IMPS 2019

International Meeting of the Psychometric Society

Santiago de Chile

Abstract Book: **Talks**

Updated on July 12, 2019

Contents

Tuesday, July 16

Keynote Speaker: Eric-Jan Wagenmakers

Frs-1 Bayesian multi-model inference for practical and impractical problems

Eric-Jan Wagenmakers, *University of Amsterdam, Netherlands*

Whenever a set of models is applied to data, uncertainty surrounds both the selection of the best model and the estimation of the model parameters. The coherent Bayesian answer to the model selection question is to avoid all-or-none selection altogether and instead retain model uncertainty, employing it for purposes such as prediction and parameter estimation. Advantages of this multi-model approach include a reduction of overconfidence, improved predictive performance, and an increased robustness against model misspecification. Moreover, the multi-model framework can be seamlessly integrated with the recent open-science ideals of multi-team inference and multiverse analyses. We debunk several philosophical objections to Bayesian multi-model inference and demonstrate its practical use for problems ranging from the simple to the complex.

Parallel Sessions, Tuesday Morning

Salón Fresno

Symposium: Model what you can see, not what you imagine

Frs-1 Modeling marginal networks

Benjamin Deonovic, *ACTNext by ACT, United States*
Gunter Maris, *ACTNext by ACT*
Maria Bolsinova, *ACTNext by ACT*
Lu Ou, *ACTNext by ACT*
Timo Bechger, *ACTNext by ACT*

Network models are increasingly important and popular in various scientific fields. For many of these models it is the case that the nodes of the network represent observed variables of interest. However, these observed variables are often only a small subset of all of the possible variables that could have been observed. Models for these networks which do not take this into account may lead to model misfit and/or model misinterpretation. If the relationship between the unobserved and observed variables is known we show how we can take into account these missing variables by approximating the distribution of the observed variables. Furthermore, with a completely observed set of variables, we show how we can use the same approximation to obtain closed form solutions to the first and second order moments. The derived model has connections to a variety of fields including structural equation modeling, models for deep learning, and cognitive diagnostic models.

Frs-2 Bayesian covariance structure modeling of structured response data

Jean-Paul Fox, *University of Twente, Netherlands*
Konrad Klotzke, *University of Twente*

The covariance structure of response data represents interesting phenomena such as the inter and intra-individual variability, and response dependencies due to higher-level clusters of persons, items and their interactions, among other things. In the well-known psychometric modeling frameworks (e.g., structural equation modeling, item response theory, multilevel modeling), latent variables have been used to model the covariance structures. However, latent variables have several disadvantages. They demand larger sample sizes, increase the number of model parameters, and limit the flexibility of the model to describe high-dimensional data. To avoid the problems associated with latent variables, the covariance structure can be modelled directly by defining priors for the (co)variance parameters, where a multivariate distribution is defined for the response data. The priors include restrictions on the parameter space of the covariance parameters such that any combinations of covariance parameters leads to a positive definite covariance matrix. A novel method is shown to determine the full conditional distributions of the covariance parameters to define an MCMC algorithm for parameter estimation. The advantages of the BCSM are illustrated through the joint modeling of high-dimensional response data and process data. A real data example is given where different types of clusters affect the dependency structure of the data (e.g., test takers nested in groups, testlet structures, different test modes). It is shown that these inter-dependencies can also be modeled across data sources (responses, responses times, and other process data).

Frs-3 Towards a neo-classical test theory

Timo Bechger, *ACTNext by ACT, United States*
Gunter Maris, *ACTNext by ACT*
Benjamin Deonovic, *ACTNext by ACT*
Maria Bolsinova, *ACTNext by ACT*
Lu Ou, *ACTNext by ACT*

As in the famous quote from George Box, any field that depends on statistical modeling, including psychometrics, has to deal with the fact that its models are not true. About a decade earlier, Georg Rasch noted that models need not be true to be applicable for a given purpose. Following these insights, we propose three basic principles to guide model development in educational measurement. We use PISA data to illustrate how these principles are put in practice.

Frs-4 Bayes factor testing of multiple intra-class correlations

Joris Mulder, *Tilburg University, Netherlands*
Jean-Paul Fox, *University of Twente*

The intra-class correlation plays a central role in modeling hierarchically structured data, such as educational data, panel data, or group-randomized trial data. It represents relevant information concerning the between-group and within-group variation. Methods for Bayesian hypothesis tests concerning the intra-class correlation are proposed to improve decision making in hierarchical data analysis and to assess the grouping effect across different group categories. Estimation and testing methods for the intra-class correlation coefficient are proposed under a marginal modeling framework where the random effects are integrated out. A class of stretched beta priors is proposed on the intra-class correlations, which is equivalent to shifted F priors for the between groups variances. Through a parameter expansion it is shown that this prior is conditionally conjugate under the marginal model yielding efficient posterior computation. A special improper case results in accurate coverage rates of the credible intervals even for minimal sample size and when the true intra-class correlation equals zero. Bayes factor tests are proposed for testing multiple precise and order hypotheses on intra-class correlations. These tests can be used when prior information about the intra-class correlations is available or absent. For the non-informative case, a generalized fractional Bayes approach is developed. The method enables testing the presence and strength of grouped data structures without introducing random effects. The methodology is applied to a large-scale survey study on international mathematics achievement at fourth grade to test the heterogeneity in the clustering of students in schools across countries and assessment cycles.

Aula Magna

Symposium: Advanced topics in linking and equating methods

Mag-1 Presmoothing method and model selection for kernel equating

Gabriel Wallin, *Umeå University, Sweden*
Marie Wiberg, *Department of Statistics, Umeå University*

When large-scale testing programs administer multiple test forms over time, the reported scores from different forms need to be comparable. Test score equating enables such comparison, usually by matching the scores by the percentiles of their respective distributions. Kernel equating is an equating framework that approximates the score distributions using kernel smoothing techniques. Previous research has found the estimated equating transformation to suffer less from sampling variability if the distributional approximations are conducted on presmoothed distributions. Presmoothing is thus usually the first step of kernel equating, and it is also an important ingredient in the calculation of the standard error of equating. The most common method to presmooth is by fitting a log-linear model to the data. Previous studies have found the choice of log-linear model to be influential on the equated scores. Recently, Item Response Theory (IRT) has been incorporated within kernel equating which implicitly defines another presmoothing option through the definition of the IRT model. As far as the authors know no comparison has been made between log-linear and IRT presmoothing and their respective influence on the equated scores. This study investigates such differences for various sample sizes, test lengths and data collection designs using both simulated and real test data. Preliminary results indicate that the two methods can result in different equated values that can have real life impact on the test takers. This study illustrates when these differences emerge, and give practical recommendations for when each respective method is the preferred one.

Mag-2 Discrete equated scores: a Bayesian nonparametric approach

Inés M. Varas, *Pontificia Universidad Católica de Chile, Chile*
Jorge González, *Pontificia Universidad Católica de Chile*
Fernando A. Quintana, *Pontificia Universidad Católica de Chile*

In different fields, scores from several forms of an instrument are used to make important decisions. For instance,

in areas of applied psychology such as clinical psychology, neuropsychology, aptitude testing, psychophysiological psychology, cognitive psychology, behavioral toxicology, among others, new screening instruments are in constant development, complicating the task of interpreting the performance of these new tests. Even though these instruments intent to measure similar attributes, they can differ in some aspects. In the field of educational measurement, before reporting scores using an arbitrary standardized scale, equating methods are used to make scores comparable. In other fields, however, a common feature of the forms instruments is that they are mostly defined on a discrete scale, i.e., integer numbers. It is commonly the case that score cut-offs, which are used to classify individuals, sometimes do not meet the clinical need if reported as unrounded, not integer values. Equating methods developed so far are thus not completely suitable because all of them are based on continuous approximation of scores distributions leading to equated scores that are formally defined on a continuous scale. Our approach tackles the problem of defining an equating method that considers the discreteness of the score scales. We propose a latent Bayesian nonparametric model for scores distributions considering scores as ordinal random variables. Discrete equated scores are obtained using the one to one relation between the score values and the latent variable. Different methods to evaluate the performance of our proposal are discussed.

Mag-3 Bayesian linking and equating methods

Matthew Johnson, *Educational Testing Service, United States*

Johnson (2011) and Varas, et al. (2018) presented Bayesian methods for linking the discrete valued observed scores from two tests based on the equivalent groups design and the assumption that both tests provide a perfect partial ordering of the examinees. Under the assumptions of perfect partial ordering and equivalent groups the methods obtain the joint posterior distribution of the probability masses $p_{xy} = Pr\{X = x, Y = y\}$, where X and Y denote the test scores on the two forms of the test. The linking function is then defined as an appropriate summary (e.g., the mean or median) of the posterior predictive distribution of Y given X and the observed data used to estimate the linking function. While the method is optimal as a predictive linking method it does not satisfy the symmetry requirement of equating. In this talk the Bayesian linking method is modified to produce a symmetric equating function and extended to the NEAT equating design. The method is evaluated through a Monte Carlo experiment and its utility is demonstrated with a real-data analysis.

Mag-4 Sharp bounds on the score distributions in test score equating

Jorge Gonzalez, *Pontificia Universidad Católica de Chile, Chile*
Ernesto San Martín, *Pontificia Universidad Católica de Chile*

By deriving bounds for the score distributions on a region where they are identified by the data, González & San Martín (2018) used the concept of partial identification of probability distributions to offer a solution to the inherent identifiability problem underlying the NEAT equating design. In this talk we discuss an approach where restrictions on the score probability distributions are imposed in terms of stochastic dominance so that narrower bounds can be obtained. The approach is illustrated using data from the sciences section of the Chilean university entrance test where students are assumed to select the test they believe will score better, thus imposing a stochastic dominance condition on the score distributions.

Sala Colorada

Symposium: Methodological contributions to improve policy evaluation in highly uncertain contexts

Col-1 Partial Identification to improve public policy evaluations

Trinidad González-Larrondo, *Pontificia Universidad Católica de Chile, Chile*

The evaluation of public policies is subjected to uncertainty not only by problems of design but also due to the fundamental problem of causal inference: It is not possible to observe a same statistical unit under two mutually exclusive treatments. This means that the counterfactual probabilities are non-identified. It is known that this problem is solved through a strong ignorability condition. In spite of its advantages, this solution hides this "connatural uncertainty." Therefore, a relevant question is the following: How to make explicit such uncertainty? In this talk, we intend to answer this question in the context of a Chilean public policy related to school leadership. We show that uncertainty is related to different assumptions leading to solving such an identification problem. Some solutions are discussed. Special attention is focused on the policy recommendation coherent with each solution.

Col-2 Students as raters: a multilevel PCM model to compare classrooms

Diego Carrasco, *Pontificia Universidad Católica de Chile, Chile*

The comparison of learning environments requires to resolve at least two problems, present in the literature on school climate research: the measurement problem and the endogeneity problem. In this presentation, I illustrate one approach to resolve these two problems from a generalised latent variable framework. I first discussed why a multilevel IRT model is needed to extract the variance of interest when students act as raters of their learning environment (see Stapleton, et al., 2016). The first threat at this point is to misspecify the model as a traditional compositional model, and as a consequence, to underestimate the effect of the school factor of interest. Secondly, we compare schools using a mixed model to get cluster specific differences regarding the chosen learning environment factor. To assess if these differences are robust to endogeneity with the uneven allocation of students to schools, we compare these estimates with pseudo residuals, which are purged from the socioeconomic status of the students, following Castellano, et al. (2014) proposal. The second threat at this point is to underestimate the effect of interest, due to the endogeneity bias. We use data from the International Civic and Citizenship Study from 2009, from the Latinoamerican regional model. We specify as the outcome the regional test of civic knowledge which made available the cognitive items; we generate a measure of classroom differences regarding the likelihood of discussion of political and controversial issues in the classroom, and we included socioeconomic status of student's families as a covariate.

Col-3 Improving weak impact evaluation designs

Verónica Santelices, *Pontificia Universidad Católica de Chile, Chile*

We present and discuss two frameworks for strengthening the inherently weak control group only-post test design (Shadish, Cook & Cambell, 2002) when no random assignment is feasible. We do so in the context of evaluating the impact of teacher preparation programs using Value-Added measures (VAM). We explain the rationale underlying the use of VAM as a measure of program impact and its limitations and we contribute to the ongoing discussion by making a necessary distinction between impact and assessment measurement (Belcher & Palamberg, 2018) and by applying contribution analysis (Mayne, 2002: 2012) and situational enhancement validity analysis (Eckert, 2000) using the case of a Chilean teacher initial preparation program as example.

Col-4 IRTrees aproach in personality data with multiformat response scale

Francisca Calderón, *Pontificia Universidad Católica de Chile, Chile*

In recent years, standardized assessments in educational measurement programs have become relevant worldwide, as many of the decisions made by policy-makers in the educational system are based on the information provided by theme. In addition to measuring cognitive abilities, these programs have been interested in measuring non-cognitive aspects through people's perceptions and attitudes, such as; academic motivation, school climate, healthy lifestyle, among others. Together with the Simce cognitive tests in Mathematics, Language, Natural Sciences, History, and English, Chilean students, -as well their teachers and parents-, express their perceptions and attitudes towards different non-academic aspects, through the Quality Questionnaires and Education Context. These questionnaires collect information using different Likert type items of ordinal categorical response (e.g., (1)"Strongly Disagree", "Disagree", "Agree", "Strongly Agree", and (2) "Very unsatisfied", "Unsatisfied", "Satisfied" and "Very satisfied") yielding score data defined on a polytomous scale. Ignoring response styles can result in biased latent trait estimates thus leading to invalid inferences on attitude and perceptions. Moreover, taking into account that this educational assessment is periodically applied to monitor the evolution of non-cognitive aspects, correcting for response styles and bias is crucial for appropriate comparison of these type of measurements. Recently, item response tree (IRTree) approaches are applied for modeling response style in non-cognitive measurement, but in general, using the same set of possible responses for all items on a scale. The aim of this work is to propose an adequate methodology to use IRTrees in the analysis of test with more than one response scale.

Sala Matte

e-learning

Mat-1 Measuring student's activity in MOOCs using extensions of the Rasch model

Dmitry Abbakumov, *University of Leuven, Belgium*
Piet Desmet, *University of Leuven*
Wim Van den Noortgate, *University of Leuven*

Understanding student's activity is essential for online learning platforms and course developers because the activity is directly related to the students' performance. Researchers and practitioners typically describe student's activity through a proportion of videos viewed and skipped,

assessments taken and skipped, correct and incorrect responses in assessments. These measures are simple and intuitive. However, an ordinal scale of such measures ceils understanding at an aggregated level, not a level of a unit of content. Moreover, it hampers predictions on how a student may interact with a unit of content. To cover this gap, we address to Item Response Theory (IRT) and propose extensions for the Rasch model, which include effects of content type (e.g., video lectures, reading assignments) and student-specific individual dynamics. We illustrate them on logged data from massive open online courses from the Coursera platform (N = 21,116, and >850K student-content interactions). Finally, we conduct cross-validation to evaluate the quality of predictions on how students interact with units of content. We found that (a) the probability of completing a unit of content varies among students and among units; (b) the probability that a student will actively work on content is higher for video lectures than for reading assignments; (c) in general, the probability decreases through a MOOC, but the evolution varies among students. Finally, the proposed approach shows high capacity in prediction whether students will interact actively with the units of content with the overall accuracy of .87.

Mat-2 Psychometric properties of MOCA: Digital assessment tool for learning analytics

Hongwook Suh, *Nebraska Department of Education, United States*
Jaehwa Choi, *George Washington University*
Ji Hoon Ryoo, *University of Southern California*
Sunhee Park, *University of Virginia*

It is necessitated to develop intelligence tests measuring cognitive abilities that can be utilized in the research area of cognitive science encompassing philosophy, psychology, linguistics, artificial intelligence, robotics, and neuroscience. Many tools available have been somewhat limited to be used in learning analytics. The main purpose of this paper is investigating various psychometric properties of a newly developed computerized non-verbal cognitive ability assessment, named "Measure of Cognitive Ability (MOCA)". MOCA is (1) based on the current theory of cognitive science such as Cattell-Horn-Caroll theory of cognitive abilities, (2) a non-verbal cognitive ability assessment such as Raven's progress matrix, Leiter international performance scale – revised, test of non-verbal intelligence – 4, (3) a digital assessment tool that was developed within assessment engineering framework such as Automatic Item Generation (AIG). Using AIG technology, which can instantly generate large number of parallel multimedia assessments/items in Internet environment through computerized item modeling, MOCA aims to overcome various limitations of the conventional paper-pencil type assessments. This study examines the various psychometric properties (i.e., item difficulty, reliability, and other validity evidences) of MOCA by analyzing empirical response data of participants ranged from 6 to 24 years old, focusing on parallel form generator functionality. In particular, this study includes a validation study on whether MOCA can be used as a component in learning analytics by confirming how the cognitive ability such as spatial reasoning measured by MOCA is related to mathematical achievement. Some potential limitations of the study as well as future research directions are discussed.

Mat-3 On test length in mastery tests based on learning objectives

Anton Béguin, *Cito Institute for Educational Measurement, Netherlands*
Hendrik Straat, *Cito institute for educational measurement*

In individualized learning trajectories, it could be valuable to administer small tests that focus on a specific learning outcome to determine mastery of the learning objective and to evaluate whether a student can progress to other learning objectives. For this type of application, testing time competes with direct learning time, and a large number of learning objectives could invoke a potentially large burden due to testing. Thus, it is effective to limit the number of items and to reduce testing time as much as possible. However, the number of items is directly related to the accuracy of the mastery decision and the applicability of this type of formative evaluation in practical situations. In this paper, we will apply informative Bayesian hypotheses to evaluate test lengths and cut-scores for items typically used in mastery testing, with a focus on fine-grained learning objectives. Typically, the items in assessments that focus on mastery of a learning objective are constructed in such a way that students who have mastered the learning objective will have a high probability of answering the items correctly. Students who have not mastered the learning objective will have a smaller probability of answering the items correctly. We establish guidelines for test lengths and cut-scores in three studies: a simulation study with homogeneous item characteristics, an empirical example, and a simulation based on the empirical example with heterogeneous item characteristics. In these studies response patterns are evaluated using Bayes Factors comparing the posterior distributions in line with mastery and non-mastery.

Mat-4 Item selection methods for e-learning assessments under cognitive diagnosis models

Sangbeak Ye, *University of Missouri - Kansas City, United States*

Computerized adaptive assessments are becoming a common educational medium with increasing availability of e-learning items. With items that are capable of delivering educational information and subsequently evaluating the mastery, administration of such items shall extend beyond the existing methodologies that were solely focused on accurately estimating a fixed set of attributes. Within e-learning assessments, the primary aims are twofold: i) promoting mastery of unlearned attributes and ii) detecting mastery with a minimal number of items. In this study, an item parameter that measures the pedagogical value of e-learning items is introduced and its use is showcased under the framework of cognitive diagnosis models. Pedagogical value of each item is viewed as a probability of latent transition from absence of targeted attribute(s) to any subset of newly mastered attributes. A variation of item selection methods that utilize the pedagogical value item parameter serves to promote learning and assist the detection of mastery. This study considers learning assessments where a complete mastery of multiple attributes is sought for any attribute vector pattern to start with. The performance of the item selection methods is evaluated with the number of items needed to achieve a complete mastery. Simulations are conducted to examine the properties of the item selection methods for promoting mastery under different cognitive diagnosis models. Methods for estimating the pedagogical value parameter are proposed and we examine the validity of these methods in assessing individual learning rates.

Auditorium 2
Measurement invariance and DIF I

Au2-1 Multiple cause multiple indicators model for unbalanced group designs

Felipe Valentini, *Universidade Sao Francisco, Brazil*
Nelson Hauck-Filho, *Universidade São Francisco*
Ricardo Primi, *Universidade Sao Francisco*

Methods for detecting DIF usually require big samples, which poses a serious challenge for the research in small populations, such as people with disabilities. In the current study, we investigated the capacity of the Multiple Cause Multiple Indicators (MIMIC) model in detecting DIF in conditions where the source of bias has an unbalanced distribution. Using simulated data, we specified a factor model containing 12 observed variables, three factors, and a group variable with two category possibilities, from which we generated a sample size of 500 subjects. We manipulated the proportion of cases in each group according to three conditions: (1) balanced design: ng1 = 250, ng2 = 250; (2) unbalanced: ng1 = 50, ng2 = 450; and (3) severely unbalanced: ng1 = 30, ng2 = 470. For each condition, we simulated 500 datasets, specifying six items to have their factor loadings biased by the group variable. Results indicated that the bias on parameter estimates was below .10, and that the 95% confidence interval contained the true parameter in more than 93% of replications. Nevertheless, the average standard errors for the estimates increased around 70% from the balanced to the unbalanced condition, and up to 140% in the severely unbalanced condition. Findings indicate that, despite the large standard errors in the unbalanced group conditions, the MIMIC model was able to accurately detect group bias in the target items. Because of their parsimony and robustness, MIMIC models are viable options for detecting bias in study designs with small sample groups.

Au2-2 How to select prior variance in Bayesian approximate measurement invariance?

Lijin Zhang, *Sun Yat-Sen University, China*
Junhao Pan, *Sun Yat-Sen University*

Measurement invariance (MI) is a pre-requisite for comparison between groups under the framework of latent variables. But in traditional multi-group factor analysis, scalar invariance is almost unachievable in practice (Marsh et al., 2018). The Bayesian approximate MI proposed by Muthén and Asparouhov (2013) compensates for this limitation to some extent by providing a zero-mean, small-variance prior for the differences in measurement parameters. However, two main problems in this method have hindered its application. Firstly, model evaluation remains underdeveloped. Secondly, there is no guideline for the selection of prior variance (e.g., De Bondt & Van Petegem, 2015; Fong, 2014). To address the latter problem and set a foundation for further research, four different priors were provided to recover the latent mean difference under different conditions using one-factor models. The simulation study conditions include: model size, number of groups, ratio of group size to model size and noninvariant size. Recommendations were provided based on the results to help researchers get unbiased and effective estimates in an approximate MI analysis, and a real data set was analyzed to demonstrates the validity and practical usefulness of this guideline. Moreover, future studies should develop guideline for fitting

criteria because simulation study has found that model is still well-fitting when estimates are biased. In addition, based on the results of simulation study, future studies are also suggested to estimate the noninvariant size of datasets and then choose the prior.

Au2-3 The effect of different ratios of group sizes in multi-group DIF detection

Thorben Huelmann, *Univerity of Zurich, Switzerland*

The aim of this work is to determine group size ratios that are optimal for detecting Differential Item Functioning (DIF) in multi-group scenarios by means of Lords generalized Chi-Square test. We address this question both by means of a simulation study and by investigating the theoretical properties of the test statistic. We show that the group size ratios do have an effect on the power, but so do the item difficulties and the specific DIF pattern. In the simulation study we consider a variety of scenarios including balanced and unbalanced DIF with different patterns and different amounts of DIF, varying numbers of total observations, and varying numbers of groups. For each scenario, different combinations of group sizes were sampled to investigate their effect on the hit rate of DIF detection. We will illustrate how these factors affect the results and try to derive a rule of thumb for planning multi-group DIF comparisons.

Au2-4 Multigroup factor rotation for unraveling factor loading non-invariance

Kim De Roover, *Tilburg University, Netherlands*
Jeroen K. Vermunt, *Tilburg University*

Multigroup exploratory factor analysis (EFA) has gained popularity to address measurement invariance for two reasons. Firstly, repeatedly respecifying confirmatory factor analysis (CFA) models strongly capitalizes on chance and using EFA as a precursor works better. Secondly, the fixed zero loadings of CFA are often too restrictive. In multigroup EFA, factor loading invariance is untenable if the fit decreases significantly when fixing the loadings to be equal across groups. To locate the precise factor loading non-invariances by means of hypothesis testing, the factors' rotational freedom needs to be resolved per group. In the literature, a solution exists for identifying optimal rotations for one group or invariant loadings across groups (Jennrich, 1973). Building on this, we present multigroup factor rotation (MGFR) for identifying loading non-invariances. Specifically, MGFR rotates group-specific loadings both to simple structure and between-group agreement, while disentangling loading differences from differences in the structural model (i.e., factor (co)variances).

Au2-5 Cross-level metric invariance violation explaining for multi-group scalar invariance violation

Jingdan Zhu, *Ohio State University, United States*
Paul De Boeck, *Ohio State University*

In this study, we explain that a violation of metric cross-level invariance leads to a violation of scalar multi-group invariance. We reason from a violation of cross-level invariance to violations of measurement invariance in a multi-group model, whereas other authors have focused on multi-group conditions for cross-level invariance (e.g., Jak & Jorgersen, 2017). They are two sides of the same coin. The reasoning is as follows: we can derive expected item means for level 2 units based on their position on the level-2 factors. Without cross-level metric invariance, the group differences in item means cannot be explained by the group differences in multi-group factor means. In the absence of metric cross-level invariance, level 2 implies another source of group differences in item means. This also means that the specific scalar invariance violations in a between-group model can be predicted from the level-2 factors. We plan to provide an illustration with real data from an annual employee opinion survey in 206 locations (after deleting missing data for the particular job type and job level, resulted in 175 available locations). We run a multi-level model to check metric invariance between level 1 and level 2. If cross-level invariance does not hold, we then select contrasting groups (i.e., locations) to conduct a multi-group factor model. Such analysis would show that the selected locations have no scalar invariance even when metric multi-group invariance is established (as seems to be the case in the application).

Auditorium 3
Multilevel and longitudinal data analysis

Au3-1 Multivariate longitudinal data with zero inflation: a study of intergenerational exchanges

Irini Moustaki, *London School of Economics and Political Science, United Kingdom*
Jouni Kuha, *London School of Economics*
Fiona Steele, *London School of Economics*
Haziq Jamil, *London School of Economics*

We develop and discuss methods for modelling multivariate longitudinal data that arise from data collected on pairs of subjects (dyads). The paper develops new methods for the analysis of longitudinal multivariate dyadic data, building on existing multilevel and latent variable modelling approaches. The proposed models study intergenerational exchanges in terms of support given and support received from the respondent to parents and children.

We use three waves from the UK Household Longitudinal Survey in order to study the dynamic intergenerational exchanges over time as a function of past exchanges and covariates such as employment status, gender and marital status. We propose a Markov hidden model for multivariate data and an autoregressive latent variable model to capture the bidirectional exchanges, (the support received and given is measured my multiple binary indicators) taking also into account the zero inflation (many of the respondents did not give or receive any type of support). Dyadic data provide a rich source of information on interpersonal processes, but are challenging to analyse because they require a joint modelling approach. The models are estimated in a Bayesian framework.

Au3-2 Bayesian modeling of bivariate associations using piecewise linear mixed-effects models

Yadira Peralta, *Center for Research and Teaching in Economics, Mexico*
Nidhi Kohli, *University of Minnesota*
Eric Lock, *University of Minnesota*
Mark Davison, *University of Minnesota*

Developmental processes rarely occur in isolation; often the growth curves of two or more variables are interdependent. Moreover, growth curves rarely exhibit a constant pattern of change. Many educational and psychological phenomena comprise of different phases (segments) of development over the course of study. Bivariate piecewise linear mixed-effects models (BPLMEM) are a useful and flexible statistical framework that allow simultaneous modeling of two processes that portray segmented change and investigates their associations over time. The purpose of the present study was to develop a BPLMEM using a Bayesian inference approach allowing the estimation of the association between the error variances and providing a more robust modeling choice for the joint random-effects of the two processes. This study aims to improve upon the limitations of the prior literature on bivariate piecewise mixed-effects models, such as only allowing the modeling of uncorrelated residual errors across the two longitudinal processes, and restricted modeling choices for the random effects. The performance of the BPLMEM was investigated via a Monte Carlo simulation study. Furthermore, the utility of BPLMEM was illustrated by using a national educational dataset, Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K), where we examined the joint development of mathematics and reading achievement scores and the association between their trajectories over seven measurement occasions. The findings obtained shed new light on the relationship between these two prominent educational domains over time.

Au3-3 Interrater reliability for multilevel data: A generalizability theory approach

Debby ten Hove, *University of Amsterdam, Netherlands*
Terrence D. Jorgensen, *University of Amsterdam*
L. Andries van der Ark, *University of Amsterdam*

Interrater reliability (IRR), which involves the degree to which ratings are dependent on raters, is imperative in social and behavioral research. It serves as an indicator for measurement precision and (loss of) statistical power in subsequent analyses. Current IRR coefficients ignore the nested structure of continuous multilevel data, which may result in biased estimates, and are not informative concerning the IRR at both the subject and cluster level. In this talk, we propose a generalizability theory (GT) based conceptual framework and estimation method of IRR for nested data. We explain how GT decomposes the variance of multilevel data into cluster, subject, and rater-related components, and propose a Bayesian hierarchical modeling approach to estimate these variance components. We explain how IRR coefficients for both the subject and cluster level can be derived from these variance components, and estimated as different intraclass correlation coefficients. We will provide an applied example from an educational context.

Au3-4 Quality education and economic growth: Causality or co-existence?

Bartosz Witkowski, *Warsaw School of Economics, Poland*
Ewa Witkowska, *The Maria Grzegorzewska University*

In the paper we analyze the impact of the quality of education on the GDP growth. While in their commonly cited paper Hanushek & Woessmann (2010) estimate a linear regression demonstrating the relevance of cognitive skills and basic literacy ratio for economic growth, we elaborate on their research filling in the existing gap in a few ways. Firstly, we use the augmented Solow growth model operationalized by a Barro-type regression, which allows to eliminate the potential omitted variable bias. Secondly, in the original article the quality of education is operationalized by the contemporary PISA test scores while we argue that the lagged performance measure should be applied in order to avoid reversed causality. However, due to the relatively short history of PISA test, only short lags of PISA scores could be used directly. In order to overcome this problem we extrapolate backwards the PISA scores

for the participating countries with the use of additional educational and socioeconomic variables. Thirdly, we use country-level longitudinal data in order to provide additional observations which in turn allow for robust inference and we perform panel cointegration analysis to avoid spuriousness of the results. The analysis is based on the set of countries participating in PISA tests. The results confirm the relevance of the quality of education for the GDP growth and they shade more light on the type of skills that should be considered as the key factors.

Au3-5 Multilevel and dynamical systems approaches to multiple time-scale analysis

Kristine O'Laughlin, *University of California, Davis, United States*
Emilio Ferrer, *University of California, Davis*

In understanding developmental systems, scientists are often concerned with how processes changing rapidly in the short-term relate to other processes changing slowly over the long-term. Data representing these multiple timescales can reveal important information about how people evolve across the lifespan; however, these rich and informative data can pose challenges for data analysis. The two predominant approaches for these data are multilevel modeling (MLM; Raudenbush & Bryk, 2002) and dynamical systems (Scheinerman, 1996). In the MLM approach, the faster moving process is typically conceptualized as nested within the slower moving process. Therefore, this approach presumes that the slower process is driving changes in the overall system. An alternative approach is dynamical systems, which utilize differential equations to examine the interrelations between the processes in continuous time. This talk will present a comparison of these methods using both an empirical and a simulated data example to highlight strengths and limitations of each approach to analyzing multiple time-scale data. Analysis of empirical data suggest that the choice of the analytic method can lead to potentially conflicting conclusions regarding which process is the driving mechanism in the system, and how the system evolves over time. Implications and future directions will also be discussed.

Auditorium 1
Nonparametric modeling

Au1-1 Two new nonparametric local independence tests for the Rasch model

Rudolf Debelak, *University of Zurich, Switzerland*
Ingrid Koller, *Alpen-Adria-Universität Klagenfurt*

A central assumption in most item response theory models is that of local independence. Numerous statistics have been proposed for checking this assumption, with most of them being based on item pairs. We present two quasi-exact tests based on the Q3 statistic for testing the hypothesis of local independence in the Rasch model. The presented tests do not require the estimation of item or person parameters and, because they do not rely on asymptotic theory, can also be applied to small datasets. We compare these tests with three simulation studies. Their results indicate that the proposed quasi-exact tests hold their alpha level under the Rasch model and are more powerful against various forms of local dependence than many alternative parameter and nonparametric model tests.

Au1-2 Extending the Wilcoxon-Mann-Whitney test for latent variables

Heidelinde Dehaene, *Ghent University, Belgium*
Jan De Neve, *Ghent University*
Yves Rosseel, *Ghent University*

The most popular statistical tests for comparing two groups are either the two-sample t-test or the Wilcoxon-Mann-Whitney test. The latter can be preferred over the former for various reasons: it is robust to outliers, it has superior power properties for a variety of distributions, it is applicable to ordinal outcomes and the associated effect size is also meaningful for skewed distributions. Because the t-test can be embedded in a structural equation model, it can be extended to the context of latent variables. For the Wilcoxon-Mann-Whitney test such an extension does not exist. In this presentation, we will show how the Wilcoxon-Mann-Whitney test can be modified to accommodate for a measurement model. We will discuss its main properties and we demonstrate the method using an example in R.

Au1-3 Nonparametric comparison of regression curves for DIF detection

Adéla Drabinová, *Charles University and Czech Academy of Sciences, Czech Republic*
Patrícia Martinková, *Charles University and Institute of Computer Science, Czech Academy of Sciences*

Many methods for detection of differential item functioning (DIF) are derived from comparison of item characteristic curves (ICC) including approaches based on IRT models and non-IRT methods such as logistic regression or its extensions. However, both approaches assume parametric model for probability of correct answer. In this talk

we introduce general nonparametric approach of comparison of regression curves proposed by Srihera and Stute (2010). We further adapt the method and we apply it to estimate ICCs and to test for DIF. For that purpose, we propose several weight functions which can improve DIF detection process and we compare their properties via simulation study.

Au1-4 A nonparametric method for learning in cognitive diagnosis

Hulya Duygu Yigit, *University of Illinois at Urbana-Champaign, United States*

Chia-Yi Chiu, *Rutgers University*

Shiyu Wang, *University of Georgia*

Jeff Douglas, *University of Illinois at Urbana-Champaign*

Cognitive diagnostic models are developed to partition students into several latent classes based on their mastery or nonmastery of a given set of attributes. They are beneficial for producing a fine-grained breakdown about the examinee's status; however, they simply lack information about the dynamic portion of learning. Recently, there have been studies conducted on CDM by incorporating learning models which recognize that traits may be acquired through the assessment, particularly when coupled with an intervention (e.g., Chen, Culpepper, Wang, & Douglas, 2018; Kaya & Leite, 2017; Wang, Yang, Culpepper, & Douglas, 2017; Li, Cohen, Bottge, & Templin, 2015). Parametric approaches are the standard but have a major drawback of requiring large sample sizes. On the other hand, a nonparametric approach by minimizing the distance between the observed response patterns and ideal response patterns can be a promising alternative. By only requiring a mapping between the items and attributes, it can be performed with sample size as small as 1 (for an overview of nonparametric approach on CDMs see Wang & Douglas, 2015; Chia-Yi & Douglas, 2013). In the present research, the consistency and performance of this approach are studied under several CDMs considering the longitudinal perspective of learning. In conjunction with the theoretical perspective, several simulation studies are conducted to compare the performance of the nonparametric and parametric approaches in terms of classification rates. The generating model and the distance measure are varied.

Parallel Sessions 1, Tuesday Afternoon

Salón Fresno

Symposium: Process data in international large-scale assessments: Methods and applications

Frs-1 Application of Kaplan-Meier curves for analysis of process data

Denise Reis Costa, *University of Oslo, Norway*

The Kaplan-Meier estimator is a well-known nonparametric statistic used to estimate the survival function from time-to-event data. Time-to-event data frequently appear in medical studies to measure the fraction of patients living for a certain amount of time after treatment or, in other fields, to analyze the time-to-failure of machine parts, for example. With the increase in computer-based assessments, it opens the possibility to evaluate time-to-event data in an educational scenario. By analyzing the amount of time students spent on the task until their final answer and how successful they were to the completion of the task, this work aims to explore the application of Kaplan-Meier estimator for analysis of the item-level performance on PISA 2012 computer-based assessment of mathematics. Using ten released items, Kaplan-Meier curves allowed the examination of student's rates to successful accomplishment on the tasks during the course of the assessment as well as the identification of group differences in the sample. These analyses aim to bring the possibility of the use of Kaplan-Meier estimator to the toolkit of process data from educational assessments, also shedding light on how students respond to the computer-based items and group differences.

Frs-2 Residual dependencies as a window on process data

Paul De Boeck, *Ohio State University and University of Leuven, United States*

Kathleen Scalise, *University of Oregon*

Process data can be sequential data, parallel data, or a combination of both. Sequential data refer to serial steps in the process, parallel data refer to different aspects of a response event (e.g., level of performance, number of actions, time on task), and a combination of both refers to parallel data on intermediate steps. In all these cases, dependencies can be expected because all data concern the same ongoing process. We will focus on parallel data and three types of dependencies which can easily be included in a factor model with a factor per aspect of the

response events. For example, performance level, number of actions, time on task, may each define a factor, but, conditional on the factor, the performance level may also depend on number of actions and time on task. The three types of residual dependencies are cross-loadings (e.g., loadings of performance level on the time factor), direct effects (e.g., time on task has a direct effect on performance level), and residual covariance (e.g., between time on task and performance level). These three types can be differentiated in theory and the process interpretations are different as well. An application will be presented with PISA collaborative task data: level of performance, number of actions, and time on task.

Frs-3 Investigating response processes using log files and finite state machines

Ulf Kroehne, *Leibniz Institute for Research and Information in Education, Germany*
Carolin Hahnel, *Leibniz Institute for Research and Information in Education*
Frank Goldhammer, *Leibniz Institute for Research and Information in Education*
Maria Bolsinova, *ACTNext by ACT*

Can differences in the response processes explain differences between groups and settings? In this talk, we apply a framework for analyzing log files using finite state machines to investigate the comparability of response processes with psychometric response times models. Item-level response times are operationalized for unit-structured tasks of reading competence based on a decomposition of the test-taking process into states. Indicators that can be compared between groups and administration modes are constructed from the reconstructed sequence of states and analyzed within the bivariate generalized linear IRT model framework (B-GLIRT, Molenaar, Tuerlinckx & von der Maas, 2015). A parameterization that includes an interaction between the speed and ability in the cross-relation function was found to fit the data gathered in a national add-on study to PISA 2012 best. After investigating the invariance of the measurement models, we show that the expected gender effect in reading ability coincides with a gender effect in speed in computer-based assessment. Analyzing paper-based and computer-based assessment data together we report an ability mode effect along with a difference in the latent speed factor between modes. However, while the relationship between speed and ability is identical for boys and girls, we found hints for mode differences in the estimated parameters of the cross-relation function used in the B-GLIRT model. Moreover, using data from the German National Educational Panel Study, we report results from our replication of the findings. Finally, we summarize areas of further research using finite state machines to analyze log data in the discussion.

Frs-4 Nonlinear response-level moderation models for product and process data

Dylan Molenaar, *University of Amsterdam, Netherlands*
Maria Bolsinova, *ACTNext by ACT*

When tests are presented in a computerized form, it is feasible to not only record the product of the response process, but also the characteristics of the response process. For each response many additional variables could be recorded: response times, confidence ratings, verbally reported response processes, number of actions in interactive items, number of item clicks, number of eye fixations on the areas of interest, inspection times, response changes, certainty scores, or physiological measures. These variables can be included as moderators in the measurement models for the ability of interest such that one can investigate whether the probability of a correct response is related to the moderator and whether there is an interaction effect between the measured ability and the moderator. Unlike person-level moderators (e.g., gender and SES), response-level moderators have not received much attention in the latent variable model literature. In this presentation we propose parametric nonlinear and nonparametric models for response-level moderation. Such approaches are valuable in exploring the exact form of the relationship between the moderator and the model. Furthermore, the assumption of linearity of response-level moderation might be violated in practice, and using linear models might lead to invalid conclusions about the relationship between the model parameters and the moderator. The response-level moderation models are applied to explore the relationship between response accuracy and the number of interactive actions in PISA computer-based mathematics, and the relationship between response accuracy and response times in PIAAC complex problem solving.

Frs-5 Analyzing log files from multiple items using data mining methods

Hong Jiao, *University of Maryland, College Park, United States*
Xin Qiao, *University of Maryland, College Park*

Data mining methods have been explored to analyze process data in log files. However, most studies were either limited to one data mining technique under one specific scenario or multiple data mining techniques under one

scenario. This study explores the use of four supervised techniques, including Classification and Regression Trees (CART), gradient boosting, random forest, support vector machine (SVM), and two unsupervised methods, Self-organizing Maps (SOM) and k-means, fitted to the 2012 PISA log files retrieved from multiple problem-solving items. After feature generation and feature selection, classifier development procedures are implemented. Suggestions for the selection of classifiers are presented. Interpretations for the results from both supervised and unsupervised learning methods are provided to understand students' problem-solving strategies.

Aula Magna

Symposium: Advanced topics in admission university tests

Mag-1 Who is called to produce changes to the admission system?

Leonor Varas, *Universidad de Chile, Chile*
Daniela Jiménez, *Universidad de Chile*
Constantanza Cortés, *Universidad de Chile*

Admission Systems to selective universities and high stakes tests produce social tensions and public debates. In Chile, as in many parts, the impact of university studies on future salaries is high and well known. In addition, the Chilean admission system is centralized, currently for 41 universities and according to a new law it has to increase considerably in the near future. The structural gaps sustained over time become explosive and threaten the very existence of tests-based access systems. This social irritation tends to spread against assessment in general and increases distrust in the field of psychometrics. In Chile, the accumulation of disagreement with the PSU and the admission system, strongly press for changes. There are proposals for changes including new tests and a new administrative structure will be created by law. Thus the governance of the system will change next year. Although ethical and fairness standards for testing have been established, faced with changes in the testing and admission system, new challenges arise that stress the role of technicians, policy makers, educational institutions, applicants and the general public, which in these matters expresses themself with vehemence. As never before experts are obliged to consider ethical and social aspects, and give public account of their professional decissions.

Mag-2 College admission test scoring gaps: a multilevel analysis

Paulina Perez Mejias, *Universidad de Chile, United States*

While it is widely agreed that performance in college admission tests in Chile is strongly associated with students' socioeconomic status and type of school attended, there is less agreement regarding the degree to which these variables explain differences among students. The scarce studies available rely on single-level regression methods that ignore the nested nature of data, thus potentially yielding biased estimations of parameters of interest. In the present study, we specified a two-level random intercept-and-slope model to accurately estimate the effects of both individual and school characteristics on student performance. We found that more than half of the total variance in test scores is accounted for by schools. The largest within-school gaps were associated with student GPA, followed by gender and educational track, while the largest between-school gaps were given by aggregate student income, school sector, and school average GPA. The analysis of cross-level effects indicated that student GPA and gender varied significantly across schools. The effect of student GPA was stronger in schools with higher average GPA and higher proportions of test-takers, while scoring gaps against female students increased in presence of higher proportions of female students and higher average GPA, as well as in public schools. The large proportion of variance explained by school characteristics revealed that student performance is more of a reflection of opportunities to learn provided by schools rather than individual efforts and capabilities, which may have serious implications for test validity and fairness.

Mag-3 Partial identification in predictive validity of selection tests

Eduardo Alarcón-Bustamante, *Pontificia Universidad Católica de Chile, Chile*
Ernesto San Martín, *Pontificia Universidad Católica de Chile*
Jorge González, *Pontificia Universidad Católica de Chile*

One of challenges in the validation of selection tests is the assessment of their predictive validity. In this context, it is of interest to know about the behavior of the non-selected group, if they had been selected. This "selection problem" is not far from the Chilean Higher Education context, being the predictive validity of the University Selection Test (PSU) a clear example. Predictive validity can be measured through correlation studies, which

include multiple linear regressions and correlation coefficients. In selection problems, the predictors of all individuals are completely observed, but only the outcome of those who were selected is observed leading to conditional probability distributions that are not fully identified. Several authors have proposed methodologies that tackle this problem, for example, through the Pearson correlation coefficient corrected by the restriction in the range of predictors and by modeling E(Y|X) assuming an a priori parametric structure for it and for the mechanism that generates the missing values (e.g., Heckman, 1976,1979; Marchenko & Genton, 2012). In this work we propose a new method to evaluate the predictive validity using bounds for marginal effects defined on a partially identified region. Bounds for marginal effects are computed using semiparametric regression models and semiparametric derivatives techniques. The method is illustrated using data from the Chilean PSU where the outcome variable corresponds to the performance of first semester students. The results will be compared with Stoye's (2006) method.

Mag-4 A Bayesian graphical and probabilistic proposal for bias analysis

Claudia Ovalle-Ramírez, *Pontificia Universidad Católica de Chile, Chile*
Danilo Alvares, *Pontificia Universidad Católica de Chile*

One of the main concerns in educational policies is to analyze whether a national test is fair for all students, especially for the economically and socially disadvantaged groups. In the current literature, there are some methodological proposals that analyze this problem through comparative approaches of performance by groups. However, these methodologies do not provide an intuitive graphical and probabilistic interpretation, which would be useful to aid the educational decision-making process. Therefore, the objective of this work is to bridge these gaps through a methodological proposal based on the two and three-parameter logistic models, where we evaluate the performance of each group using guessing and difficulty parameters estimated from a Bayesian perspective. The difference between parameters of each group and their respective 95% credible interval are displayed in graphical form. In addition, we also calculate the mean of the posterior probability of all the differences of each parameter for the paired groups compared. This probabilistic measurement provides a more accurate perception of intergroup performance by analyzing all items together. Our methodology is illustrated with Chilean University Selection Test (PSU) data of 2018, where the analyzed groups are students from (regular) high schools versus technical high schools. A sensitivity analysis between the two logistic models is presented. All analyzes were performed using the R language with the JAGS program.

Sala Colorada
Networks I

Col-1 An approach for controlling the false discovery rate in sparse networks

Ginette Lafit, *University of Leuven, Belgium*
Eva Ceulemans, *University of Leuven*

Gaussian Graphical Models (GGMs) are extensively used in many research areas, such as genomics, neuroimaging, and psychology, to study the partial correlation structure of a set of variables. In many applications, it makes sense to impose sparsity as it is theoretically meaningful and/or because it improves the predictive accuracy of the fitted model. However, state-of-the-art estimation approaches for sparse GGMs fall short because they often yield too many false positives (i.e., partial correlations that are not properly set to zero). In this talk, first we demonstrate through an extensive simulation study how state-of-the-art estimation approaches: the Graphical lasso, l1 regularized nodewise regression, and joint sparse regression, do not recover the true network and tend to incorporate many false positives. Next, we present a new estimation approach that allows to control the false discovery rate better. Our approach consists of two steps: first, we estimate an undirected graph using state-of-the-art estimation approaches. Second, we try to detect false positives edges, by flagging the partial correlations that are smaller in absolute value than a given threshold; the flagged correlations are set to zero. The results of an extensive simulation study, show that our two step method outperforms the Graphical lasso, l1 regularized nodewise regression, and joint sparse regression. We also illustrate our approach by using it to estimate (1) a gene regulatory network for breast cancer data, (2) a symptom network of patients with a diagnosis within the nonaffective psychotic spectrum and (3) a symptom network of patients with PTSD.

Col-2 The impact of endogeneity on network psychometric models

Teague Henry, *University of North Carolina at Chapel Hill, United States*
Ai Ye, *University of North Carolina at Chapel Hill*

The recently developed network psychometric modeling framework seeks to model psychological phenomena using networks of interacting variables, as an alternative to latent variable models. Increasingly, psychologists have been using this framework to analyze understand a variety of topics in psychology, particularly in areas such as psychological disorders and the structure of personality. While some work has been done establishing the replicability and stability of psychometric networks (e.g. Epskamp, Boorsboom & Fried 2018; Forbes, Wright, Markon & Krueger 2017), little has been done to evaluate the impact of unmodeled endogeneity on network estimation and the resulting inference. Endogeneity, where the error term in an equation correlates with a predictor, is caused by a variety of conditions, including omitted common causes, measurement error and/or unmodeled simultaneity, and can lead to biased and inconsistent parameter estimates. In network psychometric models, this is particularly concerning as biased estimates of the network parameters can have unpredictable effects on network statistics, which are commonly used numerical summaries of network topology. In this study, we examine the effect of unmodelled endogeneity on network psychometric models in a series of simulation studies based on the commonly used National Comorbidity Survey – Replication dataset. We show that unmodelled endogeneity results in biased network estimates, and this has a substantial effect on network statistics, specifically ones calculated using shortest paths. We discuss methods to detect and model endogeneity in network psychometric models and discuss implications for research on psychological disorders using network methods.

Col-3 Sex differences in subscales of coping schema with network analysis

Shirin Rezvanifar, *Allameh Tabataba'i University, Iran*
Hassan Mahmoudian, *Allameh Tabataba'i University*
Hamid Khanipour, *Kharazmi University*

Coping schemes are the result of individual efforts in challenging stress. This study compared Subscales of Coping Schema between male and female with using network analysis. In this study participated 1046 Iranian students (615 male and 441 female). The centrality indices are very close to each other and also network analysis graphs showed women in comparison with men used more coping schema to confront with stress. The most commonly schema used between two groups are self-restructuring and males used religious coping schema less than females

Col-4 Comparison of pairwise and partial correlation networks

Anna Wysocki, *University of California, Davis, United States*
Mijke Rhemtulla, *University of California, Davis*

Network models are an increasingly popular method to study the structure and characteristics of psychological constructs such as psychopathology and personality. Edges (i.e., the connections between nodes) can represent many different types of connections, including correlations or partial correlations. But researchers almost exclusively use partial correlations to estimate the relations between variables in psychological networks. However, in genetic and neural networks pairwise correlation are commonly used as well. A key difference between these networks is that partial correlations represent conditional independencies between variables, the unique relation between two variables partitioning out the variance explained by other variables in the model. Pairwise correlations, or marginal correlations, index the strength of a relation between two variables without considering other variables. Partial correlation networks emerged as the more popular choice in psychology because they can, in theory, map onto the causal structure of the underlying construct. However, we propose that a closer consideration into the features of each network is warranted. Here we simulated data from a range of population networks and compared the correlation and partial correlation approaches to estimating the network structure. We examined recovery of the network structure, and bias and precision of centrality indices. Based on these results, we discuss theoretical and statistical considerations that should inform the choice of method, such as how the research question, sample size, and set of measured variables impact the interpretability of each network and its metrics.

Sala Matte
Response styles

Mat-1 Modeling random responding behavior and extreme response style in surveys

Zechu Feng, *The University of Hong Kong, Hong Kong S.A.R., China*
Kuan-Yu Jin, *University of Hong Kong*
Jimmy de la Torre, *University of Hong Kong*

Likert-type scale are widely used in social and psychological surveys. Some aberrant responding behaviors have been pointed out in literatures. For example, respondents with low motivation attempt to go through the instrument quickly and consequently endorse the given response

options randomly. Another case is extreme response style (ERS), which means that respondents' tendency of using extreme responses would intervene the usage of response categories. The two responding behaviors were considered in this study. We aim to propose a new item response theory (IRT) model for distinguishing random respondents from attentive respondents and account for ERS of attentive respondents to improve the measurement quality of the test. The parameters were estimated with the Markov chain Monte Carlo (MCMC) method, which is available via the free software WinBUGS. The preliminary results showed the estimated parameters in the new model can be recovered very well. The results also indicate that fitting the new model to data without random responses and extreme responses did not yield seriously biased estimations of parameters. In the opposite way, ignoring random responses and extreme responses by fitting standard IRT models resulted in seriously biased estimations on the item slope parameters; and the item difficulty parameters and the threshold parameters were biased as well. The implications and applications of the new model will be illustrated by an empirical study.

Mat-2 Evaluating competing multiprocess IRT tree models in studies of response style

Nana Kim, *University of Wisconsin - Madison, United States*
Daniel Bolt, *University of Wisconsin - Madison*

Multiprocess IRT tree models have received much recent attention as a basis for response style modeling. The models require specification of a sequence of decision nodes by which respondents arrive at a categorical response for an item, with latent traits underlying the nodes often assumed to correspond to response style and substantive trait dimensions. However, interpretation of latent dimensions as response style dimensions assumes a homogeneity of response behavior that might be questioned in many applications. Using a multiprocess IRT tree model of extreme response style as an illustration, we demonstrate using real data how competing response trees can render alternative interpretations of the latent dimensions and the potential for superior fit. Such observations seemingly make mixtures of trees (see also Tijmstra, Bolsinova, & Jeon, 2018) an anticipated reality, and render measurement of response style a more uncertain process than is likely reflected in a single tree.

Mat-3 An integration of approaches to modeling response styles in the Divide-by-Total framework

Mirka Henninger, *University of Mannheim, Germany*
Thorsten Meiser, *University of Mannheim*

Many psychometric models have been proposed in order to measure and control for interindividual response tendencies in rating data. These models differ in the way they specify response styles, such as independence of latent dimensions or theoretical assumptions on response style effects. We integrate these IRT model extensions accounting for response styles into one superordinate framework in which we model response styles as person-specific shifts in threshold parameters. This integration allows us to compare the models' assumptions on response styles, and thus highlights commonalities and differences between the various approaches. For example, through the superordinate framework it becomes visible in how far models differ with respect to underlying assumptions and inherent restrictions, such as symmetry of threshold shifts around the item location or constraints on the variance-covariance matrix of random effects. In addition, the common parameterization for response style effects allows us to estimate the models in one software environment. On this basis, we use the Big Five standardization sample to illustrate the application and interpretation of the modeling approaches and derive two new modeling variants for response styles: one approach lifts a symmetry constraint of threshold shifts around the item location, and the second approach reduces the number of estimated parameters by imposing equality constraints on discrimination parameters based on item attributes. Overall, the new superordinate framework of IRT models for response styles provides guidance for model comparison, model choice and model extensions, and is essential to further investigating response styles and their influence on rating scale responses.

Mat-4 Modeling changes in response style with longitudinal IRTree models

Allison Ames, *University of Arkansas, United States*
Aaron J. Meyers, *University of Arkansas*

Beyond the latent trait of interest (TOI), response categories of Likert-type items may elicit different response behaviors depending on individual differences in respondent characteristics. For instance, extreme response style (ERS; Cronbach, 1946) is the tendency to choose the lowest and highest response categories. Midpoint response style (MRS) is the tendency to choose the midpoint response category (Plieninger & Heck, 2018). ERS and MRS can confound the interpretation of scores (Bolt & Johnson, 2009; Jin & Wang, 2014) by introducing bias in the total score with respect to the TOI (Leventhal & Stone, 2018). Preferences for extreme and midpoint responses have been found to differ across nationality,

region, ethnicity, race, language, age, gender, and educational level (summary in Thissen-Roe & Thissen, 2013). What remains unknown is whether ERS and MRS change over time and whether the change is proportional to changes in the TOI. This study proposes a longitudinal item response tree to model changes in average levels and variability in TOI, ERS, and MRS over time. An empirical example using Share Our Strength's Cooking Matters for Adults Survey will be provided. The example data was collected across the U.S. state of Arkansas at two time points: before a six-week intervention and 3-months post-intervention. The rate of change for ERS, MRS, and TOI will be estimated. Analysis results will inform researchers to the extent to which response style may be confounding the results of program effectiveness. A parameter recovery simulation study will be included to demonstrate adequate parameter estimation.

Mat-5 Multilevel item response tree for examining heterogeneity in response styles

Aaron Myers, *University of Arkansas, United States*
Xinya Liang, *University of Arkansas*
Allison Ames, *University of Arkansas*

Responses to attitude and personality survey items may be a function of both the respondents' trait of interest (TOI) and factors such as response styles—the propensity to respond to items independently or semi-independently of the TOI. Extreme response style, for example, is the propensity to select extreme categories of a response scale, and midpoint response style is the propensity to select the midpoint category. Response styles have been found to differ across geographic regions, ethnicities, and nationalities (Thissen-Roe & Thissen, 2013). The construct-irrelevant variance associated with response styles may lead to invalid inferences made from survey scores. Tree-like multidimensional item response theory models (IRTree) have been used to parse out variability in responses into variability associated with the TOI and variability associated with response styles (Böckenholt, 2012). This study extends the IRTree model to a multilevel framework to uncover heterogeneity in response styles both within- and between-countries. The proposed IRTree consists of the item response measurement model with a person-level latent trait and within-country variance at level 1 and a country-level latent trait and between-country variance at level 2. Attitudes toward social welfare data from the European Social Survey (2016) provide an empirical example. Results from the unconditional model suggest non-negligible variability in the trait of interest and extreme response styles across countries. Country-level covariates will be included to explain heterogeneity in cross-country differences in response styles. A simulation study will be performed to show parameter recovery of the multilevel IRTree model.

Auditorium 2
Scoring and estimation I

Au2-1 Factor score estimation from the perspective of item response theory

David Thissen, *University of North Carolina, United States*
Anne Thissen-Roe, *pymetrics*

The factor scores of Confirmatory Factor Analysis (CFA) models and the latent variables of Item Response Theory (IRT) models are similar statistical entities, so one would expect that their estimation or characterization would follow parallel tracks in CFA and IRT. However, historically they have not. Different procedures have been used to derive factor score estimates and latent variable estimates in IRT, and different computational procedures have been the result. In this presentation we approach factor score estimation for some simple CFA models from the perspective of IRT, with the kinds of graphics that are used to explain IRT estimates of proficiency, and the computational procedures that are used in test theory. We compare traditional "regression" and "Bartlett" factor score estimates with alternative computational approaches to likelihood-based factor score estimates, referring to the expected a posteriori and maximum likelihood estimates of IRT latent variables to clarify relations among the scores. This provides new insights into the ways in which the data are combined into factor score estimates, and reveals that likelihood-based factor score estimates can be computed in closed form for some simple CFA models. The results provide new alternative methods to compute factor scores in the presence of observations that may be missing at random for some variables. In addition, the integration of the CFA and IRT approaches to scoring offers a basis to extend to IRT scores existing results about the use of factor score estimates on the explanatory and response sides of regression analyses.

Au2-2 Penalized estimation of IRT models with multiple location parameters

Alex Brodersen, *University of Notre Dame, United States*

Sample size per response category is an important consideration when estimating polytomous or nominal item response models. When this sample size is not adequate,

common practice in the case of ordinal data provides guidelines such as "collapsing" adjacent categories. In the case of a nominal response data, one may apply a scoring rule that imposes an ordering to the discrete categories. These choices of when to and to not adjust the raw data are based on arbitrary cutoff values, and could impact eventual estimation of latent trait scores. This issue is especially pertinent when there are large differences in the category intercept parameters. One approach to avoiding this rule-of-thumb data manipulation that has been previously unexplored in the literature is to utilize a penalized estimation procedure. Previous use of penalized estimation in item response theory has been limited to dichotomous models or applied only in the context of detecting measurement invariance between groups. This study provides an alternative penalization scheme that results in a data driven approach for handling insufficient data requirements within some categories of individual items. An empirical simulation study was conducted to demonstrate the utility of this method. Results are presented with a discussion of future applications and limitations of the new procedure.

Au2-3 Variational Bayes inference for the cognitive diagnostic models

Kazuhiro Yamaguchi, *Hosei University, Japan*
Kensuke Okada, *The University of Tokyo*

In this presentation, we propose a variational Bayes inference method for the cognitive diagnostic models. We primarily focus on the Deterministic Input Noisy AND gate (DINA) model of cognitive diagnostic assessment for its simplicity and popularity, but our results can be extended to more general class of models. The proposed method, which applies the maximization-maximization algorithm for optimization, is derived based on the optimal variational posterior of the model parameters. For this purpose, we derive the analytic form of the evidence lower bound, and develop an iterative algorithm to maximize it. The proposed variational Bayes inference enables much faster computation than the existing Markov chain Monte Carlo (MCMC) method, while still offering the benefits of a full Bayesian framework. A simulation study revealed that the proposed variational Bayes estimation adequately recovered the parameter values that were used to generate the synthetic data. Moreover, an example using real data revealed that the proposed variational Bayes inference method provided similar estimates to MCMC estimation with much faster computation.

Au2-4 EM-estimation of multilevel item response theory models with nonlinear effects

Tim Fabian Schaffland, *University of Tuebingen, Germany*
Stefano Noventa, *University of Tuebingen*
Augustin Kelava, *University of Tuebingen*

The estimation of Item Response Theory models is an important tool in research and especially in social sciences and psychology. There are many different approaches to understand and calculate the model parameters (e.g., Bayesian, WLS estimation based on SEM with categorical variables, and Maximum-Likelihood Estimation). An often used frequentist possibility is the application of the Expectation-Maximization-Algorithm (EM-Algorithm) which is a Maximum-Likelihood type approach (Dempster, Laird & Rubin, 1977). Different packages in R have implemented some of these methods but in most of them only certain types of IRT-models can be estimated, e.g., lavaan (Rosseel, 2012), mirt (Chalmers, 2012), lme4 (Bates et al., 2015) or TAM (Robitzsch et al., 2017). The model of our interest, a multilevel IRT-model with nonlinear latent effects, cannot be estimated yet in this combination in most of these packages. In this talk we will discuss estimation of IRT-models via the EM-Algorithm in the literature (e.g., Bock & Aitkin, 1981) and we will give some technical insight in the application of the EM-Algorithm to these models. Further we will present some R-packages in which the estimation with the EM-Algorithm is implemented and which IRT-models are possible to estimate with these packages so far. Finally we will propose a new extension of the EM-Algorithm to estimate a multilevel IRT-model with nonlinear effects of latent variables.

Auditorium 3
Assessment

Au3-1 More bang for your buck: Matrix sampling background questionnaire items

Jonathan Weeks, *Educational Testing Service, United States*
Jonas Bertling, *Educational Testing Service*

Policymakers have, historically, viewed the variables from student questionnaires in large-scale assessments as, primarily, contextual factors to help better understand cognitive scores. More recently, questionnaire items and scales have also been viewed as measures of constructs in their own right. These questionnaires typically target a number of constructs within a relatively short assessment time (around 15 to 35 minutes). However,

data quality is likely to suffer in the presence of long questionnaires. Given the trade-off between brevity and depth, matrix-sampling approaches—where different respondents receive different sets of items—may be a viable option to reduce burden while maintaining content coverage, particularly in cases where no individual-level inferences should be made. For this study we used empirical PISA 2012 questionnaire data from seven scales comprised of eight or more items and simulated six different designs: 1) all items; 2) five fixed items; 3) three anchor and sets of two fixed, rotated items; 4) three anchor and two random, rotated items; 5) five random, rotated items; 6) eight fixed sets of rotated items. We considered the reliability of the resulting scores and the correlations between both student and country-level score estimates. The reliabilities of the reduced item sets remained above .7 for most scales. The student-level correlations ranged from around .7 to .95; country-level correlations were generally in the mid to high .9s. The strongest correlations were for the case with five random, rotated items. The results suggest that a matrix-sampled questionnaire may be a viable alternative to current approaches.

Au3-2 A cognitive diagnosis model analysis of a digital literacy assessment

Jimmy de la Torre, *The University of Hong Kong, Hong Kong S.A.R., China*
Qianru Liang, *University of Hong Kong*
Louie Cagasan, *University of Hong Kong*
Kuan-Yu Jin, *The University of Hong Kong*
Frank Reichert, *The University of Hong Kong*
Nancy Law, *The University of Hong Kong*

Digital literacy is a transversal competency needed to be successful in this technology-intensive society. To investigate the digital literacy of Hong Kong students, an assessment measuring five digital skills, namely, information and data literacy (A1), communication and collaboration (A2), digital content creation (A3), safety (A4), and problem solving (A5), was developed. We propose to use a general cognitive diagnosis model framework to examine the mastery of digital literacy skills of Hong Kong primary students. In addition, the relationship between digital skill mastery and a number of demographic variables is also investigated. This study analyzes data collected from 422 Grade 3 students. To determine mastery of various digital skills and test properties, the G-DINA model framework is fitted to the data; to determine the relationship between skill mastery and covariates, the former is regressed on the latter. Preliminary results indicate that several items with low discrimination may need to be revisited to determine whether modifying the item-skill specifications may improve the item quality. Notwithstanding the current test quality, students' skill mastery can be accurately classified – A1 and A4 have the highest attribute classification accuracy (.93), whereas A3 has the lowest (.77). The results also show that all digital literacy skills have low mastery proportions, with A1 and A4 having the lowest (.25) and highest (.43) proportions of mastery, respectively. Finally, the results indicate that language and digital device access are differentially related to the mastery of digital literacy skills, whereas gender and school religion are not.

Au3-3 The Mantel-Haenszel statistic for detecting testlet effects in cognitively diagnostic tests

Youn Seon Lim, *Donald and Barbara Zucker School of Medicine, United States*

A testlet is a cluster of items that shares a common stimulus (e.g., a set of questions all related to the same text passage). The testlet effect calls into question one of the key statistical assumptions of any test: local independence of the test items. Local dependence among test items is typically induced by the under-specification of the latent ability dimensions supposed to underlie a test. Hence, a common remedy for the testlet effect is to incorporate additional latent dimensions to compensate for the under-specification of the latent space. In cognitively diagnostic modeling, this remedy for reinstating local independence is mirrored by adding extra attributes to the K-dimensional latent attribute space of the test. Hence, evaluating whether local independence holds for the items of a given test can be used as a diagnostic tool for detecting testlet effects because the under-specification of the latent attribute space in cognitively diagnostic modeling should typically induce local dependence among the items. This study proposes the Mantel-Hanszel statistic as a tool for detecting dependencies among the items of cognitively diagnostic tests, as they might occur in the presence of testlet effects.

Au3-4 A graphical taxonomy of assessment models

Stefano Noventa, *University of Tuebingen, Germany*
Jürgen Heller, *University of Tuebingen*

In the past years, several theories and frameworks for assessment have been developed within the overlapping fields of Psychometrics and Mathematical Psychology. The most notable are Item Response Theory (IRT), Cognitive Diagnostic Modeling (CDM), and Knowledge Space Theory (KST). Yet, in spite of their common goals, these

theories have somewhat been independently developed to focus on slightly different aspects, for instance CDM was originally defined at the competence level (e.g., skills and abilities) while KST was originally conceived at the performance level (e.g., response patterns). In spite of the methodological differences, these methods have however several similarities and various attempts to bridge them can be found in literature. As an example, connections between CDM and KST (Heller et al., 2015), between KST and IRT (Stefanutti, 2006; Noventa et al., 2018), or between CDM and IRT (Junker & Sijtsma, 2001; Hong et al., 2015) have been highlighted. A particularly interesting similarity lies in the treatment of their conditional probability parameters. A two-processes model is here introduced, which separates guessing and ceiling parameters into a first process due to the effects of individual ability on item mastering, and a second process due to the effects of pure chance on item solving. Based on this model, a graphical taxonomy encompassing IRT models, CDMs, and KST is thus obtained. Consequences for both dichotomous and polytomous items are drawn.

Au3-5 The response order effect of likert scales and its influencing factors

Dongfang Zhao, *Beijing Normal University, China*

It is controversial whether the response order effect of likert scale exists in the education field. If it exists, researchers can influence the research results by manipulating the responses. In this study, 878 teachers finished a long questionaire with four scales. The questionaire was designed as format A whose responses displayed from 1 (strongly disagree) to 5 (strongly agree) and B whose responses displayed from 5 (strongly agree) to 1 (strongly disagree) to prove the existence of response order effect in education field. The scores were higher on format B than that on format A . Teachers were more likely to choose the first response, because of the primacy effect. However, in Format A, teachers were more likely to choose the response with negative meaning. At the scale level, the scales in which the measured concepts referred to "others", the scales which included more questions being expressed in the third person, the scales whose position were higher in the whole questionaire, and the scales in which fewer questions were expressed negatively were more likely to appear the response order effect. In terms of specific questions, the questions expressed positively and the questions whose positions were higher in the whole questionaire were more likely to appear the response order effect. The response order effect reflects obvious cognitive characteristics of easterners. It was influenced by holistic pattern of easterners and high-context culture. So researchers should pay attention to response order effect when designing questionnaires and scales.

Auditorium 1
Applications I

Au1-1 Investigating a new measure for children with autism spectrum disorder

Lorrie Schmid, *Duke University, United States*
Yasmine White, *Voices Together*
Carol Ripple, *Pajama Program*
Geraldine Dawson, *Duke University*

Children with autism spectrum disorder (ASD) have deficits in social and communication skills, which can limit their participation in academic and social activities (Baio et al., 2018). Music therapy is a popular invention for children with ASD, but studies on its efficacy have been inconclusive (Geretsegger et al., 2014; James et al., 2015). This study sought to develop a behavioral assessment tool to examine communication in this population. The measure was evaluated in a sample of 116 elementary school students from 10 ASD classrooms who participated in a music therapy intervention. The main goal of the study was to evaluated whether this new indicator effectively measured communication and socialization functioning by asking four questions: 1) were responses stable prior to the intervention; 2) is the measure unidimensional; 3) are responses stable over time; and 4) do the responses changes during the intervention? The instrument is composed of 20 items, ranging from "what is your name?" to "what do you want in a friend?" The responses are recoded and were coded by researchers on the team, with an inter-rated reliability ranging from .88 to .97. The measure had strong internal consistency within time ($\alpha >$.95) and across time ($\alpha >$ .8). Dimensionality was assessed using confirmatory factor analysis (CFA) within time and a bifactor CFA across time. The measure appears to reliability assess changes in communication and socialization among elementary school children with ASD, especially those with verbal speech. More work in the development of the non-verbal processes is warranted.

Au1-2 Effects of anchoring vignettes on the validity of student self-assessment

Maggie Yue Zhao, *The University of Hong Kong, Hong Kong S.A.R., China*

Anchoring vignettes (King, Murray, Salomon, & Tandon, 2004) are item batteries designed for correcting rating

data differences in subjective self-report survey data. In the present study, we investigated how reporting heterogeneity may bias rating data and how such bias can be adjusted by using the anchoring vignettes approach. Employing empirical data of a student self-assessment measuring intercultural competence, anchoring vignettes were developed representing various levels of intercultural competence. Analyses were performed to evaluate the effects of the anchoring vignette approach on the validity of the student self-assessment. Results suggest that vignette-based adjustments appear to be effective. Promises and limitations of the anchoring vignettes approach in the studied context will be discussed in the presentation.

Au1-3 A virtually new version of the simulator sickness questionnaire

Jessica Cornick, *Oculus, United States*
Richard Yao, *Oculus*

The objective of the current research was to improve upon the simulator sickness questionnaire (SSQ) for consumer virtual reality experiences (VRE). The current 16-item SSQ measures the severity of motion sickness symptoms (MSS) experienced during and after flight simulator sessions via three symptom clusters (Oculomotor Discomfort, Disorientation, and Nausea; Kennedy et al., 1993). Given that many VRE in our Oculus products do not include dramatic locomotion, we hypothesized that the SSQ would not be the most appropriate tool for measuring the severity of MSS resulting from many consumer VRE. To that end, we investigated what MSS factors would emerge when data was collected from users who engaged with non-locomotive mobile and virtual reality technology. Using SSQ data from 788 users, we conducted an exploratory factor analysis on data collected after 30 minutes of virtual reality (VR) head mounted display (HMD) and mobile device use in a variety of experiences on different hardware platforms. While our factor analysis results also revealed a three factor solution, the three factors were different than what was proposed in the original SSQ (General, Visual, and Dizziness symptoms) and there were different factor loading patterns identified for simulator sickness experienced after engaging with a mobile device versus an HMD. Additionally, three of the sixteen SSQ items did not load on any factor significantly suggesting they may not be applicable to non-locomotive VRE. Our new three factor groupings appear more relevant for non-locomotive VRE such as media viewing and social interactions which are common in VR.

Au1-4 Latent variable models for intergenerational exchanges of family support

Jouni Kuha, *London School of Economics and Political Science, United Kingdom*
Fiona Steele, *London School of Economics and Political Science*
Irini Moustaki, *London School of Economics and Political Science*

We describe models for whether individuals give help and support to their non-coresident parents, and whether they receive such help from their parents. This question arises from a project which will study the patterns in such family support between generations, the levels of reciprocity in it, and how these change over time. To represent these data, we define a model with two correlated latent variables, representing the tendencies of an individual and their parents to give help, measured by binary items on specific kinds of help. The model includes also a latent class component to represent the large number of individuals who give no help at all, and a partially non-invariant measurement model which allows for the possibility that some types of help may be differently strong signals of helping tendency for different types of individuals. The models are developed and estimated using data from the UK Household Longitudinal Study.

Au1-5 The relationship of inquiry-based teaching and science achievement in China

Liu Yue, *Beijing Normal University , China*

In 1962, Schwab and Brandwein published a far-reaching book that presented the value of inquiry-based teaching (IBT). IBT has become an important trend of science education reform in many countries. The emphasis put on IBT is based on the assumption that IBT can effectively improve the science achievement of all students (National Research Council, NRC, 1996). However, the relationship between IBT and science achievement does not come to a conclusion in both large-scale evaluation data study and well-designed experimental or quasi-experimental research. Practitioners expect IBT to promote students' science achievements (Borman et al, 2008), but there are also some studies that result on the contrary (Lavonen et al., 2009, Chi et al., 2015). Perhaps it should not limit the relationship to a linear relationship (Jiang et al., 2015, Teig et al., 2018). The present study investigates this relationship among students in the fourth grade of primary school using large scale data collected in China in 2017. A three stage stratification cluster sample design is used. Science acheivements along with background information of students and science teachers are collected. Teacher's

perspective is also included in the study to compensate for students responses (Jingoo et al., 2018). HLM analysis shows a curvilinear relationship like the letter "J". In the early stage of implementation of IBT, it has a negative impact on science achievement, it will show a positive correlation with much more implementation of IBT. The explanation and implication for science teaching in primary schools are discussed.

Invited and State-of-the-Art Speakers

Salón Fresno
Invited Speaker: Dani Gamerman

Frs-1 Dynamic generalized structural equation modeling, with application to the effect of pollution on health

Dani Gamerman, *Universidade Federal do Rio de Janeiro, Brazil*

Structural equation modeling (SEM) is a very useful tool for psychometrists as relations between latent constructs are frequently built, e.g., the effect of stress level in the student proficiency. This tool is also valuable in many other areas of Science. Important requirements for the SEM framework is to be able to incorporate effects associated with the passage of time in the so-called dynamic SEM and the generalization to handle non-Gaussian measurements. Dynamic generalized SEM is geared towards accommodating these extensions. We show how to: a) set up a model in state-space format; b) to perform inference; c) to make model selection and predictions and; d) to summarize results obtained from the analysis. Inference is performed with a Bayesian perspective, facilitating the use of MCMC methods and other useful modelling tools, including parsimony and identification. An application on the effect of pollution on health is used to illustrate many of these issues in the context of a real data set from northern Italy. Joint work with Luigi Ippoliti and Pasquale Valentini (Pescara).

Aula Magna
State-of-the-Art Speaker: Kathleen Gates

Mag-1 Assessing individual differences in non-traditional data structures

Kathleen Gates, *University of North Carolina at Chapel Hill, USA*

Time series data bring new opportunities – and problems – for psychological researchers. Examples of this data are plentiful and include: self-assessments obtained multiple times a day for numerous days; psychophysiological measures (e.g., functional MRI); and passive data continually captured from devices. Ideally, researchers can use this data to investigate underlying processes of interest in a way that previously was unattainable. New questions can be asked, and new types of information learned. In reality researchers face numerous challenges that can diminish the full potential of this data. At its most fundamental, it can be difficult to formulate research questions. Often little prior work has been done that can inform hypotheses on dynamic processes of interest. For this reason, many methods used on these data have a data-driven component. The question then becomes, which method to use? Even if one does have hypotheses, translating them into quantifiable and testable research questions can be daunting given that the available methods and measurement approaches may be unknown. The present talk provides an overview of the types of questions that are currently being answered with this data and examples of a few state-of-the-art methods.

Parallel Sessions 2, Tuesday Afternoon

Salón Fresno
Panel Discussion: How testing organizations have shaped the psychometric research and employment opportunities for psychometricians

Frs-1 Invited Panel

Alina von Davier, *ACTNext by ACT, United States* **Jorge Manzi**, *MIDE, Chile* **Anton Béguin**, *Cito Institute for Educational Measurement, Netherlands* **Matthias von Davier**, *National Board of Medical Examiners, United States* **Duanli Yan**, *Educational Testing Service, United States*

This invited panel will give graduate students a glimpse of the psychometric work and of the culture at a testing company; the panelists will discuss how what they do matters to the real world. The panelists will refer to a few major breakthroughs in psychometrics that have been driven by testing organizations and emphasize the opportunities for students and (junior) researchers to affect policies at the national and international level, and

to affect test takers lives. The participants are senior level representatives from testing companies (ACT, Cito, ETS, NBME, a Chilean organization) who explain how they hire psychometricians, how they rely on psychometricians, what is really important to them (from a psychometric point of view), how the research is organized between foundational and applied research, and how the work is funded. We will also discuss how the testing organizations may handle the collaborations with the academia, the participations in research grants, and the types of internships and awards that are offered to graduate students. The audience will have the opportunity to ask questions and request additional information about internships, employment, and collaborations.

Aula Magna

Symposium: Non-linear methods and complexity analysis: SEM, network psychometrics and time series analysis

Mag-1 Entropy fit index: a new fit measure for dimensionality assessment

Hudson Golino, *University of Virginia, United States*
Robert Moulder, *University of Virginia*
Luis Garrido, *Pontificia Universidad Catolica Madre y Maestra*
Dingjing Shi, *University of Virginia*
Alexander Christensen, *University of North Carolina at Greensboro*

Applied researchers and methodologists use fit indices in their daily activities while assessing the factor solutions of their instruments/data, mainly because it provides useful diagnostic information and enables the comparison of different dimensionality structures. Despite the usefulness of fit indices used in factor analysis to estimate the number of factors, there are two main limitations: 1) there are only a few simulation studies investigating the accuracy of the traditional fit indices; 2) their use is not strongly supported by the studies. In the current presentation a new fit index (termed Entropy Fit Index - EFI) specifically developed for dimensionality assessment will be presented. EFI is based on the non-linear technique proposed in the area of statistical mechanics termed total-correlation, and aims to identify the best structure of a given set of random variables, when comparing two or more dimensionality structures. In other words, EFI can be used to identify the best partitioning of a multidimensional space that better reflects the underlying latent factors. It can be used as a fit index for Exploratory Graph Analysis, a new dimensionality assessment approach part of network psychometrics. To verify the suitability of EFI as a dimensionality assessment index of fit, a Monte Carlo simulation controlling a number of conditions (e.g. sample size, number of factors, strength of the correlation between factors, number of variables) was implemented and the accuracy of this new fit index was compared to traditional index such as CFI, RMSEA and others, pointing to a higher accuracy of EFI.

Mag-2 Decomposing expected moments of products of variables for structural equation modeling

Steven Boker, *University of Virginia, United States*
Timo von Oertzen, *Universität der Bundeswehr*
Andreas Brandmaier, *Max Planck Institute for Human Development*

The expected covariance matrix of a Structural Equation Model composed of a linear combination of variables has long been known (Wright, 1921). While linear combinations of variables are powerful and useful, the current data-rich research environment has spurred interest in nonlinear models for applications such as mixed effects models, moderation models, nonlinear dynamical systems models, and graph analysis. This talk introduces a new method for decomposing the moments of products of variables. The method allows likelihood estimates to be optimized directly for structural equation models that include products of variables without resorting to two step procedures or algebraic constraints on residuals. The result is that many previous models fit by special purpose software and algorithms can be fit by general purpose software such as OpenMx. We expect that this method will enable discovery of novel classes of models that have been previously overlooked.

Mag-3 Investigating the stability and generalizability of dimensionality via bootstrap exploratory graph analysis

Alexander Christensen, *University of North Carolina at Greensboro, United States*
Hudson Golino, *University of Virginia*

Network psychometrics is a rapidly growing field. Exploratory Graph Analysis has emerged as a popular approach for estimating the dimensionality of data in networks. The appeal of EGA is the visualization of the relations between variables and the deterministic allocation of variables into dimensions. Notably, networks tend to be sample-specific, making reproducibility and generalizability a key issue in network psychometrics. To resolve

this issue, we've developed a bootstrap approach called, Bootstrap Exploratory Graph Analysis (bootEGA). bootEGA provides researchers with dimension and item stability statistics as well as item analyses that are akin to exploratory factor analysis loadings. First, we demonstrate via simulation the robustness of bootEGA and provide guidelines based on these results. Then, we apply bootEGA to real-world data to show its effectiveness at identifying problematic dimensions and items. In short, our results that bootEGA is a robust approach for identifying the stability and robustness of dimensionality in data.

Mag-4 Tangle: a computationally efficient measure of time series complexity

Robert Moulder, *University of Virginia, United States*

Recent advancement in data collection technologies have made time series data commonplace in psychological, biological, and behavioral research. Some of these time series behave rhythmically and/or show other simple to describe behaviors. Other time series show complicated and/or chaotic behaviors. These inherent differences in time series complexity may be informative for researchers interested in understanding time based phenomena. Many methods currently exist for estimating complexity in some form; however these methods tend to require a high degree of technical ability, large sample size requirements, or long computation times in order to be properly implemented. These drawbacks impede widespread implementation of these methods by acting as a bottleneck for researchers interested in testing hypotheses regarding time series complexity. In this presentation we introduce Tangle, a new metric for quantifying time series complexity. Tangle is an iterative method based on dimensional embedding and can be used for quantifying complexity in relatively short time series. Tangle is computationally inexpensive and relies only on a data transformation and a single matrix operation repeated until a convergence criterion is met. We demonstrate the application of Tangle to a number of time series data sets and discuss other possible applications.

Sala Colorada
Bayesian statistical inference

Col-1 Estimating testing time through Bayesian stochastic modeling with PyMC

Yi-Fang Wu, *ACT, Inc., United States*

According to the "Fact Sheet: Testing Action Plan" by the U.S. Department of Education (2015), fewer and smarter assessments are desired, with the features of "time-limited" in addition to many other necessary characteristics. In the United States, many states have worked on reducing over-testing and avoid double-testing within districts and/or schools while for a specific assessment or test, the urge on reducing testing time does not seem to come to an end. Whether it is linear testing or computerized adaptive testing, reducing testing time often is the call by test developers, and approaches to reducing the testing time could be studied on a data-driven basis. Through Bayesian stochastic modeling with PyMC (Fonnesbeck, Patil, Huard, & Salvatier, 2015), the study is to estimate testing time in the case when newly-developed tests are shorter than the original ones. We use empirical achievement test data from grades 3, 6, and 9, taking into account student testing experience and year of school learning. Test subjects include English, mathematics, reading, and science, and multiple-choice items is the solely item type in computerized-based testing. Assuming the two-parameter logistic item response model, the lognormal model and the hierarchical model by van der Linden (2006; 2007) are employed and model fit is evaluated through Alaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). Results are to model response times data to inform test developers, and to suggest the proper amount of testing time as a critical consideration in form assembly and actual test administration.

Col-2 Bayesian Multidimensional IRT: How far can it go?

Mauricio Garnier-Villarreal, *Marquette University, United States*
Ed Merkle, *University of Missouri*
Brooke Magnus, *Marquette University*

As a full information analysis, IRT presents estimation challenges. Several of these challenges are particularly relevant to multidimensional IRT (MIRT, Reckase, 2009): with increased dimensions, most maximum likelihood algorithms are limited in which conditions they converge on proper solutions. MHRM has been suggested as one method for estimating MIRT models (Cai, 2010). Another option is to use the limited information estimators that are often used in SEM, such as diagonal weighted least squares (DWLS), which has been shown to be stable for small sample sizes (DiStefano & Morgan, 2014; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). MCMC sampling estimation methods used for Bayesian statistics potentially address many challenges of MIRT (BMIRT).

Thus, the goal of this study to test MIRT estimation and inference in the Bayesian framework. With a simulation study, we test MIRT models across number of dimensions, sample size, and number of indicators for each dimension. BMIRT was estimated with the general Bayesian software Stan (Carpenter et al., 2017), and it was compared with four frequentist methods: one limited information method (DWLS) and three full information methods (EM, QMCEM, and MHRM). Method performance is compared in terms of convergence, bias, and variability of the parameter estimates. We present BMIRT as a framework that is robust to high dimensionality and sample size, allowing applied researchers to make direct inferences from more complex models without data limitations.

Col-3 Sensitivity of Bayesian quantile regression to the choice of scale parameter

Joon-Ho Lee, *University of California, Berkeley, United States*
Sophia Rabe-Hesketh, *University of California, Berkeley*

In Bayesian quantile regression, the most commonly used likelihood is the asymmetric Laplace (AL) likelihood because its maximum corresponds to classic quantile regression and because it is computationally convenient for Markov chain Monte Carlo algorithms. For easier computation, the scale parameter of the AL distribution is often fixed at a pre-estimated value or an arbitrary constant. This paper demonstrates that posterior inference in Bayesian quantile regression with an AL likelihood is highly sensitive to the choice of scale parameter. Based on sensitivity analyses using Monte Carlo simulations and a real data example, we make two claims. First, not only the posterior variance of the regression coefficients directly obtained from the posterior samples, but also the adjusted posterior variance proposed by Yang et al. (2015), are highly sensitive to the value of the scale parameter. We found the adjusted posterior variance mimicking a sandwich estimator to be even more sensitive to the scale parameter than the unadjusted one. Second, in finite samples, both conventional and Bayesian point estimators can be biased at extreme quantiles especially if the scale parameter is fixed at large values. Performance improves if the scale parameter is set to its maximum likelihood estimate for median regression as suggested by Yang et al. (2015) but even in that case, coverage probabilities can be low at extreme quantiles mainly due to biased point estimates.

Col-4 Evaluating stochastic search variable selection for applications in psychology

Sierra Bainter, *University of Miami, United States*
Thomas McCauley, *University of Miami*

Consider a challenge many researchers face in different contexts: given a set of candidate predictors, how do you narrow down to a smaller set of predictors which reliably relate to the outcome? For example, considering a range of vulnerabilities and potential triggers of a psychopathology in a high-risk sample. Even if the set of relevant predictors is selected on a firm theoretical basis, as the number of candidate predictors increases, it quickly becomes intractable to test all predictors simultaneously or to consider all possible sets of predictors. Stochastic search variable selection (SSVS) is a Bayesian variable selection technique which performs a targeted search through all possible models (i.e. combinations of predictors) using MCMC estimation. After sampling thousands of models, the proportion of time each predictor is selected can be used as a metric of the most important predictors that reliably relate to the outcome. SSVS has been used in a range of applications outside of psychology, but has been infrequently considered for psychological applications. Simulation studies evaluating the method have focused on applications where only a few predictors have a "true" relationship with the outcome and predictors are uncorrelated, which is unlikely to generalize to social science research. In this presentation we present results of a simulation study evaluating the behavior of the MIPs in conditions more likely to arise in psychological research. Factors of the study include sample size and reliability of the outcome. We also compare SSVS to lasso and selection of predictors based on univariate correlations.

Col-5 Bayes factors for testing order constraints on variance components

Florian Böing-Messing, *Jheronimus Academy of Data Science, Netherlands*
Joris Mulder, *Tilburg University, Tilburg, The Netherlands*

In statistical practice, researchers commonly focus on patterns in the means of multiple dependent outcomes while treating variances as nuisance parameters. However, in fact, there are often reasons to expect a certain pattern in the variances of dependent outcomes. For example, in a repeated measures study, one may expect the variance of the outcome to increase over time because subjects react differently to a certain treatment. Such expectations can be formulated as order constrained hypotheses on

variance components in the covariance matrix of the dependent outcomes. Currently, however, no methods exist for testing such hypotheses in a direct manner. To fill this gap, we develop two Bayes factors for testing order constraints on variance components of dependent outcomes. The first Bayes factor is based on an unstructured covariance matrix, where order constraints can directly be formulated on the variances on the diagonal of the covariance matrix. The second Bayes factor is based on a compound symmetry-like covariance structure with unequal residual variances, where order constraints can be formulated on these residual variances. For both Bayes factors, a prior distribution needs to be specified under every hypothesis to be tested. Here, we use the encompassing prior approach in which priors under order constrained hypotheses are truncations of the prior under the unconstrained hypothesis. The resulting Bayes factors are fully automatic in the sense that no subjective priors need to be specified by the user.

Sala Matte
Process data

Mat-1 Investigating students' testing behaviors using mixed types of process data

Yawei Shen, *University of Georgia, United States*
Shiyu Wang, *University of Georgia*

The widely used digital assessments provide a promising possibility to understand students' thoughts and behaviors with process data. Process data typically refers to students' behaviors during an assessment, such as the time spent and eye movements on an item. The challenge of analyzing process data arising from its higher dimensional and complex structure with mixed types of data including continuous and categorical variables. We addressed this challenge in a case study aiming to identify different behaviors of students using cluster analysis on process data. Without specifying probabilistic models, the core of clustering is to define distances between data points. The common distance functions, such as Euclidean distance, are not appropriate for mixed types of data and perform poorly in high dimensions. In this study, we proposed to calculate dissimilarities between students in each individual variable using Gower's coefficients, which are defined differently for categorical data and continuous data. The distance between two students over all variables is a combination of Gower's coefficients weighted by Shannon Entropy which integrating the information each coefficient conveys and its association with other coefficients. Using well-defined distances among students, agglomerative hierarchical clustering technique can discover underlying groups of students. The performances of the proposed method are compared to the results from the previous study using multiple factor analysis and Gaussian mixture models on the same data.

Mat-2 Data mining classification of math self-efficacy on large-scale assessment

Ya Zhang, *Western Michigan University, United States*
Aaron Yokonia Mapondera, *Western Michigan University*

Driven by the uncertain nature of educational research, a common question that researchers usually encounter is to identify the unknown group membership. The resulting groupings are often used to develop an understanding of the underlying individual characteristics. Exploratory data mining techniques have shown to be useful in grouping individuals into classes, which unfortunately has not been well applied to large-scale assessment dataset. The main goal of present study is to compare the effectiveness of different data mining techniques in classifying students' math self-efficacy using PISA 2012 data. The math self-efficacy scores were collected from the participants in the United States, and a three-step analysis plan is employed including: 1) Identifying cluster variables through regression; 2) Performing traditional cluster analysis (k-means) and latent class analysis; 3) Evaluating cluster stability using the bootstrap approach. The variables that are identified as the critical predictors of students' math self-efficacy are competitive-learning preference, instrumental motivation, math interest, math anxiety, socio-economic status, student-teacher relationship, and school disciplinary climate. The two-class and three-class solutions are suggested by the cluster analysis and latent class analysis respectively, and therefore, a disagreement in classification is found between the two data mining methods. Furthermore, the bootstrap evaluation of clusters using the Jaccard coefficient shows that the two-class solution is superior. The present study demonstrates the use of data mining clustering in large-scale assessment dataset and reveals the contradictory findings regarding the efficiency of clustering algorithms.

Mat-3 Understanding respondent characteristics through log data and interevent times

Susu Zhang, *Columbia University, United States*
Xueying Tang, *Columbia University*
Zhi Wang, *Columbia University*
Jingchen Liu, *Columbia University*
Zhiliang Ying, *Columbia University*

As a respondent attempts a computer-based interactive item, the sequence of actions that s/he performs can be recorded as a temporally ordered sequence of multi-type, time-stamped events. This sequence of actions, as well as the amount of time elapsed during and between actions, document the respondents' problem-solving process. They can hence entail information about the respondent's characteristics that cannot be recovered solely from the final responses. The current study explores how much we could learn about individual characteristics based on the log data and the corresponding interevent times. We analyze the log data from the Problem-Solving in Technology-Rich Environments (PSTRE) scale collected through the Programme for International Assessment of Adult Competencies (PIAAC). Using a recurrent neural net (RNN) that directly takes one's action sequence and interevent time sequence as inputs, we predict various characteristics of the respondent, including different cognitive scale scores and demographics. Besides evaluating the prediction power of each item's log and timing data, we further seek to decompose these items' prediction power by (1) interpreting the RNN-generated features, (2) identifying key actions/subsequences of actions, and (3) examining the improvement in prediction accuracy by adding timing information of specific actions (on top of action sequences without timing information). By studying these systemic behavioral differences related to underlying individual characteristics, test developers can gain an additional perspective on the evaluation and design of interactive problem-solving items.

Mat-4 Prediction of actions in process data via recurrent neural network

Zhi Wang, *Columbia University, United States*
Xueying Tang, *Columbia University*
Jingchen Liu, *Columbia University*
Zhiliang Ying, *Columbia University*

Response processes for interactive items, or process data for short, provides new routes in assessing problem-solving skills. Through a human-computer interface, process data can be recorded by log files and thus available in vast amount. Though considered more informative than a single categorical response, process moves in a much larger space due to its diverse nature and poses challenges for analysis. One cornerstone for understanding possible low-dimensional structures and phase transitions within paths is the predictability of future actions based on past observations. For this purpose, we apply a recurrent neural network model for multiple-step-ahead forecasting on the PIAAC 2012 dataset and analyze predictability under different item designs and action types.

Mat-5 Cross-item response process prediction by transformer

Xueying Tang, *Columbia University, United States*
Zhi Wang, *Columbia University*
Susu Zhang, *Columbia University*
Jingchen Liu, *Columbia University*
Zhiliang Ying, *Columbia University*

Computer-based interactive items have become prevalent in recent educational assessments. In such items, the entire human-computer interactive process is recorded in a log file and is known as the response process. One interesting question about response processes is what the response process of one item tell us about the response process of another item. If the appearance of an action pattern in one item indicates the appearance of another action pattern in another item, then the two action patterns are likely governed by some latent traits of respondents. Identifying these closely related patterns helps us understand respondents' behaviors and shed light on the latent traits that are measurable from process data. In this talk, we introduce a neural network architecture called Transformer to examine how the response processes from the two items are related. We apply the model to the process data collected in the Programme for International Assessment of Adult Competencies to demonstrate its performance. With the help of Transformer, we can identify closely related action patterns and predict the response process of one item based on the response process of another item.

Auditorium 2
Equating

Au2-1 Optimal confidence intervals for equivalence testing of test equating invariance

Ian Campbell, *University of Notre Dame, United States*

This simulation study compares different methods of constructing confidence intervals (CI) to perform TOST (Two One-Sided Tests) equivalence testing of subpopulation invariance in test equating. One of the main tenets of test equating is that the equating relationship should be invariant across different subpopulations. Demonstrating that the amount of population dependency is small enough to ignore can be done through equivalence testing. A common approach to conducting equivalence testing is the TOST method, which performs a significance test by constructing the confidence interval for the invariance statistic and then comparing the bounds of this CI to the thresholds of practical significance. However,

the sampling distribution of the test equating invariance statistic is only known asymptotically. When evaluating test equating invariance in finite samples, there are multiple ways to construct the statistic's confidence interval, and the impact of assumed normality on the performance of TOST is unknown. To answer this question, a simulation study compares TOST equivalence testing results from CIs formed by (1) assuming normality and using the empirical standard error, (2) assuming normality and using an estimated standard error from the delta method approximation, (3) constructing the bootstrap percentile CI, and (4) calculating the bias-corrected and accelerated bootstrap CI. Results are compared in terms of statistical power, Type I error rates, and computational time. For smaller sample sizes and smaller dependency levels in the population, the normality assumption is less appropriate, indicating that the more computationally demanding bootstrap approaches are better options in these situations.

Au2-2 Accuracy of IRT scale linking methods under two competing paradigms

Jaime Malatesta, *University of Iowa, United States*
Kuo-Feng Chang, *Center for Advanced Studies in Measurement and Assessment, University of Iowa*
Won-Chan Lee, *Center for Advanced Studies in Measurement and Assessment, University of Iowa*

Evident in the item response theory (IRT) literature is the philosophical divide between those who adhere to the traditional IRT paradigm (e.g., models in line with Birnbaum, 1968) and those who adhere to the Rasch (1960) IRT paradigm. Even though the Rasch model is algebraically a special case of traditional logistic models, it was derived under a different theoretical perspective and for different purposes (Andrich, 2004). To confuse matters, some researchers use the terms Rasch and one-parameter logistic (1PL) interchangeably. However, as pointed out by Thissen & Steinberg (1986), using the term "Rasch model" to describe an "equal-slope logistic" model with a specified ability distribution, is questionable. Similar to how distinct strains of research have evolved within these paradigms, so too have IRT software programs. Two programs commonly used within the traditional and Rasch IRT paradigms are flexMIRT (Cai, 2017) and WINSTEPS (Linacre, 2018), respectively. While either of these programs can be used with one-parameter models, they employ different estimation algorithms and handle IRT scale indeterminacy differently. As a result, their parameter estimates can differ. The purpose of this paper is to investigate the accuracy of several IRT scale linking methods implemented under the traditional IRT paradigm (e.g., 1PL characterized by equal slopes and a specified ability distribution) in flexMIRT and under the Rasch (1960) paradigm in WINSTEPS. Using a simulation study based on real mixed-format test data under the common-item nonequivalent design, the following additional factors are studied: sample size, group differences, proportion common-items, and linking method.

Au2-3 Optimal equating method for test equating in different test cycles

Xiao Li, *Beijing Normal University, China*
Ping Chen, *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University*

Test equating in different test cycles is essential to make scores from different cycles comparable. In this study, the optimal test equating method was explored and applied to the large-scale assessment toward basic education quality in China. To avoid item exposure, the single group design was adopted as the equating design. Four separate calibration methods (referred to as SC method, including mean/mean, mean/sigma, Stocking-Lord and Haebara) and fixed-parameter calibration method (referred to as FPC method) were considered to calculate the equating coefficients. The equating error was evaluated via the mean squared error (MSE). Simulation studies were conducted to fully compare the above methods by manipulating four factors, which are: sample size of single group, number of items administered to single group, test format administered to single group (dichotomous format or polytomous format or mixed), and examinees' ability distribution. Results showed that for all the methods above, the MSE value had a negative relationship with the sample size of single group; holding all other factors fixed, the Stocking-Lord and Haebara methods preferred over the mean/mean and mean/sigma methods; the FPC method required smaller simple size of single group than the SC method at the same precision level of test equating.

Au2-4 Comparing the accuracy of different equating methods in multidimensional tests equating

Somayeh Bahmanabadi, *Allameh Tabataba'i University, Iran*
Mohammad Reza Falsafinejad, *Allameh Tabataba'i University*

Comparability of different forms of test in large-scale tests is one of the crucial concerns of policy makers and educational planners in any country. When tests are multidimensional and the unidimensionality assumption is violated, the use of unidimensional equating methods to

equate such data causes error and bias. The aim of this study is to compare Full MIRT observed score equating (FMIRT), UIRT observed score equating (UIRTO) and unidimensionalized MIRT observed score equating (UMIRTO) methods in GRE test data in the Iranian nationwide Ph.D. entrance exam in Engineering in 2017 and 2018. The equating design was equivalent groups. Totally, through simple random sampling method, 1000 individuals were selected from the candidates as samples. The research method was descriptive and the data were analyzed with equateIRT package in R and SSMIRTeq program. The results showed that in FMIRT method, the amounts of Bias, SEE and RMSE were smaller than UIRTO and UMIRTO methods. According to the findings, the best method for equating tests that have more than one dimension is using multidimensional equating methods.

Auditorium 3
Cognitive diagnosis models I

Au3-1 A sequential exploratory approach to learning attribute hierarchies from data

Jing Lu, *Northeast Normal University, China*
Chun Wang, *University of Washington*

"Attribute hierarchies, if present, are important structural features of cognitive diagnostic models (CDMs) that provide actionable information about the nature of attributes to researchers and users of CDMs" (Templin & Bradshaw, 2014). Aside from determining attribute hierarchies by substantive experts, a data driven exploratory approach may provide useful empirical evidence. However, learning attribute hierarchies is challenging when the number of attributes is large. In this paper, we propose a divide-and-conquer sequential approach to identify attribute hierarchies using the regularized EM algorithm originally proposed by Xu and Shang (2018). The principal idea is, a subset of items measuring a subset of attributes will be analyzed first using the latent variable selection approach. Once the attribute hierarchies among the attributes in the subset are identified, these known information will be fixed in the subsequent analysis to simplify the model estimation, as reflected by some known zero parameter no longer need to be penalized. The performance of the proposed sequential approach will be demonstrated by a simulation study.

Au3-2 The impact of Q-matrix structure in multiple-strategy DINA model

Hueying Tzou, *National University of Tainan, Taiwan*
Yi-Fang Wu, *ACT, Inc.*
Ya-Huei Yang, *National Tainan Junior College of Nursing*

Most applications of cognitive diagnostic models (CDMs) defined the Q-matrix under the assumption that one problem only had one q-vector, which implying that students employed only one strategy to solve the problem. In reality, however, students might use different strategies corresponding to different q-vectors to solve a same problem. The multiple-strategy deterministic input noise "and" gate (MS-DINA) model addresses this situation, through MCMC and EM algorithm for the parameter estimation (de la Torre & Douglas, 2008; Huo & de la Torre, 2014). Based on these literature, the current assumptions under the MS-DINA model have caused two concerns: Not all problems are associated with multiple strategies, and incompleteness of the Q-matrix results from the fact that students can apply different strategies to solve for the same problem, and a greater number of strategies could be applied (e.g., three and above). To tackle the concerns, the current study is to investigate the impact of the proportion of multiple-strategy problems under the Q-matrix with specific attributes, and to examine the difference between multiple strategies (MS-DINA) and single strategy (SS-DINA) models in terms of item parameter estimation and attribute classification. The Alaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) are also used to compare the model fit between SS-DINA and MS-DINA.

Au3-3 An exploratory unified model for conjunctive processes

Auburn Jimenez, *University of Illinois at Urbana-Champaign, United States*

The unified model (DiBello, Stout, & Roussos, 1995) is a mixture of a diagnostic model and an item response model. The original model was confirmatory in that the latent structure corresponding to the diagnostic model could be pre-specified based upon content expert knowledge or cognitive theory. More generally the unified model provides a useful framework for accounting for heterogeneity in the underlying response processes, but the confirmatory assumption limits more general applicability across the social sciences. We propose an exploratory version of the unified model for conjunctive processes underlying binary response data. The exploratory unified model consists of a mixture of the exploratory reduced reparameterized unified model (rRUM) and the non-compensatory item response theory (NIRT) model. We introduce a new

Bayesian framework, propose an efficient data augmentation scheme, and describe a Gibbs sampling algorithm for approximating the posterior distribution. We apply the model to psychopathology data and compare results with the NIRT and general diagnostic models.

Au3-4 Estimation of Q-matrix with unknown number of attributes

Yinghan Chen, *University of Nevada, Reno, United States*
Steven Culpepper, *University of Illinois at Urbana-Champaign*
Yuguo Chen, *University of Illinois at Urbana-Champaign*
Ying Liu, *University of Illinois at Urbana-Champaign*

Cognitive Diagnosis Models (CDMs) are widely used for providing fine-grained classification of multidimensional collection of discrete attributes. The application of CDMs is based on the specification of attribute requirement on educational tasks in what is known as the Q matrix. Prespecified Q might be misspecified and results in biased diagnostic classifications, thus it is fundamental to estimate Q matrix. Existing methods must pre-specify the number of attribute in order to estimate Q. We will present a Bayesian strategy for jointly estimating the number of attributes (the number of columns in Q) and the elements of Q in deterministic inputs, noisy "and" gate (DINA) model. We propose the "crimp-sampler" to transit between matrices of different number of columns and estimate the underlying Q matrix and model parameters through Gibbs sampler.

Dissertation Prize: Merijn Mestdagh

Frs-1 Prepaid parameter estimation without likelihoods

Merijn Mestdagh, *University of Leuven, Belgium*

In various scientific fields, statistical models of interest are analytically intractable and inference is usually performed using a simulation-based method. However elegant these methods are, they are often painstakingly slow and convergence is difficult to assess. As a result, statistical inference is greatly hampered by computational constraints. However, for a given statistical model, different users, even with different data, are likely to perform similar computations. Computations done by one user are potentially useful for other users with different data sets. In this presentation, I recommend a pooling of resources across researchers to capitalize on this. More specifically, I propose to preemptively chart out the entire space of possible model outcomes in a prepaid database. Using advanced interpolation techniques, any individual estimation problem can now be solved on the spot. I will discuss prepaid databases created for several challenging models and demonstrate how they can be distributed through an online parameter estimation service. This method outperforms state-of-the-art estimation techniques in both speed (up to a 100,000-fold speed up) and accuracy, and is able to handle previously quasi inestimable models. Specifically I will demonstrate how this online parameter estimation technique is able to quasi instantaneously estimate the parameters of choice reaction time models without a tractable likelihood.

Wednesday, July 17

Parallel Sessions, Wednesday Morning

Salón Fresno

Panel Discussion: Stories of successful careers in psychometrics and what we can learn from them

Frs-1 Stories of successful careers in Psychometrics and what we can learn from them

Irini Moustaki, *London School of Economics and Political Science, United Kingdom*
Carolyn Anderson, *University of Illinois*
Susan Embretson, *Georgia Institute of Technology*
Jaqueline Meulman, *Leiden University*
Duanli Yan, *Educational Testing Service*

The symposium aims to bring together four women with successful careers in the academia and industry. Panel members will discuss their education and career paths and highlight their achievements as researchers and teachers. The presentations and discussion will provide an insight to the research carried out by the panel members but also give our panellists the opportunity to address strategies for success for publishing, for grant applications and promotion, as well as discuss the importance of interdisciplinary work and visibility and applicability of research in the academia and industry. The stories of our panellists will also provide a platform for discussing work environment for women and work-life balance. There will be time for questions and comments.

Aula Magna

Symposium sponsored by Agencia de Calidad de la Educación: Addressing psychometric challenges for implementing accountability policies

Mag-1 Bridge studies, the Chilean case

Myriam Lara, *Agencia de Calidad de la Educación, Chile*
Marilyn Stevenson, *Agencia de Calidad de la Educación*

Simce assessments are intended for educational policy monitoring. One of its main goals is to determine stagnation or improvement in students learning outcomes. Therefore it is of crucial importance to maintain Simce test results comparable along time. At the same time, school curriculum suffers content and form changes promoted by societal evolution, and Simce tests should also reflects those changes. Bridge studies are a valuable and valid tool for maintaining comparable Simce tests outcomes. Briefly, the methodology is to apply the old test to a student sample obtained from the student population of the new test and link outcomes from both tests by means of scale transformation (the bridge). Using such transformation a reliable score is calculated in a unique measurement scale for all the assessed student population, in order to report mean scores and for comparison purposes.

Mag-2 Addressing psychometrics challenges for implementing accountability policies

Maria de la Luz Gonzalez, *Agencia de Calidad de la Educación, Chile*
Gabriela Cares, *Agencia de Calidad de la Educación*
Marilyn Stevenson, *Agencia de Calidad de la Educación*
Myriam Lara, *Agencia de Calidad de la Educación*
Maximiliano Romero, *Agencia de Calidad de la Educación*

The Chilean assessment system has a long tradition using sound psychometric instruments for monitoring student educational achievement. Since the nineties, the national assessment uses Item Response Theory, and it yearly provides reliable information to each school, in several grades and subjects. This information has been evidence for several relevant educational policies, and it is also periodically reported to parents and public opinion. In the last decade, the country has implemented a national educational quality assurance system. It implied the creation of two new institutions and the redefinition of the responsibilities of the Ministry of Education. One foundation of this quality assurance system is the categorization of schools based on their performance. The category of each school is calculated considering results in academic achievements and indicators of personal and social development, all of them defined and mandated by law. Based on this categorization schools may receive support, guidelines and ultimately, they can lose the formal recognition from the Ministry of Education. The Educational Quality Assurance Agency (ACE), the institution in charge of implementing the school categorization, has developed several psychometrics instruments to improve the assessment of school educational quality

continually; it includes academic assessments (Reading, Writing, Mathematics, Science, and Social Science) along with questionnaires intended to evaluate indicators of personal and social development.

Mag-3 Facing methodological challenges for a multipurpose policy of non-academic assessment

Maximiliano Romero, *Agencia de Calidad de la Educación, Chile*
Miguel Traslaviña, *Agencia de Calidad de la Educación*

The Educational Quality Assurance Agency is responsible for yearly calculating non-academic educational results indicators, to every Chilean school. These non-academic indicators, measured throughout questionnaires, are Academic Self-esteem and Academic Motivation, School Climate, Healthy Lifestyle, and Civic Engagement. One of our challenges is to identify the best methodology fitting our legal and communicational demands using student's, parent's and teacher's perceptions. This presentation covers our efforts to implement an appropriate methodology to produce, evaluate, compute and report our questionnaire-based indices. On October 2018, an enhanced battery of questionnaires was tested on a student representative sample. This field trial administration included extra questions for each index and the use of three different booklets for students for addressing the Ministry of Education requirement of covering all theoretical concepts included in their definition. The methodology used to analyze the results is based on Confirmatory Analysis for checking dimensionality and identifying factorial loads and residual variance. Considering this information, we proceed to calculate indicators with the Partial Credit Model. Our short-term purpose is to achieve a pool of questions measuring key variables with different levels of difficulty, to cover all indices, emulating what other international studies do when computing questionnaire indices. The final goal is to provide reliable, and comprehensible evidence to promote better and more informed educational policies and school decisions. According to this, a mid-term task is to develop useful descriptions of the performances reached by each school.

Mag-4 Personal and social development indicators: educational quality new definition

Gabriela Cares, *Agencia de Calidad de la Educación, Chile*
Pamela Inostroza, *Agencia de Calidad de la Educación*

In Chile, across the last 30 years, educational quality definition has been associated with national assessment school's results. They are focused on traditional academic areas such as Math, Reading, and Science. At present, with a renewed institutionality in education, new areas are incorporated into the national evaluation system. This broader approach intends to promote favorable environments to the integral development of students. In 2009, a new constitutional educational law defined the general objectives for K-12 education. They are comprehensive expectations for student's development that have to be addressed by every school in the formal system. Based on these objectives, a group of eight indicators was defined, also by law, in order to be considered by each school: Academic Self-esteem and Academic Motivation, School Climate, Healthy Lifestyle, Civic Engagement, School retention, Attendance, Gender Equality, Certification of Vocational Education. The assessment of the first four uses self-reported questionnaires. These indicators are publicly known as Indicators of Personal and Social Development; by legal mandate, they have to be assessed and informed yearly to schools and parents. The results of each school are public and, together to academic performance, included for a high stakes classification. Six years after the indicators were assessed for the first time, valuable insights have been gained. We will share experiences and challenges on designing questionnaires for a diverse population, the challenges to balance legal mandates and technical considerations, and the ways to produce information to report to different audiences.

Mag-5 SIMCE psychometric history and its challenges

Marilyn Stevenson, *Agencia de Calidad de la Educación, Chile*
Myriam Lara, *Agencia de Calidad de la Educación*

In 1988 the Chilean learning outcomes evaluation system was founded aiming to institutionalize various educational evaluation initiatives that were developed from the sixties. From this starting point, several assessments were developed each addressing different purposes. In 1988, Simce tests were designed and applied in its current content and format. This presentation is about Simce assessments batteries, its timeline trajectory, the populations assessed by, as well as the instruments and the analysis and measurement models used. We also show how the Simce test outcomes are used. Finally we will talk the current challenges the evaluation system faces and our research instances aimed to address them.

Sala Colorada
Prediction and causal inference

Col-1 Machine learning algorithms for causal inference with cluster-structured observational data

Youmi Suk, *University of Wisconsin - Madison, United States*
Jee-Seon Kim, *University of Wisconsin - Madison*
Hyunseung Kang, *University of Wisconsin - Madison*

Machine learning algorithms have gained popularity in measuring treatment effects in either randomized controlled trials or observational studies, because of their advantages that include (nearly) automatic model fitting and model flexibility. Machine learning methods for causal inference, such as Bayesian Additive Regression Trees and causal forests, have been studied extensively for single-level data. Also, their performance has been scrutinized for unstructured, high-dimensional data. However, many aspects of the machine learning methods have been under-studied in the context of clustered or nested data structures. Theoretically, general machine learning methods, if flexible enough, should be able to detect data structures. However, theoretical properties of estimators do not always hold in practice. Therefore, this study investigates if combining insights from machine learning methods and multilevel models that account for data structures can provide benefits beyond the calibration of standard errors, in estimating the effects of a treatment whose assignment is not random. A simulation study is performed with various types of cluster-structured data that have different levels (e.g., two, three, cross-classified) and different effects (e.g., fixed, random). We evaluate the performance of treatment-effect estimates with respect to bias and efficiency using machine learning methods coupled with explicit cluster information, compared to machine learning methods only. This study concludes with recommendations for the proper use of machine learning methods with cluster-structured data in causal inference.

Col-2 Improving the predictive validity of psychological tests using a statistical learning perspective

Bunga Pratiwi, *Leiden University, Netherlands*
Elise Dusseldorp, *Leiden University*
Mark de Rooij, *Leiden University*

Psychological tests generally aim at 1) measuring one or more latent variable(s) and 2) predict an outcome. The first aim prefers items with high inter-correlations with each subscale. The second aim asks for items with low inter-correlations but high correlations with the outcome (Lord & Novick, 1968). Therefore to satisfy both aims simultaneously seems as an unrealizable task. As a result, it is rare to find tests that are constructed by selecting only items which correlate highly with the outcome. As a consequence, its predictive validity might suffer. When tests are used to predict an outcome, their predictive validity (i.e., the accuracy of the prediction rule evaluated on out of sample data) should be high. In this study, we focus on predictive validity of multidimensional tests, for which we investigate how to best combine the items to form a prediction rule that may improve predictive validity. We identify three different ways. First, construct subscales of the items (i.e., weighted linear combinations) and use these as predictors of an outcome. Second, create a linear combination of the items that best predict the outcome. Third, extract components from subscales or items and use them as predictors. There are two main goals of this study: 1) compare classical and statistical learning methods to construct the prediction rules (i.e., least squares, shrinkage methods, and supervised principal component analysis) and 2) investigate the influence of measurement error of the items on predictive validity. We present results from simulation experiments and two empirical data examples.

Col-3 Money is not everything: The determinants of the education quality

Ewa Witkowska, *Warsaw School of Economics, Poland*
Bartosz Witkowski, *Warsaw School of Economics*

In plentiful research the expenditures on education as a fraction of GDP are used as a proxy for the human capital quality (Anderson, Chiu, Yore, 2010; Ciccone & Papaioannou, 2010; Hanushek & Woessmann, 2010; Pelinescu, 2015; Bittlingmayer, Boutiuc, Heinemann, & Kotthoff, 2016; Komatsu, & Rappleye, 2017; Jones & Potrafke, 2014; Niemann, Martens, & Teltemann, 2017; Valero, 2017; Zagler & Zanzottera, 2018). The former stems from the common belief that the overall country-level expenditures on schooling shall be distributed appropriately and will result in leveling up the education quality. The latter, however, need not be true. In the empirical analysis we use the NUTS2 (voivodships) up to NUTS4 (counties) Polish data in order to identify the main determinants of education quality measured by the results of the national level graduation tests. While the analysis is focused on the steerable potential determinants of educational achievements, we control also for the regional socioeconomic characteristics in a developed model. However, since multicollinearity of the considered regressors disables construction of a single equation covering all the desired determinants and their subjective

selection can result in bias caused by prior assumptions, we apply Bayesian model averaging in order to identify the robust determinants of the educational effects. Equal prior relevance probabilities are assigned to each steerable schooling quality determinant. Posterior probabilities demonstrate that treating the overall expenditures as a covariate of educational effects can be highly misleading due to different relevance of different considered factors.

Col-4 Predictive inferences of bifactor models and simple structure models

Weimeng Wang, *University of Maryland, College Park, United States*
Ji Seung Yang, *University of Maryland, College Park*

Latent constructs (e.g., grit and personality traits) are often used as independent variables or covariates to predict an outcome of interest. While the one-stage estimation of latent variable models – measurement and structural models- is considered as a gold standard whenever it is feasible, the multi-stage estimation (e.g., using factor scores or plausible values) also has its own values in practices. The purpose of the current study is to investigate the performance of one-stage and multi-stage estimation of predictive models particularly when there are two competing measurement models, namely bifactor and simple structure models. Motivated by a particular construct of grit which recently received much attention in Educational Psychology, a Monte Carlo simulation study is conducted to examine how well the true model is identified between the two measurement models that are very close to each other and how well or differently the various multi-stage estimation approaches to the predictive model perform under two competing measurement models. In particular, regression coefficients recovery, standard errors of the regression coefficients, coefficient of the determination, Type I error, and power are investigated. In addition, two-step estimations based on factor scores only, plausible values, and double imputed plausible values are compared with the one-step estimation that simultaneously estimates the measurement and structural relations. Given the increased use of bifactor models in the applications, this study will provide methodological guidelines and implications for the researchers who would like to use a bifactor model as their measurement model or debating between the competing measurement models.

Col-5 A SEM-based prediction rule to assess predictive and incremental validity

Jeroen Janssen, *Leiden University, Netherlands*
Henk Kelderman, *Leiden University*
Mark de Rooij, *Leiden University*

In well-known machine learning techniques, which are most commonly used for predictive modelling, it is assumed that the predictor variables are measured without measurement error. It is shown that this assumption decreases predictive accuracy if violated, which is usually the case in social sciences. We therefore argue that this measurement error should be taken into account when entering the realm of prediction. One way of doing that is by using a latent variable model instead of a regular regression model, which will be shown in this presentation. Based on this SEM-model, we have derived a prediction rule for both a one-factor and a two-factor model. We will show the performance of this two-factor prediction rule to assess both the validation set prediction error (as a measure of predictive validity) and the difference between the validation set prediction error using one and two factors/tests (as a measure of incremental validity). We also compare this performance to that of linear, regularized (lasso/ridge) and component score regression. Results show that in most cases, the SEM-based prediction rule performed equally well or even better as compared to the regression techniques. In these cases, the SEM-based approach should be preferred, taking the measurement error into account and therefore providing more reliable estimates of the outcome variable.

Sala Matte
Response times I

Mat-1 Detecting person misfit using the diffusion modeling approach

Renske Kuijpers, *University of Amsterdam, Netherlands*
Dylan Molenaar, *University of Amsterdam*
Han van der Maas, *University of Amsterdam*

In the literature, a large number of statistics have been proposed to detect person misfit (Karabatsos, 2003; Meijer & Sijtsma, 2001). A distinction is made between group-based person-fit statistics, which count the number of response patterns deviating from the perfect pattern as expected under the Guttman (1950) model, and IRT-based person-fit statistics, which assess the distance between observed response patterns and expected response patterns given the parameter estimates (Karabatsos, 2003; Conijn, 2013). Although group-based statistics are shown to outperform IRT-based statistics (Karabatsos, 2003), both types of person-fit statistics are associated with challenges, including failing to identify potential causes of misfit, and insufficient power for relatively small tests, relatively low item discriminations and relatively small differences in item difficulties (Meijer, Niessen, &

Tendeiro, 2016). To improve upon the IRT-based person-fit approaches, Van der Linden and Guo (2008), Marianti et al. (2014) and Ferrando (2016) propose to take item response times into account in addition to the item responses. However, the joint modeling approaches still face some of the same challenges regarding test length and item characteristics as the traditional person-fit statistics. In the present study, we adopt an IRT-based diffusion modeling approach to person fit (Van der Maas et al., 2011). In this approach the responses are explicitly modeled in terms of an underlying cognitive process. As a result, the model has various theoretical advantages. In addition, the model draws strength from its theoretical assumptions which makes this approach a suitable and powerful approach to study person fit.

Mat-2 A hierarchical latent response model for inferences about examinee engagement

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

In low-stakes assessments, test performance comes with little or no consequences for examinees themselves, so that examinees may not be fully engaged when answering the items: Instead of engaging in solution behavior, disengaged examinees might randomly guess or generate no response at all. When ignored, examinee disengagement poses a severe threat to the validity of results obtained from low-stakes assessments. Statistical modeling approaches in educational measurement have been proposed that account for nonresponse or for guessing, but do not consider both types of disengaged behavior simultaneously. We bring together research on modeling examinee engagement and research on missing values and present a hierarchical latent response model for identifying and modeling the processes associated with examinee disengagement jointly with the processes associated with engaged responses. To that end, we employ a mixture model that identifies disengagement on the item-by-examinee level by assuming different data-generating processes underlying item responses and omissions, respectively, as well as response times associated with engaged and disengaged behavior. By modeling examinee engagement with a latent response framework, the model allows assessing how examinee engagement relates to ability and speed as well as identifying items that are likely to evoke disengaged test-taking behavior. An illustration of the model by means of an application to real data is presented.

Mat-3 The four-parameter normal ogive model with response time

Yang Du, *University of Illinois at Urbana-Champaign, United States*
Justin Kern, *University of Illinois at Urbana-Champaign*

In recent years, interest in the four-parameter logistic (4PL) model (Barton & Lord, 1981), and its normal ogive equivalent, has been renewed (e.g. Culpepper, 2016, 2017; Feuerstahler & Waller, 2014). The defining feature of this model is the inclusion of an upper asymptote parameter, called the slipping parameter, in addition to those included in the more common three-parameter logistic (3PL) model. The use of the slipping parameter has come into contact with many assessment applications, such as high-stakes testing (Rulison & Loken, 2009), low-stakes testing (Culpepper, 2017), and measuring psychopathology (Waller & Reise, 2010). Low-stakes tests provide a particularly interesting case in which the lack of serious decision-making consequences may affect motivation. As a result, the accuracy of locating a person on the scale of interest, may be greatly affected. It may be hypothesized that the response time (RT) of test-takers is related to the motivation of a person, so accounting for it within the test may help with overall measurement accuracy. To help with future studies examining this relationship, we intend to extend the four-parameter normal ogive model by incorporating RT into the model formulation. A Gibbs sampling approach to estimation will be developed and investigated.

Mat-4 Impact of test design change on test speededness

Huijuan Meng, *Graduate Management Admission Council, United States*

This study examines the potential impact on test speededness as a result of changing from a linear-on-the-fly test (LOFT) to a computerized adaptive test (CAT). Both LOFT- and CAT-administered exams were simulated from an item pool constructed for the predicted candidate population. Exam time was computed for each simulated test taker by summing the estimated response time (RT) across all test items in the simulated test form. In this study, a hierarchical model (Van der Linden, 2007) was used to capture the moderating effect of item difficulty and personal ability on RT. This model assumes that test takers operate at a constant speed during the exam, which generally holds true for tests that offer sufficient allotted time. When an exam is highly speeded, however, like the exam being studied in this research, candidates may undertake various test-taking strategies to answer more items correctly. For example, they may spend more

time on items that seem more easily solvable, while rapidly guessing answers to items that appear more challenging. In such a scenario, the basic assumption may be violated severely, and the model may not be robust enough to reveal the impact of a test taker's ability and item difficulty on RTs. For this reason, conditional item time intensity parameters were estimated separately for different ability groups and used in exam time projection. Findings of this study suggest that the adaptive test design may help reduce the differential speededness observed in the LOFT data.

Auditorium 2
Patient-reported outcomes

Au2-1 Evaluating the impact of measurement bias on diagnostic clinical assessments

Oscar Gonzalez, *University of North Carolina at Chapel Hill, United States*

In psychology, diagnostic assessments are used to make a decision based on an observed cut score. If measurement bias is found on assessment items across priority groups (such as those defined by ethnicity, gender, or socioeconomic status), there might be a risk of inaccuracy in diagnosis. In this case, researchers would have to decide either to drop the items, keep the items, or stop using the assessment. Here, the focus is on keeping biased items in the assessment. Millsap and Kwok (2004) developed a procedure that described the impact of ignoring measurement bias on tests for selection as changes in test sensitivity and specificity. Millsap and Kwok's procedure makes two assumptions to estimate test sensitivity and specificity – that item responses are continuous, and that observed assessment scores, conditional on the factor score, are normally distributed. In clinical settings, these two assumptions might not be tenable. In this paper, Millsap and Kwok's procedure is extended to more realistic scenarios by acknowledging the discrete nature of the items and using simulation methods to approximate the conditional distribution of the observed assessment scores given the factor score. The procedure is illustrated using an applied example and future directions are discussed.

Au2-2 The incremental value of LCA-based mixture CAT for PROMIS depression

Jan van Bebber, *Charité Universitaetsmedizin Berlin, Germany*

In the assessment of patient reported outcomes (PROs), it is routinely assumed that all respondents interpret item content equally, so that a measurement model can be constructed that is applicable to all individuals. LCA aims at partitioning the sample in clusters of respondents who respond in a comparable way to items, thereby reducing sample heterogeneity with respect to the measurement model specified. This paper investigates the potential value of LCA for adaptive testing of PROs. We use the PROMIS Depression item bank in our study. Respondents are sampled from the general population and from specific clinical populations in the US and in various European countries. We split the sample in more homogenous clusters by means of LCA. Latent classes are then used in multigroup GRM to identify those items that display substantial DIF. Multinomial regression is used to determine how well latent class membership can be predicted by information regarding the demographic and clinical background of respondents. Those items flagged for DIF are used in Real Data Simulations to compare the performance of mixture parameters estimated against performance of the official PROMIS item parameters. Using standardized scores derived from those items that did not exhibit DIF as criterion, we investigate whether mixture CAT scores exhibit bias, and how precise these are relative to standard CAT scores. Our study will investigate the feasibility and incremental value of mixture CAT for assessing PROMIS depression. Whether mixture parameters should be adopted will depend on consensus between researchers and practitioners worldwide that apply PROMIS.

Au2-3 Patient identity: Test design and empirical measurement-equivalence findings

Matthew Kerry, *Zürich University of Applied Sciences, Switzerland*

The value-based compensation systems of modern-organized medicine compels new provider competencies enabling team-based, coordinated care delivery. Toward this goal, the current paper reports on a newly developed self-report instrument for assessing a health provider's patient identity (PI). PI is defined as: A provisional perspective for regulating professional competence with personal values, interests, and beliefs. After overviewing its nomological net, theoretical conceptualization, and instrumentation, empirical evidence from three measurement studies is reported for PI. In study one, the instrument was piloted among American medical (4th-year) and nursing (2nd-year) students (N = 119). Observed-score analyses indicated preliminary factorial (exploratory factor analysis) and predictive (hierarchical linear regression) validity evidence for team attitudes. In study two,

a German-translated instrument was deployed in an equivalent sample frame of medical (4th-year) and nursing (2nd-year) students (N = 134). Latent variable analyses using structural equation modeling extended factorial validity evidence (confirmatory). Item response theory analyses was used to benchmark measurement properties, as well as test measurement equivalence of PI (minimal differential-item functioning) across American and German samples. A third study further compared German translations across culturally distinct Swiss-health providers (N = 243). Overall measurement equivalence findings (N = 496) and implications for further test development and future scale usage are discussed.

Au2-4 Developing a mental-health screening tool in Mozambique using Lasso regression

Melanie Wall, *New York State Psychiatric Institute and Columbia University, United States*
Cale Basaraba, *Research Foundation for Mental Hygiene*
Kate Lovero, *Columbia University*
Milton Wainberg, *Columbia University*

This project aimed to develop a mental health screening tool that can be regularly administered by community health workers in Mozambique. Such a tool must be relatively short while still providing good sensitivity and specificity for a wide array of mental health disorders. Nine existing mental health measurement scales for various psychiatric disorders and overall disability were assembled and administered along with a gold standard psychiatric diagnostic interview to n=988 individuals sampled within a general health care setting. Psychiatric disorders were grouped into four treatment-oriented categories: Common (e.g., depression, anxiety), Severe (e.g., psychosis), Suicide Risk, and Alcohol Abuse. This presentation will describe how a two-step screening tool was developed from this data. The first step aimed to provide high sensitivity for any psychiatric diagnosis category using a few items, while the second step aimed to place positive individuals from the first step into treatment-oriented categories using a limited number of items while maintaining good sensitivity and specificity. LASSO regression was used to reduce 107 items from all 9 instruments into 3 items that best predicted any psychiatric disorder then was used again to reduce all items to 3-5 that best predicted each of the four treatment-oriented categories. Simulation results demonstrated the influence of prevalence of disorder on the sensitivity and specificity obtained from the method. A final screener with 3 plus 10 items exhibited good diagnostic ability and was subsequently validated in a separate dataset also collected from Mozambique.

Au2-5 A multidimensional zero-inflated graded response model for ordinal symptom data

Brooke Magnus, *Marquette University, United States*
Mauricio Garnier-Villarreal, *Marquette University*

Zero inflation is common on measures of psychopathology. Psychometric researchers typically accommodate multivariate zero inflation by including a "non-pathological" class of respondents who endorse zero for all items (Magnus & Liu, 2018, Wall, Park, & Moustaki, 2015). While a non-pathological class accounts for test-level zero inflation, this approach may be restrictive on questionnaires comprising items of differing severity. For example, an item about suicide ideation likely exhibits a higher degree of zero inflation than an item about energy levels. Test-level zero-inflated models do not account for this variability. We propose a more flexible approach using a zero-inflated graded response model (ZIGRM), where two sets of parameters are estimated for each item. One latent variable underlies the probability of responding with a structural zero; a 2PL model is used for this component. A second latent variable underlies the sampling zeros and non-zero responses; a GRM is used for this component. Importantly, we do not assume that the same latent variable underlies both response processes; rather, we address this question empirically by estimating the correlation between the two latent variables. As motivating examples, we fit the ZIGRM to data from the Patient Health Questionnaire-9 and Generalized Anxiety Disorder-7 using the general Bayesian analysis software Stan. Results indicate that the ZIGRM is better able to capture the proportion of zero responses across items than a model that only accounts for test-level zero inflation. Further, we find support for a multidimensional model that allows different but correlated latent variables to underlie each response process.

Auditorium 3
Multivariate analysis

Au3-1 Extended redundancy analysis via generalized estimating equations

Sunmee Kim, *McGill University, Canada*
Sungyoung Lee, *Seoul National University Hospital*
Yongkang Kim, *Seoul National University*
Heungsun Hwang, *McGill University*
Taesung Park, *Seoul National University*

Extended redundancy analysis (ERA) is a statistical method for relating multiple sets of predictors to response variables. This method aims to extract a weighted composite or component from each set of predictors in such a way that it explains the maximum variation of response variables. Thus, ERA performs linear regression and data reduction simultaneously, providing a simpler description of directional relationships among many sets of variables. In this talk, we propose to combine ERA with generalized estimating equations (GEE) for the analysis of correlated responses. The proposed method has several advantages over ERA. Firstly, it can handle various types of response variables (e.g., continuous, binary, or count) that are assumed to follow an exponential family distribution. Secondly, it estimates model parameters and their robust standard errors, taking into account different covariance structures of responses. This can ensure valid inference regardless of the covariance structure specified. Lastly, it adopts ridge-type regularization to address potential overfitting when a large number of predictor variables per component are considered, or a large number of components influence response variables. We minimize a penalized least squares criterion for estimating parameters. We provide two applications of the method, in which response variables are repeated measures of cognitive decline among older adults in a longitudinal panel study and correlated metabolic syndrome phenotypes in a genetic pathway study.

Au3-2 Procrustes penalty function for matching matrices to targets with its applications

Naoto Yamashita, *Osaka University, Japan*

Penalized estimation is widely used for obtaining sparse solutions, which facilitates easier interpretation compared with ordinal estimation procedures. In this research, as a generalized form of penalized estimation, we propose a new penalty function. The proposed function shrinks solutions to a prespecified target matrix which possesses a certain simple structure. The resulting solution is therefore simple and easy to interpret, and its simplicity is controlled by some tuning parameters. We present the two applications of the proposed method in sparse principal component analysis and three-way component analysis. The effectiveness of the proposed method is demonstrated by real data examples.

Au3-3 A cross validation index for generalized structured component analysis

Gyeongcheol Cho, *McGill University, Canada*
Kwanghee Jung, *Texas Tech University*
Heungsun Hwang, *McGill University*

The recent onset of replication crisis in the behavioral and social sciences has led researchers to attend to the generalizability of their empirical research. Cross validation is a useful way of comparing generalizability of theoretically plausible a priori models in terms of prediction and has been utilized in structural equation modeling (SEM). A number of overall or local cross validation indices have been proposed for existing factor-based and component-based approaches to SEM, including covariance structure analysis and partial least squares path modeling. However, there is no such cross validation index available for generalized structured component analysis (GSCA), which is a component-based approach. We thus propose a cross validation index for GSCA, called Out-of-bag Prediction Error (OPE), which estimates the expected prediction error of a model over replications of so-called in-bag and out-of-bag samples constructed through the implementation of the bootstrap method. The calculation of this index is well-suited to the estimation procedure of GSCA, which uses the bootstrap method to obtain the standard errors or confidence intervals of parameter estimates. We empirically evaluate the performance of the proposed index through the analyses of both simulated and real data.

Au3-4 Analyzing cognitive similarities among occupational categories by distance-radius asymmetric MDS

Akinori Okada, *Ricky University, Japan*
Takuya Hayashi, *Nara Women's University*

A matrix consists of cognitive similarities among 10 occupational categories, which were analyzed in Okada & Hayashi (2017) by the asymmetric MDS based on singular value decomposition, were analyzed by another asymmetric MDS based on the distance-radius model (Okada & Imaizumi, 1987). The similarity from category j to the other categories is obtained from respondents whose occupations belong to category j. Okada & Hayashi (2017) derived a three-dimensional configuration. The configuration shows two kinds of effects of autonomous (authority, discretion) on the asymmetry of similarities among categories. But the three-dimensional configuration has shortcomings. The asymmetric similarity relationships among categories are represented by the sum of signed areas in the planer configuration along each of three dimensions, and thus the three planer configurations do not give a configuration where the asymmetric similarity relationships among categories are clear at a glance. The analysis by the asymmetric MDS based on the distance-radius model resulted in a two-dimensional configuration where each category is represented as a point and a circle centered at the point. The radius of the circle represents the

asymmetry of similarities among categories. The configuration suggests that the asymmetric similarity relationships among categories depend on two factors; the prestige and the female rate, and the asymmetry is closely related to the number of workers belong to the category per 100,000 people. While the recovered similarities of the two configurations given by two different MDS procedures are very close, the two configurations disclosed different aspects of asymmetric relationships among categories.

Au3-5 Time profile similarity indices with synchronization in nearest neighbor classification

Qimin Liu, *University of Notre Dame, United States*

One of the existing approaches to time series classification exploits the time profiles using the original, instead of model-implied, data with synchronization. Synchronization aligns inter-individual data of different time points to account for potential phase offsets and nonstationarity in the data. Such synchronization has been recently applied in psychology, e.g., for coordinated motion between two individuals engaged in inter-individual information exchange were used as predictor/outcome of psychological processes. Synchronization also affords better classification outcome, as discussed in the data mining community, through aligning the data to reveal the maximally shared profile underlying two compared data sequences. For inter-individual comparison of univariate time series data, existing similarity indices include Euclidean distances and squared correlations. For synchronization, we introduce dynamic time warping and window-crossed lagging. The current study compares the Euclidean distance and the squared correlation before and after synchronization using window-crossed lagging and dynamic time warping in applications to one-nearest-neighbor classification tasks. Discussion, limitations, and future directions are also provided.

Keynote Speaker: Burr Settles

Frs-1 Improving language learning and assessment with data

Burr Settles, *Duolingo, USA*

As scalable learning technologies become more ubiquitous, student data can and should be analyzed to develop new instructional technologies, such as personalized practice schedules and data-driven assessments. I will describe a few projects at Duolingo — the world's largest language education platform with more than 200 million students worldwide — where we combine learner data with machine learning, computational linguistics, and psychometrics to improve learning, testing, and engagement outcomes.

Thursday, July 18

Parallel Sessions, Thursday Morning

Salón Fresno

Symposium: The lavaan ecosystem: Past, present, and future

Frs-1 equaltestMI: Equivalence testing for measurement invariance

Ge Jiang, *University of Illinois at Urbana-Champaign, United States*
Yujiao Mai, *University of Notre Dame*
Ke-Hai Yuan, *University of Notre Dame*

equaltestMI is an open-source, freely-available R package for researchers to examine measurement invariance under the new equivalence testing framework. Traditionally, measurement invariance is evaluated using multigroup structural equation modeling through a sequence of chi-square and chi-square-difference tests. However, under the conventional null hypothesis testing framework one can never be confident enough to claim measurement invariance even when all chi-square statistics are not significant. Equivalence testing, as an alternative to null hypothesis testing, informs researchers a size of possible misspecification that quantifies the degree of measurement noninvariance. The size of misspecification is further linked to RMSEA (root mean square error of approximation) with adjusted cutoff values for labeling the degree of measurement invariance. We present the equaltestMI package (Jiang, Mai, & Yuan, 2017) that automatically calculates the sizes of misspecification and labels the degrees of measurement invariance under the new equivalence testing framework. By leveraging existing codes in the packages lavaan and semTools, the equaltestMI package enjoys the great easiness and flexibility in model specification offered by lavaan. In addition, equaltestMI outputs easy-to-interpret summary tables for a variety of equality constraints (e.g., configural, metric, scalar invariances). Beyond measurement invariance, the next step is often to compare the means of the latent constructs across groups. Therefore, we also implement an innovative projection-based method to make it possible to compare the means of the latent constructs even when the equality of intercepts (scalar invariance) does not hold.

Frs-2 Pairwise likelihood estimation for structural equation modelling in lavaan

Irini Moustaki, *London School of Economics and Political Science, United Kingdom*
Myrsini Katsikatsou, *London School of Economics and Political Science*

Pairwise likelihood estimation is a special case of composite likelihood that has been developed and proven to work well in reducing the computational complexity in structural equation models (SEM) for ordinal variables with missing values. Pairwise likelihood utilizes information from lower order margins and give estimates that are consistent. A theoretical framework exists for estimating and testing complex models. In particular, likelihood ratio tests for overall goodness of fit and for comparing nested models as well as model selection criteria have been developed for SEM models. Methodology for handling missing data under a missing at random missing data mechanism has been developed. The pairwise likelihood estimation and testing for single and multi-group models have been implemented in the R package lavaan and has been found to provide a unified framework for estimating and testing complex models as well as to be a strong competitor to existing estimation methods.

Frs-3 Model-implied instrumental variable estimation with MIIVsem

Zachary Fisher, *University of North Carolina at Chapel Hill, United States*

MIIVsem is an R package for estimating and evaluating structural equation models (SEMs) using model-implied instrumental variables (MIIVs). The MIIV approach differs from system-wide estimators such as maximum likelihood in that not all model parameters are estimated simultaneously. This feature leads to estimation routines which are computationally efficient and parameter estimates which are robust to many commonly encountered types of model misspecification. MIIVsem includes functions for identifying instrumental variables within a system of equations, fitting observed and latent variable models and conducting local tests of model fit. Analyses can be performed using either the raw-data or the sample moments. Furthermore, MIIVsem supports constrained estimation and a variety of model types including higher order factor analyses, time series models, categorical measures, and nonlinear latent variables. A variety

of computationally efficient bootstrap procedures and instrument diagnostics are also available. This presentation will provide an overview of MIIVsem development and introduce the most important aspects of the MIIV approach using the package functionality.

Frs-4 blavaan: Merging lavaan with JAGS and Stan

Ed Merkle, *University of Missouri - Columbia, United States*
Ellen Fitzsimmons, *University of Missouri - Columbia*

R package blavaan is intended to be an interface between lavaan and open MCMC software (such as JAGS and Stan), allowing users to specify a lavaan model that is automatically translated to the MCMC software syntax. In this presentation, we will provide an overview of blavaan, focusing on recent estimation issues that have arisen during package development. These issues include methods for estimation in Stan, comparisons across JAGS and Stan, and future extensions to multilevel, discrete data. We will also consider the standing of blavaan as compared to related Bayesian software, including package brms.

Frs-5 Automated selection of robust individual-level models using gimme

Stephanie Lane, *Netflix, United States*
Kathleen Gates, *University of North Carolina at Chapel Hill*

The gimme R package allows for the discovery of robust individual-level structural equation models that characterize temporal processes within time series data. The gimme R package combines the capabilities of lavaan (Rosseel, 2012) with a community detection procedure available in igraph (Csardi & Nepusz, 2006) to estimate group-, subgroup-, and individual-level relations in time series data from within a structural equation modeling framework. Common applications of the gimme R package include daily diary data and functional magnetic resonance imaging data. For example, Gates, Lane, Varangis, Giovanello, Guskiewicz (2016) used gimme to characterize heterogeneity in functional connectivity among a sample of retired NFL players. In this presentation, we will demonstrate how to use the gimme package and discuss our vision for how gimme can help answer research questions within psychological science.

Aula Magna

Symposium: Theory and assumptions underlying psychometric practice

Mag-1 Psychometrics' inherited ontologies: nomological networks, causal structures, and measurement

Keith Markus, *John Jay College of Criminal Justice, United States*

Ontology involves basic assumptions about what kinds of things populate the universe. Psychometrics has historically adopted an agnostic attitude toward ontology, focusing on solving applied problems drawing on whatever resources point to a practical solution. This agnostic approach has produced what might be described as inherited ontologies: Implicit assumptions incorporated into psychometric theory without critical reflection. Laws and causation offer two key examples. Nomological nets, constituted by laws, play a role in construct interpretation, selection, validation, assessment of interventions on constructs, and identifying formative models. Causation plays a role in causal theories of measurement, explanatory IRT, explanatory theory of validity, DIF/MI, and causal indicators. Seventeenth century assumptions about lawmaking and law-following that accompanied the popularization of laws of nature lack scientific credibility today, creating explanatory and logical gaps when the idea of laws is adopted piecemeal into contemporary psychometric theory. Likewise, the widespread adoption of a general-to-specific conception of general causal relationships determining specific behavior is not easily reconciled with the expectation that causes explain. A specific-to-general conception of individual causal powers determining causal generalities offers an under-appreciated alternative. Ontology also impacts the aims of methodologies by constraining what can be derived from other things and what can only be mapped out empirically having no underlying basis for explanation. There is probably little harm in adopting the agnostic approach as a short term strategy. At some point, however, it is important to revisit unanalyzed inherited ontologies to avoid conflicting assumptions, circular reasoning, or incoherent methodological advice.

Mag-2 Turning models upside down: a causal theory of error scores

Riet van Bork, *University of Amsterdam, Netherlands*

In modern test theory the function that relates the common factor to item scores is a probabilistic function. It is therefore essential that the model includes error scores; without error scores the model is deterministic. While

theories on the nature of the common factors in the model are part of validity research, the error scores typically lack a substantive interpretation and are instead described as 'residuals', that is, as whatever is unaccounted for by the common factors. In this talk we turn this around and start with a causal theory of error scores. In this causal theory of error scores we assume that error scores result from all unique determinants of the response variables. We distinguish between unreliable determinants, which we call circumstance variables, and reliable determinants, which we call characteristic variables. We show that different assumptions about error scores (1) result in different psychometric models, (2) have different implications for the chance experiment that accounts for the randomness in the response variables, and (3) have different implications for the reliability of item scores, item bias and local homogeneity. We argue for development of substantive theories on the nature of the variables that constitute the error scores, similar to how validity research focuses on the nature of the variable that explains the shared rather than unique variance in item scores.

Mag-3 The continuing story of coefficient alpha and the need for closure

Klaas Sijtsma, *Tilburg University, Netherlands*
Julius Pfadt, *University of Ulm*

Although the merits and limitations of coefficient alpha have been known for a long time and the limitations have been explained at great length, the psychometric literature continues producing contributions to the study of coefficient alpha. With these new contributions, new misunderstandings arise. One misunderstanding is that alpha is only useful when the items in the test satisfy the essential tau-equivalence model, thus denying the extreme usefulness of a lower-bound quality index. I will argue that alpha derives its usefulness precisely from the fact that it is a lower bound, and explain what this means for practical test construction and test use. Another misunderstanding is that alpha can have a positive discrepancy with respect to the reliability coefficient. This remarkable result is at odds with the theorem that alpha is a lower bound to the reliability. Because the correctness of the theorem is beyond doubt, statements suggesting the opposite have the effect of confusing both the psychometrician and the test constructor. I will analyze the origin of the positive discrepancy fable and show that it is true in a model that explicitly denies one of the assumptions on which the lower bound theorem is based. I will also argue that this is a useful model, yet a model different from the model that is at the basis of the lower bound theorem and as such does not impact the validity of the lower bound theorem.

Mag-4 A general framework for response dynamics with auxiliary information

Joost Kruis, *University of Amsterdam, Netherlands*
Han van der Maas, *University of Amsterdam*
Dylan Molenaar, *University of Amsterdam*
Gunter Maris, *ACTNext by ACT*

Traditional psychometric models, like the Rasch model (Rasch, 1960), are primarily based on desirable statistical properties or measurement-theoretic assumptions (van der Maas et al., 2011), and have a limited connection to the psychological process that generated the item responses. One way a response process to an item can be visualised is as a graph, where the nodes correspond to the cue (the question), alternatives (the response options), and auxiliaries (external information), and the edges between nodes describe the relationship between these. Endowing such a graph with a specific distribution we obtain the Ising model from statistical physics. We discuss how a Markov choice process that, for particular choice conditions, has several traditional psychometric models as its invariant distribution. Furthermore, we discuss how the introduction of the auxiliaries in the model, allows an interesting explanation for the occurrence of DIF.

Sala Colorada
Causal inference and mediation I

Col-1 The effect of differential measurement error on treatment effect estimation

Heather Harris, *Inteleos, United States*
S. Jeanne Horst, *James Madison University*

In applied settings, educational researchers are able to control for selection bias via propensity score matching through the creation of a comparison group that is qualitatively similar to program participants (Austin, 2011; Stuart, 2010; Stuart & Rubin, 2008). Prior to conducting propensity score matching, however, a researcher must first decide how to measure important covariates and which covariates to include in the model (Guo & Fraser, 2014; Millimet, 2011; Rudolph & Stuart, 2017; Steiner, Cook, & Shadish, 2011). Recommendations in the literature include to administer instruments that result in reliable covariate scores (Guo & Fraser, 2014; Steiner et al., 2011). However, little is known about the accuracy of treatment effect estimates when covariate measurement error varies systematically by treatment group. A

simulation study was conducted via R version 3.4.3 (R Core Team, 2017), and four propensity score matching methods were evaluated based on their ability to create qualitatively-similar treatment-comparison groups (nearest neighbor matching, nearest neighbor matching with a 0.2 caliper width, optimal matching, and Mahalanobis distance matching). Both error-prone (i.e., covariate scores simulated with measurement error) and error-free (i.e., covariate scores simulated without measurement error) sets of covariate scores were employed. The level of simulated measurement error varied by condition differentially for the treatment and comparison groups. As the level of measurement error increased for the comparison group, bias in the treatment effect estimate also increased. Implications for applied researchers are offered.

Col-2 Application of complier-average causal effect (CACE) model and issues

Jenn-Yun Tein, *Arizona State University, United States*
Will Pelham, *Arizona State University*
Chung Jung Mun, *Johns Hopkins University*

This presentation will discuss the application of Complier-Average Causal Effect (CACE; Little & Yau, 1998) analysis in intervention studies and present a simulation study that shows how the CACE analysis is affected by non-normality and compliance rates. Randomized controlled trials (RCT) and intent-to-treat analyses are gold standards for minimizing allocation bias and establishing causal effects. However, maintaining randomization is difficult in large scale longitudinal intervention studies due to issues related to treatment non-compliance, attrition, interferences, and so on. CACE analysis of intervention effects is based upon Rubin's causal inference model which incorporates engagement/compliance status such that CACE analysis examines the difference between (a) the intervention group participants who did receive the treatment (i.e., complied) and (b) the control group participants who would have received the treatment (i.e., would have complied) if offered. The premise for CACE analysis is to compare the participants in the intervention condition with a similar group of individuals from the control condition who would have participated had they been assigned to the treatment condition. The finding reflects causal effects of the receipt of treatment. CACE assumes a normal approximation to the outcome distribution. We conducted a simulation study that varied sample size, degree of skewness, regression coefficient (effect size), and compliance rate. The results showed that for studies with low compliance rates and low effect sizes, CACE might not be appropriate to evaluate intervention effects for data with even mild deviation from the normal distribution for the outcome variable.

Col-3 Assessing change of knowledge in a pretest-posttest educational design

Jairo Navarrete, *Universidad de O'Higgins, Chile*

One important goal in educational research is assessing the change of knowledge between two points in time since it allows to explain change in terms of cognitive variables, treatment effects and treatment interactions. Change of knowledge is usually assessed by using "gain scores" i.e. the difference between posttest and pretest measures. Although the unsuitability of gain scores has long been discussed (Cronbach & Furby, 1970), they are still nowadays employed and even preferred over more suitable alternatives due to its simplicity: two separate measures are transformed into an unidimensional score which is interpretable as gains of knowledge and usable in subsequent analysis. Issues of gain scores are that they (1) have low reliability (2) cannot control for the initial status of knowledge—they correlate negatively with pretest scores, and (3) important cognitive variables such as intelligence and memory influence them only sparingly. Hence, gain scores are not an adequate estimator for the change of knowledge. In this presentation, a simple statistical model for the change of knowledge is developed in the setting of an educational intervention. The model provides a theoretical rationale to the criticisms of gain scores mentioned above. In the light of these results, this work proposes a new estimator for change that has the appealing features of gain scores but none of its drawbacks. Classical test theory and Taylor series expansions are used to estimate its theoretical reliability. Further work along this line might path the way to develop novel statistical tools useful for analysing educational data.

Col-4 Small sample criterion for covariate balance in re-randomization

Jiaxi Yang, *Columbia University , United States*
Jiaqing Zhang, *Columbia University*
Rui Lu, *Columbia University*

Rerandomization (Morgan & Rubin, 2012) is proposed to eliminate covariate imbalance in the design stage. When covariates are correlated with the outcome, rerandomization will help to get more precise estimates (i.e. narrower confidence intervals) of the average treatment effect (ATE). When the sample size is small, it is hard to get balanced covariates and the criteria to evaluate the balance are obscure. Traditional criteria to measure covariate balance, such as Mahalanobis distance and standardized mean difference in propensity score matching (SMD),

require stable variances and large sample size. This paper identifies the best covariate balance criterion to ensure the most appropriate randomized assignment through rerandomization procedure in the context of the small sample. First, we simulate a large data set with multiple covariates that are only correlated with the potential outcomes. Second, generalizable small samples (n=100, 50 and 20) are taken from the large data, then the rerandomization procedure is performed multiple times. Third, different criteria are used to measure the covariate balance, including Mahalanobis distance, SMD, Bayesian-based criterion (Gelman et al., 2008) and Hansen-Bower's test (2008). Finally, according to each criterion, we select assignments with the best 2.5% covariate balance scores and analyze the estimated ATEs and their precision. We found that Hansen-Bower's test and Bayesian-based criterion perform best in small sample setting. Besides, the performance of these criteria is influenced by the level of the association between covariates and outcomes.

Sala Matte
Item response theory I

Mat-1 Developing a concept map for Rasch measurement theory

George Engelhard, *University of Georgia, United States*
Jue Wang, *University of Miami*

The purpose of this paper is to describe a concept map that reflects the key components of Rasch measurement theory (Rasch, 1960/1980). There have been several taxonomies described for item response theory (Kim, et al., 2018; Thissen and Steinberg, 1986), and this paper modifies and applies this approach to Rasch measurement theory. Rasch measurement theory reflects a key milestone in the paradigmatic shift from classical test theory to item response theory. Rasch models are most popularly used in Psychometrika articles among all item response models based on a document analysis by Kim, et al. (2018). A systematic analysis of Rasch measurement theory and its recent developments can facilitate an understanding of objective measurement in social science. The particular aspects of the concept map will include a categorization of members of the Rasch family of models (e.g., dichotomous, rating scale, partial credit, many-faceted, mixture, multilevel, explanatory, and behavioral Rasch models), methods of parameter estimation related to Rasch models (e.g., pairwise, minimum Chi-square, joint maximum likelihood, conditional maximum likelihood, marginal maximum likelihood, and Bayesian estimation methods), principles of invariant measurement for different facets (e.g., item calibration, person measurement, and rater evaluation), and the application of Rasch measurement theory in solving various measurement problems (e.g., differential item functioning, test equating, standard setting, person fit, rater effects, and computer adaptive testing). Concept maps and taxonomies provide useful didactic tools for understanding progress in measurement theory in the human sciences.

Mat-2 Fixed common item parameter calibration with fixed guessing 3PLM

Sung-Hyuck Lee, *Graduate Management Admission Council, United States*
Kyung T. Han, *Graduate Management Admission Council*

The three-parameter logistic model (3PLM) has one of the most popular choices of item response theory (IRT) models because it takes the probability of guessing behavior associated with the multiple-choice item format into account. One of the important issues with using the 3PLM in practice, however, is that the estimation of c-parameter (i.e., pseudo-guessing parameter) is often unstable even with a large sample and it could severely affect the accuracy of the other item parameters as well. The negative impact of this issue can be even worsened when new items are placed on the same scale as the old items using a linear transformation method (e.g., Stocking and Lord, 1983) because the c-parameter values are still left untouched after the linear transformation. Han (2012 & 2015) suggested a practical and the theoretical framework of the fixed guessing three parameter logistic model (FG3PLM), and this model combined with the fixed common item parameter (FCIP) calibration may offer a useful solution to tackle aforementioned practical issues associated with the 3PLM. In this paper, a practical framework for the FCIP with the FG3PLM is proposed and evaluated using both empirical and simulation data. The study compares outcome from the proposed framework with traditional methods. The preliminary results of the study suggest that the combination of the FCIP and the FG3PLM dramatically reduces the sample size required to stably estimate calibrate the item parameters and outperforms the other traditional methods studied in terms of recovering the true scales for item and person parameters.

Mat-3 Validation of an IRT model accommodating item complexity

Daniel Bolt, *University of Wisconsin - Madison, United States*
Sora Lee, *University of Wisconsin - Madison*

James Wollack, *University of Wisconsin - Madison*
Carol Eckerly, *Educational Testing Service*
John Sowles, *Ericsson, Inc.*

Asymmetric IRT models have been proposed as a way of capturing variability in item complexity. In this paper, we provide a real data validation of such models using items administered under a discrete-option multiple choice (DOMC) format. Under the DOMC format, the scheduled number of distractor response options to which an examinee is exposed varies randomly, thus manipulating the psychometric complexity of the item. In this context we demonstrate the applicability of Samejima's logistic positive exponent (LPE) model to a real DOMC dataset and discuss its advantages in accommodating the effects of this manipulation. Application of the LPE in the context of DOMC items is shown to (1) provide a superior comparative fit relative to a two-parameter logistic model, and (2) yield a latent metric with reduced shrinkage. Broader implications for the use of IRT models in addressing item complexity are considered.

Mat-4 An application of the continuous response model for subtest data

Weldon Smith, *California State University Channel Islands, United States*
James Bovaird, *University of Nebraska Lincoln*
HyeSun Lee, *California State University Channel Islands*

Traditional Item Response Theory (IRT) models rely on item-level data and large sample sizes, which may be limiting factors in applied research. Researchers may also find themselves limited to subtest or booklet scores if relying on secondary data, or if test security or data storage does not allow for item-level information. The Continuous Response Model (CRM; Samejima, 1973) offers a way to obtain IRT estimates of ability with subtest or booklet data. The current study investigated whether CRM estimates of ability using subtest scores were accurate compared to estimates obtained using a more traditional IRT model applied to item-level data. An empirical study was performed first estimating ability using a two-parameter IRT model with item-level data, next partitioning items into various subtests and applying the CRM, and finally comparing estimates between models. A simulation study was also performed to investigate factors that may influence the performance of the CRM compared to a traditional IRT model including sample size, number of subtests, and items per subtest. Results from the empirical study showed a high degree of concordance between traditional IRT and CRM estimates of ability. The simulation study showed that the CRM did not perform as well as more traditional IRT models, but again showed high concordance across all conditions. The current study highlighted similarities between applying the CRM and parceling in factor analytic models. The CRM offers researchers and practitioners a way to obtain accurate IRT estimates of ability when item-level data are unavailable but subtest or booklet scores are.

Mat-5 The direction of measurement in multidimensional IRT models

Wes Bonifay, *University of Missouri - Columbia, United States*

An important yet underreported consideration in multidimensional item response theory (MIRT) modeling is the direction of measurement (DOM). Reckase & McKinley (1991) demonstrated that directionality is essential in characterizing key descriptive features of a multidimensional response probability surface (e.g., its steepness). This topic is especially critical in MIRT applications, because distinct DOMs among test items may suggest markedly different substantive interpretations. The present study demonstrates how the orientation of a multidimensional item response surface can be expressed in ordinary trigonometric terms, and how the DOM is reckoned by the univariate slope parameters of each latent trait that is measured by a given item. The DOM is then investigated in the context of common dichotomous and polytomous MIRT models and compared to various item-level statistics in both IRT (e.g., item fit) and factor analytic (e.g., communalities) modeling. A particular focus of this study is the role of the DOM in bifactor measurement models; specifically, the DOM is compared to several metrics of general factor strength, including essential unidimensionality, explained common variance, and the percentage of uncontaminated correlations. Further issues are also briefly discussed, including overall test-level directionality, estimates of uncertainty in the DOM coordinates, and the implications of the DOM in interpreting person parameter estimates within a multidimensional latent trait space. Overall, this work establishes that the DOM is an informative aspect of MIRT modeling that should be more frequently reported and discussed in MIRT applications.

Auditorium 2
Measurement invariance and DIF II

Au2-1 Applying bootstrap to the odds ratios methods for DIF detection

Henghsiu Tsai, *Academia Sinica, Taiwan*
Ya-Hui Su, *National Chung Cheng University*

Differential item functioning (DIF) analysis is an essential procedure for educational and psychological tests to identify items that exhibit varying degrees of DIF. DIF means that the assumption of measurement invariance is violated and that test scores are incomparable for individuals of the same ability level from different groups, which substantially threatens test validity. In this study, we proposed two odds ratios (OR) methods to detect uniform DIF for the Rasch model through a series of simulation studies. For the first OR method, the empirical sampling distribution of the OR for an item is constructed based on the bootstrap method. For the second OR method, the limiting distribution of the OR for an item is based on the limiting standard error of the OR. The performance of these two methods was compared, and an application was also conducted in the study.

Au2-2 Evaluation of missing and country effects on gender DIF

Luc Le, *Australian Council for Educational Research, Australia*

The Graduate Australian Medical School Admissions Test (GAMSAT) is a cognitive test developed by the Australian Council for Educational Research (ACER) for the Consortium of Graduate-entry Medical Schools in Australia, Great Britain and Ireland. GAMSAT consists of two writing tasks and two multiple-choice (MC) sections: Reasoning in Humanities and Social Sciences (75 items), and Reasoning in Biological and Physical Sciences (110 items). This study is designed to investigate effects of country (Australia and Great Britain) and (omit) missing responses on gender differential item functioning (DIF) in GAMSAT 2018 September. The data includes 1575 males and 2169 females from Australia, and 491 males and 931 females from Great Britain, respectively. Item response theory (IRT) Rasch model is used to detect gender DIF together with three different approaches to treat missing responses: ignored, wrong, and listwise. Results show that the rate of missing is very small (about 0.6%) and contributes a small (negligible) impact on gender DIF shown by the three missing treatments. However, the country factor shows a bigger effect on gender DIF. About 50% of items with substantial gender DIF (DIF magnitude $>$0.3 logits and DIF significant test $<$0.05) are consistently detected from both country groups, while others are flagged in only one country but not in both. Discussion on common relative strengths and weaknesses of males and females, as well as those by specific countries, in each GAMSAT MC section is given based on the contents and formats of the DIF flagged items found in the study.

Au2-3 Employing divide-by-total and divide-by-distractors differential distractor functioning methods to explain DIF

Burhanettin Ozdemir, *Siirt University, Turkey*

This study aims to investigate the behavior of distractors of English proficiency test (EPT) with divide-by-total and divide-by-distractors approaches to explain DIF effects. For this purpose, both DIF and DDF methods were employed to detect potential DIF items and DDF effects. Three different DIF methods that are three-parameter (3PL-NLR), four-parameter nonlinear logistic model (4PL-NLR), and Mantel-Haenszel Delta-DIF methods were used to detect DIF items. Moreover, both multinomial log-linear (MLR) and 2PL nested logit model (2PL-NLM) DDF methods were applied to examine the distractors behavior and to determine whether DIF in correct option contributed to DDF effect or distractors caused DIF. According to MLR DFF statistics, five items were detected as exhibiting DIF, in which four of them were detected as DIF items with moderate to large DIF effect. Thus, DIF effect might have contributed to significant DDF effect of these items. Among these four DIF items, two DIF items (st3 and ca7) had significant DDF effect obtained from both MLR and NRM DDF methods indicating that distractors functioned differently across gender along with correct options. Thus, one can conclude that distractors contributed to DIF effect for these two items and they need to be examined and revised by content experts. Unlike these two DIF items, DDF effect of other two items (rc21 and rc22), were not statistically significant based on 2PL-NLM, indicating that DIF effect present in correct options contributed to DDF effect rather than distractors themselves. In other words, stem or correct option caused DIF rather than distractors.

Au2-4 Comparing methods for detecting mode effect between PBA and CBA

Yi Dai, *Beijing Normal University, China*
Ping Chen, *Beijing Normal University*

International large-scale assessments such as Program for International Student Assessment (PISA) have been moving from paper-based assessment (PBA) to computer-based assessment (CBA) to provide a platform for new item types and for collecting additional process and timing data. Although CBA offers many advantages over traditional PBA, the researchers found that there might be "mode effect" between different test modes, which would undermine the fairness of the tests. Mode effect refers to the observation that tasks presented in one test mode (e.g., paper-based) may function differently when

presented in another test mode (e.g., computer-based; OECD, 2017). Two schemes, one-step method and two-step method, have been proposed to detect mode effect. In the two-step method, the unknown parameters (e.g., item parameters) for a same set of items in PBA and CBA are estimated separately, and then ANOVA or confirmation factor analysis (CFA) or differential item functioning (DIF) can be used to examine whether there are significant differences between the two test modes. In the one-step method, there are three representative models (OECD, 2017) in which the "mode effect" parameters are defined in the model specification, and "mode effect" parameters and other parameters are estimated simultaneously. However, until now no reference has publicly become available about the full comparison of these two kinds of methods. Real data of the reading, math and science from PISA 2015 were analyzed here. The result showed that Model 2 and Model 3 in one-step method generated smaller AIC and BIC than the other methods.

Au2-5 Stability of Rasch item difficulty by test delivery modes

Van Nguyen, *The Australian Council for Educational Research, Australia*

uniTEST has been developed by the Australian Council for Educational Research to assist universities with the process of student selection. The test has been developed to assess the generic reasoning and thinking skills required to successfully complete studies at the tertiary level. The test consists of five components of 28 multiple-choice items each– Quantitative Reasoning, Critical Reasoning, Verbal-Plausible Reasoning, Scientific Reasoning, and Interpersonal Reasoning, in a mixed order. In 2018, 4777 Danish candidates sat an online test (4285 at campuses, 217 by remote Proctoring), and 275 candidates sat a paper-based test.The aim of this study was to explore the effect of the three test delivery modes on the item difficulty in the test. Initial results showed that the data from three test delivery modes fit well to the Rasch model. In general, item difficulty patterns from the delivery modes functioned similarly to each other. There were some evidences to suggest that items at the end of the test were relatively easier for test-takers who sat online test on campus than other who sat by paper-based test and by remote Proctoring. This suggests that test-taker group on campus might have managed test time better than other groups. Detailed discussion on the possible relative advantages for each delivery mode will be provided in relation to item content or formats and their position in the test. Gender and study background will be included in a further study when the sample sizes of paper-base or in remote proctoring are large enough.

Auditorium 3
Mathematical modeling

Au3-1 Modeling risk behavior by the censored generalized finite mixture model

Nienke Dijkstra, *Erasmus University Rotterdam, Netherlands*

Risk behavior can have substantial consequences on health, well-being, and functioning. Consequently, it is studied extensively in various scientific fields, such as psychology and economics. Previous studies have shown an association between real-world risk behavior and risk behavior on experimental tasks, such as the Balloon Analogue Risk Task (BART) and the Columbia Card Task (CCT). However, modeling risk behavior with these tasks is challenging for several reasons. The first reason concerns censored observations, as most experimental tasks may randomly end prematurely (e.g., by popping the balloon in the BART). Second, certain risk outcomes seem to be more attractive to participants than others. For example, in the CCT, some participants create a geometric pattern when turning over cards, such as a diamond or complete row or column, and stop when this pattern is completed. These patterns lead to inflated values for some outcomes. Last, there can be differences between a priori unknown groups of participants in their attractiveness to certain risk-levels. So far, none of the existing studies have provided a statistical model that accommodates these issues. Here, we propose the Censored Generalized Finite Mixture Model (CGFMM), which models risk behavior while handling censoring, experimental conditions, and attractiveness to certain patterns and risk-levels. This model is applied to an exceptionally large data set with n>3000 participants that each completed 16 rounds of the CCT. Background variables on socio-economic status and other individual characteristics are available. We discuss the main results of the CGFMM applied to these data.

Au3-2 POT-MIRT: Psychometric modelling of a cognitive theory of intelligence

Kristof Kovacs, *ELTE Eovos Lorand University, Hungary*

Performance on diverse cognitive tests always correlate positively. This is called the positive manifold, which can be statistically accounted for by a general factor, g. g

is usually identified with a domain-general cognitive ability. An alternative explanation, process overlap theory (POT) assumes that any item or task requires a number of domain-specific as well as domain-general cognitive processes. Domain-general processes involved in executive attention are activated by a large number of test items, alongside with domain-specific processes tapped by specific types of items only. Besides the positive manifold, the theory accounts for a number of other phenomena, such as the higher across-domain variance in low ability groups (differentiation). This is explained by executive processes acting as a bottleneck and at low levels masking individual differences in domain-specific processes. Process overlap theory is translated to a multidimensional item response model (M-IRT) that is partially compensatory (within domains) and partially non-compensatory (across-domains). POT thus abridges psychometrics and cognitive psychology: linking item response theory in general with a cognitive theory is unusual in the field of intelligence. Also, an item-level approach is arguably appropriate because we are trying to explain why someone provides a correct or incorrect answer to a test item and the cognitive processes we are trying to identify are the ones responsible for the answer. Besides the model itself simulations will also be presented that demonstrate that data generated on the basis of the M-IRT model (i.e. without positing a unitary ability) fit a higher-order general factor model.

Au3-3 The leaky integrating threshold and its impact on evidence accumulation models

Stijn Verdonck, *University of Leuven, Belgium*
Tim Loossens, *University of Leuven*
Marios Philiastides, *University of Glasgow*

Choice RT experiments are the dominant experimental paradigm for investigating elementary decision processes. A common modeling assumption is that after evidence accumulation has reached a certain threshold, an independent motor process is initiated, leading to the physical response. However, neurophysiological findings suggest that motor preparation partly overlaps with evidence accumulation, and is not independent from the stimulus difficulty level. In this talk, I propose to model this entanglement by expanding existing evidence accumulation models with a secondary, Leaky Integrating Threshold process (LIT) that represents the motor preparation ramping alongside the decision process. This additional process takes the accumulated evidence of the original decision process as a continuous input, and triggers the actual response when it reaches its own threshold. Applied to the simple constant drift diffusion model, the LIT outperforms the current standard of diffusion modeling in terms of cross-validation based on a diverse collection of choice RT datasets. Additionally, its central leak parameter offers a better explanation of the extensively studied speed/accuracy tradeoff than the traditionally used boundary separation parameter. Finally, for more advanced models of decision making like the Leaky Competing Accumulator or the Ising Decision Maker, the LIT results in effective decision boundaries that take the shape of the underlying potential function, which constitutes a more natural decision convergence criterion.

Au3-4 A novel approach to estimate the approximate number system

Victor Koleszar, *Universidad de la República Uruguay, Uruguay*
Camila Zugarramurdi, *Universidad de la República Uruguay*
Mario Luzardo, *Universidad de la República Uruguay*

The approximate number system (ANS) has been proposed as the basis of numerical representations, usually evaluated through the ability to estimate cardinality differences between two sets of objects. The simplest model that links this aptitude with the ability to discriminate is a probabilistic one, in which the response is modeled as a Bernoulli random variable worth 1 if the subject discriminated correctly, conditional on w —the value corresponding to the ability to estimate (ANS)— and dependent on R —the ratio between the cardinal number of sets presented. Estimation of w is usually obtained through a two step method (Halberda, Mazzocco & Feigenson, 2008). First, the conditional probability of discriminating is estimated via the proportion of correct answers in response to the presentation of several stimuli with the same ratio. In the second step, least squares is used to estimate w. This approach has the disadvantage that long series of repeated stimuli are needed, and thus few distinct ratios are used, which leads to an increased estimation error for w. In the present work we present a novel estimation method based on maximum likelihood without the need for repeated stimuli. Estimation of w is obtained from the conditional joint distribution and via the Newton Raphson algorithm. This approach improves the accuracy of the estimate, allows for a decreased number of stimuli and can be used in adaptive contexts.

Auditorium 1
Applications II

Au1-1 A comparison of hierarchical and bi-factor approaches in a short trait-emotional-intelligence measure

Pablo Pérez-Díaz, *University College London, United Kingdom*
K.V. Petrides, *University College London*

There is little doubt that currently, trait EI theory and their measures have been found valid and reliable in several research and application settings. With more than 35 validations and several replication studies, backing trait EI over poorly psychometrically supported EI constructs. This talk focuses on the suitability of choosing Bi-factor models instead of hierarchical when performing factor analysis in large validation samples. Although Bi-factor models have proved its suitability firstly in cognitive assessment research, they are nowadays undergoing a renaissance beyond cognitive intelligence. Thus, Bi-factor models were tested vis-à-vis with hierarchical models in two validation samples in Chile with an adapted and validated short measure of trait emotional intelligence (Spanish-Chilean-TEIQue-SF). These samples corresponded to general (n = 335) and clinical population (n = 120), respectively. We tested the two validation samples through Confirmatory Factor Analyses (CFA) and Exploratory Structural Equation Models (ESEM) with hierarchical and Bi-factor models. The results highlighted the greater fit and appropriateness of Bi-factor in comparison to hierarchical models, especially when tested through ESEM instead of the more stringent and classical CFA approach. We discussed the implications for theory and practice regarding short personality measures, such as the TEIQue-SF. Keywords: TEIQue-SF, Bi-Factor, Hierarchical, Validation, Psychometrics, Chile.

Au1-2 Multilevel analysis of perceived cybercrime risk in European Union

Ana Gomes, *University Institute of Lisbon and Portuguese Air Force Academy, Portugal*
José G. Dias, *University Institute of Lisbon*

This study addresses the perception of cybercrime risk in the 28 countries of the European Union. It aims to identify how the perception of risk of cybercrime varies across the EU taking demographic characteristics of respondents and country effects into account. Multilevel data structures, in which units of analysis are within macro units (e.g., countries) and share common characteristics, are very common in the social and behavioral sciences. Multilevel analysis has been developed to account for specific statistical characteristics of this type of data. In the context of factor analysis, multilevel factor analysis (MFA) takes this multilevel structure into account (Kim et al., 2016). The MFA considers not only the respondent level (level 1), but also a country level (level 2) that defines a nesting or hierarchical structure (Varriale & Vermunt, 2012). In particular, an MIMIC structure can be added to explain the latent variable in the multilevel structure. The measurement of the perception of risk of cybercrime is based on ten items (e.g., Identity theft, receiving fraudulent emails or phone calls asking for your personal details, Being asked for a payment in return for getting back control of your device). The data set comes from the Eurobarometer 87.4/2017 (TNS Opinion & Social, 2017) and contains information on respondents from the 28 countries of the European Union (sample of 27812 respondents reduces to 21657 for users of internet). The model shows a good fit and most of the covariates are significant in explaining the perception of risk of cybercrime.

Au1-3 Propensity to guess, self-confidence and risk-aversion of student in a test

Paula Fariña, *Universidad Diego Portales, Chile*

The presence of guessing in multiple choice tests (MCT) is a relevant topic of the educational measurement research. The main problem generated by guessing is the estimation bias of student's ability and item's difficulty, see Andrich et al. (2012, 2014, 2016). I present an empirical study where the links between propensity to guess, self-confidence and risk-aversion of students are explored. A specially prepared MCT were applied to 3rd grade students of the Econometric course in the career of Industrial Engineering of Universidad Diego Portales. The test includes for each item a Likert type scale to measure self-confidence; and it also ask the students to choose between 4 possible correction scenarios. The correction scenarios offer an extra score if an answer is correct and a penalty for an incorrect answer. The scenario selection is calibrated to measure risk aversion.

Au1-4 An exploration on the development of composite and domain scores

Danhui Zhang, *Beijing Normal University, China*
Mark Wilson, *Univeristy of California, Berkeley*
Perman Gochyyev, *Univeristy of California, Berkeley*

Hierarchically structured tests assessing across different domains are quite common in large-scale assessment settings. The needs for integrating the information from

both perspectives and for generating both the meaningful composite and domain scores are increasing. Such goal could be achieved by re-thinking about the relationship between measures, domain scores, and composite score, as well as the interpretation of these scores. A new model, composite model, proposed by Wilson and Gochyyev (2018) shows certain advantages in considering educational and psychological properties from both perspective of individual dimension and combination of multiple dimensions. It is also a new attempt for reconcile the unidimensional and the multidimensional perspectives. A large-scale national standardized mathematics assessment conducted in China was used as a real example in this study. This mathematics test consisted three domains, namely, Algebra, Geometry, and Statistics. In total, 59 items were administered, with 35 items from the domain of Algebra, 19 from Geometry, and 5 from Statistics. This composite model not only allows the correlation among dimensions, but also specifies a predicted model for an overall score based on the sub-domain scores. Therefore, different weighting schemes were applied to illustrate how the decisions on the weights would influence the final score estimation. Further discussions about the interpretability of the models in the specific usage or "policy" content were also provided.

Au1-5 Reliability and structure validity of a teacher pedagogical competencies scale

Juan Ignacio Venegas-Muggli, *Universidad Tecnológica de Chile INACAP, Chile*

The following presentation examines the reliability and structure validity of a quantitative observational teacher pedagogical competencies scale implemented at one of Chile's largest higher education institutions. In a context in which new students accessing post-secondary education are challenging traditional teaching methods, the evaluation of this instrument is presented as a relevant case study for those interested in promoting teachers' pedagogical competencies. Reliability analyses considered the KR-20 coefficient and corrected item-total correlations. Structure validity was measured through an exploratory factor analysis in which the concept's theoretical and latent structures were compared. The results suggest that the scale has high levels of internal consistency. Additionally, although the scale's theoretical and latent structures do not match exactly, relevant common elements are found. The considerations for applying these types of educational measurement instruments are discussed.

Invited Speakers

Salón Fresno
Invited Speaker: Gunter Maris

Frs-1 The wiring of intelligence

Gunter Maris, *ACTNext by ACT, USA*

The positive manifold of intelligence has fascinated generations of scholars in human ability. In the past century, various formal explanations have been proposed, including the dominant g-factor, the revived sampling theory, and the recent multiplier effect model and mutualism model. In this article we propose a novel idiographic explanation. We formally conceptualize intelligence as evolving networks, in which new facts and procedures are wired together during development. The static model, an extension of the Fortuin-Kasteleyn model, provides a parsimonious explanation of the positive manifold and intelligence's hierarchical factor structure. We show how it can explain the Matthew effect across developmental stages. Finally, we introduce a method for studying growth dynamics. Our truly idiographic approach offers a new view on a century-old construct, and ultimately allows the fields of human ability and human learning to coalesce.

Aula Magna
Invited Speaker: Minjeong Jeon

Mag-1 A latent space modeling approach to unveiling respondents' and items' dependence structures in item response analysis

Minjeong Jeon, *University of California, Los Angeles, USA*

I will present a novel statistical framework for analyzing item response data. The proposed framework leverages ideas and tools from state-of-art latent space modeling approaches and aims to capture unknown, complex item and respondent dependence structures that may be undetectable with existing methods. Specifically, the proposed method understands item response data as a function of the distances between items and respondents, between items, or between respondents. The positions of individual items and respondents are displayed in a in a low-dimensional Euclidean space, showing sub-groups of items and respondents that may be too nearby or distant from each other. Since similarities and differences are explicitly modeled, the traditional (local) independence assumptions for items and for respondents are no longer needed in the proposed framework. I will first present a

simple latent space Rasch model that explicitly incorporates the effects of item and respondent latent positions on the probability of a correct response and then explain a more flexible approach that directly estimates similarities and differences between items as well as between respondents. Lastly, I will describe a hierarchical extension of the latent space item response model that accommodates hierarchical data structures and captures dependence structures for higher-level units, such as classrooms and schools. Empirical examples are provided to illustrate the use of the proposed models in practice.

Spotlight Speakers

Salón Fresno
Spotlight Speaker: Marjolein Fokkema

Frs-1 Prediction rule ensembles: Balancing interpretability and accuracy in statistical prediction

Marjolein Fokkema, *Leiden University, Netherlands*

Prediction Rule Ensembles (PREs) are a relatively new statistical prediction method. PREs aim to strike a balance between the high predictive accuracy of methods like random forests and boosted tree ensembles, as well as the high interpretability of sparse regression methods and single decision trees. As PREs generally consist of a small number of rules, they may be easier to interpret for human decision makers than full decision tree ensembles. As such, PREs may provide a promising statistical method for bridging the gap between clinical research and practice in psychology. The current presentation will focus on R package 'pre', which implements the algorithm for deriving PREs as originally proposed by Friedman and Popescu (2008). Package 'pre' offers several potential advantages over the original implementation, like greater flexibility, the use of unbiased rule induction algorithms, and support for different types of response variables (i.e., continuous, binary, count, multivariate, survival, and multinomial). Examples based on psychological research data will be presented to illustrate the potential of PREs, with a focus on topics and features that may be particularly useful for decision-making problems in psychology: the analysis of multivariate response variables, inclusion of confirmatory rules, the application of (non-)negativity constraints and the analysis of multilevel data.

Aula Magna
Spotlight Speaker: Thorsten Meiser

Mag-1 IRTree mixture models for decomposing trait-based responses and response styles

Thorsten Meiser, *University of Mannheim, Germany*

Lale Khorramdel, *National Board of Medical Examiners*

Multiprocess IRTree models provide a flexible approach to analyzing and controlling response-style effects like mid-scale or extreme responding. Most applications of IRTree models are based, however, on the inherent assumptions (a) that decision nodes reflect distinct response processes, so that the nodes are parameterized in terms of unidimensional IRT models with separate random effects, and (b) that the model parameters are homogeneous across respondents in a population, so that only one vector of discrimination and threshold parameters is estimated. These restrictive assumptions have recently been released in IRTree models with multidimensional node parameterizations and finite mixture components, respectively. IRTree models with multidimensional node parameterizations allow researchers to specify and test the hypothesis that decisions are affected by multiple judgment processes which may have similar or opposite effects on the observed response depending on the given tree structure (Meiser, Plieninger & Henninger, in press). Mixture-distribution IRTree models allow for parameter heterogeneity across subpopulations and can disentangle subpopulations that are susceptible to response style effects to varying degrees (Khorramdel, von Davier & Pokropek, under review). In the present study, we combine multidimensional node parameterizations and finite-mixture components in order to decompose subpopulations with different response behavior. Using mixture IRTree models with uni- and multidimensional nodes, we separate subpopulations that engage in trait-based response processes for fine-grained ordinal ratings from subpopulations that employ a simplified response process in which the trait only affects directional (dis-)agreement judgments. The new approach is illustrated with rating data from PISA 2015 and validated with extraneous covariates.

Invited and State-of-the-Art Speakers

Salón Fresno
Invited Speaker: Ernesto San Martin

Frs-1 How to broker the evaluation of public policies? A proposal based on partial identification

Ernesto San Martin, *Pontificia Universidad Católica de Chile, Chile*

In this talk, we propose a way to broker the evaluation of public policies. Educational researchers, psychometricians or statisticians become honest brokers when they clarify all policy options and their associated uncertainties. In the context of public policy evaluation, we argue that a partial identification analysis is a way to make explicit different non-testable assumptions. Each of these assumptions combined with observational evidence will provide different policy recommendations. These recommendations imply different volitional attitudes of the policy maker, which can be considered as the final position of scientific research in the sense of Neyman, that is, the inductive behavior. In this context covariates are no longer viewed as controls, but as contexts in which the policy eventually works. In passing, it is shown that the ignorability conditions typically used in the evaluation of public policies are logically strong but incredible. (This is joint work with Trinidad González-Larrondo.)

Aula Magna
Spotlight Speaker: Leah Feuerstahler

Mag-1 Characterizing uncertainty in item response model metrics

Leah Feuerstahler, *Fordham University, United States*

In item response theory, item parameter standard errors are used to characterize the uncertainty associated with individual parameter estimates. These standard errors also can be used to construct confidence bands (Thissen & Wainer, 1990) around estimated item response functions. Whereas early approaches to constructing confidence bands were based on Fisher information, Yang, Hansen, and Cai (2012) recently suggested a multiple imputation (MI) approach that can be used with any approximation to the item parameter covariance matrix. In both approaches, confidence bands are constructed by treating the latent variable theta as fixed and plotting the variability of response probabilities conditional on theta. However, theta can also be understood as an artifact of the fitted model such that the theta metric is determined with error. Specifically, the latent trait metric can be defined as a multidimensional random vector of conditional response probabilities (Ramsay, 1996). Because these multidimensional random vectors will lead to somewhat different predictions across calibrations, item parameter estimation error implies uncertainty about the location the metric. In this talk, I describe how an MI approach can be used to visualize and quantify the variability of the theta metric in item response theory. I also clarify the relationship of this method with respect to other IRT model evaluation outcomes (e.g., standard errors of estimated theta, model fit). Overall, I argue that metric stability provides unique information that aids in a holistic approach to model evaluation.

Parallel Sessions 1, Thursday Afternoon

Salón Fresno
Symposium: Advances in process data analysis

Frs-1 Learning & measurement of teamwork

Alina von Davier, *ACTNext by ACT, United States*
Benjamin Deonovic, *ACTNext by ACT*
Michael Yudelson, *ACTNext by ACT*
Pravin Chopade, *ACTNext by ACT*
Lu Ou, *ACTNext by ACT*

The focus of this presentation is to present an approach to measuring learning in collaborative problem solving (CPS) environments (games), educational simulations, and intelligent tutoring systems (ITS). We adapt known methodologies to modeling student performance individually and on the team level. These methodologies include Bayesian Knowledge Tracing (BKT) and Additive Factors Model (AFM) – two approaches from the field of Educational Data Mining (EDM) that have found frequent use in learning analytics. We propose an extension of the two methods to modeling the team's ability and team's learning, as in "collective intelligence" introduced by Woolley, Chabris, Pentland, Hashmi, and Malone in 2010. In many learning support systems, students interact with problem items that are tagged with the-so-called knowledge components (KCs). Upon the success/failure of attempts to solve the problem or problem step, models update their estimate of individual student's mastery

of relevant skills. We extend the traditional approaches from individuum's to team's skills. We base the extension on the hypothesis that there exist team-level latent abilities that influence team performance tangibly. Examples of KCs for the team performance are "sharing," "negotiating," "maintaining a positive collaboration," and so on (see Hao, Liu, von Davier & Kyllonen, 2016).

Frs-2 Measurement of complex problem-solving ability – a lesson from classical psychometric theories

Yunxiao Chen, *London School of Economics and Political Science, United Kingdom*

Complex problem-solving (CPS) ability has been recognized as a central 21st century skill. Simulation-based tasks have emerged in recent years for the measurement of CPS ability. Unlike traditional tests that usually contain large numbers of items, the measurement of CPS ability usually involves a much smaller number of tasks, while each task may extract more information about a test-taker through his/her problem-solving processes. For this new measurement problem, classical psychometric paradigms, including the classical test theory (CTT) and item response theory (IRT), are no longer suitable. In this talk, we propose a new psychometric framework from a statistical learning perspective that is much more flexible than the CTT and IRT frameworks. Under this framework, we demonstrate how test-takers' problem-solving process data, as well as other information, including their cognitive assessment data, can be used for the measurement of CPS ability. As will be discussed, this framework is closely related to CTT and IRT and can be regarded as a nature extension. Illustrative examples from PISA 2012 will be discussed.

Frs-3 Exploring action sequence-based approaches in process data analysis

Qiwei He, *Educational Testing Service, United States*

Identifying sequential patterns in process data can be useful for discovering, understanding, and, ultimately, scaffolding test takers problem solving behaviors. Two sequence-based approaches were explored using process data in problem solving items in the Programme for the International Assessment of Adult Competencies (PIAAC). The first approach features in disassembling long sequences into mini-sequences by n-grams. These n-gram based features could be further selected as the robust classifier to distinguish different subgroups. This approach is more favorable to the item-level analysis given the action sequences are often dependent on the particular task being analyzed. The second approach takes the sequence as a whole. It focuses on identifying the longest common subsequence between the individual action sequence and the pre-defined ones by calculating the distance in between. It provides the possibility to generalize factors that are associated with test takers' problem-solving behaviors across multiple items. These two approaches both hold promise in process data analysis from different perspectives. The possible applications in detecting items with differential item functioning (DIF) will also be discussed.

Frs-4 An exploration of process data in computer-based assessment

Jingchen Liu, *Columbia University, United States*

In classic tests, item responses are often expressed as univariate categorical variables. Computer-based tests allow us to track students' entire problem solving processes through log files. In this case, the response to each item is a time-stamped process containing students' detailed behaviors and their interaction with the simulated environment. The key questions are whether and how much more information are contained in the detailed response processes additional to the traditional responses (yes/no or partial credits). Furthermore, we also need to develop methods to systematically extract such information from the process responses that often contain a substantial amount of noise. In this talk, I present several exploratory analyses of process data.

Aula Magna

Symposium sponsored by DEMRE: A major change of the national college admission system in Chile: opportunities to improve

Mag-1 Assessing the predictive validity of an admission test using item level information

Mónica Silva, *Pontificia Universidad Católica de Chile, Chile*
Nancy Lacourly, *Universidad de Chile*

An exercise is presented using data from the Chilean national admission system to assess the extent to which a test's ability to predict college first year grade point average is affected by the deletion of subgroups of items. Using multiple regression, PSU Mathematics item characteristics–content area (numbers, geometry, algebra and statistics), level (basic or advanced)—were used to predict first year college grade point average. For many

academic programs, the deletion of advanced items did not translate into diminished prediction. The use of item level information can allow for a more effective tailoring of tests to maximize prediction.

Mag-2 Rurality gaps in the access to higher education: initial estimations

Valentina Giaconi, *DEMRE, Universidad de Chile, Chile*
Leonor Varas, *DEMRE, Universidad de Chile*

In Chile and many other countries, there is a lack of information regarding rurality gaps in higher education. Understanding these gaps is vital because of equity and diversity issues. Regarding equity, there is a historical association between lower socioeconomic conditions and access to educational opportunities in rural areas. Regarding diversity, rural youth has tremendous potential for the development because rural territories are fundamental to sustainable development and they are related to food development, natural environments, indigenous and traditional culture. Although, estimating rurality gaps is complicated because different entities have different definitions of rurality. For example, the Chilean Ministry of Education considers a definition based on the distance to an urban center; the National Institute of Statistics definition is based on population density, and the United Nations Development Program definition is based on the association with agricultural, fishing and forestry activities. In this study, we will present the initial estimations of rurality gaps in the access to higher education in Chile considering the multiple definitions of rurality given above. The indicators of access to higher education considered in this study are the score in the admission tests, the score based in the school GPA, and a ranking score that indicates the position of the student among students of the same context.

Mag-3 The complexity of linking over time in college admission

María Inés Godoy, *DEMRE, Universidad de Chile, Chile*
Constanza Cortés, *DEMRE, Universidad de Chile*
Gabriela Toledo, *DEMRE, Universidad de Chile*

The Chilean unified college admission system has to deal with contextual and technical problems regarding the linking between processes. On the one hand, the system regulations allow the students to use their score for two consecutive admission processes, on the other hand, most items are "released" in social networks right after the test is taken and students that take the test for a second time usually participate in special trainings to improve their scores. Finally, the actual system regulations define the scaling procedure using CTT. In this context, a good technical option to link the scores is by moving to IRT and incorporating the pilot studies to anchor items between years. In this study, we will present the item bank calibration design developed for the Chilean unified college admission system. The design links operative applications through the pilot studies, that is, including operative items in the pilot study, so that they can work as anchor items. Regarding the scaling, after trying different models, 1 PL was chosen, but grouping items according to their discrimination parameter. For the Spanish language test, three groups of items were identified. For each of the other tests, two groups were identified. The technical decisions made through the process and results will be discussed.

Mag-4 The development of core content instruments for college admission

Daniela Jiménez, *DEMRE, Universidad de Chile, Chile*
Sandra Cruz, *DEMRE, Universidad de Chile*
Nora Córdova, *DEMRE, Universidad de Chile*

The composition of the Chilean applicants to the university has changed dramatically in the last ten years, and in order to explore the potential incorporation of new approaches to the assessment instruments in the admission system, the department responsible for the test development has been running a research project which aims to develop competence based instruments. The present study will compare the properties between the current college admission test battery (PSU) and the new tests developed in the research project in mathematics and science from different perspectives: (1) the content emphasis in both tests, by comparing the tests specifications and assembly; (2) the psychometric properties by the analysis of a joint calibration with students that took both tests and (3) the achievement gaps between interest groups (gender, school track, SES) addressed by standardized effect sizes. The differences found suggest improvements for the current admission system.

Sala Colorada
Model uncertainty and robustness

Col-1 An approach to addressing multiple imputation model uncertainty using Bayesian model averaging

David Kaplan, *University of Wisconsin - Madison, United States*
Sinan Yavuz, *University of Wisconsin - Madison*

This paper considers the problem of imputation model uncertainty in the context of missing data problems. We argue that so-called "Bayesianly proper" approaches to multiple imputation, although correctly accounting for uncertainty in imputation model parameters, ignore the uncertainty in the imputation model itself. We address imputation model uncertainty by implementing Bayesian model averaging as part of the imputation process. Bayesian model averaging accounts for both model and parameter uncertainty, and thus we argue is fully Bayesianly proper, in the sense of Schafer (1997). We apply Bayesian model averaging to multiple imputation under the fully conditional specification approach. An extensive simulation study is conducted comparing our Bayesian model averaging approach against normal theory-based Bayesian imputation not accounting for model uncertainty. Across almost all conditions of the simulation study, the results reveal the extent of model uncertainty in multiple imputation and a consistent advantage to our Bayesian model averaging approach under MCAR and MAR over normal-theory multiple imputation in terms of Kullback-Liebler divergence and mean squared error. A small case study is also presented. Directions for future research are discussed.

Col-2 Analyzing extremely unbalanced and correlated data with hierarchical linear models

Hsiu-Ting Yu, *National Chengchi University, Taiwan*

Hierarchical Linear Models (HLM) have been widely used as the data analytical methods in psychological research to deal with the dependency naturally existed in data with a multilevel structure. However, several important methodological issues in HLMs are still lack of systematic and comprehensive investigations. This study examined two structural properties in multilevel data possibly found in empirical research: extremely unbalanced and highly correlated data structure. Systematic Monte Carlo simulation studies were conducted to examine the accuracy of parameter estimates in fixed- and random-effects. Factors examined in this study included the number of groups (level-2 units), the mean group-sizes (level-1 units), patterns of group-sizes (Equal and Unbalanced), and degrees of dependency in data. Preliminary results suggested that the unbalanced data structure compensate the inaccuracy in the estimates of fixed-effects parameters under the same total sample size. The unbalanced data structure also has more impact on the stability of estimates for random-effects parameters. Moreover, the number of groups also affected the accuracy of parameter estimation than the sizes of the group. Higher degrees of data correlation also lead to more underestimation of model parameters. The effects of extremely small samples are also analyzed and compared. Complete findings and possible implications will be reported and discussed in the presentation.

Col-3 Comparison of methods for quantifying model misspecification in SEM simulations

Tong-Rong Yang, *National Taiwan University, Taiwan*
Li-Jen Weng, *National Taiwan University*

Simulation studies play a critical role in advancing theoretical development and facilitating applications of structural equation modeling (SEM). Model misspecification constitutes one of the key model characteristics frequently manipulated in SEM simulations and the need for quantifying the degree of misspecification in SEM simulations has been repeatedly advocated (Bandalos, 2006; Bandalos & Gagné, 2012; Fan & Sivo, 2005, 2007). This study aimed at comparing six methods that have been suggested in the literature for quantifying model misspecification in SEM simulations. Six methods including the number of nonzero parameters misspecified, the absolute size of misspecified parameters, F0, RMSEA, Gamma of an analyzed model, and statistical power for rejecting a misspecified model using the chi-square test statistic were evaluated against four proposed properties. Population study was considered so as to cover as many conditions as possible and to prevent the confounding effects from sampling fluctuations. The quantities displayed by the six methods over a wide range of misspecified models were analyzed. RMSEA, compared to other five methods, was found to be most sensitive to varying degrees of specification errors over various model conditions and thus is tentatively recommended for quantifying model misspecification in SEM simulations based on the findings of this study.

Col-4 When good loadings go bad: Robustness in factor analysis

Ken Bollen, *University of North Carolina at Chapel Hill, United States*

Structural misspecifications are all too common in factor analysis models. Such errors include the incorrect number of factors, omitted factor loadings, or left out correlated errors (unique factors). A critical question is when do these specification errors bias factor-loading estimates? This paper treats the Model Implied Instrumental Variable, Two Stage Least Squares (MIIV-2SLS) estimator and derives analytic conditions of when MIIV-2SLS is robust

to such specification errors. Previous research has established when misspecifications in a latent variable model influence the measurement model estimates and vice versa. But they do not provide robustness conditions for errors within the measurement model and their impact on measurement parameter estimates. The new analytic conditions in this paper form the foundation of individual equation diagnostic tests to help localize problems within factor analysis models. Furthermore, I discuss the pattern of diagnostic test failures for different structural misspecifications. For instance, I give diagnostic test results to expect for an omitted cross loading versus a correlated error. Finally, I establish conditions for when a factor loading is unchanged even in the face of structural misspecifications. Simulation and empirical examples illustrate the results.

Sala Matte

Cognitive diagnosis models II

Mat-1 An exploratory diagnostic model for ordinal responses

Steven Culpepper, *University of Illinois at Urbana Champaign, United States*

Diagnostic models (DMs) provide researchers and practitioners with tools to classify respondents into substantively relevant classes. DMs are widely applied to binary response data; however, binary response models are not applicable to the wealth of ordinal data collected by educational, psychological, and behavioral researchers. Prior research developed ordinal DMs, yet there are several limitations and unresolved issues that must be addressed in order to more generally apply ordinal DMs in research. Together, these limitations pose a barrier for widespread application of ordinal DMs in education, psychology, and, more broadly, the social sciences. We offer an exploratory DM for ordinal response data to address existing limitations. We present an exploratory ordinal DM, which uses a cumulative probit link along with Bayesian variable selection techniques to uncover the latent structure. We apply the model to twelve items from the public-use, Early Childhood Longitudinal Study, Kindergarten (ECLS-K) Class of 1998-99 Approaches to Learning and Self-description (ALS) questionnaire and report evidence to support a two attribute solution with 16 classes to describe the latent structure underlying the teacher and parent ratings. In short, the developed methodology contributes to the development of ordinal DMs and broadens their applicability to address theoretical and substantive issues more generally across the social sciences.

Mat-2 An empirical Q-matrix validation method for the polytomous G-DINA model

Xue-Lan Qiu, *The University of Hong Kong, Hong Kong S.A.R., China*
Jimmy de la Torre, *The University of Hong Kong*
Kevin Carl Santos, *University of the Philippines*
James Zhang, *University of Groningen*

Several empirically based Q-matrix validation methods are available in the literature. However, all the existing methods were developed for cognitive diagnosis models (CDMs) for dichotomous attributes. For some applications, classifying students into more than two categories (e.g., no mastery, basic mastery, and advanced mastery) would be more instructionally relevant. To extend the practical utility of CDMs, methods for validating Q-matrix for CDMs for polytomous attributes are needed. This study focuses on validating the Q-matrix of the generalized deterministic input, noisy "and" gate model for polytomous attributes (pG-DINA model). The pGDI, an extension of the G-DINA model discrimination index (GDI) for polytomous attributes, is proposed. The pGDI serves as the basis of a validation method that can be used not only to identify potential misspecified Q-entries, but also suggest more appropriate attribute-level specifications. The following are some of the theoretical properties of the pGDI. First, like the GDI, the correct specification for an item is the smallest q-vector among the q-vectors with the highest pGDI. And second, unique to the pGDI, misspecifying the level of a required attribute leads to a lower discrimination. The practical viability of the proposed method is also examined using a simulation study. Preliminary results show that the method can accurately identify and correct misspecified Q-entries in the pG-DINA model, particularly when high-quality items are involved. Moreover, using the pGDI in conjunction with the mesa plot allows for quantitative and graphical information to be combined, resulting in a more effective implementation of the proposed method.

Mat-3 General CDM joint attribute model formulation and selection

Jianmin Zhuo, *The University of Hong Kong, Hong Kong S.A.R., China*
Jimmy de la Torre, *The University of Hong Kong*
Kevin Carl Santos, *University of the Philippines*

Cognitive diagnosis models (CDMs) provide finer-grained information that can be used to foster teaching and learning. However, when CDMs measure a large number of attributes, or when the testing condition is less than ideal (e.g., short tests or small number of examinees are

involved), the attribute classification accuracy can suffer when the attribute joint distribution is formulated generally. With the goal of identifying the best model in non-ideal conditions, the joint attribute model with higher-order (HO) latent trait structure is used with the generalized deterministic input, noisy "and" gate model (G-DINA) model. The HO model assumes that attribute mastery is governed by a smaller number of HO abilities (typically, one), and that the attributes are conditionally independent given the abilities. To investigate the robustness of the HO and G-DINA model combination, three factors, namely, joint attribute distribution (uniform, independence, one-dimensional HO, and two-dimensional HO), item quality (low, medium, and high), and test length are examined. The HO model is compared with saturated G-DINA model in terms of the attribute classification accuracy and model fit. Preliminary results show that the HO model performs better than the saturated model across the different attribute distributions, where more obvious differences can be found in the lower item quality conditions. These findings suggest that the HO model provides a flexible and robust parametrization of the joint attribute distribution in the G-DINA model context. Guidelines on how relative and absolute fit statistics can be used in choosing between the HO and saturated models are also discussed.

Mat-4 An optimal implementation of the GDI Q-matrix validation method

Miguel A. Sorrel, *Universidad Autónoma de Madrid, Spain*
Pablo Nájera, *Universidad Autónoma de Madrid*
Francisco J. Abad, *Universidad Autónoma de Madrid*

In the context of cognitive diagnosis models, a Q-matrix reflects the correspondence between attributes and items. The construction of this Q-matrix is typically theoretically based, most of the time relying on domain experts. This approach is subjective and may lead to misspecifications in the Q-matrix. All this will negatively affect the attribute classification accuracy. In response, several methods of empirical Q-matrix validation have been developed with the aim of correcting misspecified entries in a Q-matrix. One of these methods is the general discrimination index (GDI) method proposed by de la Torre and Chiu (2016). All items with a proportion of variance accounted for (PVAF) lower than a predetermined cutoff for PVAF are modified. In this method, the Q-matrix is assumed to be correct. This assumption is likely to be violated by the problems indicated above. A possible solution to this problem is to apply the method iteratively. Hence, this study investigates the iterative application of the GDI method where only one item is modified at each step of the iterative procedure, and the cutoff for PVAF is updated considering the new parameter estimates. The performance of this implementation of the GDI method was assessed by means of Monte Carlo simulations. Results showed that the performance of the GDI method improved when the application was iterative at the item level and was used in conjunction with an appropriate cutoff point. This was more noticeable when the original Q-matrix misspecification rate was high.

Auditorium 2
Computer-based testing I

Au2-1 Robust automated test assembly

Angela Verschoor, *Cito Institute for Educational Measurement, Netherlands*
Bernard Veldkamp, *University of Twente*

In order to optimize measurement precision in a desired difficulty range, automated test assembly (ATA) models often select items based on the amount of information they provide in this difficulty range. The amount of information is calculated using item parameters that have been estimated. Usually, uncertainty in these estimates is not taken into account in ATA. Maximizing Fisher information tends to favor items with positive estimation errors in the discrimination parameter and negative estimation errors in the guessing parameter. This is also referred to as capitalization on chance. Not taking the uncertainty into account might be a serious threat to both the validity and viability of ATA. Previous research showed quite an effect on the resulting test information function (TIF), whereby the TIF was overestimated by as much as 20%. In this paper, robust ATA is presented as an alternative method that accounts for uncertainty in the item parameters. The method consists of two phases. In the first phase, the parameters of the item pool are transformed into robust parameters, while in the second phase ATA is performed in the usual way, but using those robust parameters instead. In a simulation study, the effects of robust ATA are shown. The overestimation of the TIF turned out to be very small, usually well below 1%. Some theoretical considerations are shared. Finally, the implications are discussed.

Au2-2 The constraint-weighted procedure with the continuous a-stratification Index in CAT

Ya-Hui Su, *National Chung Cheng University, Taiwan*

Computerized adaptive testing (CAT) have been increasingly applied to many fields, including educational assessment, psychological testing, personnel recruitment, and clinical diagnosis. To ensure the validity of test scores, preventing items from being overexposed is very important to practical consideration because these items might be shared with current and future examinees. One popular method for item exposure control is a-stratification. When old or overexposed items would need to be removed from the item bank or new items would need be added into the item bank, the simulation studies would need to be conducted to determinate the optimal strata for a-stratification. It is undesirable for practitioners whenever the item bank is replenished frequently, such as high-stakes testing. Recently, a continuous a-stratification index (CAI) was proposed to incorporate item exposure control into the item selection index itself, and thus no need to partition the item bank into the fixed and discrete strata. However, the CAI still produced a high percentage of overexposed items that many practitioners couldn't view as negligible. Besides, many constraints are considered to assemble tests in practice, it is important to integrate the maximum priority index (MPI) procedure with the CAI for item selection in CAT. Therefore, the study was to investigate the efficiency of the MPI procedure with the CAI for item selection in CAT.

Au2-3 Developing multistage tests using D-scoring method

K. Chris Han, *Graduate Management Admission Council, United States*
Dimiter Dimitrov, *National Center for Assessment*
Faisal Al-Mashary, *National Center for Assessment*

The D-scoring method for scoring and equating tests with binary items proposed by Dimitrov (2016) offers some of the advantages of item response theory (IRT), such as item-level difficulty information and score computation that reflects the item difficulties, while retaining the merits of classical test theory (CTT) such as the simplicity of number correct score computation and relaxed requirements for model sample sizes. Because of its unique combination of those merits, the D-scoring method has seen quick adoption in the educational and psychological measurement field. Because item-level difficulty information is available with the D-scoring method and item difficulties are reflected in test scores, it conceptually makes sense to use the D-scoring method with adaptive test designs such as multistage testing (MST). In this study, we developed and compared several versions of the MST mechanism using the D-scoring approach and also proposed and implemented a new framework for conducting MST simulation under the D-scoring method. Our findings suggest that the score recovery performance under MST with D-scoring was promising, as it retained score comparability across different MST paths. We found that MST using the D-scoring method can achieve improvements in measurement precision and efficiency over linear-based tests that use D-scoring method.

Au2-4 Modeling multistage and targeted testing data with item response theory

Paul Jewsbury, *Educational Testing Service, United States*
Ru Lu, *Educational Testing Service*
Peter van Rijn, *ETS Global*

Both multistage (MST) and targeted testing (TT) involve routing test takers to one of a small number of possible test forms, in order to more closely match the difficulty of the test to the proficiency of the test taker. Some recent research has shown that Item Response Theory (IRT) item parameter estimation procedures may fail to recover the true item parameters in datasets generated by MST or TT designs (e.g., Wu & Xi, 2017; Lu, Jia, & Wu, 2018). MST involves routing test takers based on information internal to the test (the routing item responses), while TT involves routing test takers based on information external to the test (an auxiliary variable related to proficiency). Both designs involve missing data as not all items are observed by all test takers, but the missing data mechanism is different in each design. The consequence of this distinction to IRT estimation and model assumptions is explored with reference to missing data theory (e.g., Rubin, 1976; Mislevy & Sheehan, 1989), and selection theory (Meredith, 1993). Mathematical and simulation results show that valid item parameter estimation is possible for both designs with marginal maximum likelihood, but the use of internal versus external information in routing requires distinct analysis models to ensure the missing data is modeled appropriately. Selection of the appropriate model is critical to ensure valid item parameter estimation. General principles are derived for modelling data involving routing based on information internal or external to the test, or a mixture of the two types.

Au2-5 The asymptotic distribution of average test overlap rate in CAT

Edison Choe, *Graduate Management Admission Council, United States*

The average test overlap rate is often computed and reported as a measure of test security risk or item pool utilization of a computerized adaptive test (CAT). Despite the prevalent use of this sample statistic in both literature and operations, its sampling distribution has never been known nor studied in earnest. In response, a proof is presented for the asymptotic distribution of a linear transformation of the average test overlap rate in fixed-length CAT. The theoretical results enable the estimation of standard error and construction of confidence intervals. A practical example demonstrates the statistical comparison of average test overlap rates between two CAT designs with different exposure control methods. Moreover, a series of simulations examines the convergence in distribution under dependent processes.

Auditorium 3
Statistical methods

Au3-1 Model-based bootstrapping of the chi-square test in structural equation models

Raul Ferraz, *University of South Carolina, United States*
Alberto Maydeu-Olivares, *University of South Carolina*
Dexin Shi, *University of South Carolina*

Standard methods for assessing the fit of structural equation models estimated using maximum likelihood include, under normality assumptions, the likelihood ratio (LR) test statistic, and mean or mean and variance corrections to the LR to robustify it against non-normality. However, p-values for these procedures are based on large sample theory, and the asymptotic approximation is known to fail not only in small samples but also in large models that include a large number of observed variables. It has been suggested that bootstrapping techniques may provide more accurate p-values for the LR statistic than asymptotic methods. In this study, we report the results of a simulation study performed to compare the accuracy for p-values for the LR statistic obtained using the most widely used bootstrapping procedure (Bollen & Stine, 1992) and asymptotic methods (ML and MLMV estimators). A number of conditions were obtained by manipulating sample size, the number of observed variables and degree of non-normality.

Au3-2 Synergized bootstrapping: the whole is faster than the sum of its parts

Tim Loossens, *University of Leuven, Belgium*
Stijn Verdonck, *University of Leuven*
Francis Tuerlinckx, *University of Leuven*

In this talk, an adaption of the differential evolution (DE) global optimization heuristic is proposed to speed up the process of bootstrapping for problems with non-convex objective functions. In the context of bootstrapping, the optimization algorithm has to be run once for each re-sampled dataset. This can be a time consuming task. However, since the objective functions associated with the re-sampled datasets all look very similar (they are derived from the same original dataset), efficiency gains can be made by exchanging information between the different search processes. Such exchange is possible if all objective functions are simultaneously optimized, in a single DE search. With this approach the total number of function evaluations required to complete an entire bootstrap is significantly reduced, thus alleviating the computational burden.

Au3-3 Machine learning for estimation in IRT models

Mariana Curi, *University of São Paulo, Brazil*
Benjamin Deonovic, *ACT, Inc.*
Pravin Chopade, *ACTNext by ACT*
Gunter Maris, *ACTNext by ACT*

High dimensional latent space is still a challenge for usual estimation methods in Item Response Theory (IRT) models, like MCMC or maximum likelihood. In this work, we propose a Variational Autoencoder (VAE) architecture, a kind of unsupervised deep neural network, for a multidimensional IRT model parameter estimation. Our approach allows us to model high latent trait dimensions, overcoming some of the limitations concerned to "big data" analysis. The simulation studies show that, given enough data, the proposed method is competitive with the state-of-the-art ones with respect to predictive power and is much faster in runtime performance. The new approach is applied to a real data set to illustrate the usefulness of the proposed method in the context of educational assessment.

Au3-4 Standardized regression coefficients and new estimates for R^2 in multiply imputed data

Joost van Ginkel, *Leiden University, Netherlands*

Whenever statistical analyses are applied to multiply imputed datasets, specific formulas are needed to combine the results into one overall analysis, also called combination rules. In the context of regression analysis, combination rules for the unstandardized regression coefficients, the t-tests of the regression coefficients, the F-tests for testing R^2 for significance have long been established. However, there is still no general agreement on how to

combine the point estimates of R^2 in multiple regression applied to multiply imputed datasets. Additionally, no combination rules for standardized regression coefficients seem to have been developed at all. In the current paper, two sets of combination rules for the standardized regression coefficients are proposed. Additionally, two improved point estimates of R^2 in multiply imputed data are proposed, which in their computation use the pooled standardized regression coefficients. Simulations show that the proposed pooled standardized coefficients produce only small bias. Furthermore, the simulations show that the newly proposed pooled estimates for R^2 are less biased than two earlier proposed pooled estimates.

Au3-5 Replicability in psychology: The problem of familywise Type II error

Raymond Luong, *University of Toronto Scarborough, Canada*
Ken Butler, *University of Toronto Scarborough*
Robert Cribbie, *York University*

In modern psychological research, researchers have increasingly employed preregistration, multi-measure studies, multi-study investigations, and conventionally adequate statistical power. However, the common but problematic procedures for making conclusions from such converging metrics, such as vote-counting, and the implications of these practices in tandem have seldom been addressed. Specifically, multi-measure studies and multi-study investigations introduce a problem of multiple comparisons—either within a study or across a family of studies—that inflates type II error analogously to familywise type I error. This problem of familywise type II error has received little quantification in psychology, yet at conventional test-wise levels of power, the probability of committing at least one type II error drastically increases in as few as three independent tests (e.g., three replications). If this occurs, researchers may misinterpret results, critique (un)problematic methodology, or abandon an investigation entirely, despite the error falling within statistical expectations. However, when the number of measures or studies is known in advance, such as in preregistration and registered reports, familywise type II error can be predicted and accounted for. We will discuss two approaches for doing so. First, we will review recommended meta-analytic approaches for synthesizing results within multi-study investigations. Second, we will introduce a new, assumption-free application of the Bonferroni correction to simultaneously control familywise Type I and Type II error within a multi-measure study or across a multi-study investigation. We will also contextualize the limitations of both approaches, including heterogeneity within small meta-analyses and the feasibility of the Bonferroni correction.

Parallel Sessions 2, Thursday Afternoon

Salón Fresno

Symposium: Meaningful interpretation of measurement results: challenges in applied psychometrics

Frs-1 Interpreting psychometric results when model and attributes are defined by legislation

David Torres Irribarra, *Pontificia Universidad Católica de Chile, Chile*

This paper discusses how statistical issues interact with policy and practical constraints, analyzing the challenges of assessing and justifying the use of psychometric models when the models are defined by legislation and/or administrative procedures, as opposed to solely statistical criteria (such as the examination of model fit) or a theory guided approach. Definitions from legislation or regulation can influence psychometric work in several levels: (a) they can define how the construct is characterized, specifying which topics or existing instruments are to be considered as part of the attribute that is being assessed; (b) they can define the way in which the model should aggregate the information (specifying specific weights or formula scoring) and in this way tacitly defining the dimensionality of the construct(s); (c) they can prescribe specific models that are to be used; and (d) they can define specific features of the model such as predetermined scales in which the results should be conveyed or prescribe specific cut points to be used. Each one of these legal or regulatory definitions (and their interactions) can constrain in different ways the possibilities for conducting traditional validation studies, or at the very least, can limit the options of making improvements based on their findings. However, despite these limitations, assessment programs that are bound by legal and regulatory definitions can have important consequences attached to their results. This paper discusses how to reconcile from a psychometric perspective the analysis, technical justification and interpretation of assessments results in these contexts.

Frs-2 How to interpret a guessing parameter? A strategy based on identifiability

Ernesto San Martin, *Universidad Diego Portales, Chile*
Paula Fariña, *Pontificia Universidad Católica de Chile*
Jorge Gonzalez, *Pontificia Universidad Católica de Chile*

Using the well-known strategy in which parameters are linked to the sampling distribution via an identification analysis, we offer an interpretation of the item parameters in the one-parameter logistic with guessing model (1PL-G) and the nested Rasch model. The interpretations are based on measures of informativeness that are defined in terms of odds of correctly answering the items. It is shown that the interpretation of what is called the difficulty parameter in the random-effects 1PL-G model differs from that of the item parameter in a random-effects Rasch model. It is also shown that the traditional interpretation of the guessing parameter in the 1PL-G model changes, depending on whether fixed-effects or random-effects versions of both models are considered.

Frs-3 Objectivity and intersubjectivity of measurement across the sciences

Mark Wilson, *University of California, Berkeley, United States*
Andrew Maul, *University of California, Santa Barbara*
Luca Mari, *Università Cattaneo – LIUC*

One critical condition for the quality of measurement results is that they provide information for a specific property of the object under measurement, and hence, independent of any other property of the object or surrounding environment—that is they should display objectivity. A second critical condition is that the measurement results should be interpretable in the same way by everyone, even though they may have been obtained in different contexts by different individuals using different instruments: in other words, they should be subject-independent, or intersubjective. For both physical properties and psychosocial properties, objectivity applies if the definition of the property intended to be measured is sufficiently detailed to distinguish it from other properties (called "influence properties"), and if one can provide an account of how the measurement process works, i.e., how variation in the property causes variation in the outcomes. Similarly, for both physical properties and psychosocial properties, intersubjectivity can be secured by establishing the metrological traceability of the measurement results to a measurement unit, and more generally to a set of reference properties, though at present such solutions are less commonly found in psychosocial applications. In this paper we describe traditional and newer solutions to the problems of objectivity and intersubjectivity in the physical sciences, and then explore how these and other solutions can apply to non-physical measurement as well. We see that presenting objectivity and intersubjectivity in a single framework is a significant step towards the development of a conception of measurement across the sciences.

Frs-4 Establishing invariant and substantive units in psychometric modeling

Joshua McGrane, *University of Oxford, United Kingdom*
Derek Briggs, *University of Colorado, Boulder*

Establishing psychological units of measurement was a central concern of early psychometricians, stemming from foundational work in psychophysics, which would then primarily go on to influence the early mathematical psychologists and axiomatic theorists of measurement. In contemporary psychometric modelling, proponents of the Rasch model have espoused the invariance property of the model, whereby item/person estimates are represented on a scale with a common unit of a fixed size, or what is commonly termed an 'interval scale'. However, this invariant unit, i.e., the logit, is mathematically derived and, in the absence of theoretical substantiation, magnitude differences in person/item parameter estimates cannot be interpreted in terms of a substantive dimension. This issue is more broadly reflected in statistical modelling in the social sciences, where modellers attempt to overcome the lack of substantive units through recourse to standard deviation units or transformation to units of some other known dimension, e.g., time. Such solutions only serve to obscure rather than resolve this central problem. Nonetheless, increasing numbers of psychometricians, particularly from the Rasch tradition, have stressed the importance of modelling in terms of theoretical expectations of item/test content and complexity, where determinations of model fit are grounded in theoretical hypotheses rather than fit statistics. This modelling approach provides the opportunity to examine whether the fixed unit of Rasch model estimates may also be linked to the theoretical explanations of item complexity. We explore this possibility with an assessment of relational reasoning from cognitive psychology, and a science knowledge assessment based on a learning progression.

Aula Magna

Symposium: Latent variable modeling for intensive longitudinal data

Mag-1 Latent markov factor analysis for exploring longitudinal measurement invariance

Leonie Vogelsmeier, *Tilburg University, Netherlands*
Jeroen K. Vermunt, *Tilburg University*
Kim De Roover, *Tilburg University*

Drawing valid inferences about daily dynamics of psychological constructs (e.g., depression) requires the measurement model (MM)—indicating how items relate to constructs—to be invariant within persons over time. However, the MM might be affected by time- or situation-specific artefacts (e.g., response styles) or substantive changes in item interpretation. To efficiently evaluate (violations of) longitudinal measurement invariance for multiple subjects simultaneously, we proposed latent Markov factor analysis (LMFA; Vogelsmeier et al., in press), which combines a discrete-time latent Markov model with mixture exploratory factor analysis (FA). The Markov model captures changes in MMs over time by clustering subjects' observations into a few states and state-specific FAs reveal what the MMs look like. In the present project, we tackle two remaining challenges. First, LMFA is tailored to discrete-time data with equal measurement intervals. However, longitudinal data often contains unequal intervals to capture random snapshots of continuously evolving processes. To accommodate such data, we replace the discrete-time (DT) with a continuous-time (CT) Markov model in LMFA. Second, we make the complex LMFA more accessible for applied researchers by using the following three-step approach (which is already widely used for other latent class methods) to estimate the model in three intuitive, manageable steps: (1) obtain states with mixture FA while treating repeated measures as independent, (2) assign observations to the states, and (3) pass these states to a DT or CT latent Markov model. In my talk, I will introduce LMFA with a motivating example and discuss simulation results of the two extensions.

Mag-2 Dynamic models of intraindividual variability with varying coefficients

Sy-Miin Chow, *Pennsylvania State University, United States*
Meng Chen, *Pennsylvania State University*

Dynamic models with varying coefficients provide a natural platform for testing certain intraindividual variability processes of interest. Imposing the assumption of time invariance on a system that does show varying coefficients can lead to erroneous change-related conclusions that disregard how the system changes differently depending on e.g., contextual factors or other additional covariates. We present one possible approach for examining time-dependencies in modeling parameters in the context of a dynamic factor model with lagged factor loadings in the measurement model, and vector autoregressive moving average (VARMA)-type relations among a set of latent variables in the structural (or dynamic) model. We introduce one estimation approach commonly used in the state-space literature to estimate the varying coefficients jointly with other latent variables, and review the conditions under which the varying coefficients are uniquely identifiable. We further discuss viable semiparametric and nonparametric functions for representing changes in the varying coefficients. A set of previously published data from the Affective Dynamics and Individual Differences (ADID) study is used to illustrate the consequences of ignoring the varying coefficients in the data.

Mag-3 Using latent state-trait theory to analyze intensive longitudinal data

Sebastian Castro-Alvarez, *University of Groningen, Netherlands*
Laura F. Bringmann, *University of Groningen*
Jorge N. Tendeiro, *University of Groningen*
Rob R. Meijer, *University of Groningen*

Traditionally, researchers have used time series and multilevel models to analyze intensive longitudinal data. However, these models do not directly address traits and states which conceptualize the stability and variability implicit in longitudinal research, and they do not explicitly take into account the measurement error. An alternative to overcome these drawbacks is to apply the models encompassed in the Latent State-trait Theory (LST) developed by Steyer and colleagues. These are longitudinal measurement models that are focused on distinguishing between traits and states. Yet, traditional LST models can be problematic when the number of measurement occasions increases because the models quickly become overparameterized and they might not converge. For these reasons, multilevel versions of LST models have been proposed, which facilitate the application of these models to intensive longitudinal data. In this study, we aim to identify when it is preferable to use the multilevel version of an LST model instead of the respective original version and to exemplify how to analyze intensive longitudinal data with multilevel LST models using simulated and empirical data. In order to do this, three models were selected:

The multistate-singletrait model, the common and unique trait-state model, and the trait-state-occasion model. We argue that multilevel LST models are valuable and interesting alternatives when analyzing intensive longitudinal data. Moreover, we show that the multilevel LST models are encompassed within the Dynamic Structural Equation Modeling (DSEM) framework, and, therefore, they can be further extended within this framework.

Mag-4 Approaching process-outcome research with piecewise latent growth curve models

Robin Wester, *University of Wuppertal, Germany*
Nils Frithjof Toepfer, *University of Jena*
Rolf Steyer, *University of Jena*
Gabriele Wilz, *University of Jena*

Understanding the processes underlying psychotherapeutic change is essential for improving the effectiveness of psychotherapy, yet outcome studies have been preferred over process-outcome studies, partly due to their complexity. Latent growth curve (LGC) models are a flexible tool taking on some of the difficulties of process-outcome research. The first important decision when applying LGC models is choosing a functional form that describes the overall change process. In most applications, a continuous curve with no breaks is assumed. Especially in the process of an intervention, such a smooth development seems unlikely for a lot of therapeutic elements. Piecewise trajectory models prove particularly useful to account for the transitions clients undergo in therapy, yet researchers meet certain difficulties when specifying the model, like deciding on the exact point of transition from one phase of therapy to the next. In this talk, I will explore the application of piecewise LGC models for modeling nonlinear change in process-outcome studies by drawing from empirical data on the course of resource activation in an intervention for family caregivers of people with dementia. Previous studies have used LGC modeling primarily to analyze the development of therapeutic variables or to identify predictors of the between-person differences in the change trajectories. Using the differences in the change trajectories themselves to predict therapeutic outcome has been merely done. I will discuss the usefulness as well as the limits of piecewise LGC models for analyzing phase-specific effects of the development of change mechanisms on outcomes.

Sala Colorada
Networks II

Col-1 Network analysis: A literature review and related R packages

Chang Che, *University of Notre Dame, United States*

Human beings are surrounded by network relations in an incredibly wide variety of forms and contexts. From interperson communication to inter-protein reaction, from physical inter-location transportation to virtual information exchange on webpages, many disciplines have recently shown growing interest in network analysis in recent decades. Starting from the purpose and the history of network analysis, this presentation aims to scratch the surface of the quantitative analytical methods on networks for researchers who have no previous network analysis experience. It will introduce the representation of networks, descriptive methods, and mathematical or statistical models such as Erdos-Renyi Model, Exponential Random Graph Model, Stochastic Block Model, and Latent Space Model. It will also give a glance at other topics such as community detection, network dynamics, network with covariates, applications of network analysis, and related R packages.

Col-2 Permutation test on logistic regression coefficients with social network data

Wen Qu, *University of Notre Dame, United States*
Haiyan Liu, *University of California, Merced*
Zhiyong Zhang, *University of Notre Dame*

In social and behavioral sciences, researchers need to test hypotheses on the relationship between the individuals' attributes and the social network. One way to conduct the inference of a social network is through the logistic regression (Wasserman & Pattison, 1996). However, the nature of network data (e.g. small size, non-normality, and dependence) violates the assumptions of the logistic regression, which can lead to an unreliable inference. To remedy the consequences of these violations with a normal-based hypothesis test, we present the permutation test procedure in the social network framework. The permutation test on the significance of a regression parameter can improve the accuracy of the hypothesis decision. In this study, we conduct a simulation to compare the performance of the permutation test and asymptotic likelihood-ratio test under various conditions.

Col-3 Network analysis of answer key matches for test security investigations

Joseph Grochowalski, *College Board, United States*

Haberman and Lee (2017) introduced a method for detecting unusually frequent exactly and nearly matching responses from test examinees. Their method focused on identifying response vectors (keys) that were observed in a test administration with unusually large frequency, such that agreement between responses is unlikely to emerge from a natural exam situation, thus providing possible evidence of coordinated testing among examinees matching the keys. In this paper, we modify the method and expand its scope to provide inference about networks of answer key sharing. We introduce a logic of applying Haberman and Lee's exact and near matching tools to maximize the chance of finding complete networks of cooperation, which could otherwise result in fragmented networks that appear unassociated. This modification would be useful to test administrators who are interested not only in having more statistical power to detect matching answer keys, but are also interested in knowing the possible extent and nature of key sharing in a testing administration. Our principle contributions are to redefine what is considered a suspected shared key, provide a new iterative method for comparing test takers' responses to suspected keys, and an updated method for controlling false positive associations between test takers and keys. In this paper, we review Haberman and Lee's (2017) method, identify practical limitations of their approach when applied directly to assessment data, and we illustrate our updated method and how it overcomes these limitations.

Col-4 Joint modeling of social networks and item responses

Shuo (Selena) Wang, *Ohio State University, United States*

Researchers are interested in whether and how students' friendship structures influence their school adjustment. However, flexible statistical models that can jointly analyze students' social networks and their school adjustment outcomes are still lacking. In this study, we develop a latent space model for heterogeneous and multimodal networks (LSMH), which combines the framework of the latent space modeling from network analysis with item response theory from psychometrics. LSMH simultaneously analyzes students' friendship networks and their self-report item responses by combining the two types of information in the same latent space. We applied LSMH to identify students with potential difficulties adjusting to the school life by jointly analyzing how well they are connecting to other students and what their self-report responses are to adjustment-related survey questions. We developed a variational Bayesian expectation-maximization algorithm to perform posterior inference on the item and person parameters of the model. A simulation study is conducted to evaluate the efficacy of LSMH. The results of a student adjustment study are also reported to illustrate the usefulness of LSMH.

Col-5 Hierarchical network model for peer effects: A hierarchical spatial model

Pei-Hua Chen, *National Chiao Tung University, Taiwan*
An-Shun Tai, *National Chiao Tung University*
Stephen Raudenbush, *University of Chicago*

Social network influence has become an emerging topic in education and labor economics studies. The data from social economics usually contains spatial and multilevel structure. Existing models for studying peer effects either ignore the multilevel structure or cannot use covariates to explain the variant at group level. Introducing multilevel spatial autoregressive model to capture the nested nature of data in social network is necessary. Using the friendship network matrix, this study proposed a multilevel spatial error model (MSEM) with spatial error on individual level to study peer effects. The MSEM model will be compared with multilevel linear model (MLM) and spatial error model (SEM). Using wave I data from the National Longitudinal Study of Adolescent Health, the authors analyzed how individual socioeconomic status, demographic information can predict high school GPA for MLM and SEM. In MSEM, the friendship network matrix is also included and a peer effect parameter will be estimated in the model. We applied Laplace approximation to the estimation procedure of MSEM for calculating maximum likelihood estimates. For model comparison, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) goodness of fit indices will be used. Preliminary findings suggested that the multilevel spatial error model is a better model comparing to spatial error and multilevel linear models.

Sala Matte
Scoring and estimation II

Mat-1 GLMM Scores in lme4: Derivations and applications

Benjamin Graves, *University of Missouri - Columbia, United States*
Ting Wang, *The American Board of Anesthesiology*
Edgar Merkle, *University of Missouri - Columbia*

Generalized linear mixed model scores (casewise first derivatives of the model likelihood) are useful for a variety of purposes, but their computations can be difficult due to the need for integral approximation methods. This is especially true of models estimated via R package lme4, which uses estimation methods that generally avoid derivative calculations. In this presentation, we show how these quantities can be obtained in a systematic manner and demonstrate the computations as an extension of the R package merDeriv. The derivatives have the potential to be applied to many methods of interest, including robust standard errors, comparisons of non-nested models, and score-based tests. We illustrate the many uses of these scores, focusing on logistic models that include explanatory item response models. Finally, we discuss some remaining issues surrounding the use of scores in GLMMs.

Mat-2 Estimating linear and polynomial one-factor models using conditional expectations

Elissa Burghgraeve, *Ghent University, Belgium*
Jan De Neve, *Ghent University, Belgium*
Yves Rosseel, *Ghent University, Belgium*

We propose a two-step procedure to estimate linear and polynomial one-factor models. In a first step, the latent variable is replaced by its conditional expectation given the observed data. We consider the James-Stein estimator for this conditional expectation. The second step consists of regressing the indicators on this estimator leading to consistent estimators of the factor loadings. The two-step procedure is flexible in the sense that it can easily be extended to accommodate for polynomial factor models.

Mat-3 Non-linear transformations and their effects on the comparability of transformed scores

Matthias von Davier, *National Board of Medical Examiners, United States*

The research presented in this presentation explores the effects of using non-linear functions for equating or transforming fallible, that is not perfectly reliable, test scores. A basic assumption of any valid scale transformations would be that if a perfectly reliable test score is transformed, it should be mapped on the true score of the target scale. This corresponds to the true (but unknown) equating function. For linear transformations, the true score on the new scale, the target of the transformed observed scores, equals the transformed true score of the untransformed score. However, this is no longer true if non-linear transformations are considered and tests are not perfectly reliable. The present paper derives an approximation of this bias in conditional expectations of equated observed scores and provides illustrations based on transformations used in large scale applications.

Mat-4 Score scale stability of six scoring methods

Stella Kim, *University of North Carolina at Charlotte, United States*
Won-Chan Lee, *University of Iowa*

Many large testing programs such as ACT, SAT, and Advanced Placement examinations, administer multiple forms of a test to provide examinees with ample testing opportunities. In practice, raw scores (i.e., number-correct scores) on each form are not reported, and raw scores are often transformed to scale scores developed to facilitate the interpretability of scores (Kolen & Brennan, 2014). Scale scores are typically established for the very first form of a testing program, and maintained for subsequent forms administered at a later time by identifying the relationship between the previous and subsequent forms. Maintaining the interchangeability of scores over time is a central feature of many large-scale testing programs. Although number-correct scoring is widely used, some other scoring methods often are considered especially with IRT. The impact of choice of IRT estimators (i.e., scoring method) has been investigated in the context of vertical scaling (Tong & Kolen, 2007; 2010) and multistage testing (Kim, Moses, & Yoo, 2015). However, little attention has been given to the long-term impact of using a particular scoring method on scale maintainability. The primary objective of this study is to compare the extent to which different scoring methods maintain scale scores and to evaluate six scoring methods in terms of score maintainability. Specifically, a simulation study will be conducted to assess the performance of the six scoring methods under various conditions including 1) the number of chains (the number of test forms to be linked in sequence), and 2) the pattern of change in examinee group ability.

Auditorium 2
Item response theory II

Au2-1 A Bayesian multidimensional item response theory model for small samples

Ken Fujimoto, *Loyola University Chicago, United States*
Sabina Neugebauer, *Temple University*

Nested-dimensional item response theory (IRT) models, such as the bifactor IRT model and the two-tier IRT model, are effective at determining whether general and specific dimensions of a measured trait are represented in

data, and if so, to what extent each type of dimension is associated with the item responses. These models are highly parametrized and thus could require larger samples than conventional IRT models. Unfortunately, many educational and psychological studies are conducted on smaller scales. Data from these studies, then, may not be suitable to be analyzed by nested-dimensional IRT models, thereby limiting the capabilities of researchers associated with these smaller-scale studies to confirm general and specific aspects of the measured traits represented in their data. For this presentation, a Bayesian multidimensional IRT model for smaller samples is introduced. The two-tier, bifactor, and between-item multidimensional IRT models can be obtained from the proposed Bayesian multidimensional IRT model. The proposed model, then, provides researchers conducting smaller-scale studies with the means to investigate a wide range of multidimensional structures in their data. Results from a simulation study are also reported. The simulation study shows that this model can confirm a two-tier structure (with eight dimensions), a bifactor structure (with seven dimensions), and a between-item multidimensional structure (with six dimensions) in rating data provided by as few as 100 individuals. Along with the simulation findings, details of an analysis confirming a two-tier structure in real rating data provided by 121 students is reported.

Au2-2 Facing innovation in national testing program: Rasch item-banking applications

Marta Desimoni, *INVALSI, Italy*
Cristina Lasorsa, *INVALSI*
Donatella Papa, *INVALSI*
Antonella Costanzo, *INVALSI*
Rosalba Ceravolo, *INVALSI*

Since the late 1960s, a number of scholars highlighted the advantages of applying Rasch item-banking in educational contexts. In the current years, Rasch item-banking may prove to be especially well adapted for facing the challenges of innovation in large-scale assessment programmes, such as the transition to Computer Based Assessment (CBA) and the improvement in score reporting practices. This study focuses on the application of Rasch item-banking to the national testing program on students' achievement carried out by the Italian Institute for the Evaluation of the Educational System (INVALSI). INVALSI survey undergoes two main changes from the 2018. The first one is the transition from paper and pencil to CBA. The second change is about the test-score reporting: test scores are not yet presented only as numerical proficiency score, but also as described proficiency levels, and an individual feedback is given to each student attending grade 8 (and 13 from 2019). Transition to CBA poses a number of psychometric challenges, e.g. the needs for interchangeable multiple test forms in order to ensure test security. Furthermore, item bank validity, as well as the efficiency of individual scores, need to be guarantee, in order to describe proficiency levels and to provide a reliable and valid individual feedback. Our application of Rasch item-banking to deal with this issues in the INVALSI CBA assessment will be described, and data from the INVALSI-2018 will be reported, focusing on the intertwined phases of multiple-test form design, level description and item bank evaluation and improvement.

Au2-3 Identifiability and nonparametric estimation of marginal distributions of latent variables

Danny Avello, *Pontificia Universidad Católica de Chile, Chile*
Ernesto San Martín, *Pontificia Universidad Católica de Chile*

Latent variables (LV) used in Item Response Theory (IRT) models are specified by the axiom of local independence (ALI). By doing this LV explain the non random variations in the observed scores, thus latent variable represents what is meaningful to the researcher. High impact decisions are made based on the estimates of them. However, because the marginal distributions of LV are not identified in IRT models, they are assumed. Our goal is to explore how much it can known about LV when they hold the ALI. We frame all IRT models that can be formulated as generalized linear mixed model in a Hilbert space. In this framework the ALI is replaced by conditional orthogonality, a weak version of ALI (WALI). We found that under WALI the only element for which the Empirical Bayes (EB) estimator is zero, is de zero element. Hence the WALI resolve an indeterminacy problem, because if there was more information in there, it would be impossible to recover it using EB. We found that is possible to have LV "living" among all the possibles observed scores, thus they are unobserved. Also, using the identification result of Székely & Rao (2000), we identify the marginal distributions of the LV underlying the observed scores, including the errors terms. Using the nonparametric estimation process proposed by Bohonomme & Robin (2011) we made some simulations. Finally, we used real datas and found that there are marginal distributions different from the normal distribution.

Au2-4 Measurement bias and error correction in a two-stage estimation

Xue Zhang, *Northeast Normal University, China*

Chun Wang, *University of Washington*

Currently the state-of-art estimation methods for multilevel IRT models are one-stage expectation-maximization (EM) algorithms or Bayesian MCMC algorithms. In this paper, we argue that the two-stage divide-and-conquer strategy has practical advantages, such as clearer definition of factors, convenience for secondary data analysis, convenience for model calibration and fit evaluation, and avoidance of improper solutions. However, under the two-stage framework, various studies have shown that ignoring measurement error in the dependent variable in stage II leads to incorrect statistical inferences (e.g., Fox & Glas, 2001, 2003; Rabe-Hesketh & Skrondal, 2004). To this end, we proposed a novel method to correct both measurement bias and measurement error from stage I in the stage II estimation. In this article, the higher-order item response theory (HO-IRT) model was considered as the measurement model, and a linear mixed effects (LME) model on overall abilities was considered as the structure model. The performance of the proposed correction method was illustrated and compared via a simulation study. The manipulated factors included sample size, covariance matrix of random effects and the latent coefficient. Results indicated that the structural parameters could be recovered better after correcting measurement bias and error. Finally, a real data example was given to demonstrate the efficiency and utility of the proposed method using the National Educational Longitudinal Survey data (NELS 88). Keywords: HO-IRT models; two-stage estimation; measurement bias; measurement error

Au2-5 Bayesian IRT equating: An alternative for small sample and complex design

Hyun-Woo Nam, *Soonchunhyang University, South Korea*

The purpose of this study is to find a suitable hyper priors approach that can estimate the parameters and perform equating in a stable manner, even when the examinee's ability level assigned to a particular block is not the same, and the test items are mixed with question types and difficult to assume a single dimension, and the sample size is too small to adopt the MMLE approach. The research data are from the 11th grade English test of the Korean NAEA (National Assessment of Educational Achievement) from 2016 to 2018. The K-NAEA is designed that the scoring items are administered to all students, but the anchor items distributed in 4 blocks are exposed to only some number of students assigned to each block. It will be examined whether the hierarchical hyper priors are more effective in the case of unusual data but the hyper parameter fixing method with small variance works well if the hyper priors is available and appropriate (Sheng, 2013; Nam, 2017) through the empirical data of K-NAEA.With assumption of the two-parameter logistic model for dichotomous items and the grade response model for polytomous items, the IRT parameters were estimated and equated using the WinBUGS program. The results are evaluated by calculating the root mean squared difference of parameter estimates, and by checking proportions of each achievement levels and proportion of school variance in ICC. Keywords: Bayesian IRT Equating, Hierarchical Hyper Priors, Fixed Hyper Parameters, K-NAEA, Matrix-sampled Anchor Items Design, Multidimensionality, WinBUGS

Auditorium 3
Misfitting response patterns

Au3-1 Application of person fit index to detection of faking responses

Hassan Mahmoudian, *Allameh Tabataba'i University, Iran*
Noorali Farrokhi, *Allameh Tabataba'i University*
Shirin Rezvanifar, *Allameh Tabataba'i University*

One of the main problems in the self-report personality tests is the probability of intentional deviance of responses. The purpose of this study was to apply person fit index to detect faking responses in polytomous data of personality test. In the present study, 902 students in two groups from Allameh Tabataba'i university (Iran) completed NEO Big Five Personality test. Subjects were assigned into two groups: honest and faking (honest= 462, faking= 440). Data were analyzed using Rasch Partial Credit Model (RPCM). The U3 person fit index was used to identify the faking responses' patterns. The results showed that the accuracy of U3 person fit index in Iranian samples was low in identifying the pattern of faking responding. Key words: Person fit index, Faking response, Polytomous data

Au3-2 Identifying persons who become inattentive: A dynamic modeling approach

Holger Brandt, *University of Kansas, United States*
Augustin Kelava, *University of Tuebingen*
Zachary Roman, *University of Kansas*
Mark Anderson, *University of Kansas*

An important question that has been asked repeatedly over the last decades is if participants actually respond to items or questions thoughtfully and in line with the

instructions. If they do not, and instead show an inattentive behavior by answering randomly or by using otherwise instruction-adverse behavior, results will be seriously biased for any subsequent analysis including factor models or even simple mean comparisons. The issue becomes more important when large test batteries are administered because participants might become tired or bored during testing; or when the testing is conducted using online platforms such as Amazon's Mturk where participation for some people is motivated by monetary incentives only. Here, a dynamic latent class structural equation model is presented that enables researchers to identify and account for such inattention when it occurs at some (unknown) time point during the testing. The model is based on a Bayesian framework that has been suggested recently (Kelava & Brandt, 2019). Its performance will be investigated in an intensive simulation study that models different types of instruction-adverse behavior. The results will be compared to a traditional "static" latent class model that assumes inattention throughout the testing procedure (e.g., Jin et al., 2018). Data from an online experiment will be used to illustrate the dynamic model. In this experiment randomly selected persons were instructed to answer instruction-adversely after a random number of items had been presented. Finally, advantages and challenges of the dynamic framework will be discussed.

Au3-3 Influential analysis for detecting aberrant school performances in high-stakes assessments

Andrés Christiansen, *University of Leuven, Belgium*

This study proposes a method for the detection of aberrant school performances in large-scale assessments using influential analysis under a Bayesian approach. Assessments that have high-stakes for school principals and teachers may be prone to various types of dishonest behavior at the school level (Jacob & Levitt, 2003) leading to invalid school results. Considering the scope and the scale of educational assessments it becomes necessary to screen aberrant school results that could indicate which schools may be cheating. School performances on a given year were modeled with a beta inflated mean regression model using Gibbs sampling (Bayes & Valdivieso, 2016) using as dependent variables the proportions of low and high achieving students in a school and the school's performance in previous years as predictors. The general measure of f-divergence proposed by Peng and Dey (1995) was used to determine aberrancy. This measure establishes the degree of influence a case has on a posterior distribution; observations with higher levels of divergence are considered aberrant. A simulation study revealed that it was possible to recover previously distorted school performances as aberrant using this method. The proposed method was applied to a high-stakes Peruvian national assessment in which 760 schools participated, and it was able to identify schools that showed an atypical performance increase given their previous results and conditions. Although the model does not pretend to certify that all of these schools had been cheating during the test administration, convergence was found between these results and other methods of cheat detection.

Au3-4 Does modeling wording effects help recover uncontaminated person scores?

María Dolores Nieto, *Universidad Autónoma de Madrid, Spain*
Luis Garrido, *Pontificia Universidad Católica Madre y Maestra*
Francisco J. Abad, *Universidad Autónoma de Madrid*

The item wording (or keying) effect consists of logically inconsistent answers to positively and negatively worded items that tap into similar (but polarly opposite) content. Previous research has shown that this effect can be successfully modeled through the Random Intercept Item Factor Analysis (RIIFA) model, as evidenced by the improvements in model fit in comparison to models that only contain substantive factors. However, little is known regarding the capability of this model in recovering the true uncontaminated person scores. To address this issue, the current study analyzed for the first time the performance of the RIIFA approach across three types of wording effects proposed in the literature: carelessness, item verification difficulty, and acquiescence. In the context of unidimensional substantive models, four independent variables were manipulated using Monte Carlo methods: type of wording effect, amount of wording bias, sample size, and test length. The results of the simulation corroborated previous findings by showing that the models that included the RIIFA method factor were consistently able to account for the variance in the data, attaining almost perfect fit regardless of the amount of bias. Conversely, the models without the RIIFA factor produced increasingly poorer fit with greater amounts of wording bias. Surprisingly, however, the RIIFA models were not able to better estimate the true uncontaminated person scores for any type of wording effect in comparison to the substantive unidimensional models. These apparently paradoxical findings are explained in light of the properties of the factor models examined.

Au3-5 How the omitted missingness reflects test-taking motivation? A new method

Yuan Liu, *Southwest University, China*
Kit-Tai Hau, *Chinese University of Hong Kong*

Test-taking motivation refers to the motivation to perform well on a given test (e.g., Arvey, Strickland, Drauden, & Martin, 1990; Baumert & Demmrich, 2001; Wise & Kong, 2005). It has been widely used in detecting test-takers' attitude towards the test, and could be influential to academic performance (e.g., Liu, Li, Liu, & Luo, 2019; Weirich, Hecht, Penk, Roppelt, & Hme, 2016). Among the numerous ways to detect the test-taking motivation, the nonresponse or missingness is particularly used in modern psychometric models to eliminate the non-effortful responses and to achieve more precise item estimations (e.g., Holman & Glas, 2005; Rose, 2013; Rose, von Davier, & Nagengast, 2017). In the present study, we proposed a new way to analyze the test-taking motivation by directly focus on the omitted missingness, using a zero-inflated Poisson model, and jointly analyzed the influential factors that may lead to non-effortful responses. The Programme for International Student Assessment (PISA) 2015 data was used as an example to illustrate such a method. We selected eight representative organizations from two cultures. Results indicated that the zero-inflated Poisson model could reflect test-taking motivation. We found that the Confucian students were less likely to have omitted missing values irrespective of the item positions compared to the western students. For western students, omitted missingness is more likely to be found at the end of the test, indicating a strong position effect.

Friday, July 19

Parallel Sessions, Friday Morning

Salón Fresno

Symposium: Modeling heterogeneity with time series data

Frs-1 Estimation and prediction of the Hawkes process with random effects

Peter Halpin, *University of North Carolina at Chapel Hill, United States*

Temporal point processes are usefully applied to time-stamped user-activity logs obtained from interactive education technologies such as games and simulations. Temporal dependence on these types of tasks is heterogeneous across task components (e.g., game levels), and across respondents. A random effects approach is proposed to describe this heterogeneity, with estimation discussed from maximum likelihood and Bayesian perspectives. Emphasis is given to a model specification that yields closed form expressions for the precision of the predicted values of the random effects characterizing heterogeneity over respondents. It is shown how these expressions can play a role analogous to the test information functions used in conventional item response theory. Implications for the use of interactive educational technologies to measure individual differences are discussed.

Frs-2 Dynamic mixture modeling: identifying unobserved groups in dynamic processes

Siwei Liu, *University of California at Davis, United States*
Lu Ou, *ACTNext by ACT, Inc*
Emilio Ferrer, *University of California at Davis*

Mixture modeling is commonly used to model sample heterogeneity by identifying unobserved subgroups of individuals with similar characteristics. Despite abundant evidence suggesting that individuals are often characterized by different dynamic processes underlying their physiology, psychological states, and behavior, applications of dynamic mixture modeling are surprisingly lacking. We present here a proof-of-concept example of dynamic mixture modeling whereby subgroups of individuals were identified based on distinctive dynamic patterns. Our sample consisted of 192 men who were in a heterosexual relationship. They were asked to complete a daily survey on their emotional experiences for at least 90 consecutive days. Two latent groups were identified based on the strength of association between their relationship-specific positive affect (e.g., loving) and relationship-specific negative affect (e.g., doubtful). Men in the group characterized by a strong negative association (beta = -.64) tended to be younger and showed higher levels of anxiety towards their relationship than men in the other group, which was characterized by a weaker negative association (beta = -.32). The model was estimated using "dynr", an R package capable of handling a broad class of linear and nonlinear discrete- and continuous-time models with regime-switching properties in the state-space modeling framework.

Frs-3 A methodological review on qualitative heterogeneity in quantitative changes

Lu Ou, *ACTNext by ACT, United States*

Advancements in technology make it easier to collect multimodal time series data in real time. While the data afford quantitative within- and between-subject variability at desirable resolutions for each subject, researchers often need to summarize single time series into phases or group subjects into subgroups to make inferences of the whole population. The process of categorization is essential for drawing qualitative heterogeneity from the complex data, and can occur at either the manifest or latent level. Despite its popular utility, methods for categorizing multivariate intensive longitudinal time series data are still nascent to the field of social and behavioral sciences and require examination and demonstration. In this project, we review different parametric and non-parametric approaches for identifying clusters and/or phases, including functional data analysis, hidden Markov models, and regime-switching dynamic models, and examine through simulations their respective strengths and limitations. With an empirical example, we illustrate their usage in characterizing real-time tracking data in learning environments.

Frs-4 Clustering of idiographic factor structures

Cara Arizmendi, *University of North Carolina at Chapel Hill, United States*
Kathleen Gates, *University of North Carolina at Chapel Hill*

Idiographic measurement models (i.e., P-technique) allow for the modeling of intraindividual variability. However, there is no consensus on how to arrive at nomothetic generalizations from replications of individual-level

measurement models, largely due to a lack of principled methods for deciding when latent constructs are similar enough between individuals to be considered the same construct. Methods such as the idiographic filter and LV-GIMME (Latent Variable – Group Iterative Multiple Model Estimation) allow for non-invariance at the level of the factor. A major challenge, however, is developing methods for comparison and grouping of individuals with similar measurement models. Here, we review different similarity measures for clustering of individuals based on similar factor structures. We focus on the congruence coefficient, as well as measures traditionally used for comparing cluster partitions: the Rand Index (RI), the Adjusted Rand Index (ARI), and the Jacard Index. We end with an empirical example demonstrating the utility of clustering individuals based on factor structure in LV-GIMME.

Frs-5 Multivariate generalized autoregressive conditional heteroscedasticity models for within-person variability research

Philippe Rast, *University of California, Davis, United States*

Modeling the variance in time-series data has a long tradition in econometric applications. A well known and frequently used set of models are generalized autoregressive conditional heteroskedasticity (GARCH) models for single time-series and the multivariate extension, the MGARCH. The main work surrounding these models focuses on different parameterizations of the conditional and time-varying covariance matrix H_t. While there are numerous parameterizations that satisfy specific features of time-series variability, they are derived from three basic specifications: VECH, BEKK, and DCC. While VECH and BEKK both model H_t directly, DCC operates on the decomposed H_t matrix into conditional standard deviations and a conditional correlation matrix. The focus here is mainly on the BEKK and DCC and its potential use in psychological applications. We illustrate its use and interpretation in two different intensive repeated design datasets covering approximately 90 consecutive days. In one dataset we fit different MGARCH models on several daily measured personality variables and on physical variables from fitbit devices. The other dataset contains dyadic data from up to 200 romantic couples. We use the time-series from each partners to estimate the interplay among the conditional variance parameters of affect over time. Specifically, the MGARCH enables one to estimate, in its basic form, an average error variance, a moving average and an autoregressive component for the variance – and with it, covariances and/or variance spill over effects from one time-series to the other. We close with a critical discussion.

Aula Magna

Symposium: A future for psychometric theory

Mag-1 Bridging the psychometric disciplines of individual differences and individual dynamics

Maarten Marsman, *University of Amsterdam, Netherlands*

Lourens J. Waldorp, *University of Amsterdam*

The psychometric literature comprises two distinct scientific approaches, one that is focused on individual differences and is based largely on models for cross-sectional data, and one that is focused on the individual and is based largely on models for individual dynamics. These two approaches exist in isolation and because of this it is often unclear when we can use cross-sectional results to inform individual dynamics, or vice versa. An important first step towards bridging the two aforementioned approaches is found in network psychometrics. A central model in network psychometrics is the Ising model, which can be used to describe the multivariate distribution of binary random variables, and can be estimated from cross-sectional data. This group-level Ising model can be formulated as a group-level average of idiographic networks, networks that are unique to the individual. This direct relation illustrates one way to bridge the two psychometric approaches. One way to view the idiographic approach is from ergodic theory. An ergodic trajectory can, in principle, end up everywhere in the state space. Yet in practice most of these trajectories end up close to each other. Most IQ scores tend to be larger than 60 but lower than 140, for example. This suggests the use of a weaker form of ergodicity. The difference between these two forms of ergodicity can be analyzed using Boltzmann entropy and Gibbs entropy, which tend to differ under weak ergodicity, and assess an individual trajectory or an ensemble, respectively.

Mag-2 A look into the future of psychometrics

Lisa Wijsen, *University of Amsterdam, Netherlands*

Over the last few years, I have interviewed 20 Psychometric Society presidents, with the aim of getting a better grasp of the history of psychometrics and the role of psychometrics in relation to other fields. In the final part of the interview, I asked the presidents to take a look

into the future of psychometrics and to identify the challenges ahead. As it turns out, the presidents' views on and expectations of their own field's future vary strongly. Some are afraid of psychometrics being erased by upcoming fields (data science, statistics or artificial intelligence), whereas the unique expertise of psychometricians is still very much needed in psychological research, and should not be forgotten. Others are confident psychometrics will be able to find its place and use among these other disciplines, and believe we should focus on what we can contribute outside the boundaries of psychometrics. Interestingly, many presidents agree on one thing: psychometrics is still very relevant, but psychometricians will have to take some action so as not to be forgotten, whether that is to re-establish a connection with psychology and protect our unique expertise, or to find a way to get along with other data-oriented disciplines. The testimonies paint a nuanced picture of the expectations and fears psychometricians have regarding their own field, but most of all, are a warning not to get left in the dust.

Mag-3 Why network psychometrics blocks reductionism in psychopathology research

Denny Borsboom, *University of Amsterdam, Netherlands*

In the past decades, reductionism has dominated research directions and funding policies in clinical psychology and psychiatry. However, the intense search for the biological basis of mental disorders has not resulted in conclusive reductionist explanations of psychopathology. Recently, network models, in which mental disorders arise from the causal interplay between symptoms, have been proposed as an alternative psychometric framework to understand the relation between symptoms and disorders. I will argue that this framework can help understand why reductionist approaches in psychiatry and clinical psychology are on the wrong track. First, symptom networks preclude the identification of a common cause of symptomatology with a neurobiological condition, because in symptom networks there is no such common cause. Second, symptom network relations depend on the content of mental states and as such feature intentionality. Third, the strength of network relations is highly likely to partially depend on cultural and historical contexts as well as external mechanisms in the environment. This means that neither psychological nor biological levels can claim causal or explanatory priority, and that a holistic research strategy is necessary for progress in psychopathology research.

Mag-4 Personality, resilience, and psychopathology: Slow and fast interacting network processes

Gabriela Lunansky, *University of Amsterdam, Netherlands*
Claudia van Borkulo, *University of Amsterdam*
Denny Borsboom, *University of Amsterdam*

Network theories have been put forward for psychopathology (in which mental disorders originate from causal relations between symptoms) and for personality (in which personality factors originate from coupled equilibria of cognitions, affect states, behaviors, and environments). We connect these theoretical strands in an overarching Personality-Resilience-Psychopathology (PRP) model. In this model, factors in personality networks control the shape of the dynamical landscape in which symptom networks evolve; for example, the neuroticism item "I often feel blue" measures a general tendency to experience negative affect, which is hypothesized to influence the threshold parameter of the symptom "Depressed Mood" in the psychopathology network. Thus, slow changing personality traits regulate the resilience of faster changing psychopathology symptom networks. Conversely, events at the level of the fast-evolving psychopathology network (e.g., having a depressive episode), can influence the slow-evolving personality variables (e.g., increasing neuroticism or decreasing extraversion). We apply the theory to neuroticism and Major Depressive Disorder (MDD). Through simulations, we show that the model can accommodate important phenomena, such as the strong relation between neuroticism and depression, and individual differences in the change of neuroticism levels and development of depression over time. The simulations can be replicated in an interactive, online tool. The implications for future methodological research will be discussed, namely, a novel approach to integrating different psychological constructs into one model (where the constructs vary over different time scales) and on simulating individual trajectories from population data.

Sala Colorada
Structural equation modeling

Col-1 Dealing with artificially dichotomized variables in meta-analytic structural equation modeling

Hannelies de Jonge, *University of Amsterdam, Netherlands*
Suzanne Jak , *University of Amsterdam*
Kees Jan Kan, *University of Amsterdam*

Meta-analytic structural equation modeling (MASEM) is a relatively new method in which effect sizes of different

independent studies between multiple variables are first pooled into a matrix and next analyzed using structural equation modeling. While its popularity is increasing, there are issues still to be resolved, such as how to deal with primary studies in which variables have been artificially dichotomized. To be able to advise researchers who apply MASEM and need to deal with this issue, we performed a simulation study using random-effects two stage structural equation modeling. We simulated data according to a full mediation model and systematically varied the size of one (standardized) path coefficient (βMX = .16, βMX = .23, βMX = .33), the percentage of dichotomization (25%, 75%, 100%), and the cut-off point of dichotomization (.5, .1). Next, we analyzed the simulated datasets in two different ways, namely by using (1) the point-biserial and (2) the biserial correlation as effect size between the artificially dichotomized predictor and continuous mediator. The results of this simulation study indicate that the biserial correlation is the most appropriate effect size to use, as it provides unbiased estimates of the path coefficients in the population.

Col-2 A bias-corrected limited-information estimator for small scale multilevel/categorical SEMs

Ben Kelcey, *University of Cincinnati, United States*
Kyle Cox, *University of Cincinnati*
Nianbo Dong, *University of North Carolina Chapel Hill*

Maximum likelihood estimation of parameters of multilevel structural equation models and structural equation models with categorical indicators is a common approach to operationalize sophisticated theories involving multiple latent variables. Although maximum likelihood has many desirable properties including consistency, a major limitation in these contexts is that estimated structural parameters often incur significant bias when implemented in studies with small to moderate sample sizes (e.g., 100 or fewer clusters if using a multilevel model or several hundred or fewer individuals if considering categorical indicators). To address similar limitations in the context of single-level studies with continuous indicators, recent literature has developed a two-stage limited-information estimator that draws on a method-of-moments type correction that has been shown to yield unbiased estimates with small to moderate sample sizes. In this study, we develop extensions to this bias-corrected limited-information estimator for multilevel structural equation models and structural equation models with categorical indicators. We then probe the degree to which the estimator can produce unbiased estimates of structural coefficients in these contexts with small to moderate but common sample sizes. The results suggest the promise of the estimator—the proposed estimator outperforms maximum likelihood in terms of bias, precision, and power in a wide variety of contexts. The proposed estimators are implemented in R and illustrated through several different types of multilevel structural equation models, structural equation models with categorical indicators or explanatory item response models.

Col-3 Sparse estimation for SEM via R package lslx

Po-Hsien Huang, *National Cheng Kung University, Taiwan*

Sparse estimation via penalization is now a popular approach to learn the associations among a large set of variables. In this talk, we introduce an R package called lslx that implements several sparse estimation procedures for structural equation modeling (SEM). Users can flexibly specify (1) which model parameters should be free/fixed for theory testing; (2) which parameters should be set as penalized for exploration; (3) which fitting function (maximum likelihood or least squares) and penalty method (minimax concave penalty, elastic net penalty, or stepwise search) are considered; (4) which selector for penalty level (AIC, BIC, or their extensions) should be used. Package lslx minimizes the penalized criterion through a quasi-Newton method. The algorithm conducts line search and checks the first-order condition in each iteration to ensure the optimality of the obtained solution. The current package also supports several advanced functionalities, including (1) a two-stage method with auxiliary variables for missing data handling, (2) a reparameterized multi-group SEM to explore population heterogeneity, (3) least squares methods for ordinal data SEM, and (4) valid post-selection inference methods.

Col-4 On the use of pairwise maximum likelihood estimation for clustered data

Mariska Barendse, *Erasmus University Rotterdam and Ghent University, Netherlands*
Yves Rosseel, *Ghent University*

Social and behavioural research frequently involves multilevel data, with individuals and groups defined at separate levels. Multilevel analysis within the structural equation modeling framework often leads to the use of models with a large number of (latent) variables (i.e., random slopes, random intercepts, and hypothetical constructs) on different levels. The analysis of categorical multilevel data requires the evaluation of high-dimensional integrals. Cur-

rent full-information approaches typically involve computationally intensive numerical methods (e.g., adaptive Gauss-Hermite quadrature or Markov chain Monte Carlo procedures). Alternatively, in the pairwise likelihood (PML) approach, the full likelihood is replaced by a sum of (bivariate) pairwise likelihoods, which are easier to handle. PML 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 show various possibilities of PML estimation for clustered data. Our approach is an extension of the so called 'wide' or 'multivariate' format approach that has been investigated by Bauer (2003), Curran (2003), and Mehta & Neale (2005) for continuous data with a multilevel structure.

Col-5 A mode-jumping algorithm for Bayesian factor analysis

Albert Man, *University of Illinois at Urbana-Champaign, USA*
Steven Culpepper, *University of Illinois at Urbana-Champaign*

Exploratory factor analysis is a dimension-reduction technique commonly used in psychology, finance, and economics. Advances in computational power have opened the door for fully Bayesian treatments of factor analysis. One open problem is enforcing identifability of the latent factor loadings, as the loadings are not identified from the likelihood without further restrictions. Nonidentifability of the loadings can cause posterior multimodality, which can produce misleading posterior summaries. The positive-diagonal, lower-triangular (PLT) constraint is the most commonly used restriction to guarantee identifiability, in which the upper m x m submatrix of the loadings is constrained to be a lower-triangular matrix with positive-diagonal elements. The PLT constraint can fail to guarantee identifiability if the constrained submatrix is singular. We demonstrate that though the PLT constraint addresses identifiability-related multimodality, it introduces additional mixing issues. We introduce a new Bayesian sampling algorithm that efficiently explores the multimodal posterior surface and addresses issues with current approaches.

Sala Matte
Thurstonian IRT

Mat-1 Pseudolikelihood person parameter estimates for MIRT-models of forced-choice-data

Safir Yousfi, *German Federal Employment Agency, Germany*

The Thurstonian IRT Model of Brown & 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. It is shown that multivariate integration is necessary to determine the likelihood of the response pattern exactly. The common practice of ignoring local stochastic dependencies can be understood as a pseudo-likelihood approach that will lead to similar estimates in most applications. However, Fisher information (standard errors) and Bayesian estimation techniques based on the plain pseudo-likelihood are generally distorted. However, these distortions can be amended almost completely by a correction factor of the pseudo-likelihood for MLE, MAP and EAP estimation. Moreover, unbiased weighted (pseudo-)likelihood estimation becomes feasible without facing the (often prohibitive) computational burden of weighted-likelihood estimation with the proper likelihood based on multivariate integration.

Mat-2 Observed-score reliability and its approximate index in paired-comparison Thurstonian IRT

Kensuke Okada, *University of Tokyo, Japan*
Kyosuke Bunji, *University of Tokyo*

Accumulating evidence suggests that forced-choice questionnaires are more resistant to response biases than the conventional Likert-scale ones. Accordingly, the Thurstonian item response theory (IRT) model has attracted much attention as an appropriate means of modeling paired-comparison forced-choice questionnaire. In the IRT framework, unlike linear models, reliability at the latent and manifest levels needs to be distinguished. In this study, in contrast to previous studies that have focused on the former, we explore the latter, that is, reliability at the scale of the observed variables. Following the existing single-stimulus IRT literature, we first define the exact reliability measure at the manifest level in Thurstonian IRT, which is the well-known proportion of true variance to observed variance. Then, we propose an easy-to-calculate index to approximate reliability, which is derived based on a Taylor series expansion. The proposed reliability measures are applicable both at the item-block level (item-block reliability) and at the questionnaire level (total reliability). The performance of the proposed index is next assessed through a simulation study. It is found that the values of the proposed Taylor series-based index tend to be closer

to true reliability than other conventional measures. Finally, an application to Big Five personality questionnaire data is presented.

Mat-3 Evaluation criteria for measurement invariance tests in Thurstonian IRT model

HyeSun Lee, *California State University Channel Islands, United States*
Weldon Z. Smith, *California State University Channel Islands*

We investigated whether commonly used cutoffs for fit indices can be utilized to assess model fit and measurement invariance for the Thurstonian item response theory (IRT) model. Regarding model fit, CFI $\geq$ .95 and RMSEA $\leq$ .06 from Hu and Bentler (1999) were employed. For measurement invariance tests, three evaluation criteria in absolute changes (Δ CFI $>$ .01, Δ $\hat{\gamma}$ $>$ .001, and Δ NCI $>$ .02) from Cheung and Rensvold (2002), one evaluation cutoff (Δ CFI $>$ .002) from Meade et al. (2008), and two RMSEA cutoffs (Δ RMSEA $\geq$ .015 from Chen [2007] and Δ RMSEA $\geq$ .010 from Rutkowski and Svetina [2017]) were employed to determine whether more constrained invariance models fit significantly worse than less constrained models. Power and Type I error rates were examined to evaluate performance of the six evaluation criteria under 32 manipulated conditions. The results showed that the two fit indices used for the evaluation of model fit performed well. Among six cutoffs for changes in model fit indices, only Δ CFI $>$.01 and Δ NCI $>$.02 detected metric non-invariance when the medium magnitude of non-invariance occurred and none of the cutoffs performed well to detect scalar non-invariance. Based on the generated sampling distributions of fit index differences, we suggested Δ CFI $>$.001 and Δ NCI $>$.004 for scalar non-invariance and Δ CFI $>$ .007 for metric non-invariance. Considering Type I error rate control and power, Δ CFI was recommended for measurement invariance tests for forced-choice format data.

Mat-4 Test and profile reliability of social and emotional learning assessment

Xiaohong Gao, *ACT, Inc., United States*
Sien Deng, *ACT, Inc.*

There is a growing consensus in education and workforce that social and emotional learning (SEL) skills are just as important as cognitive abilities for educational and workplace success. Different methods have been used in assessing SEL skills as each method has its merits and limitations. When multiple methods are used in an assessment, empirical evidences are needed to validate the uses of these methods including measurement errors and reliability. Generalizability theory provides a comprehensive framework for identifying, disentangling, and estimating multiple sources of measurement error associated with the assessment scores derived from multiple methods on multiple scales and for defining and estimating reliability accordingly. The purposes of the study were to (1) disentangle multiple sources of measurement error associated with items, scales, item formats, etc.; (2) estimate disattenuated correlations as well as correlated errors among the scales; (3) estimate reliability of scale scores as well as profile scores by and across the item formats; and (4) evaluate invariance of these estimates between gender groups. Both univariate and multivariate generalizability analyses were conducted. An item response theory (IRT) approach was also used to estimate reliability of different scores. Finally, profile reliability were estimated. Data from a comprehensive assessment designed to provide a holistic picture of individuals' SEL skills were used in the study. 1,768 test takers took the test online that include six scales assessed by three methods: Likert-type scales (SR), situational judgement tests (SJT), and forced choice (FC). The results of the study may be used for future assessment improvement.

Auditorium 2
Reliability in latent variable models

Au2-1 Reliability issues in high-stakes educational tests

Cees Glas, *University of Twente, Netherlands*

High-stakes tests and examinations often give rise to rather specific measurement problems. Though nowadays item response theory (IRT) has become the standard theoretical framework for educational measurement, in practice, number-correct scores are still prominent in the definition of standards and norms in educational measurement. Therefore, methods are developed for relating standards on the number-correct scale to standards on the latent IRT scale. The examinations considered here often consist of several topics giving rise to a between-items multidimensional IRT model. Next, the focus is on two related issues. The first issue is estimating the size of standard errors when equating older versions of a test to the current version. The second issue is estimating the global and local reliability of number-correct scores and the extra error variance introduced through number-correct scoring rather than using IRT proficiency estimates. It is shown that these issues can be solved in the framework of Bayes modal estimation of the IRT model, combined with maximum a posteriori (MAP) and

expected a posteriori (EAP) estimation for proficiency parameters. The examples that are given are derived from simulations studies carried out for linking the nation-wide tests at the end of primary education in the Netherlands, both over time, and between different test providers. Finally, there will be a short introduction to the public domain software used for the simulations and the actual equating procedure.

Au2-2 Acquiescence and attitude-achievement paradox in PISA 2012

Ricardo Primi, *Universidade São Francisco, Brazil*
Jonas Bertling, *Educational Testing Service*
Patrick Kyllonen, *Educational Testing Service*
Nelson Hauck-Filho, *Universidade São Francisco*
Felipe Valentini, *Universidade São Francisco*

International large-scale assessments like PISA make use of self-reports to asses socio and emotional characteristics. However, paradoxical results have been observed for cross-cultural comparisons. For instance, in PISA 2012, perseverance and openness were positively related with math achievement at the within-country level. However, when available scores from all countries were aggregated, the means on these constructs were negatively related with the average of math achievement. Hence, results from the country level analyses are counterintuitive and contradict the expected positive relationship between attitudes and achievement. We investigated whether acquiescence correction could reverse these untrustworthy country level patterns. To do so, we reanalyzed twelve PISA 2012 scales. From a selection of semantic opposite items, we computed an acquiescence index, and then used this index to re-center the scale scores, as this method partials out acquiescence variance from the scale scores. Previously to acquiescence control, eight out of 12 original scale scores had a negative correlation with achievement at the country level. All of them turned back to positive when we analyzed the acquiescence-controlled scores correlation with achievement at the country level. Acquiescence is a confounder that is negatively correlated with achievement, positively correlated with socioemotional characteristics, and differentially distributed across countries. We also discuss alternative ways to compute acquiescence indexes (like via anchoring vignettes) and model this bias using multidimensional IRT.

Au2-3 Assessing general factor reliability in exploratory bi-factor modelling.

Eduardo Garcia-Garzon, *Universidad Autónoma de Madrid, Spain*
Francisco J. Abad, *Universidad Autónoma de Madrid, Spain.*
Luis Garrido, *Pontificia Universidad Católica Madre y Maestra*

As the number of bi-factor modelling applications continues increasing, several researchers have questioned how to correctly approximate reliability estimation in such models. Although omega hierarchical has been characterised as a suitable estimator of general factor reliability, few studies have analysed its statistical behaviour under realistic conditions where bi-factor exploratory factor analysis is required (i.e., bi-factor models presenting cross-loadings and factors with low loadings). This study examines, for the first time, how the choice of the bi-factor exploratory factor analysis algorithm (i.e., bi-quartimin, bi-geomin, BIFAD and SLiD) affect the recovery of omega hierarchical. Results from a Monte-Carlo simulation evidenced that SLiD resulted in the most precise estimation of omega hierarchical. Additionally, bi-quartimin and bi-geomin systematically introduced positively biased estimation, especially if relevant cross-loadings were present. Lastly, BIFAD showed an unsatisfactory performance under all studied conditions. The re-analysis of the seven empirical datasets commonly found in the bi-factor literature confirmed the simulation results. Additionally, it was illustrated how the algorithm choice could lead to different judgments with regards to general factor reliability. To conclude, a comprehensive set of recommendations regarding how to measure general factor reliability in bi-factor modelling is provided.

Au2-4 Measuring sub-dimensions and composites

Mark Wilson, *University of California, Berkeley, United States*
Perman Gochyyev, *University of California, Berkeley*

Among psychometricians, there has been much interest in psychometric models that deal with multicomponent data. Such models are often used for modeling and scoring data from large-scale surveys and assessments with complex structures. Despite their long history in the factor analysis tradition, two models that recently been increasingly in the spotlight—bifactor models and second-order models. This renewed attention on these models and on what they have to offer is partly due to the developments in estimation techniques in recent years, which allow more efficiency in estimating complex models. However, this focus on estimation has obscured important differences between the meanings of the implied factor structures and interpretation of the factors in these two models. This has resulted in an unscrutinized acceptance

and use of the bifactor and second-order models. In this work, we propose an alternative solution to those offered by the bifactor and second-order models. This we refer to as the derived measures model—it allows clear interpretations for estimates from both the composite and dimension scores. The proposed model obtains dimension scores similar to the multidimensional item response model, and the composite score is derived from these dimensions by exploiting the correlation between dimensions as well as the dimension scores. Using Mplus software, we will use simulated data and an empirical dataset to demonstrate what the proposed model has to offer. We will discuss how one can obtain reliability indices and other relevant testing statistics usually obtained from item analysis.

Au2-5 Ability estimation accuracy under varying noneffortful responding types and rates

Joseph Rios, *University of Minnesota, United States*
Qinjun Wang, *University of Minnesota*

As group-based low-stakes accountability testing contexts grow in popularity (e.g., PISA), one major validity concern is that test-takers may noneffortfully respond to items, and thereby provide a score that underestimates their true ability. To date, much of the research conducted on noneffortful responses has been based on applied data, and thus, the biasing effects on ability estimation using traditional item response theory (IRT) models is not clearly understood. To address this concern, the objectives of this simulation study were to: (a) determine how various noneffortful factors interact to bias aggregate ability estimates, and (b) investigate the utility of the Effort-Moderated IRT (EM-IRT) model in mitigating these potential biases. Data were generated using the unidimensional two-parameter logistic (2PL) model, and the following four independent variables were manipulated resulting in 72 total conditions: (a) the relationship between true ability and noneffortful responding, (b) noneffortful responding type (random, location-based, and difficulty-based), (c) percentage of unmotivated simulees, and (d) within-unmotivated simulee noneffortful responding rates. Results demonstrated that aggregate-level ability under the 2PL model was underestimated by more than .20 standard deviations when noneffortful responding was related to ability and the overall percentage of noneffortful responses in the data matrix exceeded 10%; however, no significant differences between noneffortful responding types were observed. In contrast, the EM-IRT model was found to possess negligible biasing of aggregate-level estimates across all conditions when the overall percentage of noneffortful responses in the data matrix was below 35%, which is well within levels observed in operational settings.

Auditorium 3
Response times II

Au3-1 Detecting rapid guessing behaviors in testlet items

Po-Hsi Chen, *National Taiwan Normal University, Taiwan (R.O.C.)*
Chia-Ling Hsu, *The Education University of Hong Kong*
Kuan-Yu Jin, *University of Hong Kong*

Rapid guessing refers to quickly answering an item without thoughtful consideration during tests. However, rapid guessing is not a consistent status for a respondent; instead, a respondent can rapidly guess on any items due to non-motivation, speededness, and responding strategy, etc. By the popularization of computer-based testing, response time information can help to identify if a respondent responses an item thoughtfully or unreflectively. In addition, many assessments are administered by testlet items, especially in language testing, a reading passage is followed by several testing items, for example. In this study, a mixture testlet item response theory (IRT) model including item responses and response time is proposed to detect rapid guessing behaviors in testlet items. We conducted a simulation study to evaluate the parameter recovery of the new model by using the Markov Chain Monte Carlo estimation in JAGS. Results showed that the parameters in the new model can be recovered fairly well. Ignoring rapid guessing behaviors by fitting a standard testlet IRT model would overestimate item difficulties and underestimate item intensities seriously; and overestimate the precision of respondent latent estimates and underestimate respondent latent speed parameters. Finally, a computer-based language test was analyzed as an example for demonstration.

Au3-2 Joint modeling of responses and response time for subdomain diagnosis

Hong Jiao, *University of Maryland, College Park, United States*
Weimeng Wang, *University of Maryland, College Park*

A joint modeling approach of response and response times is often used to understand the trade-off between accuracy and speed (e.g., van der Linden, 2007; van der Linden, Klein Entink, & Fox, 2010). Recently, researchers (Klein Entink, Kuhn, Hornke, & Fox, 2009; Minchen &

de la Torre, 2016; Zhan, Jiao, & Liao, 2017) have extended joint modeling of responses and response times for cognitive diagnosis. Most often a cognitive diagnostic model is used for item response modeling. However, such diagnostic models are limited in the number of attributes that can be included in modeling for cognitive diagnosis. Further, the diagnosis at the domain level cannot provide actionable information. To facilitate subsequent learning, the subdomain level information can be directly utilized to come up with actionable learning plans. This study proposes a joint modeling approach of responses and response time for cognitive diagnosis at the subdomain level by including a higher-order structure to model the relation between the probability of an attribute mastery to the higher-order latent continuous trait. More specifically, a third-order model (Rijmen, Jeon, von Davier, & Rabe-Hesketh, 2014) assuming one highest-order continuous latent trait and a second-order model assuming multiple continuous latent domain traits (Wang, Jiao, & Sun, 2019) underlies the probability of mastery of a subdomain respectively. Both simulation study and real data analysis will be conducted to investigate the model parameter recovery and the model performance under study conditions. It is expected the higher-order structure will increase flexibility in estimating a larger number of subdomain abilities.

Au3-3 The differentiation of three types of conditional dependence

Zhaojun Li, *Ohio State University, United States*
Paul De Boeck, *Ohio State University*
Zhujing Dan, *Ohio State University*

Conditional dependence is defined as the dependence between items that cannot be fully explained by latent variables and their correlation(s). We focus on parallel data (e.g., response times and responses for same items) with a multidimensional model in which each type of data has a latent variable. There are three possible types of conditional dependence. Assume item Y1 loads on factor F1 and item Y2 loads on factor F2, the first type of conditional dependence derives from the effect of the expected values of Y1 on Y2. In this case, the dependence can be explained through a cross loading of Y2 on factor F1. The second type is based on the effect of the observed values of Y1 on Y2, which implies a direct effect of Y1 on Y2. The third type comes from the effect of the residual of Y1 on Y2, in which case the dependence is captured through a residual correlation between Y1 and Y2. Relevant derivations show that the three types of dependence will result in different correlations among variables, indicating the possibility to differentiate the types of dependence. An empirical study was first conducted to compare models with different types of conditional dependence. Even though the model with direct effects has the best goodness of fit and is theoretically reasonable, its model-fit criterion values are close to those of the model with residual correlations. We have also conducted a simulation study to find under which conditions the three types of dependence can be differentiated.

Au3-4 Bivariate change-point analysis with response time and item responses

Daniella Rebouças, *University of Notre Dame, United States*
Ying Cheng, *University of Notre Dame*

With the widespread of computerized assessments in psychology and education, item-level response times of tests and surveys have become increasingly available. Information derived from analyzing response time data may help survey or test administrators define appropriate assessment length, detect abnormal responses due to lack of motivation, fatigue, or speeded behavior, and produce more accurate estimates of the latent trait by removing speeded item responses. Previous change-point analysis methods have relied mainly on item responses (Shao, Li & Ying, 2016). Although appropriate power was found in identifying speeded responses under most simulation conditions, the change-point estimates were on average biased and highly variable across replications. In this study, we propose the application of change-point analysis through a hierarchical framework for responses and response times (Molenaar, Tuerlinckx & van der Maas, 2015). By leveraging information from both item responses and response times, we can obtain higher power in detecting speeded responses, especially for shorter tests, and more accurate estimates of the point of change. Preliminary results and current challenges are discussed.

Spotlight Speakers

Salón Fresno
Spotlight Speaker: Hyeon-Ah Kang

Frs-1 Detecting item parameter drift online using response and response times

Hyeon-Ah Kang, *University of Texas at Austin, United States*

When tests are administered continuously or at frequent time intervals, some items may become known to prospective examinees or may undergo changes in the statistical properties. The purpose of this study is to present a sequential monitoring procedure that regularly checks on the quality of items across the span of time the items are in operation. The procedure is based on a sequential generalized likelihood ratio test, which evaluates the likelihood of the currently estimated item parameters against the likelihood of the pre-calibrated item parameter values. The test is designed to integrate information from the response and response time data, and detect a change-point as soon as an item exhibits parameter drift within the hierarchical framework (van der Linden, 2007). For estimating the item parameters, we perform continuous online calibration based on moving samples. The suggested procedure provides a powerful automated tool for maintaining the quality of an item pool by conducting a series of hypothesis testing on the individual items under the parametric model that capitalizes on two sources of information. The effectiveness of the proposed method is evaluated through extensive simulation studies and an application to a large-scale high-stakes computerized adaptive test. All evaluations are made in comparison with the existing statistical quality control procedure (e.g., Veerkamp & Glas, 2000).

Aula Magna

Spotlight Speaker: Adrian Quintero

Mag-1 Selecting the number of factors in Bayesian factor analysis

Adrian Quintero, *Colombian Institute for Educational Evaluation, Colombia*
Geert Verbeke, *University of Leuven*
Emmanuel Lesaffre, *University of Leuven*

When implementing factor analysis, the selection of the number of factors is challenging in both frequentist and Bayesian approaches. The validity of the likelihood ratio test in the frequentist setting strongly depends on the assumption that the factor loadings matrix is of full rank. However, such is not the case when fitting models with more latent components than the true (unknown) number of underlying factors. This invalidates the regularity conditions necessary for LRT, and the method retains too many factors in practice. Information criteria such as AIC and BIC may also be affected by the regularity conditions. On the other hand, conventional Bayesian methods present two serious drawbacks. Firstly, implementation of the procedures is highly computationally demanding, and secondly, the ordering of the outcomes influences the results since a lower triangular structure is generally assumed for the factor loadings matrix. Therefore, we propose a Bayesian method without imposing such a structure in order to overcome ordering dependence. Our approach considers a relatively large number of factors and includes auxiliary multiplicative parameters which may render null the unnecessary columns in the factor loadings matrix. The underlying dimensionality is then inferred based on the number of non-null columns in the factor loadings matrix. We show that implementation of our approach is simple via an efficient Gibbs algorithm. The advantages of the method in selecting the correct dimensionality are illustrated via simulations and using real data sets from ICFES, the Colombian Institute for Educational Evaluation.

Career Award for Lifetime Achievement: Susan Embretson

Frs-1 Modeling cognitive processes, skills and strategies in item responses: implications for test and item design

Susan Embretson, *Georgia Institute of Technology, Atlanta, Georgia, USA*

Interpretations of test scores typically involves the assumption that examinees are applying the same cognitive processes, skills and strategies in their item responses. If so, then test scores indicate examinee's standing on a single latent trait. Results from several studies using mixture and explanatory item response models will be presented to show that this assumption is often not met. The implications of these results for test scoring, interpretations and external correlates, as well as for item and test design will be discussed.

Parallel Sessions, Friday Afternoon

Salón Fresno

Symposium: Recent developments in school value-added modeling

Frs-1 Cohort varying, temporally dynamic, value-added models

Ernesto San Martin, *Pontificia Universidad Catolica de Chile, Chile*
Garritt Page, *Brigham Young University*
Joniada Milla, *Saint Mary's University*
Edgar Valencia, *Pontificia Universidad Católica de Chile*

We aim to estimate school value-added dynamically in time. The motivation is to determine the persistence of the school effectiveness across the time when the school treats different cohorts. We propose two methods of incorporating temporal dependence in value-added models. The first models the random effects that are commonly present in value-added models with an AR(1) model. The second correlations value-added parameters by introducing performance from the previous cohort: by so doing, we intend to incorporate how a shock of information eventually modifies the school effect. We show through simulation that ignoring temporal dependence when it exists results in diminished efficiency in value-added estimation and that incorporating (even when temporal dependence is weak) results in improved estimation. Finally, we consider two cohorts from the SIMCE data and show utility in estimating value-added.

Frs-2 Exploring complete school effectiveness via quantile value-added

Garritt Page, *Brigham Young University, United States*
Ernesto San Martin, *Pontificia Universidad Católica de Chile*
Javiera Orellan, *Pontificia Universidad Católica de Chile*
Jorge Gonzalez, *Pontificia Universidad Católica de Chile*

In education studies value-added is by and large defined in terms of a test-score distribution mean. Therefore, all but a particular summary of the test score distribution is ignored. Developing a value-added definition that incorporates the entire conditional distribution of student's scores given school effects and control variables would produce a more complete picture of a school's effectiveness and as a result provide more accurate information that could better guide policy decisions. Motivated in part by the current debate surrounding the recent proposal of eliminating co-pay institutions as part of Chile's education reform, we provide a new definition of value-added that is based on the quantiles of the conditional test score distribution. Further, we show that the quantile based value-added can be estimated within a quantile mixed model regression framework. We apply the methodology to Chilean standardized test data and explore how information garnered facilitates school effectiveness comparisons between public schools and those that are subsidized with and without co-pay.

Frs-3 Augmenting multidimensional value added with non-cognitive skills

Joniada Milla, *Saint Mary's University, Canada*
Ernesto San Martin, *Pontificia Universidad Catolica de Chile*

Schools and teachers contribute to their students' human capital by developing their skills in multiple domains. The literature to date has focused only on the use of students' test scores to evaluate the performance of schools or teachers, upon which high-stakes decision making is based. Examples include pay-for-performance programs, hiring and firing decisions for the teachers. School closures have also depended on such evaluations. However, the conventional approach lacks a complete representation of the school or teacher contribution, and at the same time induces incentives for schools or teachers that can be detrimental to the social capital of the students in the long-run. As an example consider shifting time allocation from extracurricular and social activities to preparing for the test whose scores are used in evaluations. In this paper we augment the Multidimensional Value Added (MVA) indicators (proposed in Milla, San Martin and Van Bellegem, 2016, "Higher Education Value Added Using Multiple Outcomes," Journal of Educational Measurement, 53(3), p. 368-400.) with non-cognitive student skills. We then analyze whether incorporating these non-testable outcomes along with other cognitive skill measures delivers a significantly different result in school ranking. We find that it does. We analyze the consequences of the conventional ranking within the context of a group performance-pay program that has been implemented in Chile since 1996. We show that disadvantaged schools in particular miss out on the positive effects of program that can push these schools out of the non-winners' vicious cycle.

Frs-4 How stable are value-added indicators across time? an empirical analysis

Edgar Valencia, *Pontificia Universidad Católica de Chile, Chile*

While the general opinion sustains the idea that school effect differences should be large, stable over time, and consistent across subject areas, empirical evidence does not support this claim (Marks, 2015). Correlation coefficients of value-added school effects across cohorts are often small and decrease over time. In this panel, we present correlations of school effects across time and subject-matters on a panel of student achievement data from Chile using the time-dependent value-added model (San Martín, Page and Torres, 2019) which explicitly accounts for the relationship between school effects between cohorts. We compare the results with correlation coefficients obtained using traditional VAM approaches. Based on a definition of school effect and value-added, we discuss the observed differences between tdVA and tra-

ditional approaches and the meaning of small the correlations of school-effects over time along with the role of value-added in educational evaluation and policy making.

Aula Magna

Symposium: Recent advances in assessing small group collaborations

Mag-1 Estimating an individual's contribution to small group performance

Patrick Kyllonen, *Educational Testing Service, United States*
Jiangang Hao, *Educational Testing Service*
Michelle Martin-Raugh, *Educational Testing Service*

This research involves developing and evaluating approaches to quantifying individual contributions in small group tasks, which we call collaborative problem solving (CPS) skill. We define CPS skill across a range of tasks including (a) problem solving, (b) decision-making/hidden profile, and (c) negotiation. The main questions are (a) can we measure an individual's contribution to collective performance? (b) can we do so by evaluating contributions to process variables that are related to collective outcomes? (c) are there background determinants (e.g., ability, personality) of individuals' contributions? (d) how generalizable are individual contributions to collective outcomes across task variants and types? and (e) are there differences in two- versus four-person teams and online versus face-to-face contexts? We review preliminary findings from a set of 2-party online negotiation tasks. Participants were given a set of tasks comprising distributive (zero-sum), integrative (trade-off), and compatible (aligned) issues, so that each party received different payoffs (points) depending on how issues were settled. Participants communicated through a chat window. We found that first, there was a correlation between number of points accrued on two distinct tasks at both the individual and joint level, indicating generalizability of negotiation skill. Second, a principal components analysis of annotated chat turns suggested four tactic types—empathic, assertive, felicitation, and impasse. Third, we found that empathic tactic use was associated with joint negotiation success. We discuss challenges in this kind of work, and prospects for developing CPS skill assessments generally.

Mag-2 Validating measures of small group collaboration: a process perspective

Nafisa Awwal, *University of Melbourne, Australia*

Recent employment trends indicate that there has been a shift in workplace requirements for more non-routine and interpersonal skill. Workplaces demand employees to be equipped with complex skills such as problem solving, creativity, critical thinking, communication, collaboration and teamwork. Collaboration has gained extensive attention over the past decade in both education and workforce as one of the key skills to become productive citizens in 21st century. There is a need to research how to measure and assess such complex skills as it is believed that the knowledge and skills exposed by traditional education are no longer seen as adequate preparation for success in life. Some initiatives have been undertaken to assess how well people collaborate with one another. However, the group measures in collaborative problem solving context have not been explored adequately. Researchers have argued for importance of these measures as an increasing number of tasks are now conducted in teams. In particular, they have indicated that viewing the group as the unit of analysis could produce qualitatively different conclusions about collaboration. However, most research on collaborative problem solving has so far focused on individuals instead of the group as a whole. Thus, this presentation will explore these measures of collaborative problem solving and examine the validity of their measures.

Mag-3 Further findings from modeling data in collaborative assessments

Mark Wilson, *University of California, Berkeley, United States*
Perman Gochyyev, *University of California, Berkeley*

This presentation investigates the assessment of cognitive skills through collaborative tasks, using field test results from assessments based on the "ICT Literacy—Learning in digital networks" learning progression, from the Assessment and Teaching of 21st Century Skills (ATC21S) project. After a brief description of the development of the learning progression, as well as examples and the logic behind the instrument construction, the presentation will focus on the use of a multilevel item response model to find estimates for randomly-assigned small groups of students using the demonstration digital environment, as well as individual estimates for each person. The modeling employed unidimensional and multidimensional item response models. The results indicated that, based on this data set, the models that take grouping into consideration in both the unidimensional and the multidimensional analyses fit better. In earlier research, the group-level variances were substantially higher than the individual-level variances. This has several important consequences.

This presentation summarizes new results concerning this issue. Implications are discussed in the results and conclusions.

Mag-4 Designing and modeling "new" item types for assessments involving small groups

Peter Halpin, *University of North Carolina at Chapel Hill, United States*

Research in cooperative learning has addressed how to design tasks that structure small group interactions among students. This presentation considers how these task designs can be used to "collaborify" conventional test items. The basic idea of these tasks designs is to systematically induce statistical dependence among the responses of individuals working together in a small group, such that the probability of the group arriving at a correct response is greatly increased when the members work together in the intended way. One common approach is to induce dependence by distributing resources (i.e., item content) in such a way that no single member can provide a correct response working individually, which is referred to as a jigsaw or hidden profile design. Three additional item designs are proposed in this presentation. It is argued that this approach to collaborifying conventional test content has three main of advantages. First, it bridges research in cooperative learning, social and organizational psychology, and psychometrics. Second, it provides a high level of specificity about the collaborative skills to be measured. Third, working from the template of conventional test items lends itself to psychometric modeling. In particular, it is shown how the new item types can be modeled using the social-combination item response theory (SC-IRT) model proposed by Halpin and Bergner. Item designs and statistical results are illustrated with an empirical example in which dyads work together on a twelfth-grade-level mathematics assessment.

Sala Colorada
Models for dynamics and learning

Col-1 IRT models for learning with item-specific learning parameters

Albert Yu, *University of Illinois at Urbana Champaign, United States*
Jeff Douglas, *University of Illinois at Urbana-Champaign*

With more and more online learning opportunities for an ever expanding number of domains, items or groups of items may be tools for learning as well as assessments, especially when they are coupled with interventions. This can be seen on many popular educational sites like IXL and Khan Academy. Accordingly, there is a psychometric advantage of not only providing a chance to assess learning as a student practices, but also foster learning if we can identify the items or item clusters that have the greatest instructional benefit. Models for item-level and item bundle-level learning are proposed ranging from simple models associating the item with a constant increase in latent ability, to more general models that consider learning as dependent on ability and position of the item in an assessment. These models may be fitted by an MCMC algorithm simultaneous with the IRT measurement model parameters of common IRT models like the two-parameter logistic. Several models are fitted and compared on a dataset obtained through an intervention to train spatial rotation ability. The DIC was used to select the best fitting among several competing models and chose one that included a learning constant for each item-bundle scaled by a measure of agreement between latent ability and difficulty, and including exponentially decaying growth in exam length. Simulations were conducted both to assess accuracy of recovering item parameters, and to study how learning can be made more efficient by selecting according to item level learning parameters.

Col-2 Four-dimensionalism and the measurement of evolving psychological attributes over time

Tyler Matta, *Pearson, United States*

Measuring change in psychological attributes that are sensitive to learning environments has been a major focus of psychometricians working in education. Such work relies on measurement invariance techniques to assess the psychometric equivalence of a construct across time (Widaman, Ferrer, & Conger, 2010). The use of such techniques assumes that the psychological attribute being measured is unchanging. Recent work in cognitive science suggests, however, the psychological attribute being measured is likely to evolve over a specified temporal region. While methods have been proposed to overcome such issues (Martineau 2004), the heart of the issue is ontological, not epistemological. There is a great deal of thinking to support philosophically realist commitments regarding psychological attributes being measured (e.g., Borsboom, 2005; Maul 2013). However, little emphasis has focused on the measurement of psychological attributes over time. This paper applies the ontological position of four-dimensionalism (Sider, 1997) to psychological attributes. In this context, four-dimensionalism describes a psychological attribute that exists in time as having temporal parts in the various subregions of the total region of time it occupies. Such a position enables

the psychological attribute to extend across time as a four-dimensional causal series of three-dimensional "time-slices." That is, the psychological attribute would remain numerically identical over the total region of time while allowing individual time-slices of it to differ from each other. The paper applies four-dimensionlism to a strand of elementary mathematical knowledge, and concludes with considerations for instrument design and administration, as well as calibration and measurement invariance using empirical data.

Col-3 Detecting item effects with cognitive diagnostic model for learning trajectories

Anqi Li, *University of Illinois at Urbana-Champaign, United States*
Yanglei Song, *University of Illinois at Urbana-Champaign*
Georgios Fellouris, *University of Illinois at Urbana-Champaign*
Steven Culpepper, *University of Illinois at Urbana-Champaign*

The increasing popularity of online and computer-based learning provides new opportunities to track the student learning process. While recent research has made efforts to detect and model students' learning trajectories, additional strategies and tools are needed to accelerate students' learning process. Therefore, we focus on differentiating effects of various learning tools. In particular, different item types may offer varying benefits to students. In adaptive learning settings item types that are more likely to promote learning should be given priority. In this study, we proposed a cognitive diagnosis model to measure the effects of item types while tracking students' learning trajectories. Specifically, the proposed model focuses on training students on a certain skill with different item types across time. By fitting this learning model, the effects of different item types could be obtained, which will then provide instructional information on item assignment rules to accelerate students' learning process. A Bayesian approach was developed for parameter estimation, and parameter recovery of the proposed model was evaluated through a Monte Carlo simulation study.

Col-4 A nonlinear dynamic latent class structural equation model

Augustin Kelava, *University of Tuebingen, Germany*
Holger Brandt, *University of Kansas*

In this talk, we propose a nonlinear dynamic latent class structural equation model (NDLC-SEM; Kelava & Brandt, in press). It can be used to examine intra-individual processes of observed or latent variables. These processes are decomposed into parts which include individual- and time-specific components. Unobserved heterogeneity of the intra-individual processes are modeled via a latent Markov process that can be predicted by individual-specific and time-specific variables as random effects. We discuss examples of sub-models which are special cases of the more general NDLC-SEM framework. Furthermore, we provide empirical examples and illustrate how to estimate this model in a Bayesian framework. Finally, we discuss essential properties of the proposed framework, give recommendations for applications, and highlight some general problems in the estimation of parameters in comprehensive frameworks for intensive longitudinal data.

Col-5 A unified framework of longitudinal models to examine reciprocal relations

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

Inferring reciprocal effects or causality between variables is a central aim of behavioral and psychological research. To address reciprocal effects, a variety of longitudinal models that include cross-lagged relationships have been proposed in different contexts and disciplines. However, the relationships between these cross-lagged models have not been systematically discussed in the literature. This lack of insight makes it difficult for researchers to select an appropriate model when analyzing longitudinal data, and some researchers do not even think about alternative cross-lagged models. The present research provides a unified framework that clarifies the conceptual and mathematical similarities and differences between these models. The unified framework shows that existing longitudinal models can be effectively classified based on whether the model posits unique factors and/or dynamic residuals, and what types of common factors are used to model changes. The latter is essential to understand how cross-lagged parameters are interpreted. We also present an example using empirical data to demonstrate that there is great risk of drawing different conclusions depending on the cross-lagged parameters used in different models.

Sala Matte

Causal inference and mediation II

Mat-1 Causal effects based on Poisson regression models

Axel Mayer, *RWTH Aachen University, Germany*
Christoph Kiefer, *RWTH Aachen University*

For estimating the effects of a treatment or an intervention on a count variable, we can for example use a Poisson regression model. In order to then compute the average effectiveness of the treatment based on the regression coefficients, various effects can be considered such as unconditional treatment effects and marginal or average treamtent effects. The effects can either be defined as ratios of (conditional) expectations or as differences between (conditional) expectations, and they can either be on the on the original count scale or on the log scale. We consider these effects from a causal inference perspective, provide clear definitions of the effects, and identify conditions under which we can estimate the various types of causal effects from empirical studies. In particular, we discuss which of the effects are collapsible and can be estimated in randomized experiments without further adjustment for covariates. For illustration, we use data from a randomized controlled trial on problem drinking to study the effects of a mobile messaging interventions on the weekly sum of standard drinks. In addition, we present a new way to analytically compute average effects based on Poisson regression models and stochastic covariates and develop new formulas to obtain standard errors for the average effect. In a simulation study, we evaluate the statistical performance of our new estimator and compare it to the traditional approach.

Mat-2 Sensitivity analysis in longitudinal mediation model

Davood Tofighi, *University of New Mexico, United States*
Yu-Yu Hsiao, *University of New Mexico*
Eric S. Kruger, *University of New Mexico*
David P. MacKinnon, *Arizona State University*
M. Lee Van Horn, *University of New Mexico*
Katie Witkiewitz, *University of New Mexico*

Multiple mediation models with two mediators have become more common. For a multiple mediation model, like any another mediation model, even with random treatment assignment, a critical but untestable assumption for valid and unbiased estimates of the indirect effects is that there should be no omitted variable that confounds indirect effects. One way to address this untestable assumption is to conduct sensitivity analysis to assess whether the inference about an indirect effect would change under varying degrees of confounding bias. We developed a sensitivity analysis technique for a multiple mediator model with two mediators. We compute the biasing effect of confounding on point and confidence interval estimates of the indirect effects in a structural equation modeling framework. We illustrate sensitivity plots to visually summarize the effects of confounding on each indirect effect. We present an empirical example to illustrate the application of the sensitivity analysis along with computer code.

Mat-3 Maximum likelihood analysis of mediation models with treatment-mediator interaction

Kai Wang, *University of Iowa, United States*

This research concerns a mediation model where the mediator model is linear and the outcome model is also linear but with a treatment-mediator interaction term. Assuming the treatment is randomly assigned, parameters in this model are shown to be partially identifiable. Under the normality assumption on the residual of the mediator and the residual of the outcome, explicit full information maximum likelihood estimates of model parameters are introduced assuming the correlation between the error for the mediator and the error for the outcome is known while the variances of these two errors are not. A consistent variance-covariance matrix of these estimates is derived. Currently the coefficients are estimated using the iterative feasible generalized least squares (IFGLS) method that is originally developed for seemingly unrelated regressions (SURs). We argue that the standard errors of the IFGLS estimates are inconsistent as this model is not a system of SURs. Our results are demonstrated by simulation studies and an empirical study.

Mat-4 Two-Step BART: Estimate average treatment effects when treatment is latent

Jiaqing Zhang, *Columbia University, United States*
Rui Lu, *Columbia University*
Jiaxi Yang, *Columbia University*

The motivating aim of this study is to estimate unbiased treatment effect when the treatment is latent in the potential-outcomes framework. The existing methods, including 1-step approach (Kang & Schafer, 2010) and 3-step approach (Schuler et al., 2014), are not free from restrictions of the functional forms and dimensionality. We propose a novel 2-step approach to remove these limitations by combining latent class model with Bayesian Additive Regression Tree (BART). It is proved to be much more stable, consistent and efficient than all existing approaches. In the first step, all covariates and indicators for latent classes are used to build the measurement model. We exclude the distal outcome to preserve the temporal ordering of the causal relationship. In the second step, we use BART to address confounding issue. BART flexibly fits all response surfaces and finds interactions and

nonlinearity terms automatically. BART, which only uses covariates to predict response surfaces, has shown to outperform propensity score approaches. All approaches are applied to 100 simulated datasets with different response surfaces and levels of confounding effect. We found that the new approach has overall the best performance. It has the smallest average percentage of bias, standard error, and root mean square error. As the confounding effect becomes larger, all statistical criteria show a general growth trend, among which the 2-step approach with BART has the smallest amount of increase.

Mat-5 Random forests versus matching methods for estimating heterogeneous treatment effects

Jee-Seon Kim, *University of Wisconsin - Madison, United States*
Youmi Suk, *University of Wisconsin - Madison*

This study investigates the properties of the random forest algorithm for the estimation of heterogeneous treatment effects and compares them to finite mixture matching strategies for causal inference with multilevel observational data. Latent-class multilevel models have been implemented to account for heterogeneous selection or outcome mechanisms in the evaluation of treatment effects in clustered data. Alternatively, machine learning algorithms such as random forests have also been used for estimating treatment effects without distributional assumptions. This study examines and compares the procedures and implications of random forests and multilevel matching methods in terms of removing selection bias and providing a proper estimation of treatment effects when subpopulations may have distinctive treatment effects and the data are clustered. Based on analytic comparisons and simulation studies, the strengths and limitations of the two different methods will be summarized with recommendations and guidelines for practice. An ensemble learning method of random forests and multilevel modeling of matching will be applied to TIMSS data to demonstrate analysis steps, interpret findings, and compare the similarities and differences of results based on the two competing methods.

Auditorium 2
Measurement invariance and DIF III

Au2-1 Multidimensional DIF, part A: A theoretical analysis of fixed-effects DIF

Edward Ip, *Wake Forest School of Medicine, United States*
Terry Ackerman, *University of Iowa*
Tyler Strachan, *University of North Carolina at Greensboro*
Jake Cho, *University of North Carolina at Greensboro*

Differential item functioning (DIF) is a relatively well-defined concept in unidimensional IRT. However, the meaning of DIF is less clear when multidimensionality exists. One way to think about DIF in higher dimension is to set one of the dimensions as a target and conduct DIF analysis on the targeted dimension. For example, in a 2D multidimensional IRT (MIRT), if theta1, which represents the latent construct on the first dimension, is set as target, then it is possible to study the effect of applying unidimensional IRT DIF detecting procedure when DIF only arises in the second dimension (theta2) but not the first (theta1). Using 2D MIRT as illustration, we first provide an overview of how DIF arises in the multidimensional latent space. We will then focus on biases and misidentifications of DIF/non DIF when one applies a unidimensional IRT (UIRT) model for assessing DIF in multidimensional latent space under different scenarios in which DIF arises. DIF arising from differences in the latent distributions of the focal and reference groups will be discussed in a separate abstract (Ackerman et al.). For this presentation, fixed-effects DIF due to a small number of items functioning differently across groups is discussed. Our first set of analysis makes use of theoretical tools, which are related to the composite direction of the latent abilities, sheds light onto how false positive and false negative DIF could occur. The second set of analysis offers data examples to further illustrate the different scenarios as delineated by the theoretical analysis.

Au2-2 Multidimensional DIF, part B: Examining two-dimensional DIF using projective IRT modeling

Terry Ackerman, *University of Iowa, United States*
Edward Ip, *Wake Forest School of Medicine*
Ye Cheryl Ma, *University of Iowa*
Jinmin Chung, *University of Iowa*

Several researchers have shown that multidimensionality is one of the main causes of differential item functioning, DIF, (Ackerman, 1992, Shealy & Stout, 1993). These researchers have demonstrated that when items are measuring irrelevant dimensions and groups of examinees differ in their latent ability distributions on these invalid dimensions the potential for DIF exists. In such cases, spurious DIF might be detected using standard unidimensional procedure. Ip (2010) and Ip & Chen (2012) developed an IRT model called the Projective IRT, (PIRT) model. The motivation of this model was to eliminate

unwanted dimensionality in test response data. The assumption was that in a multidimensional latent ability space, the $\theta 1$-dimension was the valid skill of interest and thus they derived a 2PL unidimensional dependent model in which the nuisance dimensions were integrated out. In this research we examine if DIF, created from a two-dimensional perspective, can be eliminated using the PIRT transformation. This will be a simulation study in which we will examine how relevant factors (sample size, correlation, angular composite of DIF item, and Ref-Foc mean ability differences) affect DIF analyses using the PIRT model. Two-dimensional calibration will be conducted using flexMIRT. Then the two-dimensional item parameter estimates will be transformed using PIRT formulation. The amount of DIF will then be computed using Raju's ICC area difference method and Lord's Chi-square approach. For comparison purposes the nonparametric SIBTEST approach will also be conducted.

Au2-3 Detection of differential item functioning under small sample size conditions

Chansoon (Danielle) Lee, *National Council of State Boards of Nursing, United States*
David Magis, *University of Liège*
Doyoung Kim, *National Council of State Boards of Nursing*
Sonya Sedivy, *University of Wisconsin – Madison*

Differential item functioning (DIF) methods have been extensively studied for use with large groups of test takers. These studies have often found that most DIF methods have suitable power and Type I error rates when sample sizes are sufficiently large. However, in practice, testing organizations may be faced with small sample sizes and the need to effectively perform DIF analyses under these conditions persists. The purpose of this study is to investigate the performance of commonly used as well as newly proposed DIF methods under the condition of extremely small sample sizes (e.g. 20 people in the focal group). A simulation study was conducted comparing four DIF methods under five factors: the focal and reference group sizes, test length, item impact, proportion of DIF items, and DIF size. The performance of the DIF methods was evaluated using Type I error rate and power. The simulation study highlights which methods were found to be reliable for small groups and which factors have a meaningful impact on detecting DIF items. The simulation results showed that there was a baseline difference among the DIF methods, depending on the group size, the test length, or item impact. In the presence of DIF, both the Type I error and power rates were best explained by the group size, test length, and DIF size. The presentation slides include comprehensive simulation results and the analysis of two case studies. The research findings will offer useful practical guidelines for successful implementation of small sample DIF methods.

Au2-4 A multi-dimensional approach to lack of invariance in measurement invariance

Hye-Eun Seok, *Ewha Womans University , South Korea*

In the field of psychology, the multiple-group analysis is frequently used to test measurement invariance. The determination of the measurement invariance using multiple-group analysis is generally judged using the χ^2 difference test and the alternative fit index. In addition to application in psychology area, various simulation studies have been conducted in the field of psychometrics for the purpose of evaluating the performance of the fit index for the past 15 years in terms of the lack of invariance (LOI). In previous studies, there are two methods for setting various conditions of LOI in the metric invariance. One is to control the factor loading size of the corresponding items between groups, and the other is to combine the various attribute levels of the item. Metric invariance is verified by equally constraining factor loadings across groups. Equally constraining factor loadings among groups imply to identify the hypothesis that the corresponding factor loading matrix is the same. The multiple-group analysis is an expanded analytical model of the factor model developed in the multivariate analysis framework. For this reason, it is necessary to explore the various levels of LOI in the structural dimension of the factor loading pattern. The purpose of this study is to systematically design various conditions that induce LOI with the viewpoint of multivariate analysis. And we identify the limitations that can be overlooked from the traditional approach and ensure the validity of a multidimensional factor loading pattern between groups.

Au2-5 A relation between multidimensionality and uniform DIF

Saemi Park, *Ohio State University, United States*
Paul De Boeck, *Ohio State University*

This study explores a relation between multidimensionality and uniform DIF. We revisit multidimensional model of DIF (MMD) proposed by Shealy and Stout to elaborate the way uniform DIF and multidimensionality are related and to assist to foresee uniform DIF. Most of DIF detection methods compare item parameters across groups to identify DIF, assuming group membership is a third

factor that accounts for how individuals respond to an item. However, the group membership per se is an observable variable that classifies individuals into groups and manifests between-group differences of item parameter estimates. The approach ignores what the fundamental reason is and, in fact, DIF occurs because unwanted latent trait is measured which is associated with the group membership. MMD explicates how an additional latent trait plays a role in inducing uniform DIF by considering three factors: a group mean difference in the primary dimension, a group mean difference in the additional dimension, and the correlation between two dimensions. An interrelation among the three factors defines a function termed DIF potential. We investigate in what systematic way DIF potential and item discrimination affect three properties of uniform DIF- occurrence, magnitude, and direction. It is found that an interaction between DIF potential and the item discrimination of primary dimension can cause a false positive of unidimensional items, that null DIF potential suppresses uniform DIF of multidimensional items, and that DIF potential is a more dominant factor that the item discriminations of both dimensions to predict the three properties of uniform DIF.

Auditorium 3
Computer-based testing II

Au3-1 Multi-stage testing of mathematical competence in NEPS

Claus Carstensen, *University of Bamberg, Germany*
Timo Gnambs, *Leibniz Insitute for Educational Trajectories, Bamberg, Germany*

Within the National Educational Panel Study (NEPS) in Germany different competences are measured coherently across the life span and most of the competence tests are scaled using item response theory models (IRT). In this paper results for a Multi-Stage-test (MST) for mathematical competence which was administered to three German populations, 1) young adults, 2) representative adults and 3) tertiary students in are presented. After introducing the mathematics framework, design and construction of the MST are illustrated. Results of quality control and performances of the three populations will be presented. Overall, the test administration worked out and results can be meaningfully be analysed. The data can be made public to the scientific community as this is one of the important goals of the NEPS. Interestingly, the reliability of the MST wais not higher than it may have been with a linear test administration format, as was expected. This may be due to a higher amount of missing responses observed in the MST administration. The test targeting met the intermediate and lower levels of competence very well, for higher levels of competence the number of items was rather limited. So far, most of the competence tests in the NEPS used a linear design with single booklets for 30 minutes of working time. In this MST design, the respondents took more time for each item and thus completed less items as usually in our linear testing designs, what might explain the lower reliability of this MST compared to our usual linear tests.

Au3-2 An automated method to detect enemy items using NLP approach

Haruhiko Mitsunaga, *Nagoya University, Japan*

In developing a practical approach to constructing computer adaptive testing (CAT) or multistage testing (MST) system, item selection rule is one of the most important issues. To make CAT system more effective and practical, it is important that the context of test items of the CAT system does not contain clues to the answer of other test items in the same test. We call these pairs of test items 'enemy items', and claim that we monitor which pair of items can potentially be the enemy items in an item pool. In particular, we should avoid administrating enemy items in a vocabulary test. However, if we have multiple items in the item pool, the number of pairs exceeds as many as we can examine and record manually. In this study, an automated enemy item detecting method is presented, which uses various kinds of word embedding techniques, enabling us to compare similarity of words or phrases under the assumption of mapping words on the vector space in natural language processing (NLP), and its effectiveness is probed. Two vocabulary test forms (English and Japanese) which contain enemy items were used, as well as a benchmark form. Standardized item difficulty and qualitative evaluation of item developers were compared between focal and benchmark forms. Results indicated that the proposed method can detect meaning-based enemy items with the same accuracy as experts, but the accuracy depends on how the indices of similarity are integrated.

Au3-3 Stabilizing measurement precision through scale transformation and adaptive testing

Dongmei Li, *ACT, Inc., United States*

Conditional standard error of measurement (CSEM) provides important information for score interpretation. Having a constant CSEM across all score levels not only simplifies score reporting and score interpretation but also

contributes to fairness. There are two fundamentally different approaches to achieve constant CSEM in operational test programs. One is through scale transformation using methodologies as described by Kolen (1988), Li, Woodruff, Thompson, & Wang (2014), or Moses and Tim (2017), and the other is through computer adaptive testing (CAT) with the fixed precision stopping rule (Wainer, 2000). The purpose of this study is to illustrate the differences and similarities of these two approaches and to explore the best solutions for transitioning from linear tests with constant CSEM to CAT. Specifically, the study is intended to investigate the following research questions: 1. What are the differences and similarities for tests whose CSEM is made constant through scale transformation versus those through fixed precision CAT? 2. When transitioning from linear testing to CAT, if the linear forms have been scaled to have constant CSEM, how can the CAT have scores that are interchangeable with linear forms but maintain the same constant CSEM property? These research questions will be investigated through simulations of test scores from a linear test with 50 items that is scaled to have constant CSEM and those from a CAT administration using an item pool containing 600 items. Results from this study can inform decisions for test scaling in both linear testing and CAT.

Au3-4 Impact of IPD on pretest-item parameters using different calibration methods

Meichu Fan, *ACT, Inc., United States*
Xin Li, *ACT, Inc.*

The change in item parameter statistics over time, which is referred as item parameter drift (IPD; Goldstein, 1983), can have negative impacts on pretest item calibrations. Fan, Li & Cho (2017) conducted a study to investigate the potential impact of IPD on pretest item calibrations under different CAT conditions using the concurrent calibration method, and found that the proportion of items with IPD showed the largest effect on both a- and b-parameter estimates. The purpose of this study is to extend their work to include the use of fixed operational items in calibration, and compare the results with their 2017 results. The CAT design includes using the maximum item information with the Sympson-Hetter exposure control procedure and content balance control as the item selection method, expected a posterior as the provisional ability estimation method, and the maximum likelihood estimation method for the final ability estimate. Factors to be investigated are IPD type/magnitude/direction/percentage, and the size of calibration samples for pretest items. Ten thousands simulees will be generated from a normal distribution for each replication. BIOLOG-MG will be used for item parameter calibration using both calibration methods, and Stocking-Lord scale transformation (Stocking & Lord, 1983) will be used as the IRT linking method in the concurrent calibration. Correlation, root mean square error, and bias will be assessed for each replication, and their averages over 50 replications will be used to evaluate and compare the impact of IPD on the accuracy of pretest item calibrations under various conditions.

Au3-5 Comparing methods for calibrating pretest items with fixed operational forms

Hyung Jin Kim, *University of Iowa, United States*
Won-Chan Lee, *University of Iowa*

Most operational tests administer an additional section of pretest items (i.e., experimental section) to examinees. For SAT, a 20-minute pretest block is administered at the end of testing; and, beginning in September 2018, ACT started administering a 20-minute experimental block between the science and writing tests. Among examinees administered the same operational form, multiple different blocks tend to be administered to different but possibly random groups of examinees. This design of administering pretest blocks along with a fixed operational form can be considered as one form of linking designs in pretest item calibration. Although pretest items do not count toward students' scores, it is important that their parameters are estimated accurately for their possible uses in constructing future test forms. Hanson and Beguin (2002) and Kang (2009) have investigated different calibration methods such as to link item parameters to a common. Moreover, various studies have compared different calibration methods for on-line pretest items (Ali & Chang, 2014; Ban, Hanson, Wang, Yi, & Harris, 2001; Segall, 2003; Zheng, 2014; Zheng & Chang, 2017). However, the author could not locate a study which investigated calibrating pretest items under such a linking design. Therefore, this study considers multiple procedures for pretest item calibration: (1) concurrent calibration, (2) separate calibration block by block, (3) fixed item parameter calibration. A simulation study is conducted considering several factors regarding characteristics of pretest blocks, operational forms, and examinees. This study compares results of accuracy in parameter estimates and addresses practical issues that each procedure possesses.

Early Career Award: Dylan Molenaar

Frs-1 Beyond Simple main effects: challenges to the substantive interpretation of higher-order statistical effects

Dylan Molenaar, *University of Amsterdam, Netherlands*

Using traditional psychometric models like the linear factor model and the two-parameter logistic item response theory model, the simple main effects of items and subjects can be inferred from the first two moments of the matrix of observed scores. The resulting parameters are statistically useful for various practical issues including item calibration, test equating, and adaptive testing, and for various substantive issues like establishing group differences in IQ and personality. However, some substantive hypotheses predict statistical effects that go beyond the first two moments of the data. Examples of these higher-order effects include non-linear effects, non-normal effects, heteroscedastic effects, and mixtures of different effects. Studying phenomena like these is substantively interesting but statistically challenging as the distributional assumptions underlying common psychometric models may distort the modeling results once violated. In this presentation, the challenges of studying higher-order statistical effects are illustrated using cases from the fields of intelligence, personality, and response time research.

Presidential Address: Francis Tuerlinckx

Frs-1 Things I have learnt so far

Francis Tuerlinckx, *University of Leuven, Belgium*

In this presentation I will discuss three important insights that I have recently acquired regarding modeling and statistical analysis: (1) It is not about these data but other data; (2) it is not about the parameters but about the data; (3) it is not only about this model but also about other models. I will illustrate these insights with examples of recent work carried out by us or others.

Author Index