa. We demonstrate that the former centering approach proves much more efficient and robust against misspecification of cross- and upper-level effects, compared to the latter.

MULTI 2b: Inter-Industry Wage Differentials: A Structural Equation Modelling Approach
Maria de Fátima Salgueiro, Instituto Universitário de Lisboa; Ricardo Freguglia, Universidade Federal de Juiz de Fora; Marcel D.T. Vieira, Universidade Federal de Juiz de Fora

Understanding the pattern of wage differentials among individuals with similar characteristics and jobs has been a longstanding issue in labor economics; several studies have been conducted in various countries. Classical approaches to modeling longitudinal data have often been considered, namely Fixed-Effects (FE) and Random-Effects (RE) models. Such models provide a way to control for all the time-invariant unmeasured variables (whether known or unknown) that influence the dependent variable. However, these models are not flexible enough (Bollen & Brand, 2010) and have some limitations: in FE models it is not possible to estimate the effect of observed time-invariant variables; estimated coefficients are fixed over time; no effects from the lagged dependent variables are allowed; residual error variances are equal across all waves; regarding unobserved heterogeneity, latent time-invariant variables either freely correlate with all time-varying covariates, as in the FE model, or they must be uncorrelated with all covariates, as in the RE model. There is a vast literature in econometrics suggesting models and estimators that, separately, try to address some of these limitations. This paper proposes using the flexibility and potentialities of the structural equation modelling framework to parameterize and estimate models, thus simultaneously overcoming most of these limitations. Brazilian data from the Ministry of Labor and Employment (RAIS Identificada), available for the period 2006 to 2013, is used to model inter-industry wage trajectories. Latent growth curve models with different types of restrictions are presented. The advantages of the proposed approach over the classical econometric models are discussed.

MULTI 2c: Testing the Upper-Level Exogeneity Assumption in Random Slope Models
Tom Loeys, Ghent University; Wouter Talloen, Ghent University; Beatrijs Moerkerke, Ghent University

It may often be unrealistic in practice to assume that the random intercept is independent from the predictors in a random intercept model. For example, when studying the effect of intimacy on daily mood in a longitudinal diary study, there may be unmeasured personality traits that are both associated with the predictor and the outcome. Treating the intercept as fixed rather than random in linear random intercept models solves the issue of omitted variable bias inherent to such upper-level endogeneity. Contrasting the Fixed Effect estimator (FE) with the Random Effect (RE) estimator, as is done in the Hausman test, may therefore yield evidence against the upper-level exogeneity assumption. However, unmeasured upper-level heterogeneity can also be present in the effect of the predictor on the outcome, and a random slope effect is then required. In our example, the effect of intimacy on mood may depend on unmeasured personality traits that are associated with both the predictor and the outcome. One can easily extend the FE-approach to allow for such upper level endogeneity as well, but the standard Hausman test can no longer be used to contrast the FE-estimator and RE-estimator. In this presentation we therefore propose an extension of the Hausman test that is based on bootstrapping the difference between the FE-estimator and RE-estimator under the null. We find that it performs equally well as the standard Hausman test in the random intercept model, but outperforms the robust Hausman test in the random slope model.

MULTI 2d: Derivative Computations for Linear Mixed Effects Models with Applications

Ting Wang, The American Board of Anesthesiology; Edgar C. Merkle, University of Missouri

While robust standard errors and related facilities are available in R for many types of statistical models, the facilities are notably lacking for linear mixed models estimated via lme4. This is because the necessary statistical output, including the casewise gradient and Hessian of random effect parameters, is not immediately available from lme4 and is not trivial to obtain. These quantities calculations involve first (casewise) and second derivatives of the marginal log-likelihood with respect to all parameters (fixed effects, random effect (co)variances, and residual variance). In this presentation, we show how these derivatives can be theoretically obtained and demonstrate their computation in a newly developed R package. We also illustrate the derivatives' applied uses, including robust standard errors, standard errors of random effect estimates, and score-based statistical tests.

MULTI 2e: Estimating Individual-Level Effects in Data Streams with Binary-Outcomes

Lianne Ippel, Tilburg University

Recently, it has become increasingly easy to collect data from individuals over long periods of time. Examples include web-log data tracking individuals' browsing behavior, or longitudinal (cohort) studies where many individuals are monitored over extensive periods of time. Such datasets cover a large number of individuals and collect data on the same individuals repeatedly, causing a nested structure in the data. Moreover, the data collection is never "finished" as new data keep streaming in. It is well known that predictions that use the data of the individual whose individual-level effect is predicted in combination with the data of all the other individuals, are better than those that just use the individual average. However when data are nested and streaming, and the outcome variable is binary, computing these individual-level predictions is computationally challenging. In this presentation, we introduce four estimation methods which do not revisit "old" data and deal with nested data. The methods are based on existing shrinkage factors: the James-Stein estimator, (approximate) maximum likelihood estimator, beta binomial estimator, and a heuristic estimator. A shrinkage factor predicts an individual-level effect by weighing the individual mean and the overall mean. In a simulation study, we compared the performance of existing and newly developed shrinkage factors. We find that the existing methods differ in their prediction accuracy, but the differences in accuracy between our novel shrinkage factors and the existing methods are small. Our online implementation of the well-known shrinkage factors are however computationally feasible in the context of streaming data.

FMM 2: 1:10 PM – 2:40 PM

Finite Mixture Models & Eye Tracking

Chair: Sabrina Giordano, University of Calabria

FMM 2a: Procedures for Multi-Factor Latent Class Parameter Estimation and Classification Accuracy

Hsu-Lin Su, National Taiwan Normal University

Multi-factor structure exists in many tests (or inventories) conditions such as the Primary Mental Abilities Test (Thurstone, 1941) and NEO Personality Inventory (Costa & McCrae, 1978). Relatively, heterogeneity exits in populations, too. Still, research may focus on deciding to which of the subpopulations a participant most likely belongs according to one's factor pattern. It is the context of "multi-factor latent class" this study is devoted. Thus, we proposed three procedures based on factor mixture models (FMM) with a semi-confirmatory approach to analyze data in the context of multi-factor latent class models. Simulations were manipulated with different procedures, class separation and class number, factor number and factor correlation. Results showed that the procedures of "factor structure first then class number" (procedure 1) and "factor structure and class number considered simultaneously" (procedure

3) most often detects the correct model, using the Bayesian Information Criterion (BIC). As to parameter estimation and Classification Accuracy (CA), when class separation was large enough, CA increased to above 0.8, and the Mean Difference (MD) and Root Mean Square Difference (RMSD) of estimated parameters were reduced. However, the procedure of "class number first then factor structure" (procedure 2) could only detect the correct model in half of the conditions, the CA ranged around 0.5 - 0.8. Procedure 2 was obviously influenced by the factor number and correlation; when the factor number increased to three, with lower correlation, the true model could not be detected. It also had higher MD and RMSD compared with procedure 1 and 3. Conclusions implied that when strong measurement invariance was assumed in the context of multi-factor latent classes, it may be appropriate to choose procedure 1 and 3.

FMM 2b: The Contribution of Count Items to Health-Related Symptom Assessment

Brooke Magnus, Marquette University

Items that elicit count responses are sometimes found on psychological questionnaires. Such items often assess symptom severity, asking respondents to recall the frequency of various thoughts or behaviors over a certain number of days. These types of responses can pose analytic challenges and may require specialized IRT models that can account for phenomena such as zero inflation, maximum inflation, and heaping (Magnus & Thissen, in press; Wang, 2010). To justify their use, it is helpful to understand what information count items can provide beyond more traditional item types (e.g., Likert responses). This research uses data from the National Comorbidity Survey - Replication (NCS-R) to evaluate the contribution of two count items to a scale comprising eight Likert items. We adapt a latent class IRT model to accommodate a mixture of item types that exhibit zero inflation and heaping, testing its software implementation with simulated data. We then fit the model to the NCS-R data, evaluate item parameter estimates, and compute scale scores and posterior standard deviations from two different versions of the measure: one excluding the count items and one including the count items. The contribution of the count items is measured as the change in posterior standard deviations (i.e., improvement in measurement precision) between the two versions of the measure. Results suggest that while these count items do increase computational burden, they substantially reduce the posterior standard deviations of scale scores, particularly for individuals at high (poor health) levels of the latent variable. Implications for scale development are discussed.

FMM 2c: Misuse of Complier-Average Causal Effect (CACE) Analysis with Non-Normal Data

Jenn-Yun Tein, Arizona State University; William Pelham, Arizona State University; Chung Jung Mun, Arizona State University; David P. MacKinnon, Arizona State University; Thomas Dishion, Arizona State University

Randomized Controlled Trials (RCT) often include non-compliance (i.e., participants not receiving the intended treatment) making it difficult to estimate the causal effect of the treatment received. The Complier Average Causal Effect (CACE) analysis, based upon Rubin's (1974) causal model, incorporates compliance in the analysis of treatment effects. CACE models often use mixture modeling to identify control group participants who are latent compliers (those that would have engaged in the intervention given the opportunity) and latent non-compliers (those that would not have engaged, even if given the opportunity) (Little & Yau, 1998). Mixture modeling is vulnerable to over-extraction and misidentification from a skewed distribution (Bauer & Curran, 2003). We conducted a simulation study to investigate how skewed data might affect the performance of mixture modeling approaches to CACE estimation in RCTs with non-compliance. The simulation conditions varied by skewness, effect size, compliance rate, sample size, and use of covariates to predict compliance status. Results indicated that when the compliance rate is low (20-40%), the models can have dramatically elevated false positive rates at even modest levels of skewness (e.g., skewness = 1.5). Results also indicated that when the compliance rate is less than 50%, the mixture models are likely to extract the tail of a skewed outcome distribution as one component and identify this component as compliers. We conclude that mixture modeling approaches to the estimation of the CACE should be employed only with caution, especially with skewed measures and when overall compliance is low.

FMM 2d: Circular Distribution Models for Saccade Directions
Ingmar Visser, University of Amsterdam; Maartje Raijmakers, University of Amsterdam; Daan van Renswoude, University of Amsterdam

Eye movement data are frequently used to measure cognitive processes because they provide a window into attentional processes that are involved in solving cognitive tasks. Such data can be used to validate cognitive theories that are otherwise hard to test. In category learning, attentional processing of different aspects of stimuli is essential to different theories. Also, in analogical reasoning, different patterns of eye movements are associated with different strategies in solving the task. Eye movements can be particularly relevant in participants whom are otherwise hard to test, such as young children and infants. In infants, eye movement are one of the best sources of information that we have to learn about their cognitive processes. Here we study free-viewing data from infants looking at natural scenes. Free-viewing patterns are characterized by a sequence of saccades and we model the distributions of saccades. Saccades are modelled here using Von Mises mixture distributions (Hornik & Gruen, 2014). An interesting result is that infants, like adults (Foulsham et al, 2008), have a bias for showing more horizontal than vertical or oblique saccades (Van Renswoude, 2016). The mixture distribution approach clearly identifies this bias. Moreover, it quantifies the variance of the saccade distributions accurately and shows a developmental trend towards more accurate eye movements.

FMM 2e: Gazepath: An Eye-Tracking Analysis Tool That Accounts for Measurement Error
Daan van Renswoude, University of Amsterdam; Maartje Raijmakers, University of Amsterdam; Arnout Koornneef, Leiden University; Scott Johnson, University of California, Los Angeles; Sabine Hunnius, Radboud University Nijmegen; Ingmar Visser, University of

Eye-trackers are a popular tool to study cognitive, emotional and attentional processes in different populations (e.g., clinical and typically developing) and participants of all ages, ranging from infants to elderly. This broad range of processes and populations implies there are many inter- and intra-individual differences that need to be taken into account to accurately measure gaze behavior. An important step in the analyses process is to parse raw data into interpretable events such as fixations and saccades. Standard parsing algorithms supplied by the eye-tracker manufacturers are typically optimized for adults and do not account for these individual differences. In this talk we presents gazepath, an easy-to-use R-package that comes with a Graphical User Interface (GUI) implemented in Shiny (RStudio, Inc, 2015). The gazepath R-package combines solutions from the adult and infant literature to provide a data-driven eye-tracking parsing method that reduces measurement error on an individual level. Although gazepath is a suitable tool for both adult and infant data, in this talk we highlight its usefulness on infant data. Measurement error is common in infant data, we show how the gazepath method reduces this error and is able to pick up, an otherwise obscured, developmental pattern in the data.

Keynote: 3:00 PM – 3:45 PM

Keynote Speaker- Methods for Resolving Measurement Error Challenges in a Two-Stage Approach
Early Career Award: Chun Wang, University of Minnesota
Chair: Hua-Hua Chang, University of Illinois at Urbana-Champaign

A rapidly developing outcome-based culture among policymakers in education recognizes the need to use standardized test scores, such as item response theory (IRT) scaled θ scores, along with other outcome measures, to make high-stakes decisions. However, there exist potential errors in estimating the latent θ scores (and other outcome measures), and ignoring the measurement errors will adversely bias the subsequent statistical inferences. With the growing computational power nowadays, a recommended approach to address the measurement error challenge is to use an integrated multilevel IRT model. Despite the statistical appeal of the one stage approach, we advocate that a “divide and-conquer” two-stage approach has its practical advantage. In the two-stage approach, an appropriate measurement model is first fitted to the data, and the resulting θ scores (or its distributions) are used in subsequent analysis. Three different methods are introduced within the two-stage framework that actively take the measurement error into consideration. They are compared with the integrated modeling approach via simulation studies. Results have shown that little precision is lost when the new methods are used. Practical guidelines, future studies and potential challenges are discussed in the end.

Keynote Speaker- The Role of Conditional Likelihoods in Latent Variable Modeling

Presidential Address: Anders Skrondal, Norwegian Institute of Public Health & University of Oslo & University of California, Berkeley
Chair: Cees A. W. Glas, University of Twente

When applicable, constructing a conditional likelihood is one way of handling incidental parameters (whose numbers increases in tandem with the observations) in statistical models. In measurement, the use of conditional likelihoods has a long history associated with the Rasch model. In this talk I will argue that conditional likelihoods (and their approximations) may have an even more important role to play in more general latent variable models. In particular, such «fixed-effects» approaches can allow protective estimation under common challenges such as (1) unobserved confounding, (2) heteroskedasticity, (3) endogenous sampling, (4) cluster-size dependence, and (5) missing data. I will also discuss some limitations of conditional likelihood estimation.