Book of Abstracts (Talks, Posters)

11 July 2022 - 15 July 2022

IMPS 2022 International Meeting of the Psychometric Society

Bologna, Italy

Table of Contents

Assessing measurement invariance for longitudinal data through latent Markov models

Tuesday, 12th July - 09:50: Invited Speaker: Alessio Farcomeni (Room B) - Individual Oral Presentation

Prof. Alessio Farcomeni [1]

1. *University of Rome "Tor Vergata"*

Central assumptions in psychological, social, and economic evaluations are that all items in a questionnaire have the same discrimination power, unidimensionality of the latent trait, and measurement invariance of the scale. We briefly revise latent Markov models for modeling data arising from repeatedly measured items. We then focus on the concept of measurement non-invariance. We define different notions of differential item functioning in the context of panel data. A simple model selection approach based on the Bayesian information criterion (BIC) is argued to successfully identify the correct measurement structure. We show the practical relevance by means of an extensive simulation study, and illustrate its use on two real-data examples from the social sciences.

Latent variable models for social network data

Tuesday, 12th July - 09:50: Invited Speaker: Tracy Sweet (Room A) - Individual Oral Presentation

Dr. Tracy Sweet [1]

1. *University of Maryland, College Park*

To accommodate the ill-defined dependence structure among network ties, one class of network models uses latent variables. Conditionally independent tie or dyad (CID) models are also relatively simple to fit, allowing for a number of interesting extensions. In this talk, I will present two examples from my own research: a multilevel model mediation model and a latent variable model for social influence as well as discuss some future directions.

Improved estimation of autoregressive models through contextual impulses and robust modeling

Tuesday, 12th July - 10:55: Symposium: New lights on dynamic models for intensive time series and panel data (Room B) - Symposium Presentation

Dr. Janne Adolf [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven*

The aim of the dynamic paradigm of affect research is to characterize daily life affective processes by means of intensive longitudinal data and dynamic, typically autoregressive-type models. The contextual conditions accompanying and potentially influencing affective processes obviously form an important part of the picture. Especially distinct contextual events – ranging from salient daily events to major life events – are regularly assessed and their effects modelled. Such an explicit approach to studying context can however reach its limits, if relevant events are hard to define, measure or model. In that case one finds oneself in a situation where contextual events play out as hidden contaminators possibly obscuring the affective process of interest. Interestingly, such contamination can also have beneficial effects in that it can leverage autoregressive dynamics and improve their estimation. In this talk we take a closer look at this phenomenon. We also demonstrate that robust autoregressive models deal especially well with contextual contamination as they not only capitalize on positive leverage effects but also mitigate the negative obscuring effects contextual events might have. We then turn to a comprehensive simulation study that compares the performance of different robust and classical autoregressive models under different contamination settings.

Including context and serial dependence in regression models of time series

Tuesday, 12th July - 11:10: Symposium: New lights on dynamic models for intensive time series and panel data (Room B) - Symposium Presentation

Mr. Sigert Ariens [1], ***Dr. Janne Adolf*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven*

To model the dynamics present in time series of psychological data, researchers often appeal to autoregressive structures. In recent years, the need to take into account contextual influences, e.g. emotionally relevant events in the study of affect dynamics, has become recognized. A straightforward way of doing so is by including covariates in the model equation. A first option is to include covariates in autoregressive models. Alternatively, one can include covariates in classical linear regression models, allowing for autocorrelation in the residuals. Researchers in the behavioral sciences have proposed situations in which one of the two approaches should be preferred, e.g. depending on whether contemporaneous or lagged effects are of prime interest. However, evidence for these propositions remains inconclusive. In this talk, we delineate the differences between the two approaches, and hope to provide the information needed for researchers to make an informed choice on this decision. Specifically, we show that the residualized approach can impose implicit restrictions on the model-implied relationships between the variables, which can result in biased estimates of the model parameters if these restrictions do not hold for the data at hand. We also touch on how the restrictions can be tested, relying on a simple likelihood-ratio test procedure, also for multilevel versions of the considered models. We present the results of a simulation study confirming these insights, and provide a real-data example showing that misleading results can be obtained when the residualized approach is used without testing the restrictions invoked by the model structure.

Dynamics in large scale online leaning

Tuesday, 12th July - 11:25: Symposium: New lights on dynamic models for intensive time series and panel data (Room B) - Symposium Presentation

Dr. Charles Driver [1], ***Prof. Martin Tomasik*** [1]

1. *University of Zurich*

Using intensive longitudinal data from the large scale online educational testing system 'Mindsteps' offers interesting possibilities in comparison to classical, infrequent educational testing approaches, but also poses substantial computational hurdles. I will first discuss the development of new item response theory software designed to handle very large numbers of both items and students, as well as some empirical results from vertical scaling approaches employed on the Mindsteps data. I then turn to the problem of incorporating the ability estimates for different aspects of student performance (such as Maths or German) in a hierarchical continuous-time dynamic systems model, to better understand the relations between different student abilities, and how these relations may change with age.

Mixture multilevel vector-autoregressive modeling

Tuesday, 12th July - 11:40: Symposium: New lights on dynamic models for intensive time series and panel data (Room B) - Symposium Presentation

Dr. Anja Ernst [1], ***Prof. Marieke Timmerman*** [1], ***Mr. Feng Ji*** [2], ***Dr. Bertus Jeronimus*** [1], ***Prof. Casper Albers*** [1]

1. *University of Groningen,* ***2.*** *University of California, Berkeley*

The modeling techniques for intensive longitudinal data are increasingly focused on individual differences. In my talk I will present mixture multilevel vector-autoregressive modeling, which extends multilevel vector-autoregressive modeling by including a mixture, to identify individuals with similar traits and dynamic processes. This exploratory model identifies mixture components, where each component refers to individuals with similarities in means (expressing traits), autoregressions, and cross-regressions (expressing dynamics), while allowing for some inter-individual differences within the components on these attributes. I will illustrate the model using affective data from the COGITO study. These data consist of samples for two different age groups of over 100 individuals each who were measured for about 100 days. I will demonstrate the advantage of exploratory identifying mixture components by analyzing these heterogeneous samples jointly.

Latent Markov factor analysis for evaluating measurement model differences in intensive longitudinal data

Tuesday, 12th July - 10:55: Symposium: Measuring the moment: Psychometric considerations and applications (Room A) - Symposium Presentation

Dr. Leonie Vogelsmeier [1]***, Prof. Jeroen Vermunt*** [1]***, Dr. Kim De Roover*** [1]

1. *Tilburg University*

When studying intensive longitudinal data (e.g., with Experience Sampling Methodology), drawing conclusions about dynamics of psychological constructs (e.g., well-being) over time requires the measurement model (MM; indicating which items measure which constructs) to be invariant between subjects and within subjects over time. However, there might be heterogeneity or "non-invariance" in the MM, for instance, due to subject-specific differences and changes in item interpretation or response styles. Mixture modeling approaches have proved to be powerful tools to detect unobserved heterogeneity, but methodology to evaluate MM differences for multiple time-points and subjects simultaneously was lacking. To fill this gap, we built upon common mixture modeling approaches and proposed latent Markov factor analysis (LMFA). LMFA combines a discrete- or continuous-time latent Markov model (that clusters observations into separate states, according to state-specific MMs) with mixture factor analysis (that evaluates which MM applies for each state). In my talk, I describe this novel methodology, illustrate it with an empirical example, and introduce the new user-friendly software package "lmfa" that allows researchers to easily investigate MM differences in their own intensive longitudinal data in the freely available software R.

Evaluating dynamics in affect measurement with latent Markov factor analysis

Tuesday, 12th July - 11:15: Symposium: Measuring the moment: Psychometric considerations and applications (Room A) - Symposium Presentation

Ms. Leonie Cloos [1], ***Dr. Leonie Vogelsmeier*** [2], ***Prof. Peter Kuppens*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven,* ***2.*** *Tilburg University*

Affect is a universal but inherently subjective experience involving the ongoing evaluation of how someone feels or expects to feel. The dynamic nature of affect is studied with intensive longitudinal methods (e.g., experience sampling; ESM), by repeatedly administering a self-report questionnaire to participants. The time and situation-specific observations provide insight into substantive effects but also challenges to measurement. The measurement model (MM) is likely to change across occasions, violating the assumption of measurement invariance (MI). The difference may be caused by artefacts (extreme response styles, item interpretation) but can also be explained by substantive theory. For example, in response to important events, people tend to perceive positive and negative affect as two opposites, while other times they can co-occur. The recently proposed latent Markov factor analysis (LMFA) evaluates MI by classifying observations into latent states according to the underlying MM. Observations that belong to the same state are invariant and comparable but observations that belong to different states are not. In this study, we explore the extent to which MMs differ across observations in ESM data and how person and occasion-specific variables may explain the states and transitions between them. First, we explore which MM underlie which parts of the data and determine latent subgroups of persons that differ in their transition patterns. Second, we determine whether neuroticism can explain differences between subgroups and whether stressful events trigger transitions between states. Our goal is to apply LMFA as a tool to test MI and investigate substantive research questions.

Manageable intensive longitudinal measurement(?)

Tuesday, 12th July - 11:35: Symposium: Measuring the moment: Psychometric considerations and applications (Room A) - Symposium Presentation

Dr. Joran Jongerling [1]

1. *Tilburg University*

Experience sampling and related methods (e.g., ecological momentary assessment and ambulatory assessment) have tremendous potential for uncovering within-person processes as they unfold over time. This high-resolution view of processes and the high ecological validity of these intensive longitudinal methods make them valuable tools for many different research topics, ranging from development of individualized treatment of cancer patients to understanding social media use and its correlates.

While carefully incorporating measurement (error) into analyses has become common in cross-sectional research, proper measurement modeling is not yet routinely employed in intensive longitudinal studies, which often work with single-item measures or averages/sum-scores. While this clearly threatens the validity and undercuts the promise of these methods, including a measurement model in intensive longitudinal analyses is not easy and poses unique challenges. For example, it increases sample-size requirements, which is already a problem in many intensive longitudinal studies.

Therefore, we take an extensive look at different methods to include (or fail to include) measurement to determine which approaches are feasible at realistic sample sizes, and to determine the impact of ignoring measurement on accuracy of results. We focus on a “compromise” method introduced by Schuurman & Hamaker (2019) that adds a measurement-error term to composite scores, and we evaluate how well this method compares to analyzing uncorrected averages, analyzing only the best indicator of a factor, and modeling the true measurement model. The effect of missing data and characteristics of the true measurement model on the performance of the different measurement methods is also investigated.

Assessing the reliability of single-item momentary mood measurements in experience sampling

Tuesday, 12th July - 11:55: Symposium: Measuring the moment: Psychometric considerations and applications (Room A) - Individual Oral Presentation

Prof. Francis Tuerlinckx [1], ***Dr. Egon Dejonckheere*** [1], ***Ms. Febe Demeyer*** [1], ***Ms. Birte Geusens*** [1], ***Mr. Maarten Piot*** [1], ***Dr. Stijn Verdonck*** [1], ***Dr. Merijn Mestdagh*** [1]

1. *KU Leuven*

Experience sampling methods (ESM) are commonly used to study how, for example, mood fluctuates in everyday life. To reach valid conclusions, confirming the reliability of momentary mood measurements is essential. However, to minimize participant burden, ESM researchers often use single-item measures, preventing a reliability assessment of people's mood ratings. Furthermore, because mood constantly change, checking reliability via conventional test-retest procedures is impractical, for it is impossible to separate measurement error from meaningful mood variability. In two complementary ESM protocols (n's = 91 and 96), we overcome these challenges and explicitly evaluate the reliability of single-item momentary mood measurements. In Protocol 1, we randomly repeat one emotion item within the same momentary survey and evaluate the discrepancy between test and retest ratings to determine measurement error. In Protocol 2, we vary the time interval between consecutive surveys, and extrapolate the size of this discrepancy to the hypothetical instance where no time between assessments would exist. Crucially, both protocols converge on the amount of measurement error they establish: The size of random distortions in people's mood ratings corresponds with almost a 10th of the measurement scale, comprising around 23% of the total variability in participants' affective responses. Although our results are more optimistic than model-based reliability estimations, we demonstrate that disregarding measurement error in ESM is not inconsequential.

Regularized parameter estimation for probabilistic knowledge structures

Tuesday, 12th July - 10:55: Symposium: Connecting Knowledge Space Theory to other formal theories (Room C) - Symposium Presentation

Prof. Matthias Gondan [1], ***Mrs. Alice Maurer*** [1]

1. *University of Innsbruck*

If Q is a non-empty set of items to a given knowledge domain, a knowledge structure defines a family of knowledge states $K \subseteq Q$, together with some dependencies that reflect prerequisite relations between the states. The probabilistic measurement model (i.e., basic local independence model) assumes a probability distribution on the knowledge states, as well as item-specific probabilities for lucky guesses and careless errors. In practical applications with potentially large knowledge structures, the number of free parameters can grow rapidly, which raises difficulties in obtaining reliable estimates for the parameters. Matters get worse in generalized models that entail skills and misconceptions, as well as processes of learning and/or forgetting. In our study we use classical regularization techniques (e.g., LASSO) that are well-known from other areas such as neural networks. We investigate the consequences of regularization on the bias and variance of the parameter estimates, as well as on the accuracy of knowledge assessment in simple toy examples and more complex realistic knowledge structures. The results are compared to those obtained from established procedures, e.g., minimum discrepancy maximum likelihood.

On the identifiability of 3 and 4 parameters Item Response Theory models from the perspective of Knowledge Space Theory

Tuesday, 12th July - 11:10: Symposium: Connecting Knowledge Space Theory to other formal theories (Room C) - Symposium Presentation

Dr. Stefano Noventa [1], ***Dr. Sangbeak Ye*** [1], ***Prof. Augustin Kelava*** [1], ***Prof. Andrea Spoto*** [2]

1. *University of Tuebingen,* ***2.*** *University of Padova*

In spite of its relevance, the identifiability issue in Item Response Theory models is still a topic under investigation and a general solution has so far proved to be elusive. We show that the identifiability issue of some IRT models can be grounded in the notion of identifiability in Knowledge Space Theory. Specifically, that the identifiability of the three- and four-parameters Item Response Theory models (e.g., logistic models) are sub-cases of the more general issues of forward- and backward- gradedness of a knowledge structure. As a consequence, the identifiability problem is split into two parts: a first one, which is the result of a trade-off between the guessing/ceiling parameters and the parameters within the Rasch model or the two-parameters logistic model; and a second one, which is the already well-known identifiability issue of the 2PL model itself.

Constructing, improving, and shortening tests with competence-based test development methodology

Tuesday, 12th July - 11:25: Symposium: Connecting Knowledge Space Theory to other formal theories (Room C) - Symposium Presentation

Dr. Pasquale Anselmi [1]***, Prof. Jürgen Heller*** [2]***, Prof. Luca Stefanutti*** [1]***, Prof. Egidio Robusto*** [1]

1. *University of Padova,* ***2.*** *University of Tübingen*

Competence-based Knowledge Space Theory and Cognitive Diagnostic Models both provide a theoretical framework for assessing the latent set of skills an individual has available in a certain knowledge domain (called the "competence state" in Competence-based Knowledge Space Theory) from the responses to test items. After inferring the subset of mastered items (called the "knowledge state") from the observed item responses, the set of underlying skills is derived. A good test ensures that the uncertainty about the competence state underlying the knowledge state is as small as possible. Competence-based Test Development is introduced as a novel approach for constructing, improving, and shortening tests for skill assessment. Concepts originally introduced in Rough Set Theory are exploited to develop tests that are as informative as possible about individuals' competence states and minimal (no item can be deleted without altering the assessment). For a fixed collection of competence states, tests can be obtained that differ from one another in certain desired features (e.g., the number of skills assessed by each item), but are equivalent with respect to informativeness and minimality. The development of conjunctive (for each item, there is a unique set of skills) and disjunctive (each of the skills assigned to an item is sufficient for solving it) tests is considered. Various applications of the proposed methodology are presented.

A competence-based knowledge space theory learning platform for promoting quantitative thinking in higher education

Tuesday, 12th July - 11:40: Symposium: Connecting Knowledge Space Theory to other formal theories (Room C) - Symposium Presentation

Dr. Andrea Brancaccio [1]***, Dr. Debora de Chiusole*** [1]***, Prof. Luca Stefanutti*** [1]

1. *University of Padova*

An adaptive and personalized learning platform, called QHelp (Quantitative Higher education learning platform) is presented. The system is based on the competence-based knowledge space theory (CbKST) and it is aimed at promoting quantitative thinking and skills of higher education students. Its architecture essentially consists of an assessment module and a learning module that continuously exchange information about the current competence state of a student and provide personalized suggestions for learning. Following a bottom-up pedagogical principle, some competence structures in the domain of quantitative psychology were developed by groups of university students supervised by experts in the respective fields. Results concerning the efficiency and accuracy of knowledge assessments through the competence structures that currently populate the system are illustrated in a series of simulation studies and in a number of different scenarios. Finally, a concrete example of an adaptive assessment session resulting in the final report and of the subsequent learning session of the student is illustrated.

Modeling preference with the basic local independence model

Tuesday, 12th July - 11:55: Symposium: Connecting Knowledge Space Theory to other formal theories (Room C) - Symposium Presentation

Prof. Luca Stefanutti [1], ***Dr. Andrea Brancaccio*** [1], ***Dr. Debora de Chiusole*** [1]

1. *University of Padova*

In recent years, KST have been generalized and extended to areas of application that are different or even far from the context for which it was originally conceived and designed, namely the assessment of knowledge. One of these areas of application is preference. Through a formal approach named media theory, it was targeted by Jean-Claude Falmagne (1997), one of the two founders of KST. In this talk I will show that, curiously enough, a formal equivalence can be established between a restriction of one of the most well-known and studied probabilistic models in KST, named the basic local independence model (BLIM), and a probabilistic model of preference in discrete choice experiments, developed by Birnbaum (2008), and named the true-and-error model (TEM). It is shown that the unrestricted BLIM provides sound assumptions for modelling preference, and is slightly more flexible than the TEM. A few examples of application of the BLIM to preference data are illustrated.

Understanding Students' test engagement for feedback

Tuesday, 12th July - 10:55: Symposium: Advances in use of response process data in measurement (Room D) - Symposium Presentation

Dr. Hongwen Guo [1], ***Dr. Matt Johnson*** [1], ***Dr. Kadriye Ercikan*** [1], ***Dr. Luis Saldivia*** [1]

1. *Educational Testing Service*

Students can obtain the same scores on assessments through different ways of test engagement. Process data may provide insight on what path a student took. For example, Student A may have started the test with much effort, spending long time on the first few items and playing with the available tools, but then rapidly guessed the rest items – a possible sign of frustration; Student B may have spent test time evenly throughout the test, using the tools when necessary – a possible sign of regulated time management; Student C may have quickly gone through most of the items without trying – a possible sign of lack of motivation. This study uses both students' test engagement and performance sequential data on the release NAEP item blocks and applies ML algorithms to mine useful information on why students did not perform well, and whether and where students were frustrated on the assessment. Insights from the analyses can provide potentially useful feedback regarding students' test engagement and content areas for further learning and development.

Examining numeric sequence similarity measures with dynamic time warping method

Tuesday, 12th July - 11:10: Symposium: Advances in use of response process data in measurement (Room D) - Symposium Presentation

Dr. Qiwei He [1], ***Dr. Elizabeth Tighe*** [2], ***Dr. Marcia Davidson*** [2], ***Dr. Gal Kaldes*** [2]

1. *Educational Testing Services,* ***2.*** *Georgia State University*

Sequence similarity measures have been increasingly used in process data analyses with educational assessments to extract meaningful patterns from time-stamped action sequences or navigation path. Such similarity measures are commonly employed on categorical variables (e.g., actions, pages, words) to compute the distance between pairs of sequences, for instance, by the common subsequence method and edit distance method. Besides the information learned from respondents' problem-solving action sequences, the time intervals between each action by item, the visit-revisit pattern and the time allocated on each item and section could also provide information on respondents' strategies to solve digital tasks. In this study, we present a novel approach to measure the similarity among the numeric sequences (i.e., sequence of time allocation through the item section) using the dynamic time warping (DTW) method. Specifically, we will showcase the application of DTW on adults' digital literacy skills with the 2012 U.S. PIAAC sample. We extracted individuals' time allocation sequences and sequence of time for the first action on the literacy items and we will compute the pairwise distance among all these sequences. Based on the DTW sequence distance matrix, we will conduct a cluster analysis to understand typical patterns of time allocation for low-skilled adults and examine whether the identified clusters predict literacy performance and vary by background characteristics (e.g., demographics, reading habits). The method presented in this study holds promise for a broader application to compute distance similarity of numeric sequences in different digital assessment settings.

Psychometric considerations for the joint modeling of response and process data

Tuesday, 12th July - 11:25: Symposium: Advances in use of response process data in measurement (Room D) - Symposium Presentation

Dr. Matt Johnson [1], ***Dr. Xiang Liu*** [1]

1. *Educational Testing Service*

The increasing availability of process data from educational and psychological assessments has led to a number of new latent variable models for the joint analysis of the process and item response data. In this presentation we will discuss examples of these joint process/response models and examine the psychometric implications of different modeling choices and will demonstrate our findings on data from a large-scale educational assessment.

Modeling student problem solving behavior using mixed types of response process data

Tuesday, 12th July - 11:40: Symposium: Advances in use of response process data in measurement (Room D) - Symposium Presentation

Dr. Caitlin Tenison [1], ***Dr. Burcu Barslan*** [1]

1. *Educational Testing Service*

The use of interactive computer tasks (ICTs) within large-scale assessments provides the opportunity to capture authentic student problem solving. Prior research modeling response process data has primarily focused on identifying student problem solving strategies by identifying patterns in student action sequences. While the observed actions of students can help us distinguish differences in student decision outcomes, the pauses between actions can provide an additional view into the cognitive processing that drives these decisions. In this study, we present the use of heterogeneous hidden Markov models (HHMMs) to model mixed data types (i.e., actions and the pauses between actions) generated from student problem solving within a science inquiry ICT. Using HHMMs we capture the probabilistic transition between latent states in sequential timesteps. Unlike discrete and gaussian HMMs, which only use a single data type, HHMMs can use different distributions to estimate the emission probabilities of observed data. We fit an HHMM to three data streams generated from student's interaction with the task. This model produced a concise representation of student's problem-solving strategy that accounts for the order in which actions occur and differences in the time it takes to produce those actions; two important indications of the cognitive processes underlying those actions. Our initial findings suggest this method distinguishes between distinct problem-solving states and navigational behaviors. We discuss the implications this approach has for understanding test taker behavior, generating response process indicators, and improving assessment design.

A speed-accuracy response model with conditional dependence

Tuesday, 12th July - 11:55: Symposium: Advances in use of response process data in measurement (Room D) - Symposium Presentation

Dr. Peter van Rijn [1], ***Dr. Usama Ali*** [1]

1. *Educational Testing Service*

Conditional independence assumptions can be problematic in modeling process data from educational tests such as response times. For this reason, a speed-accuracy response model with conditional dependence is developed. It is a generalization of the model developed by Maris and van der Maas (2012) in which a scoring rule that incorporates both accuracy and speed of item responses is assumed to be a sufficient statistic for the latent proficiency variable. The assumption of conditional independence is dropped in a similar vein as in the interaction model developed for dichotomous item responses by Haberman (2007). Recently, Verhelst (2019) discussed similar models in the context of exponential-family models for continuous item responses. Both conditional and marginal maximum likelihood approaches can be developed to estimate item parameters. The model is illustrated by an application to data from a recent experimental study on the impact of scoring instructions and timing on quantitative reasoning.

Estimating effect heterogeneity in rare events meta-analysis with nonparametric mixture models

Tuesday, 12th July - 10:55: Meta Analysis (Room G) - Individual Oral Presentation

Prof. Heinz Holling [1], ***Ms. Katrin Jansen*** [1]

1. *University of Münster*

In this talk, we explain how nonparametric mixture models can be used as an approach to deal with heterogeneity in meta-analysis of count data. These models are applicable to meta-analyses of studies which report the occurrence of an event in a treatment and a control group. In particular, they are well-suited for meta-analysis of rare events, since they avoid assuming a normal distribution within studies and instead assume the counts to arise from Binomial or Poisson distributions, which enables them to naturally incorporate evidence from zero-count studies. Above that, these models replace the assumption of a normal distribution between-studies by a nonparametric mixture distribution and thus, provide a flexible way to model effect heterogeneity.
We provide an extension of the simulation study by Holling et al. (2022), which focused on the performance of nonparametric mixture models for rare events and found that these models perform well in estimating both the pooled effect and effect heterogeneity in numerous situations. However, previous research shows that models for rare events meta-analysis often suffer from serious performance issues when events are very rare (i.e., when event probabilities are 0.01 and below). Thus, we investigated the performance of nonparametric mixture models in conditions with very rare events in a new simulation study, the results of which will be presented at the conference.

Synthesizing research on complex sampling surveys: Two-stage meta-analysis with individual participant data

Tuesday, 12th July - 11:10: Meta Analysis (Room G) - Individual Oral Presentation

Mr. Diego Campos [1], ***Prof. Mike W.-L. Cheung*** [2], ***Prof. Ronny Scherer*** [1]

1. *University of Oslo,* ***2.*** *National University of Singapore*

Meta-analyses typically synthesize summary statistics from multiple studies to draw conclusions from the combined information. This method is helpful when the original data used in the previous analyses are not available. However, the growing availability of individual participant data (IPD) shared by researchers and larger institutions offers new ways to synthesize research results via IPD meta-analysis. When the available datasets are based on complex sampling designs—that is, designs involving multi-stage sampling, clustering, and stratification—generating the effect sizes of interest and synthesizing them across studies or datasets face several challenges. These challenges include accounting for different sample procedures and weights across datasets and dealing with the hierarchical structures of the primary and meta-analytic data.

We propose a two-stage IPD meta-analytic approach for synthesizing data from complex sampling surveys to address these challenges. In this approach, the data from larger sampling units (e.g., countries, study cycles) are analysed independently, and then, the parameter estimates are combined via meta-analysis. To illustrate this approach, we estimated the so-called Big-Fish-Little-Pond-Effect (i.e., the negative effect of classroom achievement on students' self-concept; BFLPE) across five cycles of the Trends in International Mathematics and Science Study (TIMSS). In stage one, we performed multi-group multilevel structural equation modeling to obtain the BFLPE effect sizes per country, ensuring measurement invariance. In stage two, we synthesized the effect sizes and quantified their heterogeneity via multilevel random-effects models. Our study illustrates how researchers can use two-stage IPD meta-analysis to synthesize research evidence using complex sampling survey data.

Checking the inventory: Currently available methods for raw data MASEM

Tuesday, 12th July - 11:25: Meta Analysis (Room G) - Individual Oral Presentation

Mr. Lennert Groot [1]

1. *University of Amsterdam*

Researchers conducting Meta-Analytical Structural Equation Modeling (MASEM) may, whether through personal effort or favorable circumstances, have at their disposal the raw data of the studies they wish to meta-analyze. The present study identifies, illustrates and compares a range of possible analysis options for researchers to whom raw datasets are available. This study describes techniques based on summary statistics (such as correlation-based MASEM) and techniques that directly analyze the raw data (such as multi-level and multi-group SEM) and discusses differences in requirements, procedures, outcomes, and limitations. This is done using a collection of 39 raw datasets that were previously published in Hagger et al. (2022). A path model reflecting the Theory of Planned Behavior (TBP; Ajzen, 1991), is fitted to these datasets using SEM. The goal of this study is to aid researchers in finding a readily available approach that allows them to make the most of the data that is available to them, depending on the characteristics of the datasets and the research question.

Dependent effect sizes in MASEM: The current state of affairs

Tuesday, 12th July - 11:40: Meta Analysis (Room G) - Individual Oral Presentation

Ms. Zeynep Bilici [1], ***Dr. Suzanne Jak*** [1]

1. *University of Amsterdam*

The current practice when it comes to carrying out a meta-analytic structural equation modeling (MASEM) analysis cannot properly deal with cases where there are multiple effect sizes available from the same study for the same relation. Existing applications either treat these effect sizes as independent, randomly select one effect size amongst many or create an average effect size. None of these approaches can deal with the inherent dependency in effect sizes, and either leads to biased estimates or loss of information and power. An alternative technique is to use univariate three-level modeling in the two-stage approach to model these dependencies (Wilson et al., 2016). These different strategies have not been previously compared in a simulation study in terms of their performance when it comes to dealing with dependent effect sizes in varying research conditions. The aim of this study is to compare these strategies to establish the performance of the current strategies before moving on to establishing new and better methods to tackle the problem of dependent effect sizes. We assessed the performance of these strategies across conditions of changing number of studies, number of dependent effect sizes within studies, sample size, the magnitude of the correlation between the dependent effect sizes, within and between studies variance to see under which combination of conditions the current practices are the most harmful in terms of the results they point to.

Using meta-analytic structural equation modeling to synthesize randomized controlled trials

Tuesday, 12th July - 11:55: Meta Analysis (Room G) - Individual Oral Presentation

Ms. Hannelies de Jonge [1], ***Dr. Kees-Jan Kan*** [1], ***Dr. Suzanne Jak*** [1]

1. *University of Amsterdam*

Meta-analytic structural equation modeling (MASEM) is an increasingly popular statistical technique as it allows for the meta-analytic investigation of multiple relations among variables simultaneously. The great advantage of MASEM is that one can include effect sizes of primary studies even when they do not include all variables of interest. Currently, it is not evident if and how one can include data from randomized controlled trials (RCTs) in MASEM. A possibility is to transform Cohen's *d* based effect sizes into point-biserial correlations. Such a transformation, however, is surrounded with confusion as conversion formulas differ across publications, statistical software, and online conversion tools. In this paper we show how Cohen's *d* values can be transformed to the point-biserial correlation under various assumptions. This yields four expressions or approximations, of which three figure prominently in the literature. As it is not clear to what extent the use of the different conversion formulas may impact the results in MASEM, we conducted a Monte Carlo simulation study. Preliminary results show that when one uses the most general conversion formula, bias in the path coefficients (and their standard errors) can be considered negligible. The other conversion formula's, which are often used in practice, can result in serious bias in the path coefficients. Concludingly, MASEM can be applied to synthesize the results of RCTs by transforming Cohen's *d* to a point-biserial correlation, but researchers should be aware of the different conversion formulas and are advised to use the most general conversion formula.

Propensity score analysis with latent variables: Choices of factor scores and data mining methods

Tuesday, 12th July - 10:55: Causal Inference I (Room E) - Individual Oral Presentation

Dr. Ge Jiang [1], ***Ms. Jiye Kim*** [1], ***Dr. Catherine Corr*** [1], ***Ms. Tianshu Qu*** [2]

1. *University of Illinois at Urbana–Champaign,* ***2.*** *Rutgers University*

When random assignment is not feasible, propensity score methods (PSM) are commonly used to balance groups and reduce bias in estimating treatment effects. Like any other statistical method, PSM depends on certain assumptions. One is that all covariates and outcomes are free of measurement errors (Steiner, Cook, & Shadish, 2011). For variables that contain measurement errors, structural equation modeling (SEM) can be used to model them as latent variables. Another key assumption is that the relationship between propensity scores and the covariates is correctly specified. Existing studies almost exclusively use a main-effects logistic regression model, which overlooks higher-order effects, interaction effects, and/or more complicated patterns. Data mining methods like Random Forests (RF) and Generalized Boosted Models (GBM) have been considered as alternatives but they have not been evaluated for estimating propensity scores with latent variables.

The goal of the current paper is to adapt RF and GBM for estimating propensity scores in SEM models. The proposed approaches will be evaluated using a simulation study that considers a broad range of scenarios and their performances will be compared to that of the traditional logistic regression. The outcome measures are the bias and efficiency of treatment effect estimates as well as the average standardized absolute mean difference. Preliminary simulation results showed that the two new approaches perform slightly worse than logistic regression when the relationship is linear but can significantly reduce the bias with non-linear relationships. We will also illustrate the new approaches using a large-scale real dataset.

Bayesian mediation analysis with power prior distributions

Tuesday, 12th July - 11:10: Causal Inference I (Room E) - Individual Oral Presentation

Dr. Milica Miocevic [1]

1. *McGill University*

Bayesian analysis is exact for small samples and is often suggested as a solution for convergence issues and low power (e.g., Depaoli, 2013; Lee & Song, 2004; Yuan & MacKinnon, 2009). However, as was pointed out in a recent systematic review (Smid, McNeish, Miočević, & van de Schoot, 2020), Bayesian methods have advantages over classical methods almost only with accurate informative prior distributions, and consequences of using inaccurate informative priors can be detrimental for the statistical properties of parameters of interest (see e.g., Depaoli, 2014; Holtmann, Koch, Lochner, & Eid, 2016; Miočević, Levy, & MacKinnon, 2020).

This talk describes how power prior distributions (Ibrahim & Chen, 2000) based on a historical data set can be used to increase power to detect the indirect effect in Bayesian mediation analysis (Miočević & Golchi, 2021). The only requirements for implementing the procedure are that the data from the current study constitute a representative sample from the population of interest, and that the historical and current data sets contain measures of the same covariates and independent variable, mediator, and outcome. The simulation study findings show that the proposed method leads to appropriate amount of borrowing from the historical data set, which leads to increases in precision and power when the historical data and current data are exchangeable and does not induce bias when the historical and current studies are not exchangeable.

Synthesizing data from pretest-posttest-control-group designs in meta-analyses of mediating mechanisms

Tuesday, 12th July - 11:25: Causal Inference I (Room E) - Individual Oral Presentation

Dr. Zijun Ke [1], ***Ms. Zhiming Lu*** [1], ***Dr. Rebecca Cheung*** [2], ***Dr. Qian Zhang*** [3]

1. *Sun Yat-sen University,* ***2.*** *The Education University of Hong Kong,* ***3.*** *Florida State University*

After a treatment effect has been quantified using meta-analyses over data from pretest-posttest-control-group designs, researchers are likely to resort to meta-analytic structural equation modeling (MASEM) techniques to investigate how the treatment effect takes place. A typical MASEM analysis in this context would use the grouping variable (treatment vs. control) as the independent variable, the amount of change in the proposed mediator as the mediating variable, and the amount of change in the outcome variable as the dependent variable. Given that MASEM techniques now allow missing data, studies on the treatment effectiveness with pretest-posttest-control-group designs can be (and should be) included as data for the meta-analysis of mediating mechanisms. A convenient way to do this is to convert the effect sizes for pretest-posttest-control-group designs (i.e., Cohen's d) to correlation coefficients, which are then used as part of the data for MASEM analyses. The problem of this approach is that the textbook ways of calculating effect sizes for pretest-posttest-control-group designs are derived as if between-subject designs were used. This causes the converted correlations to deviate from the actual bivariate correlations. We thus present a new approach to the synthesis of data from pretest-posttest-control-group designs in the MASEM analysis. The approach is based on a "standardized mean change difference" measure. We use a simulation study to examine the finite sample performance of the presented approach. Analyses are illustrated using results of the effectiveness of mindfulness-based interventions on health outcome variables. A discussion and practical guidelines are also included to conclude the study.

Further remarks on evidence and inference in educational assessment

Tuesday, 12th July - 13:30: Keynote: Robert Mislevy (Room B) - Individual Oral Presentation

Prof. Robert Mislevy [1]

1. *University of maryland*

My 1994 Presidential Address “Evidence and inference in educational assessment” examined the interplay of probability-based reasoning and psychological perspectives in educational measurement through the lens of evidentiary reasoning (ER). Since that time there have been rapid developments in areas related to assessment—in technology, psychology, learning domains, and analytic methods. I begin here by recapping basic tenets of ER, its relationship to between-persons educational measurement models, and the complex adaptive sociocognitive systems view that can undergird assessment arguments. I then note insights and support that this framework provides for tackling current issues in educational assessment, such as the following:

- Assessments and measurement models in educational systems
- Assessment design
- Strengthening the connection between assessment and learning
- More complex forms of assessment such as those including simulations and interactivity
- Integrating psychometric and data-analytic concepts and methods in complex assessments
- Assessing high-level / 21st Century skills
- The situated meanings of models, variables, probabilities, and measurements
- The situated meanings of validity, reliability, comparability, generalizability, and fairness

What changes in diffusion IRT model parameters can tell us about the speed-accuracy tradeoff

Tuesday, 12th July - 14:50: Symposium: Advances in using response times for understanding between-person differences and within-person changes in test-takers' behavior (Room B) - Symposium Presentation

Mr. Tobias Alfers [1], ***Mr. Georg Gittler*** [2], ***Ms. Esther Ulitzsch*** [3], ***Prof. Steffi Pohl*** [1]

1. *Freie Universität Berlin,* ***2.*** *University of Vienna,* ***3.*** *IPN - Leibniz Institute for Science and Mathematics Education*

The performance of test takers responding to items of a cognitive ability test is subject to the infamous speed-accuracy tradeoff (SAT): Accuracy is traded for an increased speed. To detect and understand the SAT underlying psychological processes of individuals need to be revealed. A promising model approach that is gaining popularity in explaining underlying psychological processes in cognitive ability tests is the Diffusion IRT model. Specifically, this model jointly considers information from both item responses and response times and delineates test takers' response behavior as an information accumulation process. In order to investigate the SAT it would be of particular interest applying a Diffusion IRT model to test settings that experimentally provoke a change in response speed and thereby allow capturing processes behind individual adaptations in response accuracy. A possible experimental manipulation is to introduce conditions that either emphasize speed or accuracy within the instruction of a task. In the present study we provide empirical data from a high-stakes test situation, where items from two spatial ability tests were administered to over 1300 applicants for an air traffic controller training program. In particular, every applicant had to solve item sets of both tests under a self-paced (accuracy emphasized) and time-pressured (speed emphasized) instructional condition. This test setting enabled us to uncover interpretable changes in parameters of the Diffusion IRT model that refer to mechanisms of the SAT. We discuss in which way the diffusion IRT model parameters and changes thereof may provide relevant parameters for diagnostic purposes.

A sequential Bayesian changepoint detection procedure for aberrant behaviors in computerized testing

Tuesday, 12th July - 15:05: Symposium: Advances in using response times for understanding between-person differences and within-person changes in test-takers' behavior (Room B) - Symposium Presentation

Dr. Jing Lu [1], ***Dr. Chun Wang*** [2], ***Dr. Jiwei Zhang*** [3], ***Ms. Xue Wang*** [1]

1. *Northeast Normal University,* ***2.*** *University of Washington,* ***3.*** *Yunnan University*

Changepoints are abrupt variations in a sequence of data in statistical inference. In educational and psychological assessments, it is pivotal to properly differentiate examinees' aberrant behaviors from solution behavior to ensure test reliability and validity. In this paper, we propose a sequential Bayesian changepoint detection algorithm to monitor the locations of changepoints for response times in real time and, subsequently, further identify the types of aberrant behaviors in conjunction with response patterns. Two simulation studies were conducted to investigate the efficiency and accuracy of the proposed detection procedure in terms of identifying one or multiple change points at different locations. In addition to manipulating the number and locations of changepoints, two types of aberrant behaviors were also considered: rapid guessing behavior and cheating behavior. Simulation results indicate that ability estimates could be improved after removing responses from aberrant behaviors identified by our approach. Two empirical examples were analyzed to illustrate the application of proposed sequential Bayesian changepoint detection procedure.

Using item response times to explain group differences in item parameters

Tuesday, 12th July - 15:20: Symposium: Advances in using response times for understanding between-person differences and within-person changes in test-takers' behavior (Room B) - Symposium Presentation

Dr. Dylan Molenaar [1], ***Mr. Thijs Carrière*** [2], ***Dr. Remco Feskens*** [2]

1. *University of Amsterdam,* ***2.*** *Cito Institute for Educational Measurement*

Measurement invariance is the well-known prerequisite that item parameters should be invariant across groups if those groups are compared on the latent variable underlying a psychological or educational test. Various factor analysis and item response theory approaches are available to test for measurement invariance. Ideally, violations that are found are theoretically interpretable in terms of the item content and the groups under consideration. However, often, items seem theoretically sound, and it is ambiguous why a given item violates measurement invariance. This is problematic as in such situations, it is unclear what to do with the item in the group comparison: If the item is theoretically sound, why omit it? In this talk, it is demonstrated how item response times can help in interpreting the results from a measurement invariance analyses and separate between meaningful violations and (still) uninterpretable violations.

Modeling the process underlying solution and non-solution behavior with a non-linear ballistic accumulator model

Tuesday, 12th July - 15:35: Symposium: Advances in using response times for understanding between-person differences and within-person changes in test-takers' behavior (Room B) - Symposium Presentation

Mr. Sören Much [1], ***Dr. Jochen Ranger*** [1], ***Mr. Augustin Mutak*** [2], ***Dr. Robert Krause*** [2], ***Prof. Steffi Pohl*** [2]

1. *Martin-Luther-Universität Halle-Wittenberg,* ***2.*** *Freie Universität Berlin*

Non-solution behavior may result e.g., in rapid guessing or item omission. So far, while the occurrence of this behavior has been modeled in previous research, the process underlying it has less been considered. We propose a non-linear ballistic accumulator to model and understand both solution and non-solution behavior in computerized achievement tests. In contrast to other accumulator models that involve a race either between several response options (e.g., Brown & Heathcote, 2008) or between a solution process and a separate stopping process (e.g., Ranger, Kuhn & Gaviria, 2015), our model is based on only one process. If enough information is accumulated to solve the task, a correct response is given. If instead the information accumulation stagnates, non-solution behavior is shown. Both solution and non-solution behavior are generated by the same process. It can be interpreted as an incorporation of a motivational factor that is part of both solution and non-solution behavior. This results in a slightly more parsimonious and more plausible model regarding the response process. Another benefit compared to similar sequential sampling models is a tractable likelihood function. However, the calculation is computationally intensive. We discuss different methods of model fitting, e.g., Bayesian estimation based on extensive data simulation and training a likelihood approximation network (LAN) (Fengler et al., 2021). In its simplest form, the model can account for typical test-taking phenomena like rapid guessing. It can be extended to incorporate ability-based guessing and omissions.

SAT in psychometric assessments: a latent growth approach

Tuesday, 12th July - 15:50: Symposium: Advances in using response times for understanding between-person differences and within-person changes in test-takers' behavior (Room B) - Symposium Presentation

Mr. Augustin Mutak [1]***, Dr. Robert Krause*** [1]***, Ms. Esther Ulitzsch*** [2]***, Mr. Sören Much*** [3]***, Dr. Jochen Ranger*** [3]***, Prof. Steffi Pohl*** [1]

1. *Freie Universität Berlin,* ***2.*** *IPN - Leibniz Institute for Sciene and Mathmaics Eduation,* ***3.*** *Martin-Luther-Universität Halle-Wittenberg*

The speed accuracy tradeoff (SAT) has been one of the most well-documented phenomena in psychology. Stemming from the early research in psychophysics, it was noticed how the increase in speed of performing a task usually leads to a decrease in task performance. SAT has so far been more studied in the area of cognitive psychology, but less so in the domain of psychometric assessments. One of the possible reasons for such situation is that the construction of a function which describes the relationship of speed and accuracy in a single individual requires several data points of both speed and accuracy for the said individual, whereas in the current settings each test taker usually receives a single score for each of these traits. In order to examine the natural shifts in the ability and speed during the course of the test, and thus relate them to each other, we have developed a latent growth IRT model by extending van der Linden's (2007) speed-accuracy model with growth terms for ability and speed. We present the development of the model, its specification, and technical details about the model estimation. We discuss difficulties encountered during model estimation, as well as overall model performance. Lastly, we apply the model to empirical data collected in PISA tests and present the findings, specifically focusing on SAT.

Playing HAVOK with the chaos caused by internet trolls

Tuesday, 12th July - 14:50: Symposium: Information warfare: quantitative methods to understand online manipulation campaigns (Room A) - Symposium Presentation

Ms. Elena Martynova [1]

1. *University of Virginia*

Strategic misinformation has been spiraling out of control with intranational and international actors attempting to manipulate social media users for political and commercial reasons. Fortunately, social media platforms keep records that can be aggregated into publicly available datasets and subjected to quantitative analysis. Modern quantitative techniques can be a key to extracting, understanding, and predicting misinformation for regulation and prevention. One example comes from Internet Research Agency (IRA) linked Russian state-sponsored Twitter accounts that manipulated US voters around the 2016 US presidential elections. A large database containing almost 3 million Twitter posts, from 2013 to 2018, from 2843 unique IRA-linked accounts is available online. Modern quantitative techniques, such as HAVOK (Hankel Alternative view of Koopman; Brunton et al., 2017), can yield insights into temporally relevant communicative strategies when qualitative text data is converted into time series of topical word frequencies. HAVOK is a powerful tool for modeling nonlinear and chaotic time series by decomposing them into intermittently forced linear systems. A forcing parameter allows HAVOK to demarcate regions where a time series is approximately linear from those that are nonlinear, with the forcing extrema often preceding the shifts and aligning with the extrinsic events. Sociohistorical events that incited the cascades of topical twitter posts may be identified by alignment with the nonlinear forcing activity. Through HAVOK, the dynamics between left-wing, right-wing and fearmonger accounts can be structurally defined, and compacted into mathematical and visual representations using the R package *havok* (Moulder, Martynova & Boker, 2020).

Dynamic exploratory graph analysis: Russian trolls and the US elections

Tuesday, 12th July - 15:05: Symposium: Information warfare: quantitative methods to understand online manipulation campaigns (Room A) - Symposium Presentation

Dr. Hudson Golino [1]

1. *University of Virginia*

The past few years were marked by increased online offensive strategies perpetrated by state and non-state actors to promote their political agenda, sow discord, and question the legitimacy of democratic institutions worldwide. In 2016, the US congress identified a list of Russian state-sponsored Twitter accounts that were used to try to divide voters on a wide range of issues. In the current presentation, we introduce a new method to estimate latent topics in texts from social media termed Dynamic Exploratory Graph Analysis (DynEGA). We applied DynEGA on a large dataset with Twitter posts from state-sponsored right- and left-wing trolls during the 2016 US presidential election. DynEGA revealed topics that were pertinent to several consequential events in the election cycle, demonstrating the coordinated effort of trolls capitalizing on current events in the U.S. This example demonstrates the potential power of our approach for revealing temporally relevant information from qualitative text data, above and beyond traditional methods such as Latent Dirichlet Allocation.

Metric invariance: Differences between Left and Right-wing Russian trolls and the US elections

Tuesday, 12th July - 15:20: Symposium: Information warfare: quantitative methods to understand online manipulation campaigns (Room A) - Symposium Presentation

Ms. Laura Jamison [1], ***Dr. Hudson Golino*** [1], ***Dr. Alexander Christensen*** [2]

1. *University of Virginia,* ***2.*** *University of Pennsylvania*

In this presentation we will introduce an approach for testing metric invariance in the exploratory graph analysis framework. We will then apply this test to test if the latent topic structure differs for left and right-leaning Russian troll accounts. In our approach, we run exploratory graph analysis for each group (in this case, for left and right-leaning accounts) and ensure that the same structure is identified for each group. Then, we calculate network loadings and use a permutation test to test whether or not network loadings are equivalent across groups. In this way, we are able to see which variables specifically have a different relationship to their overall communities. This will allow us to see what specific differences exist in the type of language used by left and right-leaning accounts and how it relates to the overall latent topics being addressed by both political leanings. When differences arise, it shows how words function differently for topics based on what political leaning is using them.

A general Monte Carlo method for sample size analysis in the context of network models

Tuesday, 12th July - 14:50: Symposium: Advances in methods for sample size determination in the social sciences (Room C) - Symposium Presentation

Mr. Mihai Constantin [1], ***Dr. Noémi Schuurman*** [2], ***Prof. Jeroen Vermunt*** [1]

1. *Tilburg University,* ***2.*** *Utrecht University*

The network approach to psychology is an increasingly popular framework for studying pairwise interactions among variables. As the field matures and psychological network modeling becomes more prevalent, there is an increasing need to aid researchers with a network approach in mind that plan to collect data. In this talk, I introduce a general method for performing sample size analysis in the context of network models. The method takes the form of a three-step recursive algorithm designed to find an optimal sample size value given a model specification, an outcome measure (e.g., sensitivity), and a statistic of interest (e.g., power). It starts with a Monte Carlo simulation step for computing the outcome measure and the statistic at various sample sizes. It continues with a monotone non-decreasing curve-fitting step for interpolating the statistic. The final step employs stratified bootstrapping to account for the uncertainty around the interpolated curve. In the first part of this talk, I provide an overview of the method and discuss its validation and performance. In the second part, I illustrate, in the form of a tutorial, how the method can be applied to a network model. The tutorial showcases the open-source implementation of the method as an **R** package called **powerly**.

Power analysis methods for mediation and moderated mediation models

Tuesday, 12th July - 15:05: Symposium: Advances in methods for sample size determination in the social sciences (Room C) - Symposium Presentation

Dr. Jessica Fossum [1], ***Dr. Amanda Montoya*** [1]

1. *University of California, Los Angeles*

Many factors are important to consider in power analysis for mediation and moderated mediation models. Mediation models require researchers to choose which test of mediation to use (e.g., bootstrapping, Sobel test). But how similar are the power recommendations and sample size estimates among these tests? For example, if the percentile bootstrap confidence interval requires 500 participants to achieve 80% power of the indirect effect, does the joint significance test come up with a similar power for 500 people using the same effect sizes? Understanding these similarities is advantageous for researchers wanting to use the more computationally efficient joint significance test for power analysis where the test must be performed thousands of times, then still use the recommended yet more time consuming percentile bootstrap confidence interval for the data analysis. We compared the power estimates from six commonly used inferential tests with respect to recommended sample sizes, and found that the joint significance test performs similarly to the percentile bootstrap confidence interval and the Monte Carlo confidence interval. Additionally, we explore the impact of assumption violations (e.g. nonnormality and heteroskedasticity) on the similarity of sample size recommendations among the tests. We discuss how to use the power analysis tool pwr2ppl for mediation models, including several moderated mediation models (i.e., where the mediation models include interaction effects with moderator variables), and provide suggestions for when the model is complex enough that tools are not yet available and simulation methods are necessary for power analysis.

Fast power computations in multilevel models for intensive longitudinal designs

Tuesday, 12th July - 15:20: Symposium: Advances in methods for sample size determination in the social sciences (Room C) - Symposium Presentation

Dr. Ginette Lafit [1], ***Dr. Richard Artner*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven*

To unravel how within-individual psychological processes fluctuate in daily life, and how these processes differ between persons, intensive longitudinal (IL) designs have become popular. In IL research, hypotheses are usually tested using multilevel regression models. An important question in the design of IL studies is how to determine the number of participants needed to conduct statistical inference for model parameters with an accuracy that is sufficiently high to detect realistic effect sizes with high probability. Recent advances in computational methods and software have put forward simulation-based approaches for conducting power analysis for testing fixed regression coefficients derived from a wide variety of multilevel models. Unfortunately, this approach is highly computationally intensive. We, therefore, derive analytic formulas to compute statistical power for many commonly used multilevel models. This is done by obtaining closed-form expressions of the information matrix of the fixed effect parameters for multilevel models that are popular in intensive longitudinal research, including models with auto-correlated errors and Longitudinal Actor-Partner Interdependence Models. However, when no closed-form expression exists (e.g., multilevel AR(1) and VAR(1) models), we combine analytical derivations and Monte-Carlo simulations to estimate power with similar accuracy but much smaller computational effort than existing simulation-based approaches.

Predictive accuracy analysis: A new sample size planning method

Tuesday, 12th July - 15:35: Symposium: Advances in methods for sample size determination in the social sciences (Room C) - Symposium Presentation

Mr. Jordan Revol [1], ***Dr. Ginette Lafit*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven*

In the behavioral sciences, most research on sample size planning has focused on two interrelated purposes: optimizing power or optimizing the accuracy of parameter estimates. These two purposes have in common that they take an explanation perspective: they presume that estimated parameters are useful to explain observed scores and shed light on underlying processes. Meanwhile multiple authors (e.g., Yarkoni & Westfall, 2017) have pointed towards the added value of a predictive perspective. In this perspective, one evaluates how well a fitted model generalizes as a whole to unseen data. Taking this predictive perspective, it has been shown that complex time series models, such as VAR(1) models, tend to overfit the data, hampering prediction of new data as well as further interpretation of the separate model parameters (e.g., Bulteel et al., 2018; Lafit et al., 2021; Mansueto et al., 2022). We therefore argue for the added value of basing sample size planning on predictive accuracy estimates. Focusing on AR(1) and VAR(1) in a N=1 context, we propose a novel simulation-based method, called predictive accuracy analysis, to assess how many measurement occasions are required in order to optimize predictive accuracy. To this end, we present a new, easy- to-interpret predictive accuracy metric. We then showcase how the different model parameters impact predictive accuracy and thus sample size planning, based on deliberately chosen parameter values as well as parameter estimates for a large set of empirical data. Similarities and differences between the power analysis and predictive accuracy analysis results will be highlighted.

A simple method for handling reflective invariance in Bayesian IRT

Tuesday, 12th July - 14:50: Bayesian SEM (Room D) - Individual Oral Presentation

Mr. Keishi Nomura [1], ***Dr. Shiro Kumano*** [2], ***Dr. Kensuke Okada*** [1]

1. *The University of Tokyo,* ***2.*** *Nippon Telegraph and Telephone Corporation*

Reflective invariance is inherent in latent variable models. It yields multimodalities in likelihood: a mixture of 2 to the q-th power unimodal likelihoods for a q-factor model. Item response theory (IRT) models usually resolve this invariance by constraining discrimination parameters of all items to be positive. In some applications, however, negative discrimination is meaningful, and thus the constraint cannot be readily applied, which presents some difficulties for Bayesian estimation via multiple Markov chains. In this study, we investigated the capability of initial value specification to choose an IRT solution that corresponds to a unimodal density. Specifically, we propose setting initial values for thetas to standardized scores and/or setting initial values for discrimination parameters to be positive. The rationale is that under a reasonable condition of data size and hyperparameters of Markov chain Monte Carlo (MCMC) algorithm such as step size and rejection rate, individual unimodal densities are clearly separated, and MCMC sampler gets stuck at the unimodal density closest to its initial values. A simulation study of online customer ratings analogous to Su, Chang, & Weng (2020) shows that our method correctly selects a unimodal density even when there is an extremely small unbalance in the proportion of positive and negative discrimination parameters. Our method provides a handy alternative to other methods for handling reflection invariance in Bayesian IRT, such as parameter constraints and post-processing of MCMC output.

Bayesian evaluation of approximate measurement invariance

Tuesday, 12th July - 15:05: Bayesian SEM (Room D) - Individual Oral Presentation

Ms. Dandan Tang [1], ***Dr. Xin Gu*** [2], ***Dr. Caspar Van Lissa*** [1], ***Prof. Herbert Hoijtink*** [1]

1. *Utrecht University,* ***2.*** *East China Normal University*

Measurement invariance (MI) is of vital importance in multiple-group research with latent factors. Latent factors in CFA models should be under the same measurement scale to make corresponding loadings or intercepts comparable across groups. In practice, exact MI is often not met, and approximate MI offers a working solution. This article thus proposes a Bayesian approach to test approximate MI in two-group confirmatory factor analysis (CFA) models using Bayes factors. This approach fixes the product of loadings to 1 and the sum of intercepts to 0. These corresponding comparable loadings or intercepts in two groups are about equal in the hypotheses of approximate MI. To test these approximate hypotheses, researchers have to carefully specify prior distributions for loadings and intercepts, as well as tolerant differences under approximate equality. Simulation studies explore prior choices and tolerance differences. A flowchart of testing approximate MI is presented. The method is implemented in the function BMI() in the bain R-package, and a real data example was used to illustrate the procedure.

How does prior variance affect local dependence detection in CFA

Tuesday, 12th July - 15:20: Bayesian SEM (Room D) - Individual Oral Presentation

Ms. Xinyu Qiao [1], ***Prof. Junhao Pan*** [1]

1. *Sun Yat-sen University*

Local independence is a fundamental assumption in confirmatory factor analysis (CFA). However, several factors (e.g., similar expression in the questionnaire) could result in violation of it. The Bayesian structural equation model can be used to deal with local dependence. Researchers can allow the existence of small local dependence by assigning a zero-mean and small-variance prior to the residual covariance matrix. It's still unclear how prior variance affects the detection of local dependence. Therefore, two simulation studies based on CFA model were conducted with continuous and binary data, respectively. The conditions included sample sizes, number of items, the magnitude of factor loadings, and levels of local dependence. Inverse Wishart distribution was applied for the residual covariance matrix. Five levels of prior variances and the default prior setting in M*plus* for residual covariances were considered. Results indicated that models were not identified for continuous data, and the convergence rate was low for binary data under default prior setting. For continuous data, larger prior variance resulted in less Type I error. The power to detect small local dependence (e.g., residual correlation = 0.2) was lower than 0.1 regardless of prior variances. For binary data, the prior variance mainly affects the accuracy of the estimates of the residual covariance. With the increase of prior variance, the mean square error and relative bias decreased. Power for detecting small or moderate dependence was always less than 0.8 regardless of prior variances. We provided recommendations and two real data analyses to demonstrate the validity of this recommendation.

BSEM modification indices: missed parameters and how to find them

Tuesday, 12th July - 15:35: Bayesian SEM (Room D) - Individual Oral Presentation

Dr. Mauricio Garnier-Villarreal [1], ***Dr. Terrence Jorgensen*** [2]

1. *Vrije Universiteit Amsterdam,* ***2.*** *University of Amsterdam*

An intrinsic part of theoretical latent variables models is that they are simplifications of complex relations between variables. This leads to every model being incorrect to some degree, which is usually evaluated with a variety of fit measures such as exact and approximate ones. But when it is deem that the degree of misfit is too large, the next step is to identify which parameters that were previously fix at 0 should be estimated to improve the overall model fit. In frequentis SEM (f-SEM) modification indices (MI) present the expected improve in model fit, and the expected parameter change (EPC) present the approximate magnitude of the respective parameter. But this hasn't been tested in BSEM yet. We are presenting the results of a simulation testing the performance of MI and standardized EPC (SEPC) to identify meaningful missed parameters in BSEM. In f-SEM, MI and SPEC have been use based on recommended cutoffs (such as |SEPC| > 0.1). In BSEM we have the advantage evaluating MI and SEPC at each iteration of the posterior distribution, creating distributions of realized values for both MI and SEPC, evaluating uncertainty on their estimates. This way we are able to make inferences about MI and SEPC based on the magnitude of their point estimates (mean or median), and the credible interval of their realized values distributions. From our results we would present practical recommendations for applied researchers to identify which parameters should be added based on MI and SEPC accounting for uncertainty

An iterative approach to flexible multivariate prior elicitation

Tuesday, 12th July - 15:50: Bayesian SEM (Room D) - Individual Oral Presentation

Ms. Sydne McCluskey [1], ***Dr. Jay Verkuilen*** [1]

1. *CUNY Graduate Center*

Although Bayesian methods have steadily gained popularity over the past two decades, implementation of informative priors is still relatively rare. This is partly due to a dearth of literature demonstrating how to select informative priors and examine their appropriateness and impact on analytic results. We present an empirical illustration of prior selection via an iterative procedure of multivariate prior elicitation from domain experts and feedback via prior predictive checks with data from observers using the Baker Rodrigo Ocumpaugh Monitoring Protocol in college classrooms. This is followed up with a full sensitivity analysis to elucidate the impact of the chosen prior on analytic results. We focus on a flexible approach that can apply to prior elicitation in parameter or observable space, whichever is most appropriate to tapping into expert knowledge and achieving analytic objectives. Most models involve multiple parameters, but the prior elicitation literature focuses primarily on univariate priors. We consider different approaches to multivariate prior elicitation which more faithfully represent prior beliefs about the joint distribution of parameters. We demonstrate how to transparently communicate the results of a sensitivity analysis when estimates are not robust to perturbations of the prior without characterizing this finding as necessarily problematic. Rather than suggesting our approach is definitive, our goal is to stimulate discussion around key areas of prior elicitation from domain experts: eliciting multivariate priors, developing flexible elicitation techniques that can apply to parameter or observable space, and examining the consequences of prior selection through prior predictive checks and sensitivity analysis.

Evaluation of common items using DIF on longitudinal TIMSS datasets

Tuesday, 12th July - 14:50: Large-scale assessments (Room G) - Individual Oral Presentation

Dr. Youn-Jeng Choi [1], ***Ms. Yelin Gwak*** [1], ***Ms. Hunwon Choi*** [1], ***Mrs. Eunjeong Jeon*** [1]

1. *EWHA WOMANS UNIVERSITY*

The TIMSS presents an ideal opportunity to compare U.S. students' mathematics and science achievement at fourth and eighth grade with their peers worldwide. Examiners often repeatedly use a set of items to scale examinee performance on multiple test cycles. These repeatedly administrated items are common, anchor, or trend items. Common items are important because it is impossible to link and equate items across different test cycles without them. Thus, it is unlikely to explore student achievement performance changes over time. We can evaluate the quality of common items using differential item functioning (DIF). The purposes of this study are (1) to apply the mixture IRT model, which is one of the methods to determine latent classes, to common items in mathematics on the TIMSS datasets during 2007–2019, (2) to investigate latent class DIF of common items in each cycle of TIMSS, and (3) to apply the IRT model to detect DIF using manifest groups such as race, gender, English usage and so on. We will use the OpenBUGS, WINMIRA, or R computer software to detect DIF. Our contribution is to offer the initial insight into the DIF of mathematics common items on TIMSS datasets administered during 2007–2019.

Concordance for large scale assessments

Tuesday, 12th July - 15:05: Large-scale assessments (Room G) - Individual Oral Presentation

Dr. Liqun Yin [1], ***Dr. Matthias von Davier*** [1], ***Dr. Lale Khorramdel*** [1], ***Mr. Pierre Foy*** [1], ***Mrs. Ji Yoon (Jenny) Jung*** [1]

1. *Boston College TIMSS and PIRLS International Study Center*

Interest has grown recently in linking national or regional assessments to international large-scale assessments. However, commonly used equating and linking methods are not defensible for such purposes as they would make unrealistic assumptions such as construct equivalency and error free measurement, and usually only provide a point to point projection.

This paper introduces a new approach for score projections by constructing an enhanced concordance table between two large-scale assessments with one source and one target tests. More specifically, the proposed method employs predictive mean matching (PMM; Little, 1988; Rubin, 1986) method to find a set of donors with the smallest distances to the predicted mean generated by an imputation model on the source test for each concordance score level within the identified score range. Both the means and standard deviations of donors' plausible values on the target test are utilized to construct a concordance table between the two tests with the predicted conditional distributions (means and SDs) included. This approach ensures not only the score uncertainty due to measurement error and imperfect correlation between tests appropriately are taken into account, but also avoids complex statistical functional forms and linearity assumption.

The robustness of the new approach is demonstrated in a linking study to relate a regional assessment to TIMSS and PIRLS international long-standing large-scale assessments, where students take both the source and the target tests. Recommendations for educators and researchers to make inferences and interpret the concordance table are also provided.

Consequences of hierarchical data structures for the estimation of plausible values

Tuesday, 12th July - 15:20: Large-scale assessments (Room G) - Individual Oral Presentation

Ms. Eva Zink [1], ***Prof. Sabine Zinn*** [2], ***Dr. Timo Gnambs*** [1]

1. *Educational Measurement, Leibniz-Institute for Educational Trajectories,* ***2.*** *Deutsches Institute für Wirtschaftsforschung Berlin*

Educational large-scale assessments (LSAs) measure domain-specific competences and cognitive abilities to evaluate and monitor educational processes within and between countries. Thereby, competence scores are typically represented by plausible values (PVs) that allow the analysis of latent effects. So far, an unresolved challenge for estimating PVs is the multilevel structure that arises when students are clustered in different school contexts. Current practices in LSAs involve either ignoring the hierarchical data structure or including cluster-specific mean scores in the background model for estimating PVs. However, both approaches might bias analyses if the generated PVs do not appropriately reflect the multilevel structure. Therefore, a Monte Carlo simulation evaluated the consequence of either ignoring or including the hierarchical data structure in the background model of the PVs estimation. Moreover, alternatively to current practice, we compare whether including a random effect in the background model accounting for the hierarchical data structure improves the accuracy of parameters. The simulation study mimicked realistic data in LSAs and evaluated the performance of the different approaches in an experimental 3 x 2 design formed by the PV estimation model (ignoring, cluster scores, random effects) and the analysis model (ignoring, random effects). We simulated responses of respondents nested in clusters to 20 items conforming to the one-parametric item response model. The bias, relative bias, and root mean squared error of parameters resulting from regressing the PVs on a continuous and a binary predictor were our main evaluation criteria. Based on the results of these analyses, recommendations for practice are provided.

Motivation towards Mathematics from 1980 to 2015: Exploring the feasibility of longitudinal scaling

Tuesday, 12th July - 15:35: Large-scale assessments (Room G) - Individual Oral Presentation

Ms. Erika Majoros [1], ***Dr. Andrés Christiansen*** [2], ***Dr. Edwin Cuellar*** [3]

1. *University of Gothenburg,* ***2.*** *International Association for the Evaluation of Educational Achievement,* ***3.*** *Teach for Colombia*

The Trends in International Mathematics and Science Study (TIMSS) has been assessing students' attitudes towards learning mathematics and science every fourth year since 1995. However, the longitudinal scaling of certain non-cognitive constructs started only in 2011. This study explored the feasibility of establishing long-term extrinsic- and intrinsic motivational scales based on TIMSS data extended with the Second International Mathematics Study administered between 1976 and 1982. We used grade-eight data from five educational systems that have participated in every time point up to TIMSS 2015. First, cross-cultural and longitudinal comparability were evaluated with multiple-group confirmatory factor analysis and the delta plot method. Second, three methods for linking the assessments were investigated: an item response theory-, a confirmatory factor analysis-, and a market-basket approach. The three methods yielded similar results on the individual- as well as the country levels. The trend descriptions show that in Hong Kong, the average level of intrinsic motivation has been consistently higher than the average extrinsic motivation. The average extrinsic motivation shows very similar trends in Israel and England. The average intrinsic motivation has changed substantially in England. Hungary's trend descriptions of the two types of motivation show an interesting mirroring. Finally, the trends of the United States show the most stable average motivation level in this set of countries. The methodological implication of this study is that the market-basket approach may be applied to include more data in the longitudinal scales without the need for recalibrating the original scales or employing equating methods.

Bayesian dynamic borrowing with longitudinal large-scale assessment data

Tuesday, 12th July - 15:50: Large-scale assessments (Room G) - Individual Oral Presentation

Prof. David Kaplan [1], ***Dr. Jianshen Chen*** [2], ***Mr. Weicong Lyu*** [1], ***Mr. Sinan Yavuz*** [1]

1. *University of Wisconsin - Madison,* ***2.*** *The College Board*

The purpose of this paper is to extend and evaluate Bayesian dynamic borrowing originally developed by Viele (2014) for case-control trials and extended to multilevel cross-sectional large-scale educational assessments by Kaplan, et al., (2022) to the case of longitudinal data. A joint prior distribution over the historical and current data sets is specified with the degree of heterogeneity across the data sets controlled by the variance of the joint distribution. To date, this method has not been applied to longitudinal data where early cycles of longitudinal data could be used to inform prior distributions for current cycles of longitudinal. For this paper, we focus attention on the United States Early Childhood Longitudinal Study: Kindergarten cohort of 2011, and dynamically borrow from the Early Childhood Longitudinal Study: Kindergarten cohort of 1998. We focus attention on the utility of dynamic borrowing for estimation of growth parameters and evaluate its performance against other historical borrowing methods such as complete pooling, Bayesian synthesis and power priors (Ibrahim & Chen, 2000) in terms of in-sample (historical) simulation statistics based on early econometric work by Theil (1966) , as well as pseudo out-of-sample measures of prediction such as the Kullback-Liebler divergence (Kullback, 1987). Results of case studies and simulation studies reveal the utility of dynamic borrowing in terms of improved in-sample and out-of-sample predictive performance.

Latent thresholds in classification tasks: A novel statistical model

Tuesday, 12th July - 14:50: Rater Models and Related Topics (Room E) - Individual Oral Presentation

Mr. Giuseppe Mignemi [1], ***Dr. Antonio Calcagnì*** [1], ***Prof. Andrea Spoto*** [1]

1. *University of Padova*

Categorical rating scales with two or more ordered categories are widely used in many contexts such as screening and diagnostic assessment, quality control, sport refereeing and emergency. Each classification task is influenced by the threshold of the rater, that is a reference point or a criterion used by the rater to classify into the ordered categories. Recent research in biostatistics provided interesting attempts to model systematic individual variability in these tasks using GLMM, but did not provide a tailored model for this kind of processes. To fill this gap, in the present contribution we propose a novel approach to the statistical modeling of raters classification process in which each rater evaluates the belonging of an item to a certain category according to several ordered levels. The classification outcomes are described as a function of two independent latent sources of uncertainty: The first one is the extent to which the item belongs to a category; the second one is the individual threshold of each rater. Both of them can involve several covariates, as it usually happens in statistical models. Model parameters have been estimated using an ABC-based algorithm. Finally, the characteristics of the proposed method have been discussed by means of a real case study. Further developments of this conditional model might regard the improvement of inter-rater agreement estimation, the identification of different types of raters (e.g., conservative *vs.* liberal), or specific dependency structures among items (e.g., different clusters: the more *vs.* the less certain).

Optimal weights for compound scores derived from multiple raters

Tuesday, 12th July - 15:05: Rater Models and Related Topics (Room E) - Individual Oral Presentation

Prof. Cees Glas [1]

1. *University of Twente*

In the fields of psychology, sociology, health, educational measurement and epidemiology, multiple measures are often combine into an overall index, i.e., a compound, or composite score. Van Lier, Siemons, van der Laar and Glas (MBR, 2018) proposed a method for constructing indices as linear functions of variables such that the reliability of the compound score is maximized. The measures that comprise the compound score can be a combination of directly observable variables, and latent variables issued by itemized instruments.
This approach is generalized here to a situation where one or more measures comprise of itemized observations made by several observers and possibly pertaining to multiple tasks. Response data on itemized observation instruments are modelled by an IRT model. The complete model is a combination of a multidimensional IRT model and a generalizability theory model. A method is proposed for the computation of weights for the constituent measures of the compound score that maximize reliability and agreement indices. Computations are made in a Bayesian framework using Markov chain Monte Carlo (MCMC) computational methods. Optimal weights for the components are found by maximizing the posterior variance relative to the total relevant variance. An important advantage of the Bayesian method is that credibility regions for all parameters of the model, including reliability indices and indices of agreement, can be straightforwardly computed. Analysis of an instrument for the observation of teacher classroom behavior is presented as an empirical example of the approach.

Examining rater bias using IRT cross-classified multilevel modeling

Tuesday, 12th July - 15:20: Rater Models and Related Topics (Room E) - Individual Oral Presentation

Dr. Nai-En Tang [1], ***Mr. H. Daniel Edi*** [2], ***Dr. Igor Himelfarb*** [1]

1. *National Board of Chiropractic Examiners,* ***2.*** *University of Northern Colorado*

Although achieving reliability of objective structured clinical examinations (OSCEs) is of particular importance, rater bias may influence the test scores, even after extensive training. Therefore, several modeling approaches have been introduced to assess rater effects. The purpose of this study is to compare the performance of the conventional multilevel model and the cross-classified random effect model (CCrem) under the classical test theory and item response theory frameworks, respectively, in examining potential rater bias in an OSCE. Four models (i.e., a conventional multilevel model and a CCrem under each framework) were applied to a 10-station exam rated by 160 examiners for 677 examinees. The estimates of examinees' ability obtained with the conventional multilevel models were compared to the one obtained with cross-classified models in order to evaluate the effects of potential rater bias. The results showed that the variability of examinee random effect increased by 38% and 24% after rater random effect was accounted for in the model under the Classical Test Theory (CTT) and Item Response Theory (IRT) frameworks, respectively. The result also suggested that using the CCrem under the CTT and the IRT frameworks did lead to a significantly improved fit over the conventional CTT and IRT models, respectively. These results confirmed that rater effect cannot be ignored under both the CTT and IRT frameworks as they introduce bias in the ability estimates. Overall, the results provided evidence for the use of a cross-classified random effect model under the IRT framework

Evaluating quality of selection procedures in a binary classification framework: An alternative to inter-rater reliability

Tuesday, 12th July - 15:35: Rater Models and Related Topics (Room E) - Individual Oral Presentation

Mr. František Bartoš [1], ***Dr. Patricia Martinkova*** [2]

1. *Czech Academy of Sciences and University of Amsterdam,* ***2.*** *Czech Academy of Sciences*

Inter-rater reliability (IRR) has been the prevalent quality and precision measure in ratings from multiple raters. However, the selection procedures employing raters and ratings usually result in a binary outcome: An applicant is either hired or refused, a grant proposal is either funded or dismissed, and an article is either accepted or rejected. This final outcome is not considered in IRR, which instead focuses on ratings of the individual subjects or objects. In this work, we outline how to transform the ranking and selection procedures into a binary classification framework and develop a quantile approximation that connects the measurement model for the ratings with the binary classification framework. The quantile approximation allows us to obtain the probability of correctly selecting the best candidates and to assess the error probabilities when evaluating the quality of selection procedures using rankings based on ratings from multiple raters. We assess the performance of the quantile approximation in a simulation study and we draw connections between the binary classification metrics and the inter-rater reliability.

A simulation-based test to investigate interrater agreement for binary time series

Tuesday, 12th July - 15:50: Rater Models and Related Topics (Room E) - Individual Oral Presentation

Dr. Nadja Bodner [1], ***Prof. Guy Bosmans*** [1], ***Prof. Francis Tuerlinckx*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven*

In observational studies, intensive longitudinal data are often collected by coding the presence/absence of behavior across time. Having a sufficiently high interrater agreement is then quintessential. Although a host of alternatives exist, Cohen's Kappa has been most popular ever since the measure has been introduced. In many cases, the obtained interrater agreement values are interpreted using the benchmarks provided by for example Landis and Koch. This is, however, problematic because of two reasons: First, the value of many interrater agreement measures is impacted by the prevalence of the coded behavior. Second, these measures ignore the serial dependence that typically characterizes time series data. In this study, we aim to bring the focus back on a pivotal feature of agreement: the error rate committed during the coding process. We start by deriving which error rates correspond to the Landis and Koch benchmarks given a known prevalence and the absence of serial dependence. Next, we show how to assess and test the error rates of reported data, by introducing a simulation paradigm that accounts for the serial dependence and relative frequency of these data. In this paradigm, data are simulated by implementing a specific error rate, that can be chosen based on the above benchmark results. The simulations yield a sampling distribution of plausible interrater agreement values under the Null hypothesis that the error rate underlying the reported data corresponds to the implemented rate, paving the way for a one-sided significance test.

Individual dynamic models using regularized hybrid unified structural equation modeling

Tuesday, 12th July - 14:50: Dynamic SEM (Room F) - Individual Oral Presentation

Ms. Ai Ye [1]

1. *University of North Carolina at Chapel Hill*

With the development of technology to collect time series data in a fast and economical fashion, recent decades have witnessed another surge of psychological and neurological research at an individual level. One goal in such endeavors is to construct person-specific dynamic assessments using time series techniques such as Vector Autoregressive (VAR) models. Within the psychometric field, researchers have developed psychometric methods to fit variants of VAR models. However, these methods are often limited in VAR representations and/or model selection regimes. The current research aims to evaluate and overcome these limitations. Specifically, I proposed a novel modeling approach that uses LASSO regularization under the unified Structure Equation Modeling (uSEM) to estimate a more flexible VAR representation, called regularized hybrid unified SEM (or Reg-huSEM). The first simulation study has shown that the proposed approach is more reliable and accurate than alternative methods in recovering hybrid types of dynamic relations and in eliminating spurious ones. Lastly, I extended this Reg-huSEM approach to estimate individual dynamic models with latent variables (i.e., dynamic factor models). In the second simulation study, besides model selection and path recovery performance, I also examined unbiasedness and robustness under model structure misspecification of parameter estimates obtained from three estimation methods including regularization, pseudo-ML, and model implied instrumental variable using two-stage least square estimation. The present work, to my knowledge, is the first application of the recently developed regularized SEM technique to the estimation of a time series SEM, which points to a promising future for statistical learning in psychometric models.

Forecasting university student drop out in math with ILD

Tuesday, 12th July - 15:05: Dynamic SEM (Room F) - Individual Oral Presentation

Prof. Augustin Kelava [1], ***Dr. Pascal Kilian*** [1], ***Dr. Judith Glaesser*** [1], ***Prof. Samuel Merk*** [2], ***Prof. Holger Brandt*** [1]

1. *University of Tuebingen,* ***2.*** *Karlsruhe School of Education*

The longitudinal process that leads to university student drop out in math can be described by referring to a) inter-individual differences (e.g., cognitive abilities) as well as b) intra-individual changes (e.g., affective states), c) (unobserved) heterogeneity of trajectories, and d) time-dependent variables. Large dynamic latent variable model frameworks for intensive longitudinal data (ILD) have been proposed which are (partially) capable of simultaneously separating these complex data structures (e.g., DLCA; Asparouhov, Hamaker, & Muthén, 2017; DSEM; Asparouhov, Hamaker, & Muthén, 2018; NDLC-SEM, Kelava & Brandt, 2019). From a methodological perspective, forecasting in dynamic frameworks allowing for real-time inferences on latent variables based on ongoing data collection has not been an extensive research topic. From a practical perspective, there has been no empirical study on student drop out in math that integrates ILD, dynamic frameworks, and forecasting of critical states of the individuals allowing for real-time interventions. We show how Bayesian forecasting of multivariate intra-individual variables and time-dependent class membership of individuals (affective states) can be performed in these dynamic frameworks using a Forward Filtering Backward Sampling algorithm. As an illustration, we use an empirical example where we apply the forecasting method to ILD from a university student drop out study in math with multivariate observations collected over 50 measurement occasions from N=122 students. More specifically, we forecast emotions and behavior related to drop out. This allows us to predict emerging critical dynamic states (e.g., critical stress levels or pre-decisional states) 8 weeks before the actual drop out occurs.

Early response to treatment in anorexia nervosa: A dynamic study

Tuesday, 12th July - 15:20: Dynamic SEM (Room F) - Individual Oral Presentation

Mrs. Nuria Real-Brioso [1], ***Mrs. Ani Laura Ruiz-Lee*** [1], ***Dr. Bronwyn C. Raykos*** [2], ***Mr. David M. Erceg-Hurn*** [2], ***Dr. Ricardo Olmos*** [1], ***Dr. Eduardo Estrada*** [1]

1. *Universidad Autónoma de Madrid,* ***2.*** *Center for Clinical Interventions, Perth*

Early response to treatment is one of the strongest predictors of intervention outcome in people with eating disorders (EDs). A recent meta-analysis by Chang et al. (2021) concluded that there is a lack of exhaustive investigation of this phenomenon, and its operationalization differs considerably among studies. Latent Change Score (LCS) models provide a powerful tool to study early response, due to its ability to analyse such phenomenon as a dynamic system. We studied two common indicators of early response: body mass index (BMI) and global score in the Eating Disorder Examination Questionnaire (EDE-Q). They were measured in a sample of people with anorexia nervosa. Using bivariate LCS models, we found that a) the change mechanisms of each process appear to be independent from the other, b) change in BMI is mainly linear; and c) change in EDE-Q is partly linear, and partly predicted by its own rate of change in the previous weeks (i.e., an self-feedback component, ϕ_Y, which captures the influence from latent change during the previous interval (t-1) to the subsequent change (t) in the same variable). We discuss clinical implications and methodological considerations of our findings. The present work evidences the potential of dynamic models in the research of eating disorders and changes through interventions.

The growth components approach: Recent developments, extensions, and applications

Tuesday, 12th July - 15:35: Dynamic SEM (Room F) - Individual Oral Presentation

Prof. Axel Mayer [1]

1. *Bielefeld University*

The so-called growth components was originally proposed as a flexible alternative to classic growth curve models and true change models (Mayer et al., 2012). The approach is briefly introduced using a recent example to model change of behavior problems over time in children with and without specific learning disorders. But the approach is more generally applicable. In this talk, it is shown how it can extended and recent developments are discussed. In particular, its extension to analyzing main and interaction effects in multifactorial repeated measures designs is presented. This can contribute insights into interindividual differences in effects, latent variables, and estimation algorithms from structural equation modeling to (latent) repeated measures ANOVA. Also, it is shown that models with method effects, in which method effects are defined as differences between true score variables, can be considered a special case of the growth components approach. The approach can be extended to multilevel structural equation models where growth components are defined at the within- and the between-level. Recently, the growth components approach has also been combined with pattern mining techniques, i.e., subgroup discovery and exceptional model mining, to find subgroups with unusual developmental trajectories. This exploratory data mining approach can help researchers to better understand heterogeneity in growth trajectories and to generate hypotheses for future research. The subgroup latent growth component approach is applied in an empirical example to find subgroups of university students that show unusual developmental trajectories of drop-out intentions.

Bayesian estimation of restricted latent class models: Extending priors, link functions, and structural models

Tuesday, 12th July - 16:25: Dissertation Prize: James Balamuta (Room B) - Individual Oral Presentation

Dr. James Balamuta [1]

1. *University of Illinois Urbana-Champaign*

Restricted latent class models (RLCMs) provide a pivotal framework for supporting diagnostic research that enhances human development and opportunities. In earlier research, the focus was on confirmatory methods that required a pre-specified expert-attribute mapping known as a Q matrix. Recent research directions have led to the creation of exploratory methodology that is able to infer the Q matrix without expert intervention. Within this talk, we focus on novel Bayesian formulation of a less restrictive monotonicity condition when estimating the underlying latent structure and attributes. Moreover, we extend the framework to allow for using logit-link function instead of the probit and addressing the dependency structure found among attributes with a higher-order model that generalizes under an exploratory factor analysis (EFA).

Conditional dependence between response time and accuracy in cognitive diagnostic models

Tuesday, 12th July - 16:25: Dissertation Prize: Ummugul Berzihan (Room A) - Individual Oral Presentation

Dr. Ummugul Bezirhan [1]

1. *Boston College TIMSS and PIRLS International Study Center*

Computer-based tests are rapidly becoming the common practice in educational and psychological testing, thus a considerable effort has been put into incorporating process data specifically response time (RT) into measurement models. The assumption of conditional independence between response accuracy and RT, given latent ability and speed, is commonly imposed in the joint modelling framework of response and RT. Recently several studies have shown violations of the conditional independence assumption, which prompted various models that accommodate conditional dependence of responses and RTs, especially in the Item Response Theory framework. Despite the widespread usage of Cognitive Diagnostic Models as formative assessment tools, little has been done in exploring the conditional joint modelling of responses and RTs within this framework. This research proposes a conditional joint response and RT model in CDM by using an extended reparametrized higher-order deterministic input, noisy 'and' gate (DINA) model for response accuracy. The item-specific effects of residual RT are incorporated into the response accuracy model to capture the conditional dependence. The effect of ignoring the conditional dependence on parameter recovery is explored with a simulation study, and empirical data analysis is conducted to demonstrate the application of the proposed model.

Inference with cross-lagged effects – Problems in time and new interpretations

Wednesday, 13th July - 09:15: Symposium: Beyond VAR: New developments in modeling psychological time series (Room B) - Symposium Presentation

Dr. Charles Driver [1]

1. *University of Zurich*

The interpretation of cross-effects from vector autoregressive models to infer structure and causality amongst constructs is widespread, and sometimes problematic. I first detail the known but not broadly understood issue of invalid hypothesis testing and regularization when processes that are thought to fluctuate continuously in time are, as is typically done, modeled as changing only in discrete steps. I then describe an alternative interpretation of cross-effect parameters that explicitly considers correlated random changes (while interpreting, not simply while modelling) for a potentially more realistic view of how process are temporally coupled. Using an example based on wellbeing data, I demonstrate how some classical concerns such as sign flipping and counter intuitive effect directions can disappear when using this combined deterministic / stochastic interpretation. Models that treat processes as continuously interacting offer both a resolution to the hypothesis testing problem, and the possibility of the combined stochastic / deterministic interpretation.

Mapping modality in emotion time Series

Wednesday, 13th July - 09:30: Symposium: Beyond VAR: New developments in modeling psychological time series (Room B) - Symposium Presentation

Dr. Jonas Haslbeck [1], ***Dr. Oisín Ryan*** [2], ***Mr. Fabian Dablander*** [1]

1. *University of Amsterdam,* ***2.*** *Utrecht University*

The ability to measure emotional states throughout the day using mobile devices is a game-changer for emotion research and has already led to a surge of exciting new research on the temporal evolution of emotions. However, much of the potential of these data still remains untapped. In this paper, we re-analyze emotion measurements from seven Experience Sampling Methodology studies (with a total of 835 individuals) to map out the modality of emotion measurements (e.g., unimodal, bimodal). We show that a large proportion of emotion variables exhibit multiple modes. In addition, we show that modality varies both across items and individuals. The presence of multimodality implies that time series models with unimodal distributions such as the Vector Autoregressive model are only poorly capturing the data and that other approaches are needed. From a theoretical perspective, the finding that emotion measurements have multiple modes is critical because it establishes a key phenomenon that should be explained by future theories of emotion dynamics.

Machine learning for clustering ecological momentary assessment time-series data

Wednesday, 13th July - 09:45: Symposium: Beyond VAR: New developments in modeling psychological time series (Room B) - Symposium Presentation

Ms. Mandani Ntekouli [1], ***Prof. Gerasimos Spanakis*** [1]

1. *Maastricht University*

In the field of psychopathology, EMA methodological advancements have offered new opportunities to collect time-intensive, repeated and intra-individual measurements. This way, a large amount of data has become available, providing the means for further exploring mental disorders. Consequently, advanced machine learning (ML) methods are needed to understand data characteristics and uncover hidden and meaningful relationships regarding the underlying complex psychological processes. ML can also enhance the identification of similar patterns in data of different individuals through clustering. Various clustering techniques have been already developed and widely applied in time-series data. The focus of this talk is twofold: First, since clustering is an unsupervised problem, it remains a challenge to pick the preferred clustering method. Thus, several clustering methods based on different distance measures are investigated and assessed for the stability and quality of the derived clusters on real-world EMA datasets. Second, it is explored how patterns discovered among different individuals may provide complementary predictive capabilities. More specifically, it is examined how the performance of future-behavior predictive models can be improved, being applied in a group-based (nomothetic) approach. This aims to show that integrating data of several similar individuals in a single model could also accurately predict future outcomes compared to personalized models.

Equilibrium causal models: Connecting dynamical systems and cross-sectional data

Wednesday, 13th July - 10:00: Symposium: Beyond VAR: New developments in modeling psychological time series (Room B) - Symposium Presentation

Dr. Oisín Ryan [1], ***Mr. Fabian Dablander*** [2]

1. *Utrecht University,* ***2.*** *University of Amsterdam*

Increasingly, researchers have called for psychological phenomena to be characterized as dynamical systems: Processes that evolve and vary within-individuals over time. According to the principle of ergodicity, if individuals differ in the parameters governing their dynamical systems, then statistical models fit on certain types of cross-sectional data will not reflect the within-person properties of that system. This argument has in part motivated a large movement towards the use of time series data in psychology in order to gain insights into within-person dynamics.

Although (non-)ergodicity is clearly an important problem, the use of cross-sectional data is still widespread. Furthermore, it is clearly possible to study some types of systems using single-time-point data, while studying systems that evolve over a long timescale may not be feasible with time series approaches.

In this talk we introduce Equilibrium Causal Models (ECMs) for psychological phenomena. Equilibrium Causal Models capture the long-time-scale causal relationships between variables in a dynamical system and — crucially — can be learned from certain types of cross-sectional data. As such, Equilibrium Causal Models offer a tool by which we can connect dynamical systems ideas to cross-sectional data, allowing us to understand if, when, and how the latter can be used to make inferences about the former

Non-compliant survey responding: An IRTree model for dynamically changing response strategies

Wednesday, 13th July - 09:15: Symposium: Model based approaches to detecting item level non-compliant response behavior (Room A) - Symposium Presentation

Mrs. Viola Merhof [1], ***Prof. Thorsten Meiser*** [1]

1. *University of Mannheim*

Various psychometric approaches have been proposed to control self-reported trait measurements for response style effects that are assumed to be stable throughout a questionnaire. However, response style-based, heuristic responding may systematically increase over items and thereby take over from accurate, trait-based responding. Reasons for such a change in the response strategy could be, for example, fatigue, loss of motivation, and reduced test-taking effort. Here we present a dynamic IRTree model, which accounts for continuous shifts in response strategies by defining item position-dependent effects of the substantive trait and response styles. The model accurately detects and quantifies an increase in non-compliant response behavior throughout a questionnaire, and thus, is a useful tool for researchers to evaluate questionnaires with regard to the associated burden and required effort, or to compare motivational effects across measurement settings (e.g., online vs. lab survey). An empirical data set is used to illustrate the application of the dynamic model and to show the advantages over traditional IRT models.

A Bayesian latent class approach for the detection of automated survey responses

Wednesday, 13th July - 09:30: Symposium: Model based approaches to detecting item level non-compliant response behavior (Room A) - Symposium Presentation

Dr. Zachary Roman [1], ***Prof. Holger Brandt*** [2], ***Mr. Jason Miller*** [3]

1. *University of Zurich,* ***2.*** *University of Tuebingen,* ***3.*** *University of Kansas*

Online survey platforms such as Amazon's Mechanical Turk have many advantages. For example easy access to difficult to sample populations, a large pool of participants, and an easy to use implementation. A major drawback that has recently been discovered is the existence of bots that are used to complete online surveys for financial gain. These bots contaminate data and need to be identified in order to draw valid conclusions from data obtained with these platforms.

In this talk, we will provide a Bayesian latent class joint modeling approach that can be routinely applied to identify bots and simultaneously estimate a model of interest. This method can be used to separate the bots' response patterns from real human responses that were provided in line with the item content. The model has the advantage that it is very flexible and is based on plausible assumptions that are met in most empirical settings.

We will briefly present simulation results which validate model performance under realistic assumptions. Next we will illustrate the model and its capabilities with data from an empirical political ideation survey with known bots. We will discuss the implications of the findings with regard to future data collection via online platforms.

Experimental evidence for a dynamic latent class model of non-compliance

Wednesday, 13th July - 09:45: Symposium: Model based approaches to detecting item level non-compliant response behavior (Room A) - Symposium Presentation

Dr. Zachary Roman [1], ***Mr. Patrick Schmidt*** [1], ***Mr. Jason Miller*** [2], ***Prof. Holger Brandt*** [3]

1. *University of Zurich,* ***2.*** *University of Kansas,* ***3.*** *University of Tuebingen*

Non-compliant survey responders are those whom respond to surveys without considering the actual instructions or questions. Non-compliance based on inattention or careless behavior can be prominent in repeated measurements through online surveys or mobile phone apps. As such data becomes more important, we need methods that consider the dynamic nature of non-compliance and identify non-compliant survey respondents. We compare the performance of different approaches in an novel experimental study, complementing theoretical evidence and existing simulation studies. We employed a survey based experiment via the Mechanical Turk platform (MTurk) and primed half of the participants at a random point in time to stop paying attention to the content of the questions. Using this experimental data set, we compare the performance of several different approaches to analyze the data.

If non-compliance remains unaccounted, estimates are biased. Established methods of identifying non-compliant survey responders are mostly static in nature and fail if respondents become non-compliant late in the survey. A dynamic latent class model of non-compliance exhibits good model fit and accurately detects the timing of experimentally induced non-compliance. By directly accounting for non-compliance at the item level the model improves power and efficiency.

The novel experimental approach also allows us to improve estimation of non-compliant behavior in the control group of the MTurk sample. We observe strong compliance and little evidence of inattention or careless responses.

An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures

Wednesday, 13th July - 10:00: Symposium: Model based approaches to detecting item level non-compliant response behavior (Room A) - Symposium Presentation

Ms. Esther Ulitzsch [1], ***Ms. Seyma Yildirim Erbasli*** [2], ***Ms. Guher Gorgun*** [2], ***Mr. Okan Bulut*** [2]

1. *IPN - Leibniz Institute for Science and Mathematics Education,* ***2.*** *University of Alberta*

Careless and insufficient effort responding (C/IER) on self-report measures results in responses that do not reflect the trait to be measured, thereby posing a major threat to the quality of survey data. Reliable approaches for detecting C/IER aid in increasing the validity of inferences being made from survey data. First, once detected, C/IER can be taken into account in data analysis. Second, approaches for detecting C/IER support a better understanding of its occurrence, which facilitates designing surveys that curb the prevalence of C/IER. Previous approaches for detecting C/IER are limited in that they identify C/IER at the aggregate respondent or scale level, thereby hindering investigations of item characteristics evoking C/IER. We propose an explanatory mixture item response theory (IRT) model that supports identifying and modeling C/IER on the respondent-by-item level, can detect a wide array of C/IER patterns, and facilitates a deeper understanding of item characteristics associated with its occurrence. As the approach solely requires raw response data, it is applicable to data from paper-and-pencil and online survey administrations. The model shows good parameter recovery and can well handle the simultaneous occurrence of multiple types of C/IER patterns in simulated data. The approach is illustrated on a publicly available Big Five inventory data set, where we found later item positions to be associated with higher C/IER probabilities. We gathered initial supporting validity evidence for the proposed approach by investigating agreement with multiple commonly employed indicators of C/IER.

Statistical techniques for analyzing dyadic interactions

Wednesday, 13th July - 09:15: Symposium: Tackling challenges in analyzing intensive longitudinal data (Room C) - Symposium Presentation

Ms. Sophie Berkhout [1]***, Dr. Noémi Schuurman*** [1]***, Prof. Ellen Hamaker*** [1]

1. *Utrecht University*

Dynamic models are becoming increasingly popular to study the dynamic processes of interpersonal interactions, where two interacting partners, or a dyad, are measured densely in time. However, connecting the statistical concepts of dynamic models to substantive ideas remains a major challenge. We propose data simulation as a tool to guide dyadic research by exploring the data patterns that are implied by various dynamic models. For instance, prior to data collection, data simulation may help inspire theory building, explore what dynamic models are relevant, and choose a corresponding research design. After data collection, data simulation may help to test theories by examining how well the model-implied data patterns match the observed data patterns. Guided by data simulations, I discuss a selection of the following models used for analyzing dyadic intensive longitudinal data: (a) the first-order vector autoregressive (VAR(1)) model; (b) the latent VAR(1) model; (c) the time-varying VAR(1) model; (d) the threshold VAR(1) model; (e) the hidden Markov model; and (f) the Markov-switching VAR(1) model. To make data simulation for these various models accessible also to researchers who do not have the necessary programming skills, we present a Shiny web application (i.e., Shiny app) which allows researchers to easily generate data from these models through a graphical user interface. My goal is to make these dynamic models more accessible to psychological researchers, so they can make more informed decisions on what modeling approach would fit their research question and data best.

It's all about timing: Exploring the consequences of choosing different temporal resolutions for analyzing passive measures

Wednesday, 13th July - 09:30: Symposium: Tackling challenges in analyzing intensive longitudinal data (Room C) - Symposium Presentation

Mrs. Anna Langener [1], ***Dr. Laura F. Bringmann*** [1], ***Dr. Gert Stulp*** [1], ***Dr. Andrea Costanzo*** [1], ***Mr. Raj Jagesar*** [1], ***Prof. Martien Kas*** [1]

1. *University of Groningen*

Social interactions are important in our daily life and crucial for mental health. Collecting data passively (e.g., through smartphones) offers great value to track (social) behavior by measuring individuals frequently over time with a low participant burden. Even though passive measures are increasingly used to capture the social environment, there are important methodological challenges that may hamper widespread uptake. A crucial one is the issue of time scale: the temporal resolution may have an impact on how results can be interpreted. Importantly, the choice of temporal resolution is rarely justified, and the impact is rarely studied. Here we address various time-related decisions researchers face when analyzing passive measures. We use data collected during the EU-funded PRISM study as an example. The PRISM study aims to explore behavioral markers for social withdrawal in schizophrenia and Alzheimer's disease and collects passively information on smartphone usage and GPS (via the Behapp application) in these patient populations and healthy controls. These smartphone measures need to be aggregated on a specific time scale to compute certain variables. We provide a walk-through on how to study different time scales. Our results indicate that choosing a different temporal resolution can have consequences for the data quality and results, leading to changes in the interpretation. We propose to investigate the consequences of choosing different temporal resolutions, by doing a multiverse analysis. This improves the replicability of findings and may combat the replications crisis that is in part due to researchers making decisions that are left implicit.

Feasibility of personalized network models: Statistical power and missing data

Wednesday, 13th July - 09:45: Symposium: Tackling challenges in analyzing intensive longitudinal data (Room C) - Symposium Presentation

Ms. Alessandra Mansueto [1], ***Prof. Reinout Wiers*** [1], ***Prof. Julia van Weert*** [1], ***Dr. Barbara Schouten*** [1], ***Dr. Sacha Epskamp*** [1]

1. *University of Amsterdam*

The idiographic network approach is becoming increasingly popular in psychology due to its potentials to (1) understand the development of psychopathology over time in one person, and (2) investigate how complex interactions between different variables characterise this development. The Graphical Vector Autoregressive model (GVAR) can be used to analyse intensive time-series data in one individual to graph personalized psychopathological mechanisms over time (temporal network) and at the same time point (contemporaneous network), potentially allowing to personalize treatment. However, intensive time-series data collected in single patients in clinical settings are characterised by missing data and limited numbers of time points, thus it is unclear whether personalized networks can be reliably estimated. To answer this question, we ran a large-scale simulation study with two psychological datasets where we investigated GVAR performance with different sample sizes, missing data percentages, and estimation methods. Results show that sensitivity is small with sample sizes usually available in clinical settings (time points = 75 and 100). It appears that we can aim to estimate the global structure of idiographic networks, but we are unlikely to recover all existing edges. Estimating temporal networks seems especially difficult, thus with around 100 time points the number of nodes should be limited as much as possible (e.g., around 6). FIML and the Kalman filter perform well with random item-level missing data, opening the way to planned missing data designs and adaptive testing. Possible solutions to the challenges related to low statistical power uncovered in this study are proposed.

Prospectively detecting mean and variance changes through statistical process control

Wednesday, 13th July - 10:00: Symposium: Tackling challenges in analyzing intensive longitudinal data (Room C) - Symposium Presentation

Ms. Evelien Schat [1], ***Prof. Francis Tuerlinckx*** [1], ***Prof. Bart De Ketelaere*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven*

Retrospective analyses of experience sampling (ESM) data have shown that changes in mean and variance levels may serve as early warning signals of an imminent depression. Detecting such early warning signs prospectively would pave the way for timely intervention and prevention. The exponentially weighted moving average (EWMA) procedure seems a promising method to scan ESM data for the presence of mean changes in real-time. Affective ESM data, however, violate assumptions of the EWMA procedure: the observations are not independent across time, often skewed distributed and characterized by missingness. To deal with these data characteristics, we compute and monitor the day averages rather than the individual measurement occasions with EWMA. Since simulation studies showed promising results, we use a similar approach for the detection of variance changes, where we compute and monitor day statistics of variability (i.e., variances, standard deviations and the natural logarithm of the standard deviations). Simulation studies again show good performance, and allow to provide recommendations on which statistic of variability to monitor based on the type of change (i.e., variance increase or decrease) one expects. This is good news because existing exponentially weighted procedures for monitoring variability directly (e.g., EWMV, EWMA-S^2) all have important downsides, preventing easy application. Additionally, we reflect on the design choices of ESM studies, as they also influence the performance of the EWMA procedure. Specifically, the number of beeps per day (i.e., sampling frequency), the distribution of the in-control data and the number of in-control days influence the performance of EWMA.

Assessing Bayesian fit of an item response theory model for psychological time series

Wednesday, 13th July - 10:15: Symposium: Tackling challenges in analyzing intensive longitudinal data (Room C) - Symposium Presentation

Mr. Sebastian Castro-Alvarez [1], ***Dr. Sandip Sinharay*** [2], ***Dr. Laura F. Bringmann*** [1], ***Prof. Rob R. Meijer*** [1], ***Prof. Jorge N. Tendeiro*** [3]

1. *University of Groningen,* ***2.*** *Educational Testing Service,* ***3.*** *Hiroshima University*

To assess model fit in the Bayesian framework, one can use posterior predictive model checking methods (PPMC), which are flexible tools that can be customized for specific models. Previously, PPMC methods have been proposed for traditional dichotomous and polytomous item response theory models (IRT). In this talk, we propose PPMC methods for a dynamic partial credit model (DPCM). In particular, the DPCM is an extension of the partial credit model that is suitable to analyze psychological N=1 time series. With the proposed PPMC methods, we aim to assess violations of the assumptions of the model such as unidimensionality and trend stationarity. We present results from a preliminary simulation study that tested the PPMC methods under different misfit conditions. Furthermore, the proposed PPMC methods are illustrated with an empirical example about affect dynamics.

Permutation-based generalizable profile analysis for explaining DIF using item features

Wednesday, 13th July - 09:15: IRT I (Room D) - Individual Oral Presentation

Dr. Jesper Tijmstra [1], ***Dr. Maria Bolsinova*** [1], ***Prof. Leslie Rutkowski*** [2], ***Dr. David Rutkowski*** [2]

1. *Tilburg University,* ***2.*** *Indiana University Bloomington*

Profile analysis (Verhelst, 2012) is one of the main tools for studying whether differential item functioning (DIF) of items on a test can be related to specific features of the items on the test (e.g., their response format, or the depth of knowledge required to answer the item). While relevant, profile analysis in its current form has two restrictions that limit its usefulness in practice: It assumes that all items on the test discriminate equally well, and it does not test whether conclusions about the item feature effects generalize outside of the test (i.e., whether they hold in general rather than just for the specific set of items on the test). This paper addresses both of these limitations, by generalizing profile analysis to work under the two-parameter logistic model and by proposing a permutation test that allows for generalizable conclusions about item-feature effects. This latter step is especially important in practice, where test developers will be interested in determining whether item-feature effects can be expected to hold for similar future tests as well, or whether observed patterns in DIF should not be expected to generalize beyond the particular considered test.

The developed methods are illustrated using PISA 2015 Science data, which show important differences between the results obtained using 'standard' profile analysis versus those obtained using its generalizable extension. The methods are evaluated using simulation studies that assess their Type I error rate and power.

Detection of differential item functioning using residuals from item trace lines

Wednesday, 13th July - 09:30: IRT I (Room D) - Individual Oral Presentation

Mr. Youngjin Han [1], ***Dr. Ji Seung Yang*** [1], ***Dr. Yang Liu*** [1]

1. *University of Maryland - College Park*

Various methods for detecting differential item functioning (DIF) have been developed within the multiple-group item response theory (IRT) framework including likelihood ratio tests or Wald tests. However, item response data from multiple groups are not necessarily always available to researchers. For example, when some of the previously calibrated items are suspected to exhibit DIF in a new sample, researchers might not have access to the previous calibration data. The purpose of this study is to suggest a DIF detection method that utilizes the residuals from the item trace lines, which do not require item response data but only item parameter estimates from the previous calibration. When the measurement invariance fails to hold between the groups, the model-implied trace line with the pre-existing item parameter estimates and the empirical trace line constructed with item response data from the new sample are expected to show some discrepancy. The study suggests a formal significance test for the discrepancy between the two trace lines. In particular, the transformation of generalized residuals from the item trace line (Haberman, Sinharay, & Chon, 2013) is used to derive the standard errors for certain functionals of the trace line that represent the item difficulty and item discrimination. The standard error also accounts for the sampling variability contained in the pre-existing item parameter estimates that are often treated as fixed values when applied to subsequent samples.

Proposing an EIRT approach that includes linguistic characteristics of items

Wednesday, 13th July - 09:45: IRT I (Room D) - Individual Oral Presentation

Ms. Magdalen Beiting-Parrish [1], ***Ms. Sydne McCluskey*** [1], ***Dr. Jay Verkuilen*** [1], ***Dr. Howard Eveson*** [1], ***Dr. Claire Wladis*** [2]

1. *CUNY Graduate Center,* ***2.*** *Borough of Manhattan Community College*

Standardized assessments are frequently used to make definitive decisions about students' futures; however, however, performance gaps have been documented on standardized assessments between genders (Kan & Bulut, 2014), ELLs and Native English speakers (U.S. Department of Education, 2009), and low and high socioeconomic groups (Reardon, 2012). One potential source of these differences is that the language demands of the items used on these exams are likely a large source of construct-irrelevant variance (Messick, 1989). Essentially, if the test item is so linguistically complex that this is a barrier to demonstrating knowledge, this is an addressable equity and fairness issue. Previous research has shown that high word counts (Martiniello, 2009), polysemous words (Lager, 2006), second-person pronouns (Walkington et al., 2018), prepositions and complex verbs (Shaftel et al., 2006), and overall linguistic complexity (Solano-Flores & Trumbul, 2003) all negatively impact student performance on mathematics exams. The present research proposes to add to the literature by using EIRT combined with NLP techniques to examine linguistic aspects of the items more thoroughly alongside examinee demographic characteristics. This study hopes to better understand how examinee and item characteristics and their interactions predict general IRT parameters as well as differential item functioning. The larger aim of this research is to model the specific impact that different linguistic structures have on student performance on standardized items with the goal of creating a set of heuristics for how to better construct items that are not linguistically biased against examinees to measure examinee ability more accurately.

Agreement among DIF detection algorithms: A multiverse analysis

Wednesday, 13th July - 10:00: IRT I (Room D) - Individual Oral Presentation

Dr. Veronica Cole [1], ***Mr. Conor Lacey*** [1]

1. *Wake Forest University*

Differential item functioning (DIF) threatens the validity of inferences made in latent variable analyses. A perennial problem is that, absent a priori hypotheses about which items have DIF based on which covariates, researchers do not know which DIF effects to include. This issue is compounded by the adoption of frameworks such as moderated nonlinear factor analysis (MNLFA; Bauer, 2017), which allow the incorporation of many covariates. There are many algorithms to help researchers detect DIF in these contexts, including IRT-based standards (e.g., Mantel-Haenszel, IRT-LR-DIF) as well as MNLFA-specific methods such as automated MNLFA (aMNLFA; Gottfredson et al., 2019) and regularized MNLFA (Belzak & Bauer, 2020). However, it is unknown whether the results from each of these methods agree with one another in terms of the DIF identified and the factor scores obtained. The current study attacks this question using two multiverse analyses. First, we apply multiple different methods for DIF detection to an adolescent sample (N = 663), fitting an MNLFA to scales of future expectancies among adolescents. We then fit the same models to simulated data with population values based on this dataset, adding DIF of varying size. In general, algorithms disagree in the DIF items identified. However, as long as the detected DIF was accounted for, factor scores arising from different models are generally highly correlated with one another, even if the nature of the DIF itself is misspecified. These findings underscore that there is no one “correct” way to find DIF.

The gradient test in a conditional likelihood framework

Wednesday, 13th July - 10:15: IRT I (Room D) - Individual Oral Presentation

Mr. Andreas Kurz [1], ***Dr. Clemens Draxler*** [1]

1. *UMIT—Private University for Health Sciences Medical Informatics and Technology*

This talk discusses the gradient test, proposed by Terrell, as a recent likelihood-based hypothesis testing approach. It can be considered as an alternative to the well-established trinity of likelihood ratio, Rao score, and Wald tests. The gradient test has not yet entered into the mainstream of applied statistics. This is particularly true for the psychometric context. This presentation illustrates a novel application of the gradient test within the conditional maximum likelihood and the Rasch modeling framework. Some of its finite sample size properties will be compared with the classical trinity of chi square tests by conducting an extensive Monte Carlo study. The results confirm that the gradient test has its pros and cons.

Two-way outlier detection for item response data

Wednesday, 13th July - 09:15: Outlier Detection (Room G) - Individual Oral Presentation

Dr. Gabriel Wallin [1], ***Dr. Yunxiao Chen*** [1], ***Prof. Irini Moustaki*** [1]

1. *London School of Economics and Political Science*

Detection of cheaters and compromised items in large-scale assessment tests has received much attention in recent years. However, most of the existing methods target either respondent or item outlier detection, but not both. Furthermore, prior knowledge about a subset of respondents who are not cheaters or a subset of items who are not compromised is typically required. We therefore propose a flexible item response theory (IRT) modelling framework that can simultaneously classify both respondent and item outliers and that can incorporate, but doesn't require any, prior information. The proposed model combines a baseline IRT model with an outlier component where respondent and item outliers are indicated through a binary random vector each. The two outlier terms are assumed to interact with each other and to have a sparse structure. The simultaneous classification is conducted in two steps. First, an initial estimate of the model parameters based on joint maximum likelihood is retrieved. In a second, post-estimation learning step, we seek a sparse representation of the outlier component of the model by rotating the parameter solution using a sparse-pursuit transformation and hard-thresholding. An algorithm based on iterative reweighted least squares is developed to find the rotation matrix which yields a sparse solution. Our simulation results are much promising, showing a very high rate of true positive and true negative classifications for both respondent and item outliers. Our method will furthermore be illustrated using empirical data and theoretically justified under a double asymptotics regime to give a classification consistency result.

Forward Search for IRT estimation and for atypical responses detection

Wednesday, 13th July - 09:30: Outlier Detection (Room G) - Individual Oral Presentation

Mrs. Anna Comotti [1], ***Prof. Matteo Bonzini*** [1], ***Mrs. Alice Fattori*** [2], ***Prof. Francesca Greselin*** [3]

1. *Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico,* ***2.*** *University of Milano,* ***3.*** *University of Milano-Bicocca*

We implement a forward search (FS) algorithm for identifying atypical subjects/observations in item response theory (IRT) models for binary data.
FS algorithm was initially introduced as an outlier detection tool for the estimation of covariance matrices (Hadi, 1992), and regression models (Atkinson et al. 2000). It was extended to standard multivariate methods (Atkinson et al. 2004) and factor analysis (Mavridis & Moustaki, 2009); it was recently applied in meta-regression (Petropoulou et al. 2021). In the IRT framework, all methods proposed for outlier detection start upon a first inference performed on the entire dataset. Therefore, they will not be preserved from the issue of masking and swamping. They are hence potentially unaffordable, and conclusions derived from them could be misleading.
Our proposal introduces diagnostic tools, based on robust methodologies, to avoid distortion in the estimation of the model, and to detect atypical response patterns. Forward plots of goodness of-fit statistics, residuals and parameter estimates allow to identify atypical observations and reveal deviations from the hypothesized model. Methods to initialize, progress and monitor the forward search are explored. Simulation envelopes are constructed to investigate whether changes in the statistics being monitored are solely due to random variation. One real and one simulated datasets are used to illustrate the performance of the suggested algorithm. The simulated dataset explores the effectiveness of the method in the presence of outliers. Real data come from questionnaires administered to a health-care population in a large hospital in Milan (Italy), whose mental well-being was evaluated during COVID-19 pandemic.

A robust item fit assessment

Wednesday, 13th July - 09:45: Outlier Detection (Room G) - Individual Oral Presentation

Dr. Ummugul Bezirhan [1], ***Dr. Matthias von Davier*** [2]

1. *Boston College TIMSS and PIRLS International Study Center,* ***2.*** *Boston College*

Item fit analysis is an integral part of any assessment development process in order to investigate whether an item performs in an expected manner, meaning the functional form of the item characteristic curve (ICC) follows the one implied by the model. Since detection of item misfit is critical in ensuring test validity, it is still an ongoing concern in educational and psychological testing. Classical approaches to item fit detection rely on null hypothesis significance testing even though it has been shown that these approaches are affected by sample sizes. Other residual-based measures, such as the root mean square deviation (RMSD), are sample size free, but with these measures, the question becomes which threshold to use. Thus, the question of determining a threshold remains unless one is willing to declare any empirical distribution found as a definitive one and use a critical value to flag item misfit just as in significance testing. The aim of this study is to utilize a robust measure of dispersion, the median absolute deviation (MAD) to detect misfit by identifying outliers in an RMSD distribution. The MAD classifies an observation as an outlier if the difference to the median of the other absolute distances exceeds a certain boundary. A real data application using PIRLS 2016 data, along with a number of simulation studies with various conditions performed to show the effectiveness of the proposed item fit diagnostic. The new method demonstrated a higher detection rate compared to the classical cut-of-value approach.

Person misfit and person reliability in rating scale measures: The role of response styles

Wednesday, 13th July - 10:00: Outlier Detection (Room G) - Individual Oral Presentation

Ms. Tongtong Zou [1], ***Prof. Daniel Bolt*** [1]

1. *University of Wisconsin - Madison*

Person misfit and person reliability indices in IRT can play an important role in evaluating the validity of a test or survey instrument at the respondent level. In this paper, we demonstrate an important applied distinction between these methods when applied to rating scale items, specifically their varying sensitivities to response styles. Using several empirical datasets, we show contexts in which estimates of these indices are in one case highly correlated and in two other cases poorly correlated. In the datasets showing inconsistencies between indices, the primary distinction appears due to differences in normative response style behavior, whereby respondents whose response styles are atypical (e.g., disproportionate selection of 3's on a 5-point rating scale) are found to misfit using Drasgow et al's person misfit index, but often show high levels of reliability from a person reliability perspective; just the opposite frequently occurs for respondents that select the rating scale extremes. It is suggested that simultaneously attending to both types of indices may be needed to best understand the validity of measurement at the respondent level when using IRT models with rating scale measures.

Detecting aberrant behaviors of test-takers with hierarchical IRT-based Response times models

Wednesday, 13th July - 10:15: Outlier Detection (Room G) - Individual Oral Presentation

Dr. Burhanettin Ozdemir [1]

1. *Prince Sultan University*

The response time data is used to enhance test design, item calibration, detecting aberrant behaviors, item pre-knowledge in the context of computer-based testing (CBT). This study aims at examining response patterns of students to detect students with aberrant behaviors and the possible occurrence of item exposure. In this study, data obtained from the general aptitude test (GAT) administered to the 4308 high school graduates was analyzed using the Bayesian IRT-based hierarchical log-linear response time modeling. This method utilizes both students' responses to items and the response time data to calculate the person fit statistics and item parameters. Moreover, the person-fit measures based on the Bayesian joint model (l_z and l^t) were used to detect aberrant response time and accuracy patterns, respectively. According to the person-fit statistics for person speed parameter (l_z), 13.14% of the response patterns were detected as aberrant while only 2.36% of the response patterns were identified as aberrant patterns based on person ability parameters. Moreover, only 0.46% of test-takers were labeled as aberrant based on both ability and speed parameters. Moreover, there was a negligible small negative correlation between item and time discrimination parameters. Additionally, the ability and speed parameters were negatively correlated indicating that students with the higher ability worked less on the questions than students with low ability levels. Overall, this study emphasizes the importance of investigating aberrant response patterns for both response accuracy and response times data to increase the overall quality of tests by detecting the item exposure, item pre-knowledge, and cheating behaviors.

Bayesian Region of Measurement Equivalence (ROME) approach with alignment

Wednesday, 13th July - 09:15: Measurement Invariance (Room E) - Individual Oral Presentation

Ms. Yichi Zhang [1], ***Dr. Mark Hok Chio Lai*** [1]

1. *University of southern california*

Measurement invariance (MI) research has focused on identifying biases in test indicators measuring a latent trait across two or more groups. However, little is known about the practical implications of noninvariance. The recently proposed Bayesian *Region of Measurement Equivalence* (ROME) approach quantifies the impact of partial invariance on the observed composite scores across groups, and allows researchers to use this effect size index to directly support MI. Under the ROME framework, researchers first compute the *highest posterior density intervals* (HPDIs)—which contain the most plausible values—for the expected group difference on total scores—with a predetermined range of values that is practically equivalent to zero (i.e., region of measurement equivalence). However, when the number of noninvariant items is large, the partial invariant model cannot guarantee that the latent traits across groups are on the same scale. Alignment, as an alternative to the traditional confirmatory factor analysis for testing MI, automates the MI analysis based on a configural model. Thus, the current study extends the ROME method by including the Bayesian alignment method as a first step to set the scale for latent traits across groups and identify the anchor item for the Bayesian confirmatory factor analysis model. This fully Bayesian approach can be used for both continuous indicators and ordinal items. The illustrative example, which examines the MI of a mathematics-specific self-efficacy scale across gender using a nationally representative sample of tenth graders, shows the extended ROME method can efficiently assess the practical significance of the noninvariant items.

Finding clusterwise measurement invariance with mixture multigroup factor analysis

Wednesday, 13th July - 09:30: Measurement Invariance (Room E) - Individual Oral Presentation

Dr. Kim De Roover [1]

1. *Tilburg University*

Psychological research often builds on between-group comparisons of (measurements of) latent variables, for instance, to evaluate cross-cultural differences in mindfulness. A critical assumption in such comparative research is that the same latent variable(s) are measured in the same way across all groups (i.e., measurement invariance). Nowadays, measurement invariance is often tested across lots of groups. When (a certain level of) measurement invariance is untenable across many groups, it is hard to unravel invariances from non-invariances and for which groups they apply. Mixture multigroup factor analysis (MMG-FA; De Roover, 2021; De Roover, Vermunt, & Ceulemans, 2020) was recently proposed to cluster groups based on the measurement parameters, whereas the structural parameters are allowed to differ between groups within a cluster. Since MMG-FA clusters the groups according to a specific level of 'clusterwise measurement invariance' (e.g., based on factor loadings only), there are different ways to proceed when the targeted level of clusterwise invariance is two or more levels higher than the invariance level that holds across all groups. For instance, if the research question requires scalar invariance when the initial level of invariance across all groups is merely configural invariance, one may immediately determine clusters of groups with scalar invariance, or one may first determine clusters of groups with metric invariance and then continue to (clusterwise) scalar invariance within each of those clusters. In this presentation, I identify the best way to take the steps from the initial overall level of invariance to the desired level of clusterwise invariance.

A new Bayesian method for investigating, quantifying, and visualizing measurement invariance

Wednesday, 13th July - 09:45: Measurement Invariance (Room E) - Individual Oral Presentation

Mr. Miljan Jovic [1]***, Dr. Maryam Amir-Haeri*** [1]***, Dr. Stéphanie van den Berg*** [1]

1. *University of Twente*

This study aimed to present a new method for visualizing measurement invariance in terms of its practical implications.

There is a lack of optimal statistical methodology for visualizing measurement invariance, while existing methods have several important limitations. Those methods use Test Characteristic Curves to visualize expected sum scores based on maximum likelihood estimates. They do not contain (full) information about uncertainties in the item parameter estimates and expected sum scores, visualize only measurement invariance on the test (DTF) but not on item level (DIF), and use visualization only for illustrative purposes while decisions are based on various statistical tests.

The graphical method that we propose is based on the Bayesian approach and Item Response Theory. We use the posterior distributions of item parameters to visualize conditional posterior probabilities for each response category for each item for different groups given different latent trait values (DIF visualization). After that, we use those probabilities to generate posterior predictive item responses in order to compute posterior predictive densities of sum scores given the latent trait and group membership (DTF visualization).

This straightforwardly solves the problem of quantifying the uncertainties in the item parameters and expected sum scores and makes the method statistically comprehensive. The method is also psychometrically comprehensive (visualizes both DTF and DIF), easy to understand and interpret (conclusion based on sum scores distributions and practical significance), but also flexible because it is not based on a cut-off criterion and enables researchers to take into account the context of the research in decision-making.

A systematic review of measurement invariance research of the CES-D across gender: calculation and report of effect size

Wednesday, 13th July - 10:00: Measurement Invariance (Room E) - Individual Oral Presentation

Mr. Gengrui Zhang [1], ***Dr. Mark Hok Chio Lai*** [1], ***Ms. Hailin Yue*** [1]

1. *University of southern california*

The Center for Epidemiology Studies Depression Scale (CES-D) is commonly used for measuring depressive symptoms for research and screening purposes. Previous research has found and examined evidence of non-invariance in some items across various demographic characteristics. Despite the abundance of invariance research across gender groups, the literature often provided conflicting information on the level of invariance of CES-D items. In this study, our goal is to offset the ambivalence by providing both qualitative and quantitative analysis of non-invariance of CES-D items. We conducted a systematic review of 32 articles about existing measurement invariance (MI) research on the CES-D across gender groups and found 21 articles that reported non-invariance among various items. Given the interpretation of effect sizes in empirical studies should inform the magnitude of non-invariance, we decided to further compute and synthesize effect sizes by d_{MACS} (Nye & Drasgow, 2011). Comparable to Cohen's d commonly reported in meta-analyses, the d_{MACS} index indicated standardized mean differences on items between gender groups due to systematic differences in factor loadings and intercepts. However, 8 articles did not report sufficient and necessary statistics to compute d_{MACS} (e.g., unstandardized factor loadings, intercepts, factor variance, item standard deviation, etc). To fully investigate the effect sizes of non-invariance, we will present several methods to compute d_{MACS} without sufficient information and illustrate patterns of non-invariance across gender groups (e.g., pseudo R^2, Riemaan sum approximation, categorical response approximation.) The results will inform guidelines for future researchers to compute and report effect sizes of measurement non-invariance.

The effect of acquiescence bias on measurement invariance testing

Wednesday, 13th July - 10:15: Measurement Invariance (Room E) - Individual Oral Presentation

Mr. Damiano D'Urso [1], ***Dr. Jesper Tijmstra*** [1], ***Prof. Jeroen Vermunt*** [1], ***Dr. Kim De Roover*** [1]

1. *Tilburg University*

In social sciences, group differences concerning latent constructs (e.g., self-esteem) are ubiquitously investigated, and these constructs are often measured using scales composed of multiple self-report ordinal items. For these comparisons to be valid, it is fundamental that these measures function equivalently across groups, or, in technical jargon, measurement invariance (MI) must hold. Testing for MI allows one to investigate group-specific systematic biases that occur in item responses across groups. A potential source of systematic bias in self-report measures is that of response styles (RSs) or response bias, which can be viewed as a stylistic tendency in the manner respondents use a rating scale when responding to self-report items. Acquiescence response style (ARS) is a well-known response bias, which represents a tendency to agree with items regardless of their content. Failing to take into account an ARS when testing for MI may result in concluding that a measure is non-invariant while this is purely due to this stylistic tendency. In this project, by means of a simulation study, we investigated the effect of ARS on MI testing both when such tendency is taken into account by including an additional ARS factor in the measurement model (MM) or not. Based on the simulation study results, recommendations and guidelines are provided for applied researchers.

Machine learning methods for propensity score estimation in hierarchical data

Wednesday, 13th July - 09:15: Causal Inference II (Room F) - Individual Oral Presentation

Ms. Marie Salditt [1], ***Prof. Steffen Nestler*** [1]

1. *University of Münster*

In quasi-experimental studies, propensity score methods are commonly used to control for confounding when estimating the causal effect of treatment. In single-level data, the performance of these methods has been shown to improve when the propensity score is estimated using machine learning methods such as classification trees, random forests, or gradient tree boosting rather than logistic regression, especially when nonlinear relationships exist between the confounding variables and the treatment indicator. However, in many psychological studies the data have a hierarchical structure. To estimate the propensity score non-parametrically even in the context of hierarchical data, one can combine machine learning methods with the generalized linear mixed model (GLMM). We first present a general algorithm for estimating the parameters of the combination of the GLMM with a classification tree, a random forest, and gradient tree boosting. We then report the results of a simulation study in which we compared the performance of these combinations in estimating the propensity score with (i) a GLMM, (ii) logistic regression, a classification tree, a random forest, and gradient tree boosting, in which the hierarchical structure was modeled as fixed effects, and with the estimates of (iii) the respective methods that ignore the hierarchical data structure.

Beyond the mean: A flexible framework for studying causal effects using linear models

Wednesday, 13th July - 09:30: Causal Inference II (Room F) - Individual Oral Presentation

Dr. Christian Gische [1], ***Prof. Manuel Voelkle*** [1]

1. *Humboldt-University zu Berlin*

Graph-based causal models are a flexible tool for causal inference from observational data. We present a comprehensive framework to define, identify, and estimate a broad class of causal quantities in linearly parametrized graph-based models. The proposed method extends the literature, which mainly focuses on causal effects on the mean level and the variance of an outcome variable. For example, we show how to compute the probability that an outcome variable realizes within a target range of values given an intervention, a causal quantity we refer to as the probability of treatment success. We link graph-based causal quantities defined via the do-operator to parameters of the model implied distribution of the observed variables using so-called causal effect functions. Based on these causal effect functions, we propose estimators for causal quantities and show that these estimators are consistent and converge at a rate of N^(−1/2) under standard assumptions. Thus, causal quantities can be estimated based on sample sizes that are typically available in the social and behavioral sciences. In case of maximum likelihood estimation, the estimators are asymptotically efficient. We illustrate the proposed method with an example based on empirical data, placing special emphasis on the difference between the interventional and conditional distribution.

Estimating latent baseline-by-treatment interactions in statistical mediation analysis

Wednesday, 13th July - 09:45: Causal Inference II (Room F) - Individual Oral Presentation

Dr. Oscar Gonzalez [1]

1. *University of North Carolina at Chapel Hill*

In the social sciences, statistical mediation analysis is used to uncover potential mechanisms, known as mediators [M], by which a treatment [X] led to a change in an outcome [Y]. In many applications, baseline measures of M and Y are used as covariates in the statistical mediation model to control for nuisance variation of M and Y at posttest, thus increasing the precision and power to detect the mediated effect. Furthermore, it is often of interest to evaluate if baseline scores of M and Y moderate the effect of X on posttest M or Y. However, there is limited guidance on how to estimate baseline-by-treatment interactions (BTIs) when M and Y are latent variables, which involves accommodating latent-by-observed interactions in the mediation model. In this paper, two general approaches to accommodate latent BTIs in statistical mediation analysis are discussed: using structural models (e.g., latent moderated structural equations or Bayesian methods) or calculating observed scores prior to estimating the BTI (e.g., using summed scores or factor scores for posttest M and Y). Simulation results are presented on the bias and Type 1 error rate of the latent BTI and mediated effect estimates across these approaches as a function of the latent structure of M and Y and path effect sizes. Methods are also illustrated with an applied example consisting of an intervention to increase job seeking self-efficacy and in turn reduce mental health issues in unemployed individuals. General guidance and considerations for each of the approaches will be discussed.

Causal inference in latent class analysis in the presence of differential item functioning

Wednesday, 13th July - 10:00: Causal Inference II (Room F) - Individual Oral Presentation

Mr. Felix Clouth [1], ***Prof. Steffen Pauws*** [1], ***Prof. Jeroen Vermunt*** [1]

1. *Tilburg University*

Causal inference techniques such as inverse propensity weighting (IPW) are increasingly used in medical, social, and behavioral research. When data is collected with an observational study design rather than in a randomized controlled trial, treatment effect estimates will be confounded. However, causal inference provides a toolbox for accounting for these confounding effects and to estimate average treatment effects (ATE) based on observational data. IPW can be easily combined with standard statistical models such as generalized linear models or survival analysis. However, sometimes the outcome of interest is not directly observable and a measurement model is needed, e.g., when analyzing patient reported outcome measures data. Latent class analysis (LCA) and its extensions are particularly suited for analyzing such data as they explicitly model the multidimensionality of these constructs. Recently, a one-step approach (Lanza, Coffman, & Xu; 2013) and a three-step approach (Clouth, Pauws, Mols, & Vermunt; 2021) have been proposed to incorporate IPW in LCA. While these approaches work well when the latent class model is correctly specified, differential item functioning (DIF) often prohibits estimating the ATE correctly. DIF occurs when treatment or confounding variables have direct effects on some of the indicator variables which violates the assumption that indicator variables and auxiliary variables are independent conditional on class membership. This can lead to biased estimates of the ATE or even to the detection of spurious classes. In this talk, I will present an analysis strategy that allows for the correct estimation of the ATE in the presence of DIF.

Fully Latent Principal Stratification: Combining PS with model-based measurement models

Wednesday, 13th July - 10:15: Causal Inference II (Room F) - Individual Oral Presentation

Mr. Sooyong Lee [1], ***Prof. Adam Sales*** [2], ***Dr. Hyeon-Ah Kang*** [1], ***Prof. Tiffany Whittaker*** [1]

1. *The University of Texas at Austin,* ***2.*** *Worcester Polytechnic Institute*

Despite prominent potentials of randomized controlled trials (RCTs) with computer-based interventions, log data poses a challenge since it differs in structure and size from the type of data commonly encountered in studies of causal mechanisms. The current study developed a method for RCTs suitable for big data, or complex implementation data. The method is an extension of principal stratification (PS), a causal framework used for studying how treatment effects vary as a function of post-treatment or intermediate variables. To exploit the complex structure of the log data, the proposed method can incorporate latent variables or measurement models into PS, substantially extending the scope of PS modeling into scenarios with multivariate and complex implementation data. With the method development, we did a simulation study to evaluate if our proposed FLPS model worked properly under various conditions, including sample sizes, number of items, effect sizes, and response rates, with different IRT models. FLPS models will allow researchers to gain deeper and more nuanced insights into the relationship between the effectiveness of the interventions under study, and how they are used. This ability will, in turn, deepen our understanding of our rapidly-evolving and growing interaction with technology, guiding the development of more effective interventions and guiding the implementation decisions of users.

KCP-RS and statistical process control: Flexible tools to flag changes in time series

Wednesday, 13th July - 10:50: Invited Speaker: Eva Ceulemans (Room B) - Individual Oral Presentation

Prof. Eva Ceulemans [1]

1. *KU Leuven*

ntensive longitudinal studies (e.g., experience sampling studies) have demonstrated that detecting changes in statistical features across time is crucial to better capture and understand psychological phenomena. For example, it has been uncovered that emotional episodes are characterized by changes in both means and correlations. In psychopathology research, recent evidence revealed that changes in means, variance, autocorrelation and correlation of experience sampling data can serve as early warning signs of an upcoming relapse into depression. In this talk, I will discuss flexible statistical tools for retrospectively and prospectively capturing such changes. First, I will present the KCP-RS framework, a retrospective change point detection framework that can be tailored to capture changes in not only the means but in any statistic that is relevant to the researcher. Second, I will turn to the prospective change detection problem, where I will argue that statistical process control procedures, originally developed for monitoring industrial processes, are promising tools but need tweaking to the problem at hand.

Network approaches to psychological constructs: A review, an evaluation, and an agenda

Wednesday, 13th July - 10:50: Invited Speaker: Denny Borsboom (Room A) - Individual Oral Presentation

Prof. Denny Borsboom [1]

1. *University of Amsterdam*

In the past decade, network approaches to psychological constructs have rapidly gained popularity and have seen a proliferation of psychometric techniques designed to estimate such networks from data. In this talk, I will review and critically evaluate this development. Networks are clearly a psychometric success story in terms of rate at which they have been adopted by the psychological research community, and offer powerful explorative data analysis tools and visualizations of complex multivariate dependencies. However, network analysis is also showing signs of developing some of the same problems that have plagued other psychometric methodologies in the past: unrealistic expectations of what data analysis can do in the first place and a lack of formalized theory that is powerful enough to guide model applications. This hampers the full realization of the potential of the network approach. I will outline a number of ways in which statistical network models can inform network theories of psychological phenomena in connection to recent attempts to create methodologies for theory formation.

Computational aspects of reliability estimation

Wednesday, 13th July - 11:40: Spotlight Talk: Patricia Martinkova (Room B) - Individual Oral Presentation

Dr. Patricia Martinkova [1], ***Mr. František Bartoš*** [1], ***Dr. Marek Brabec*** [1]

1. *Czech Academy of Sciences*

In this work, we discuss several computational aspects of reliability estimation. We focus on two topics: We first discuss the issue of zero estimates. We then propose a flexible approach for assessing IRR in cases of heterogeneity due to covariates. The method directly models differences in variance components, uses Bayes factors to select the best performing model, implements the Bayesian model-averaging for obtaining IRR and variance component estimates accounting for model uncertainty, and it employs the inclusion Bayes factors to provide evidence for or against differences in variance components due to covariates while considering the whole model space. We focus on optimizing the estimation process in the case of higher sample sizes, higher number of covariates, and on other computational aspects connected to software implementation. Study is motivated by real data from grant proposal peer review and teacher hiring.

Incorporating intersectionality using latent class analysis within health contexts

Wednesday, 13th July - 11:40: Spotlight Talk: Melanie Wall (Room A) - Individual Oral Presentation

Prof. Melanie Wall [1]

1. *New York State Psychiatric Institute and Columbia University*

Intersectionality posits that social categories (e.g. race, gender, sexual orientation) and the forms of social stratification that maintain them (e.g. racism, sexism, homophobia) are interlocking, not discrete. An intersectionality framework considers harms and oppression and also privileges and unearned advantages. By focusing on intersectionality, we can examine axes of social power that underlie our overall health and the systems that support it with the goal of identifying levers for change. A recent systematic review (Bauer et al 2022 Social Psychiatry and Psych Epi) demonstrated a growing use of latent variable methods including latent class analysis for applications of intersectionality. Latent class analysis (LCA) has been described as a "person-centered" approach as it clusters within-individual characteristics seen to be appropriate to intersectionality. In the present talk, I will demonstrate the use of LCA for combining intersecting social positions with multiple factors characterizing an initial mental health encounter. The example comes from a study of ethnoracial disparities in coordinated specialty care for people with psychosis. Clusters were identified based on the first-contact experience (i.e., referral source, type of first mental health service contact, symptoms at referral) in combination with sociodemographic variables impacting an individual's social position (age, gender, ethnoracial group, language proficiency, sexual orientation, living situation, type of insurance, homelessness, and urbanicity). Visualizations of intersectional cluster results and comparisons between the LCA approach and analyses focused on each variable separately will be presented.

Developing an evidence-base when treatment effects vary

Wednesday, 13th July - 13:40: Keynote: Elizabeth Tipton (Room B) - Individual Oral Presentation

Prof. Elizabeth Tipton [1]

1. *Northwestern University*

There is an increasing need for evidence-based practices in a variety of fields, from education to development to medicine. Buttressing this evidence base are improvements in methods for causal inference – from advanced experimental designs to methods for observational studies. As the field of evidence-based practice has matured, however, it has become clear that there is a disconnect between these methods – which prioritize internal validity – and the needs of decision makers – which prioritize external validity. I begin by reviewing this problem and then propose how causal inference research might better incorporate these external validity concerns. This will include a review of methods for generalizing results from samples to populations and methods for exploring treatment effect heterogeneity. Much of the talk will focus on how to better design studies to meet both internal and external validity goals.

Steering player behavior in adaptive learning environments

Wednesday, 13th July - 14:45: Symposium: Psychometrics for adaptive learning environments (Room B) - Symposium Presentation

Dr. Abe Hofman [1], ***Mr. Nick ten Broeke*** [1]

1. *University of Amsterdam*

Online learning environments can be used in different ways. Prowise Learn, for example, is used inside the classroom, for homework assignments and as extra tooling bought by parents. Thus it is important that children want to use the program and keep coming back. To provide players an optimal learning experience we need to optimize a tricky balance: learning should be fun and challenging. We would like to challenge players, so that they are confronted with items from which they can learn. However, these items should not be too difficult, possibly resulting in unwanted quitting behavior.

Since the learning platforms of Prowise Learn are built on IRT models, we can use the IRT estimates to optimize the system. In this talk we will present two A/B testing experiments. In the first experiment we investigate which games a player should play next. Can we select an optimal next game for a player based on the ability estimates on a set of already observed games? In a second experiment, we focus on preventing quitting behavior. The most important variable related to quitting a session is based on streaks of incorrect responses. Can we prevent quitting by changing the item selection function such that an easier item is selected after two sequential errors?

A decade of psychometric optimisation in Prowise Learn

Wednesday, 13th July - 15:00: Symposium: Psychometrics for adaptive learning environments (Room B) - Symposium Presentation

Dr. Joost Kruis [1]***, Prof. Han L. J. van der Maas*** [2]***, Dr. Abe Hofman*** [3]

1. *Prowise Learn / University of Amsterdam,* ***2.*** *University of Amsterdam,* ***3.*** *Universty of Amsterdam / Prowise Learn*

Prowise Learn is an adaptive learning platform that was set up using explicit scoring rules and adapted versions of the famous Elo-algorithm originating from chess (Elo, 1978; Klinkenberg et. al., 2011; Maris & Van der Maas, 2012). Running the system for over a decade has provided us with a lot of information about what works well and what could be improved on the initial setup of this system. In this talk we discuss two examples of psychometric updates that we have done in the previous years to optimise the platform. The update from a high-speed high-stakes scoring rule to a fixed penalty scoring rule, and the implementation of the paired item update.

Tracing students' systematic errors in large-scale online multiplication practice

Wednesday, 13th July - 15:15: Symposium: Psychometrics for adaptive learning environments (Room B) - Symposium Presentation

Dr. Alexander Savi [1], ***Dr. Benjamin Deonovic*** [2], ***Dr. Maria Bolsinova*** [3], ***Prof. Han L. J. van der Maas*** [1], ***Dr. G.K.J. Maris*** [4]

1. *University of Amsterdam,* ***2.*** *Corteva,* ***3.*** *Tilburg University,* ***4.*** *Tata Consultancy Services*

Diagnosing students' cognitive processes, such as multiplication strategies, is a major challenge in adapting education to the individual. Systematic errors hold a key to such latent processes, as these may signal students' misconceptions. In this talk, I will discuss the various challenges involved with diagnosing misconceptions, and introduce a new method to identify and trace systematic errors. Serving as a recommendation system, this method provides probability estimates for a student's potential misconceptions. The method is derived from the Ising model originating in physics, and exploits a theoretical mapping between latent misconceptions and manifest errors. I evaluate the performance of different model configurations, using single-digit multiplication data from a large-scale adaptive practice system, and key metrics for recommendation systems. The results show that the Systematic Error Tracing (SET) model outranks a majority vote baseline model when two or more recommendations are considered. Also, for some misconceptions it improves the adaptation of recommendations to individual students. I discuss the opportunities of SET in real-world large-scale learning applications.

Urnings: meet your digital twin

Wednesday, 13th July - 15:30: Symposium: Psychometrics for adaptive learning environments (Room B) - Symposium Presentation

Dr. G.K.J. Maris [1]

1. *Tata Consultancy Services*

The purpose of learning is to change the learner. Hence the traditional statistical framework (i.e., point estimation) is of little use, as it doesn't deal with change, and a different statistical framework is needed. Approaches such as the Kalman filter can deal with change, but at the expense of having to know the learner's dynamics which are typically unknown and depend on the measurements itself.

We introduce a digital twin of the human learner that tracks learning and is used to guide instruction and practice, and thus creates the learner's dynamcis. The main contribution is the statistical theory needed to ensure that the digital twin will faithfully track the true learning in real time. We illustrate the new framework with a number of real world use cases.

Rectangular latent Markov modeling for advising students in self-learning platforms

Wednesday, 13th July - 14:45: Classification (Room A) - Individual Oral Presentation

Ms. Rosa Fabbricatore [1], ***Dr. Roberto Di Mari*** [2], ***Dr. Zsuzsa Bakk*** [3], ***Prof. Mark de Rooij*** [3], ***Prof. Francesco Palumbo*** [1]

1. *University of Naples Federico II,* ***2.*** *University of Catania,* ***3.*** *Leiden University*

In recent years, there has been a growing interest in using technology to provide adaptive learning environments. In this vein, recommender systems play a fundamental role, supporting students in their learning path with tailored feedback. To achieve this, essential steps consist in collecting students' responses and diagnosing their learning state throughout the learning process. As part of this research line, this contribution proposes a three-step rectangular Latent Markov modeling to analyse data collected via the Moodle platform during an introductory statistics course. Data collection consists of three waves, each focusing on different statistical topics. Students' ability was conceived as a multidimensional latent variable according to three Dublin descriptors: Knowledge (K), Application (A), and Judgement (J). Thus, for each wave, students were asked to respond to 30 multiple-choice questions equally divided in K, A, J. Due to the different measurement models per time point, we exploited a three-step estimation. In Step 1, we carried out a multidimensional Latent Class IRT model for each time point to detect sub-populations of homogeneous students according to their ability level. In Step 2, we computed the time-specific classification error probabilities, whereas in Step 3, modal class assignments are used to estimate the structural component of the model correcting for the classification error. Because the number of latent classes was different for the three time points, we employed the rectangular formulation of Latent Markov models (Anderson et al., 2019). Moreover, the effect of demographic and psychological variables on initial and transition probabilities will also be discussed.

A modified method to balance attribute coverage in CD-CAT

Wednesday, 13th July - 15:00: Classification (Room A) - Individual Oral Presentation

Dr. Chia-Ling Hsu [1], ***Mr. Zi-Yan Huang*** [2], ***Prof. Shu-Ying Chen*** [2], ***Prof. Chuan-Ju Lin*** [3]

1. *Hong Kong Examinations and Assessment Authority,* ***2.*** *National Chung Cheng University,* ***3.*** *National University of Tainan*

This study introduces a new attribute balancing method, namely, the modified attribute balancing index (ABI-M), for cognitive diagnostic computerized adaptive testing (CD-CAT). The new method can both yield acceptable measurement accuracy and attribute coverage with respect to test reliability and validity, regardless of the complexity of Q-matrix structure. A simulation study is carried out to evaluate the ABI-M performance compared with the original ABI and RTA (the ratio of test length to the number of attributes). The results showed that the ABI-M and ABI achieved the requirement of attribute coverage; the ABI-M was comparable to the ABI in respect to measurement accuracy. and the ABI-M was more cost-effective in item usage than the others in all simulation conditions. Overall, these results suggest the feasibility of utilizing ABI-M in CD-CAT to ensure both measurement accuracy and attribute coverage and increase item usage.

Evaluate the mastery of learning objectives

Wednesday, 13th July - 15:15: Classification (Room A) - Individual Oral Presentation

Dr. Anton Béguin [1], ***Dr. Hendrik Straat*** [2]

1. *International Baccalaureate,* ***2.*** *Cito*

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. Building on previous research we identify what item characteristics are more efficient to detect mastery and non-mastery. For this a small simulation study is carried out. Also, a comparison is made between the definition of mastery in an IRT context and using Bayesian decision making and we discuss the practical implications coming from this.

E-ReMI: Extended maximal interaction two-mode clustering

Wednesday, 13th July - 15:30: Classification (Room A) - Individual Oral Presentation

Dr. Alberto Cassese [1], ***Dr. Jan Schepers*** [1], ***Prof. Gerard van Breukelen*** [1], ***Mr. Zaheer Ahmed*** [1]

1. *Maastricht University*

In this presentation, I show a method for studying two-way interaction in row by column (i.e., two-mode) data. This method, named E-ReMI, is based on a probabilistic two-mode clustering model that yields two-mode partitions of the data with maximal interaction between row and column clusters. The proposed model allows for unequal cluster size of the row clusters. I will, briefly, discuss two parameterizations of this model and show that, for finite samples, only one of the two can be used in a conditional classification likelihood approach. I will further introduce a test statistic for testing the null hypothesis of no interaction, discuss its properties and present an algorithm to obtain its distribution under this null hypothesis. I will show the performance of the new method through a simulation study and an analysis of data from a study of person by situation interaction.

Consequences of sampling frequency for estimating dynamics in continuous time models

Wednesday, 13th July - 14:45: Intensive Longitudinal Data (Room C) - Individual Oral Presentation

Mr. Rohit Batra [1], ***Ms. Simran Johal*** [1], ***Dr. Meng Chen*** [1], ***Dr. Emilio Ferrer*** [1]

1. *University of California, Davis*

Continuous time (CT) models are a flexible approach for modeling longitudinal data of psychological constructs, where a researcher can assume one underlying continuous function for the phenomenon of interest. In principle, these models overcome some limitations of discrete time (DT) models and allow for comparison of findings across measures collected using different time intervals, such as daily, weekly, or monthly intervals. Theoretically, the model parameters for equivalent models can be rescaled into a common time interval that allows for comparisons across individuals and studies, irrespective of the time interval used for sampling. In this study, we carry out a simulation to examine the capability of CT autoregressive models to recover the true dynamics of a process, when the sampling interval is different from the time interval of the true generating process.
We use two generating time intervals (daily or weekly) with varying strengths of the autoregressive parameter and assess the recovery when sampled at different sampling intervals (daily, weekly, or monthly). Our findings suggest that sampling at a shorter time interval than the generating dynamics can mostly recover the autoregressive effects, whereas sampling at a longer time interval leads to generally poor performance overall. In addition, when the sampling and generating frequencies match, DT models perform better than CT models, even when measurement occasions are not equidistant. Based on our findings, we recommend researchers to use sampling intervals guided by theory about the variable under study and, whenever possible, sample at shorter time intervals.

Dynamic conditional network models for intensive repeated data

Wednesday, 13th July - 15:00: Intensive Longitudinal Data (Room C) - Individual Oral Presentation

Dr. Philippe Rast [1]

1. *University of California, Davis*

Longitudinal networks are typically estimated using a VAR structure, resulting in several networks: A contemporaneous, a temporal and, for multilevel models, a between-subjects network. So far, the contemporaneous network has been treated as constant resulting in a time-invariant partial correlation network that needs to hold across all time points t.

We present a network modeling approach that relaxes this requirement of a constant contemporaneous network. The model presented here estimates dynamic conditional contemporaneous networks for each time point, thus allowing changes in the partial correlations over all time-points. In other words, the model combines a VAR structure for the location and a multivariate GARCH model for the scale. Specifically, we will discuss the dynamic conditional correlation structure for the residual covariance structure. That is, the residual covariance structure of a VAR model can be defined as a partial correlation (contemporaneous) network and its values at t can be modeled conditional on the previous contemporaneous network at t-1.

This method is currently available for N=1 multivariate models (via the bmgarch package in R). In this work we will present two approaches, a dynamic conditional network for single individuals and a multilevel version that estimates fixed and random network structures. The methods are illustrated using simulated data as well as data from an intensive repeated design study with personality data.

Modelling agreement for intensive longitudinal binary data

Wednesday, 13th July - 15:15: Intensive Longitudinal Data (Room C) - Individual Oral Presentation

Dr. Sophie Vanbelle [1], ***Prof. Emmanuel Lesaffre*** [2]

1. *Maastricht University,* ***2.*** *KU*

In many scientific domains, like medical and psychological sciences, technical advances led to the development of devices collecting biological, physical, behavioral or environmental information in real-time on and real-life settings. These devices generate intensive longitudinal data (ILD) on one or more variables. ILD are characterized by many observations very close in time.

The measurement quality of these devices should be assessed through reliability and agreement studies. These studies provide information about the amount of error inherent to any diagnostic, score or measurement. For example, the CAM study was designed to validate a new accelerometry sensor to measure physical activity during the revalidation of patients with chronic organ failure. During one hour of unconstrained activity, 10 patients were videotaped in a revalidation center while their body activity was continuously recorded with the new device, worn simultaneously on the leg and on the trunk for comparative purposes. The aims were (a) to determine the validity of the new accelerometry sensor through the agreement level between device recordings and human observations of body activity on the videotape (considered as reference method); (b) to assess temporal stability of the agreement levels (e.g., they can decrease because of device shifts); and (c) to determine the influence of the body location where the device is held on the agreement levels.

To answer these research questions, we developed a general longitudinal model for sequential kappa statistics within the Bayesian framework. The model can be implemented in standard Bayesian software (e.g., Jags) and shows good statistical properties.

Using time-varying dynamic parameters to improve prediction of future outcomes

Wednesday, 13th July - 15:30: Intensive Longitudinal Data (Room C) - Individual Oral Presentation

Ms. Simran Johal [1]***, Dr. Emilio Ferrer*** [1]

1. *University of California, Davis*

The application of time series models – such as Vector Autoregressive (VAR) models – to intensive longitudinal data has allowed researchers to explore the dynamics and interactions of multiple processes over time. Previous research has shown that incorporating dynamic information can improve the prediction of distal outcomes of the system – for example, predicting future relationship quality or break-up amongst romantic couples (Castro-Schilo & Ferrer, 2013)

One limitation of VAR models, however, is that they assume that the processes under study are weakly stationary, which requires the means and the variance-covariance matrix to be constant over time. Yet such an assumption is often untenable in psychological data – for example, the association between the affect of two romantic partners likely changes over time (Bringmann et al., 2018). To account for potential non-stationarity, researchers have proposed the use of time-varying VAR models, where the autoregressive and cross-lagged parameters are modeled as changing continuously over time.

The current project investigated the question of whether VAR models that account for non-stationarity can improve the prediction of distal outcomes. We applied time-varying models to time series data on relationship affect from 117 romantic couples, and examined whether the variability in the dynamic parameters over time predicted relationship dissolution above and beyond the mean level and variability of affect, or the time-invariant dynamic parameters. The results from these analyses can help us better understand whether modeling more complex dynamics can be useful for predicting distal outcomes and future states of the system.

Machine-learning-based factor retention – the Comparison Data Forest

Wednesday, 13th July - 14:45: Factor Analysis (Room D) - Individual Oral Presentation

Prof. David Goretzko [1]

1. *LMU Munich; University of Leipzig*

Determining the number of factors is arguable the most crucial, yet difficult decision a researcher faces when performing exploratory factor analysis (EFA). Over the years, several simulation studies have shown that classical approaches such as the infamous eigenvalue-greater-one rule are not very accurate in retaining the correct number of latent factors. Goretzko and Bühner (2020) developed the so-called Factor Forest (FF) – a machine-learning-based factor retention method that outperformed common criteria such as parallel analysis. Since FF is based on comprehensive data simulation and model training (+ hyperparameter tuning), it is computationally very costly. In 2012, Ruscio and Roche introduced a different approach using comparison data (CD) sets that closely resemble the empirical data. This similarity can be exploited to develop a computationally more efficient machine-learning-based factor retention – the Comparison Data Forest (CDF). The idea of CDF is to simulate comparison data analogous to the initial CD approach and extract several features that are associated with the number of latent factors (e.g., the eigenvalues of the correlation matrix, inequality measures applied to this correlation matrix). Based on these features, a computationally efficient machine learning model is trained to predict the dimensionality of the measurement model. First simulation results show that CDF is able to outperform classical CD across different data conditions – especially when multiple factors were present.

Rotation to sparse loadings using L^p functions

Wednesday, 13th July - 15:00: Factor Analysis (Room D) - Individual Oral Presentation

Ms. Xinyi Liu [1], ***Dr. Gabriel Wallin*** [1], ***Dr. Yunxiao Chen*** [1], ***Prof. Irini Moustaki*** [1]

1. *London School of Economics and Political Science*

In this talk, we propose a family of loss functions, the component-wise L^p loss, for oblique rotations in exploratory factor analysis. The proposed loss functions take the form of the sum of the p^{th} power of the absolute loadings, for p<=1. They are special cases of the concave component-wise loss functions (Jennrich, 2006), but the cases when p<1 have been overlooked in the past. We establish the connection between the proposed rotation method and regularized estimation based on L^p penalty functions, showing that the former is a limiting case of the latter when the tuning parameter in regularized estimation converges to zero. The statistical consistency of the rotation-based estimator is established. In addition, procedures are developed for drawing statistical inference on the sparse true loading matrix, such as hypothesis testing and constructing confidence intervals. It is worth noting that since the objective function is non-smooth, classical statistical inference methods for rotation-based procedures fail (as the delta method is no longer applicable). A computationally efficient iteratively reweighted least square algorithm is developed that is suitable for the entire family of loss functions. The proposed method is evaluated via simulations and compared with the regularized estimation methods. It is found that the rotation method performs similarly in terms of model selection accuracy as the corresponding regularized estimation method but is computationally much faster. Moreover, the rotation loss function with p<1 tends to be more effective in recovering the sparse loading matrix than that with p=1 under various settings.

The current state of LASSO-Penalization within CML-Estimation for IRT-Models

Wednesday, 13th July - 15:15: Factor Analysis (Room D) - Individual Oral Presentation

Mr. Can Gürer [1], ***Dr. Clemens Draxler*** [1]

1. *UMIT - Private University for Health Sciences, Medical Informatics and Technology*

The cmlDIFlasso (Gürer & Draxler, 2021) has been proposed as an approach for the identification of covariate influence on item difficulties, selecting parameters by L1-penalized conditional maximum likelihood (CML) estimation. As within cmlDIFlasso the person parameters are eliminated from the likelihood by conditioning on the person sum scores in Rasch models for binary data, two separate extensions of the approach will be presented. Firstly, a method for DIF-detection in models with more than two response categories per item, i.e., the Partial Credit Model. The effectiveness and computational efficiency of the (extended) approach compared to a MML-based similar detection technique is discussed on the basis of simulation results and scenarios are illustrated in which the conditional approach can have advantages in DIF-detection.

Secondly, for binary data the likelihood function is modified by conditioning on person scores as well as on item scores, thereby eliminating all parameters except the ones incorporating covariate effects as discussed by Draxler and Zessin (2015). Estimation is then conducted via an MCMC-approach based on the RaschSampler (Verhelst, 2008). Challenges, problems and advantages of the approach are discussed and first results presented.

Matched-pair binary item response analysis using Bayesian adaptive Lasso factor model

Wednesday, 13th July - 15:30: Factor Analysis (Room D) - Individual Oral Presentation

Prof. Edward Ip [1], ***Prof. Joanne Sandberg*** [1], ***Ms. Lijin Zhang*** [2], ***Prof. Junhao Pan*** [2]

1. *Wake Forest School of Medicine,* ***2.*** *Sun Yat-sen University*

This work was motivated by a matched-pair design that was applied to pre- and post-tests of genetic knowledge items and in the context of informal science learning for Latinx immigrant adults with limited education. To improve knowledge and interest in genomics, the intervention used lay educators to deliver culturally and linguistically appropriate home-based instructional sessions. The primary outcome was change in genetic-related knowledge. As the overall respondent burden was already high, the study required short and efficient assessment of change in the primary outcome. The matched-pair design aimed to reduce between-item variance as well as possible learning effects, and 12 matched pre- and post-test items in genetics content were developed. Although this design is deemed efficient, matched item responses may violate the local item independence assumption and lead to biased estimates. We extended previous work on the Bayesian covariance Lasso confirmatory factor analysis (CFA) model in Pan, Ip & Dubé (2017) to (1) handle potentially locally dependent binary data by using a modeling approach for which the observed binary outcome was the result of being below or above a threshold of an underlying continuous distribution; and (2) adopt a new adaptive covariance Lasso prior procedure for parsimoniously modeling a block diagonal residual covariance structure between matched pairs. Bayesian inference was based on parameter expansion and MCMC procedures. Our simulation studies assessed the performance of the estimates of the unknown parameters of interest. Real data on genetic knowledge items were analyzed to demonstrate practical application of the proposed procedure.

Data-driven direct consensus standard setting without IRT

Wednesday, 13th July - 14:45: Test Equating (Room G) - Individual Oral Presentation

Dr. Marieke van Onna [1]

1. *Cito*

In general, the use of data enhances the quality of standard settings, as experts get realistic feedback on the appropriateness of their estimates of item or test difficulty for borderline students. Keuning, Straat & Feskens (2017) proposed a data-driven direct consensus setting method (3DC), where experts have to indicate cut scores on clusters of items. The experts get visual feedback on the alignment of their cut scores with the actual difficulty of the clusters by means of a figure. In the figure, all cluster scores are equated to the total scores. Keuning, Straat & Feskens (2017) derived these estimates from a fitted one-parameter logistic model. The R-package dexter (Maris, Bechger, Koops, & Partchev, 2022) supports the use of this method. In this paper, a more classical approach is presented. It gives experts a similar visual feedback. The relative difficulty of cluster scores is based on the observed score distribution. Several smoothing methods are explored to equate the cluster scores to the total scores. This approach results in similar estimates as the 3DC-method when only one exam form is administered, and a Rasch model fits the data. In addition, this new approach can also be used when the data are not suited for IRT-modeling. The new method will be illustrated with data that do not fit an IRT-model.

Parameter Identifiability of the linear equating transformation under the NEAT design

Wednesday, 13th July - 15:00: Test Equating (Room G) - Individual Oral Presentation

Dr. Jorge González [1], ***Dr. Ernesto San Martín*** [1]

1. *Pontificia Universidad Católica de Chile*

The equating literature is inconclusive, mixed, and ambiguous in that a distinction should be made between internal and external anchor items when conducting statistical inference on the equating transformation under the NEAT design.

By conducting a formal identifiability analysis, in this paper we show that the identified parameters needed to estimate the linear equating transformation are different depending on the choice taken for the anchor score to be either internal or external. Numerical illustrations using real data sets complement the theoretical results.

FDA meets IRT

Wednesday, 13th July - 15:15: Test Equating (Room G) - Individual Oral Presentation

Prof. James Ramsay [1], ***Dr. Juan Li*** [2], ***Prof. Marie Wiberg*** [3]

1. *McGill University,* ***2.*** *Ottawa Hospital Research Institute,* ***3.*** *Umeå University*

Functions can now be constructed using spline basis systems that are arbitrarily accurate for fitting data and that can also satisfy constraints such as being strictly monotone, as differentiable as the data and the model require, meet fixed boundary constraints, and take on legitimate probability or log-probability (surprisal) values that lie within lower dimensional and possibly curved spaces.

The fitting of these functions to data now takes only a few iterations to achieve useful accuracy, so that all option curves for 80-100 items can be computed within a second or two. Moreover, concepts such as difficulty discriminability can be expressed as functions that are much more informative than their one-dimensional counterparts in parametric counterparts.

We offer a variety of resources on functional data analysis such as R vignettes, web-based examples stand-alone analysis tools and printable material especially aimed at testers with relatively basic levels of statistical and mathematical expertise.

A framework to quantify overall errors in equated scale scores

Wednesday, 13th July - 15:30: Test Equating (Room G) - Individual Oral Presentation

Dr. Stella Kim [1], ***Dr. Won-Chan Lee*** [2]

1. *University of North Carolina at Charlotte,* ***2.*** *University of Iowa*

The current study attempts to quantify overall errors in equated scale scores as a function of standard errors of measurement (SEM) and standard errors in equating (SEE). For large-scale educational assessments, numerous sources of measurement errors are known to exist such as items, raters, occasions, and so on. SEM quantifies the amount of fluctuations in observed scores over replications of the same, or similar, measurement procedure resulting from the multiple sources of errors.

In practice, however, examinees typically receive a reported scale score which often is adjusted from the observed scores to compensate for differences in difficulty across alternate forms of a test. This statistical process called equating inevitably introduces additional errors in equated scale scores. The *Standards* (AERA, APA, & NCME, 2014) recommends documenting SEE as a standard practice if equating is involved in computing reported scale scores. SEE quantifies the variability of equated scores obtained over replications of an equating procedure based on different samples of persons.

To appraise the exact amount of errors involved in reported scale scores, consequently, both SEE and SEM should be taken into account. In the Programme for International Student Assessment (PISA; 2018), for instance, three general sources of measurement errors are considered: 1) student sampling variability, 2) the unreliability of measurement, and 3) errors attributable to a linking procedure across test administrations. This study proposes an overall index that effectively treats the major sources of error in equated scores.

Comparing presmoothing methods for kernel equating with mixed-format tests

Wednesday, 13th July - 15:45: Test Equating (Room G) - Individual Oral Presentation

Mr. Joakim Wallmark [1], ***Ms. Maria Josefsson*** [1], ***Prof. Marie Wiberg*** [1]

1. *Umeå University*

When equating test forms, it is common to *presmooth* the test score distributions before conducting the equating. In this study, the log-linear and item response theory (IRT) presmoothing methods were compared when equating mixed-format test forms using kernel equating. The equivalent group and common item non-equivalent group sampling designs were considered in both simulations and real data applications. In simulation studies it is common to generate test scores using prespecified item response theory (IRT) models or by resampling test scores from large datasets in order to obtain realistic test data. The use of IRT models to generate test data gives an advantage to methods which are themselves based on IRT models, while the resampling approach is not feasible when one does not have access to real test data or when the datasets are small. In this study, we propose a novel approach for generating realistic test data for simulation studies without the use of IRT. The results show that IRT model presmoothing results in smaller equating standard errors for high and low performing test takers when compared to log-linear presmoothing. The amount of bias of each presmoothing method in the simulation study was shown to be heavily dependent on the underlying data generating process as well as the definition of true equating function. As no true equating transformation is known in a practical setting, using IRT models for presmoothing should be preferred because of the lower equating standard errors for high and low scoring test takers.

Priors in Bayesian estimation under the three-parameter model

Wednesday, 13th July - 14:45: Bayesian Methods (Room E) - Individual Oral Presentation

Prof. Seock-Ho Kim [1], ***Ms. Ye Yuan*** [1], ***Dr. Youn-Jeng Choi*** [2], ***Prof. Allan Cohen*** [1]

1. *The University of Georgia,* ***2.*** *EWHA WOMANS UNIVERSITY*

The purpose of this study is to review various priors used in Bayesian estimation under the three-parameter model with clear mathematical definitions of the prior distributions. A Bayesian estimation method, Gibbs sampling, was compared with the marginal Bayesian estimation method using empirical data. The effects of the priors and their specifications on both item and ability parameter estimates are demonstrated. Issues in Bayesian estimation, use of priors in item response theory, and selection of item response theory models are discussed.

Model-based missing data handling for composites with missing items

Wednesday, 13th July - 15:00: Bayesian Methods (Room E) - Individual Oral Presentation

Ms. Egamaria Alacam [1], ***Dr. Han Du*** [1], ***Dr. Craig Enders*** [1]

1. *UCLA*

Composite scores (e.g., scale scores computed from questionnaire items) are widely used in the behavioral sciences, and it is common for items within scales to be missing. Among approaches that assume a missing at random (MAR) mechanism, one of the most widely used is fully conditional specification (FCS). FCS generally works well with item-level data, but convergence problems are common when the number of items is very large, and it is known to cause bias when the analysis model contains nonlinear terms such as interactions. Recently, the missing data literature has focused on model-based missing data handling procedures that tailor imputations to a specific model of interest, but these recent studies have thus far ignored the common problem of item- or component-level missing data in a composite score. The purpose of this research is to extend model-based procedures to composites with missing items. In the context of a single-level regression model that contains multiple composite predictors, we compared our new approach to the gold standard FCS method using four criteria: 1) convergence rates, 2) percent bias, 3) mean-squared error ratios, 3) and coverage rates. Simulation studies 1-3 explored this approach by implementing between-scale and within-scale constraints. Simulation study 4 explored how this approach handled categorical items. Overall, computer simulation results suggest that this new approach can be very effective, even under extreme conditions such as when there are many scales and items, sample size is small, and missing data rates are high.

Bayesian prior specification and model fitting propensity

Wednesday, 13th July - 15:15: Bayesian Methods (Room E) - Individual Oral Presentation

Dr. Sonja Winter [1], ***Dr. Wes Bonifay*** [1], ***Dr. Ashley Watts*** [1]

1. *University of Missouri - Columbia*

Competing statistical models often differ in *fitting propensity* (FP): the inherent ability to fit diverse (potentially meaningless) data patterns. Previous research has identified the number of free parameters and the functional form of the model as two factors that affect a model's FP. However, even if the number of parameters and functional forms of competing models are identical, we argue that FP will still vary depending on other modeling choices made by the researcher. In this talk, we focus on one such choice: prior specification in Bayesian modeling. We hypothesize that by specifying informative priors, a model will be less prone to fitting random data (i.e., FP will decrease). Using a Monte Carlo simulation design, we illustrate the impact of prior specification by considering the bifactor measurement model, which has demonstrated high FP in previous studies. Results suggest that, in general, a bifactor model that is estimated using informative priors will have lower FP than a bifactor model that is estimated using diffuse priors. In addition, not all informative prior specifications will result in good fit for the same data patterns. These findings suggest that informative priors could be used to set up riskier tests of the bifactor model (i.e., beyond goodness-of-fit evaluation). However, results also indicate that the impact of priors on FP depends on factors such as sample size. Implications for model complexity and recommendations for researchers are discussed.

Variable selection with missing data

Wednesday, 13th July - 15:30: Bayesian Methods (Room E) - Individual Oral Presentation

Prof. Sierra Bainter [1]

1. *University of Miami*

Methods for variable selection—identifying a subset of important predictors for a given outcome from a set of candidate predictor variables—can be usefully applied to a variety of research questions in psychology. Use of these methods, such as lasso regression and stochastic search variable selection (SSVS), has increased in recent years. Unfortunately, lasso and SSVS require complete data for the entire set of predictor variables, and suitable procedures for performing variable selection with missing data are needed. In this talk, we review current practices for handling missing data for variable selection. We describe the assumptions and limitations of two practices that are frequently used: complete case analysis and single imputation. Further, we describe the requirements, assumptions, and limitations of applying standard missing data methods, multiple imputation (MI) and full information maximum likelihood estimation (FIML), in the context of variable selection. Because both FIML and MI are model-based procedures, they must account for model uncertainty in addition to missing data uncertainty. Finally, we present results from a simulation study comparing standard approaches for variable selection with missing data to an alternative nonparametric procedure using Bayesian Additive Regression Trees.

Evaluating item parameter drift for Bayesian longitudinal item response theory models

Wednesday, 13th July - 15:45: Bayesian Methods (Room E) - Individual Oral Presentation

Dr. Allison Boykin [1], ***Ms. Nana Amma Asamoah*** [1], ***Dr. Brian Leventhal*** [2], ***Mr. Nnamdi Ezike*** [1]

1. *University of Arkansas,* ***2.*** *James Madison University*

In their presentation of the longitudinal IRT (LIRT) model, Kim and Camilli (2014) describe a model that embeds a longitudinal model for change within a measurement model. One assumption is that item parameters do not drift - they are invariant across time. Evaluating this assumption entails a systematic approach comparing a baseline model (i.e., no drift) to a model allowing one item's parameters to drift across time points. Few studies have examined model selection indices for Bayesian IRT, let alone Bayesian LIRT to evaluate the longitudinal invariance assumption. Ezike et al. (2021) report conditional likelihood approaches performed better than marginal likelihood approaches, but not in the context of LIRT drift.

We address the following research questions via simulation: Which model selection indices have the highest hit rates (and lowest false positive rates) in detecting item parameter drift for Bayesian LIRT? How do the simulation conditions affect the hit and false positive rates of model selection indices in detecting item parameter drift for Bayesian LIRT? The generating and analysis model is the Graded Response Model with five categories, the most commonly used LIRT model in the surveyed literature. Simulation conditions are based on empirical examples. They include survey length (10, 20 items); longitudinal invariance level (0%, 30% drift); time points (2, 4, 8); and sample size (500, 1000, 3000). We examine the behavior of conditional and marginal versions three Bayesian fit indices: deviance information criterion (DIC), widely applicable information criterion (WAIC), and leave-one-out cross validation (LOO).

Efficient marginal maximum likelihood estimation of longitudinal latent variable models

Wednesday, 13th July - 14:45: Estimation (Room F) - Individual Oral Presentation

Dr. Björn Andersson [1]

1. *University of Oslo*

Marginal maximum likelihood estimation of longitudinal latent variable models for categorical observed variables is challenging due to the high latent dimensionality required to specify complex models with many time points and with residual correlations for the repeated measurements. Here, we propose using second-order Laplace approximations to the high-dimensional integrals in the marginal likelihood function for longitudinal latent variable models and implement an efficient estimation method based on the approximated marginal likelihood. In the implementation, we support longitudinal generalized partial credit, graded response and nominal response models for categorical observed variables. In a simulation study with up to four time points and six observed variables at each time point, we compare the method based on second-order Laplace approximations to using the simulation-based method Metropolis Hasting Robbins Monro (MH-RM) in terms of estimation efficiency and estimation bias and precision. The results of the simulation study indicate that the proposed method is substantially more efficient in estimation than MH-RM while having equal or better estimation properties. The method is illustrated with the Montreal Cognitive Assessment, administered at four time points in a Hong Kong study of aging and well-being, where longitudinal measurement invariance is evaluated with sequential model comparisons. We discuss the advantages and limitations of the proposed estimation method and outline a potential extension to the approach that uses a dimension-reduction technique.

A deep learning approach for estimating response time models

Wednesday, 13th July - 15:00: Estimation (Room F) - Individual Oral Presentation

Dr. Rudolf Debelak [1]

1. *Universität Zürich*

The model parameters of psychometric models, such as models of item response theory, usually use maximum likelihood or Bayesian estimation methods to estimate their model parameters. A recent study by Urban and Bauer (2021; Psychometrika, 86(1), 1-29) proposed a deep learning approach based on an importance-weighted autoencoder to estimate the item parameters of item factor models, with a focus on the graded response model. Conceptually, this approach is related to marginal maximum estimation, but offers computational advantages in large datasets. In this presentation, an adaptation of this approach for response time models is presented. The accuracy of the resulting item parameters is evaluated by simulation studies. Their results indicate that this estimation method leads to accurate estimates of the item parameters. Applications of this estimation approach to further psychometric models are discussed in an outlook.

Estimating and using block information in the Thurstonian IRT model

Wednesday, 13th July - 15:15: Estimation (Room F) - Individual Oral Presentation

Dr. Susanne Frick [1]

1. *Tu Dortmund University*

Multidimensional forced-choice (MFC) tests become increasingly popular, but their construction is complex because it requires combining items based on their properties. The Thurstonian item response theory model is most often used to score MFC tests with dominance items. Currently, information in the Thurstonian IRT model is computed for binary outcomes of pairwise comparisons. This procedure neglects stochastic dependencies and item interactions.

In this talk, I will show how Fisher information on the block level can be estimated and summarized to make it usable for test construction. The accuracy of test information was evaluated in a simulation study on standard error (SE) accuracy. True, expected and observed SEs were computed for various block sizes, test lengths and estimators, both based on block information and when neglecting stochastic dependencies. The results showed that SEs based on block information were unbiased and that expected and observed SEs were comparably accurate.

To evaluate how well the block information performs in test construction, automated test assembly was simulated. In automated test assembly, blocks are selected such that information is maximized and the test design is optimal. Several information summaries and test construction targets were compared. The simulation results showed that all block information summaries performed on par, but better than random block selection.

Thus, block information can aid the construction of reliable MFC tests. Computing information on the block level will allow to capture the relative response process and to examine the extent of and influences on item interactions.

Unbiased distribution free estimator in SEM

Wednesday, 13th July - 15:30: Estimation (Room F) - Individual Oral Presentation

Dr. Han Du [1], ***Dr. Peter Bentler*** [1]

1. *University of California, Los Angeles*

In SEM, researchers conduct goodness-of-fit tests to evaluate whether the specified model fits the data well. With nonnormal data, the standard goodness-of-fit test statistic T does not follow a chi-square distribution; comparing T to $Chisq_{df}$ can fail to control Type I error rates and lead to misleading model selection conclusions. To better evaluate model fit, researchers have proposed various robust test statistics, but none of them consistently control Type I error rates under all examined conditions. To improve model fit statistics for nonnormal data, we propose to use an unbiased distribution free weight matrix estimator ($gamma\_DF^U$) in robust test statistics. First, we propose $gamma\_DF^U$ for models with mean structures. Second, we apply $gamma\_DF^U$ to robust statistics that have relatively simple forms in factor analysis and growth curve models.

Comparing the same model fit statistic using $gamma\_DF^U$ and the more widely used gamma_ADF by Browne (1984), we find that $gamma\_DF^U$ is closer to the theoretical distribution than gamma_ADF. In factor analysis, the Satorra–Bentler statistic with $gamma\_DF^U$ could control Type I error rates better than other statistics with $gamma\_DF^U$ or $gamma\_ADF^U$, followed by T_MVA2 from Hayakawa (2019) with $gamma\_DF^U$. In growth curve model, T_MVA2 with $gamma\_DF^U$ was the most stable statistic, which does not provide too many inflated or deflated Type I error rates across conditions. Additionally, $gamma\_DF^U$ provides smaller relative biases of the robust SE estimates than gamma_ADF.

Generalized Procrustes problem allows to estimate subject-specific functional connectivity in fMRI data

Wednesday, 13th July - 16:20: Symposium: Advanced methods to explore individual differences (Room B) - Symposium Presentation

Ms. Angela Andreella [1], ***Prof. Livio Finos*** [2]

1. *Ca' Foscari University of Venice,* ***2.*** *University of Padova*

The functional variability of neural brain activation between individuals is well known among neuroscientists. Hence, recently, Haxby et al. (2011) suggested a functional alignment called hyperalignment, which uses orthogonal transformations to map the brain images from fMRI into a common abstract high-dimensional space representing a linear combination of subjects' voxel activations. The individual-specific and shared functional information are modeled by high-dimensional transformations rather than transformations that rely on the 3D anatomical space. Nevertheless, hyperalignment mixes data across spatial loci. Its use to align the whole cortex is questionable since it can combine information from distant voxels to create the common abstract high-dimensional space. In addition, these high-dimensional transformations are not unique, leading to interpretability problems. Therefore, we propose the ProMises (Procrustes von Mises-Fisher) model. It returns a unique representation of the aligned images and related linear transformations in the anatomical brain space. Furthermore, ProMises allows inserting topological information into the estimation process thanks to the prior distribution - the von Mises-Fisher distribution - assumed for the orthogonal transformation parameters. The practitioners can give up the black-box concept, understand how the functional alignment acts effectively, and give a neurophysiological interpretation of the aligned images and related results. Besides, it permits analyzing the models' residuals, which describe how each individual is distant from some reference/shared matrix. Clustering methods and dimension reduction techniques can then be applied to these residuals analyzing task-related fMRI data. This permits to find groups of individuals sharing patterns of neural brain activation with respect to some stimuli.

Unbiased methods to study interindividual variability using multilayer brain networks

Wednesday, 13th July - 16:35: Symposium: Advanced methods to explore individual differences (Room B) - Symposium Presentation

Dr. Simone Di Plinio [1], ***Prof. Sjoerd Ebisch*** [1]

1. *University of Chieti-Pescara*

The brain is a complex system in which the functional interactions among its subunits vary over time. The trajectories of this dynamic variation contribute to inter-individual behavioral differences and psychopathologic phenotypes. Despite many methodological advancements, the study of dynamic brain networks still relies on biased assumptions in the temporal domain. Our study has two goals. First, we present a novel method to study multilayer networks by modeling intra-nodal connections using a probabilistic and biologically-driven approach. Our innovative modeling of multilayer networks introduces a temporal resolution based on signal similarity across time series.

We tested the performance of this original methodology through both synthetic, simulated datasets and real data from a resting-state study. The probabilistic modeling of cross-layer connections was implemented on synthetic networks by varying the number of modules and noise sources in the simulation to study the accuracy in the reconstruction of single-subject simulated modular architectures. In a resting-state fMRI experiment, we then employed probabilistically weighted multilayer networks to study the association between network dynamics and subclinical, psychosis-relevant personality traits in healthy adults.

Our findings demonstrate that the PW method for multilayer networks outperforms the standard procedure in modular detection and is less affected by increasing noise levels. Additionally, the PW method highlighted associations between the temporal instability of default mode network connections and psychosis-like experiences in healthy adults. PW multilayer networks allow an unbiased study of dynamic brain functioning and its behavioral correlates.

Unsupervised and supervised learning algorithms for accurate classification of cognitive profiles.

Wednesday, 13th July - 16:50: Symposium: Advanced methods to explore individual differences (Room B) - Symposium Presentation

Mr. Matteo Orsoni [1]***, Dr. Sara Garofalo*** [1]***, Dr. Sara Giovagnoli*** [1]***, Prof. Mariagrazia Benassi*** [1]

1. *University of Bologna*

The assessment of students' cognitive abilities in academic contexts could be informative for teachers to map their cognitive profile and program individualized learning strategies. Although several studies reported promising results in recognizing students' cognitive profiles (Yokota et al., 2015; Loehlin, 2019), effective comparisons between different clustering methods are missing in this literature.

In this study, we aim to compare the effectiveness of two clustering techniques to group students based on their cognitive abilities including general intelligence, attention, visual perception, working memory, and phonological awareness. 274 students, aged 11-15 years, participated in the study.

A two-level approach based on the joint use of Kohonen's Self-Organizing Map (SOMs) and the k-means clustering algorithm was compared with another approach based only on the k-means algorithm. The resulting profiles were then predicted by Adaptive Boosting (AdaBoost) and Artificial Neural Networks (ANN) supervised algorithms. The results showed that the two-level approach combined with the ANN algorithm gives the best solution for this problem, allowing to develop a useful instrument for predicting the students' cognitive profile.

Testing the structure of network communities using the total entropy fit permutation test.

Wednesday, 13th July - 16:20: Symposium: New strategies for dimensionality analysis using exploratory graph analysis (Room A) - Symposium Presentation

Dr. Hudson Golino [1]

1. *University of Virginia*

The correct identification of communities in networks is a long-standing problem in computer science, psychometrics, and algorithm complexity. In the field of network psychometrics, community detection algorithms are known to be a very accurate proxy for the structure of latent factors. It has been widely used in the past five years instead of traditional factor analysis. However, when a community detection algorithm identifies a specific partition of nodes (or items) into communities (or factors), there's no guarantee that the estimated structure reflects, indeed, the best partition of a multidimensional space. A possible way to search for the optimal partition of a multidimensional space into specific communities is to test all possible combinations of variables (or nodes) into communities. But this strategy is not computationally feasible, quickly escalating to hundreds of millions of possible combinations. In this talk, I will present an algorithm to test the structure of network communities that is a viable alternative to the computationally unfeasible use of all combinations of variables into communities. This new algorithm (the total entropy fit permutation test) uses a topological analysis of search space landscapes to search for a community structure in a network that minimizes the entropy fit index, giving an initial "guess" for the optimal structure. The search landscape for the total entropy fit permutation test is a collection of M permutations of the "best guess structure." A brief simulation will show how the total entropy fit permutation test works for different types of structures and initial "best guess" solutions.

A Bayesian approach for dimensionality assessment in network psychometrics

Wednesday, 13th July - 16:35: Symposium: New strategies for dimensionality analysis using exploratory graph analysis (Room A) - Symposium Presentation

Dr. Dingjing Shi [1], ***Dr. Hudson Golino*** [2]

1. *University of Oklahoma,* ***2.*** *University of Virginia*

Examining the structure and dimensions of variables is essential to understanding many psychological data. Recently, Exploratory Graphical Analysis (EGA), built upon Gaussian graphical models (GGM; undirected networks), was developed to explore dimensions and estimate the number of factors underlying multivariate data. EGA was found to be superior to traditional techniques such as parallel analysis and Kaiser's rule in assessing dimensionalities (Golino et al., 2020). In GGM, lasso (least absolute shrinkage and selection operation) regularization, associated with frequentist inference, remains the default estimation method. In this presentation, a Bayesian approach to estimating the GGM and exploring the dimensionality structure of multidimensional data (termed Bayesian EGA or BEGA) will be introduced. Instead of obtaining a fixed parameter estimate, BEGA estimates the posterior probabilities of the graphical structures to assess the conditional dependence relations among nodes (variables), using Bayesian Gaussian graphical model estimation. Monte Carlo simulations were conducted and suggested that BEGA outperformed EGA and parallel analysis in determining the dimensions when the sample size is small (e.g., N<500). The outperformance is more apparent when factor loadings are medium to small (e.g., loading < 0.55) under small data conditions. The performance of BEGA and EGA were similar in most large sample conditions (e.g., N=1000). BEGA performed worse than EGA under large item pools (e.g., nItem > 60). The study suggested the potential to extend the regularization-based EGA method to the Bayesian framework. The Bayesian methodology has the additional potential to quantify network predictability.

Optimizing Walktrap's community detection in networks using the total entropy fit index

Wednesday, 13th July - 16:50: Symposium: New strategies for dimensionality analysis using exploratory graph analysis (Room A) - Symposium Presentation

Ms. Laura Jamison [1], ***Dr. Hudson Golino*** [1], ***Dr. Alexander Christensen*** [2]

1. *University of Virginia,* ***2.*** *University of Pennsylvania*

Exploratory graph analysis (EGA) is used to estimate the structural organization of variables, uncovering latent dimensions as clusters of nodes. EGA first estimates a weighted network then uses the walktrap algorithm to detect clusters of nodes. The walktrap algorithm uses random walks to estimate the topography of a graph. The number of random walks taken is typically set statically. We will discuss a new approach to optimizing the number of steps by iteratively varying them and employing the total entropy fit index as a fit index to identify the number of steps that best fit the data. Results from a Monte-Carlo simulation varying data structures indicate that the proposed method is most effective for a higher number of variables per factor and when variables are polytomous. Varying the number of steps is important as spurious connections are introduced between communities. As these are common data structures in psychological research, using an optimization method when employing the walktrap algorithm is important for identifying the correct structure of a network.

Modeling cluster-level constructs with individual-level measures

Wednesday, 13th July - 16:20: Multilevel Modelling and Factor Analysis (Room C) - Individual Oral Presentation

Dr. Suzanne Jak [1], ***Dr. Terrence Jorgensen*** [1], ***Dr. Barbara Nevicka*** [1], ***Ms. Debby ten Hove*** [1]

1. *University of Amsterdam*

Researchers frequently use the responses of individuals in clusters to measure constructs at the cluster level. For example, student's evaluations may be used to measure the teaching quality of instructors, patient reports may be used to evaluate social skills of therapists, and residents ratings may be used to evaluate neighborhood safety.

When multiple items are used to measure such cluster-level constructs, multilevel confirmatory factor models are useful. These models allow for the evaluation of the factor structure at the cluster level (modeling the (co)variances among item means across clusters), and at the individual level (modeling the (co)variances across individuals within clusters).

If the cluster-level construct, for example teacher quality, would be perfectly measured using the responses of students, all students evaluating the same teacher would have exactly the same item scores. In that case, there will not be any systematic variance in the item scores within clusters (only sampling error), so there will be nothing to model at the individual level.

In practice, individuals do not all provide the same responses to the items, leading to systematic variance (and covariance) to be explained at the individual level. The question then arises how the variance within clusters should be modeled. In this talk, I will review some of the interpretational difficulties related to existing two-level models for cluster-level constructs and I will discuss possible alternative options.

Nonparametric IRT models for two-level test data

Wednesday, 13th July - 16:35: Multilevel Modelling and Factor Analysis (Room C) - Individual Oral Presentation

Ms. Letty Koopman [1], ***Mr. Bonne J. H. Zijlstra*** [1], ***Prof. Andries van der Ark*** [1]

1. *University of Amsterdam*

Two-level test data arise when respondents to a test or questionnaire are nested in groups. Mokken scale analysis consist of a selection of methods to evaluate whether a set of items can be used to scale respondents and items, based on Nonparametric Item Response Theory (NIRT) models. Traditional Mokken scale analysis is not suitable for two-level test data because a) the methods assume a simple random sample, which is violated in two-level test data, and b) there exist no methods to evaluate whether groups can be scaled. In this talk we present four two-level NIRT models and discuss several observable properties on the respondent- and the group-level that these models imply. In addition, we present an overview of methods that use these properties to evaluate the fit of the two-level NIRT models. We conclude with suggestions for future research.

Multilevel X-Learner: Extending meta-learners for causal inference with clustered data

Wednesday, 13th July - 16:50: Multilevel Modelling and Factor Analysis (Room C) - Individual Oral Presentation

Prof. Jee-Seon Kim [1], ***Ms. Xiangyi Liao*** [1], ***Dr. Wen Wei Loh*** [2]

1. *University of Wisconsin - Madison,* ***2.*** *Ghent University*

Efforts to estimate treatment effects and draw causal inferences based on observational data are increasingly relevant with the abundance of such data in the social and behavioral sciences. This study proposes multilevel X-learner methods by generalizing the meta-learner to account for clustered structure and dependency in multilevel data. Our algorithm is built on base learners that handle clustered data and also uses multilevel propensity score weights for the X-learner. Simulation studies were conducted to investigate the importance and effectiveness of the choice of machine learning base learners and propensity score estimation methods across various forms of clustered data and multilevel models, such as random slope, three-level, and crossed random-effects models. The empirical data analysis illustrates the analytic steps to implement the method and interpret the results. We conclude with providing recommendations and guidelines in using the multilevel X-learner and discussing its future directions.

A module selection between subtests for improving measurement precision in multidimensional multistage testing

Wednesday, 13th July - 16:20: IRT and Computerized Tests (Room D) - Individual Oral Presentation

Ms. Yi-Ling Wu [1], ***Mr. Huang Yao-Hsuan*** [1], ***Dr. Chia-Wen Chen*** [2], ***Prof. Po-Hsi Chen*** [1]

1. *National Taiwan Normal University Research Center for Psychological and Educational Testing,* ***2.*** *Centre for Educational Measurement at University of Oslo (CEMO)*

Multistage testing (MST) is an algorithm that adaptively selects a set of items (i.e. a module) maximizing the test information as building blocks for a computerized test. Empirically, a test is often divided into subtests each of which measures a unidimensional construct. For example, an English test often has the subtests for reading, listening, speaking, and writing tasks. The drawback of implementing unidimensional MST independently for each subtest is that the item selection ignored the information about the covariance between the abilities that could be obtained priorly from the empirical data prior to the current tests. This study aims to propose an MST for the subtest design with a multidimensional item selection algorithm that takes the prior information of covariance between abilities and the responses to the administered subtests into account for improving the measurement precision of ability estimation in the current subtests.

We conducted a simulation study to investigate the extent of measurement precision in applying the proposed MST approach to a test composed of two subtests compared with the multiple unidimensional MST algorithm. As expected, the result showed that compared to the unidimensional MST, the proposed MST improved the precision of ability estimation (lower root mean square error) measured by the second subtest after the first subtest had been administered. The study ended with a discussion of the implication and limitations.

Matthew effects and metric distortions due to measurement model misspecification

Wednesday, 13th July - 16:35: IRT and Computerized Tests (Room D) - Individual Oral Presentation

Ms. Xiangyi Liao [1], ***Prof. Daniel Bolt*** [1], ***Prof. Jee-Seon Kim*** [1]

1. *University of Wisconsin - Madison*

A Matthew effect refers to a tendency to see a positive correlation between baseline proficiency level and growth, and is frequently observed in the longitudinal assessment of proficiencies like reading comprehension. We theorize that such estimated correlations are highly sensitive to the possibly incorrect assumption of symmetry in the item characteristic curves (ICCs), a feature of all traditional IRT models. In this paper, we show how the observation of Matthew effects may well be attributed to the presence of negative ICC asymmetry, a form of asymmetry anticipated in the presence of proficiency-related guessing. The result of the misspecification is a compression of latent metric units at the lower end of the scale, a compression that implies that students who start lower on the metric will be credited with lesser gains than students that start higher even if they grow equivalent amounts. Simulation and real data studies demonstrate the introduction of the pseudo-correlation and the inherent difficulty in identifying the misspecification using traditional statistical goodness-of-fit criteria.

Effect of matching/weighting equating samples during the pandemic

Wednesday, 13th July - 16:50: IRT and Computerized Tests (Room D) - Individual Oral Presentation

Dr. Kyoungwon Bishop [1], ***Dr. Yoon Ah Song*** [2]

1. *WIDA,* ***2.*** *Center for Applied Linguistics*

This study investigates the impact of subgroup changes in the equating data in a large language assessment administered during the COVID-19 pandemic. Due to remote learning and student absences during testing, changes in the ethnic composition and ability distribution were observed in the data during the pandemic period. When subgroups of test takers have different levels of skills measured by the test, the equating function can vary depending on the subgroup compositions (Kolen & Brennan, 2004).
Unequally weighted samples contribute to inaccuracy in equating.
The impact of unusual characteristics of test data during the COVID pandemic to equating is investigated in this study. Specially, subgroup weighting related to equating is investigated via weighting (Lu & Kim, 2021) and matching methods. The weighting technique is intended to make equating data similar to the target population. The matching method balances the percentage of matched units by allowing control units to be matched to multiple treated units. The weighting method is that samples are weighted as a function of subgroup membership, which is related to student performance on the test (Lu & Kim, 2021).
This study compares the weighting/matching methods in equating with empirical data of a large-scale language test and examines to what extents each method impacts parameter estimation in equating. With this study, the authors hope to provide practical guidance to maintain the score stability over time with fluctuations in data.

The use of factor scores in linking depression scales

Wednesday, 13th July - 17:05: IRT and Computerized Tests (Room D) - Individual Oral Presentation

Ms. Nika Zahedi [1], ***Ms. Emma Somer*** [1], ***Mr. Nikolas Argiropoulos*** [2], ***Dr. Milica Miocevic*** [1]

1. *McGill University,* ***2.*** *Concordia University*

In depression research, the routine practice is to use a single depression scale without providing a rationale for choosing that specific instrument over alternatives (Fried 2017). To assess whether different scales would yield the same depression diagnosis, we used data available through the PROsetta Stone project which collected scores on the BDI-II, HADS, and PROMIS for the same set of participants. We harmonized these three measures using different methods: 1. z-scores, 2. factor scores, and 3. cross-walk tables (Choi et al. 2014). Using harmonized scores, we illustrate the distributions, correlations, and intraclass correlation coefficients (ICCs; capturing the ratio of between-participant variance to total variance across depression measures and across harmonization methods). Additionally, we compare the mediated effect of physical ability on depression through social ability across depression measures and harmonization methods. We found that 1. factor scores lead to the largest ICC and strongest correlations among harmonized scores from different measures, 2. BDI-II and PROMIS scores are strongly correlated regardless of the harmonization method. 3. the effect of physical ability on depression through social ability is significant in all cases, and the difference in magnitude was negligible. Therefore, each person's depression score changed as a function of the depression scale and harmonization method used, however, no difference was observed in the estimated structural paths in mediation analysis. To follow-up on the results, we will use simulation studies to examine conditions in which using factor scores for harmonizing measures leads to optimal statistical properties of the indirect effect in mediation analysis.

Theories and applications of centrality measures in psychometric network analysis

Wednesday, 13th July - 16:20: Network Psychometrics (Room G) - Individual Oral Presentation

Dr. Hsiu-Ting Yu [1], ***Mr. Chi-Yun Deng*** [1]

1. *National Chengchi University*

Centrality indices are commonly used to describe structural aspects of psychological networks. These measures are usually suggested to reflect the various aspects of nodes in a network. Most of these centrality metrics were developed in the context of social networks, it is unclear about the suitability of these measures in a psychological network context. Recent work has illustrated situations where centrality is not a good proxy for causal influence, and cases where peripheral nodes may be more important in reflecting overall network structure. Moreover, these centrality measures usually do not represent physical distances and should not be interpreted as such. Many network analysis softwares generate and output many types of centrality (e.g., closeness, betweenness, degree, strength, eigenvalue and expected influence) automatically; however, there are no clear guidelines on interpretations of each of these measures. There is lack of theoretical work and justifications of applying centrality measures in psychological networks. In this research, we systematically manipulate the weights of edges, sparseness of a network and numbers of nodes to generate various features and patterns of a network. We investigate the relations between theoretical properties of possible generating models and empirical estimates of centrality. The theoretical discussions in the usage of centrality measures, as well as practical guidelines and recommendations in empirical analyses will also be given in the presentation.

Bayesian analysis of a Markov random field for ordinal variables

Wednesday, 13th July - 16:35: Network Psychometrics (Room G) - Individual Oral Presentation

Dr. Maarten Marsman [1], ***Dr. Jonas Haslbeck*** [1]

1. *University of Amsterdam*

Responses to Likert scale questions and symptom severity indicators; ordinal variables abound in psychological measurement. Researchers commonly use Gaussian Graphical Models (GGMs) or truncated GGMs in network psychometrics to model the multivariate distribution of ordinal variables. But ordinal variables do not fit the continuous nature of GGMs, and the truncated GGM can be hard to interpret or sometimes does not match the ordinal variable's characterization. In this talk, we investigate a Markov Random Field (MRF) for ordinal variables as an alternative to the truncated GGM. Initially proposed by Arun Suggala and colleagues, we show that the ordinal MRF is related to Eiji Muraki's generalized partial credit model. The proposed model has the Ising model for binary variables as a special case and generalizes to a model for nominal variables. We will analyze the model using newly developed Bayesian methods.

Detecting redundant items for the purpose of network modeling

Wednesday, 13th July - 16:50: Network Psychometrics (Room G) - Individual Oral Presentation

Mr. Joshua Starr [1], ***Dr. Carl Falk*** [1]

1. *McGill University*

Network models of psychopathology such as the Gaussian graphical model (GGM) estimate systems of *partial* correlations between many variables. As such, network structure is liable to change when even a single variable is removed or added. One case of concern is when two items in the network substantially overlap in content and could be measuring the same symptom. Partial relationships between *that* symptom and other symptoms in the network may become distorted. Networks researchers could benefit from screening tools for identifying redundant items prior to model estimation. The *networktools* R package contains the *goldbricker* function for this purpose, whereby difference-based null hypothesis testing is used to determine whether two items have the same zero-order correlations with *other* items (i.e., same topology). But this approach appears to suffer from backwards logic and has, to our knowledge, never been formally evaluated. We adapted two equivalence testing (ET) approaches – one that tests topology and one that tests the item pair correlations directly – and evaluated performance relative to *goldbricker* in simulations. Manipulated conditions included sample size, strength of all correlations, and the correlation between target items as well as their topology. While *goldbricker* had good ability to flag redundancy across conditions, it had high false positive rates at small sample sizes. Both ET-based approaches maintained nominal false positive rates and had good ability to detect redundancy as sample size increased. ET on the item pair correlation directly may offer a good balance of true and false positives, especially at smaller sample sizes.

Targeting toward inferential goals in Bayesian nonparametric Rasch models

Wednesday, 13th July - 16:20: Estimation of Latent Traits (Room E) - Individual Oral Presentation

Prof. JoonHo Lee [1], ***Prof. Stefanie Wind*** [1]

1. *The University of Alabama*

Researchers who use measurement models (e.g., Rasch models) to estimate student achievement on educational assessments do so with different inferential goals in mind. Whereas some purposes focus on accurately estimating abilities for individual students, studies on large-scale assessments may focus on modeling distributions of latent traits. These inferential goals call for different estimation approaches.

Researchers have proposed various techniques to obtain accurate estimates of individual-specific latent traits. Such proposals tend to focus on relaxing the normality assumption to protect against model misspecification. We consider an alternative Bayesian method developed to facilitate inferences about the empirical distribution and ranks of latent variables. We explore posterior summary methods directly targeted toward inferential goals via a choice of appropriate loss functions, such as constrained Bayes or triple-goal estimators. We then examine how two strategies: (a) adopting flexible nonparametric models, and (b) choosing targeted loss functions work or fail under varying levels of test reliability.

Our results suggest that estimators that produce the best estimate of the collection of individual latent traits can produce poor estimates of the shape of latent traits and vice-versa. When test reliability is low, specifying a flexible nonparametric model is generally not effective compared to a Gaussian model. A simple parametric model combined with a posterior summary method targeted toward an inferential goal performs better. Since the flexible models are designed adapt to the data, they require large examinee sample sizes and long instruments to effectively utilize information from the data. We consider implications for research and practice.

Statistical scoring rules in person parameter inference: Some general pitfalls.

Wednesday, 13th July - 16:35: Estimation of Latent Traits (Room E) - Individual Oral Presentation

Dr. Pascal Jordan [1]

1. *University of Hamburg*

The main paradigms (Maximum-Likelihood, Bayes) of deducing statistical estimators enable the derivation of statistically efficient estimators in a broad variety of models. However, when specifically inferring person parameters, additional aspects of the inference which go beyond the mere statistical framework also become important. For example, (a) test takers may expect their ability estimates to increase when providing a correct instead of an incorrect response; (b) test takers may also expect their ability estimates not to be dominated by the performance in a small part of the test; and (c) the added value of incorporating information on the distribution of the latent ability in the population should be properly reflected in the ability estimate of a test taker. In this talk, I provide a short overview of some challenging effects which are at odds with these expectations and which occur in a variety of prominently used models - including multidimensional compensatory IRT models, factor analysis models and drift-diffusion models.

Estimation and use of ability distributions

Wednesday, 13th July - 16:50: Estimation of Latent Traits (Room E) - Individual Oral Presentation

Dr. Won-Chan Lee [1], ***Dr. Tianyou Wang*** [2], ***Dr. Hyung Jin Kim*** [1], ***Dr. Robert Brennan*** [1]

1. *University of Iowa,* ***2.*** *Consultant*

The primary concern for many IRT applications has centered around the accuracy of item parameter estimation. In some contexts (such as IRT scale linking and equating), however, the ability distribution plays an important role and often is paired with item parameter estimates that are expressed on that ability scale. Most modern IRT calibration programs implement the EM algorithm providing two options with respect to specifying a prior ability distribution: (a) updating the prior ability distribution at every EM cycle using the posterior distribution from the previous cycle and (b) fixing the prior to the standard normal distribution throughout the entire EM cycles. Furthermore, rescaling of the ability distribution to a standard scale with a mean of zero and a standard deviation of one is an additional option that affects the final posterior ability distribution and item parameter estimates. Obviously, there are a few possible ability distributions associated with different calibration options (i.e., the standard normal prior, final posterior with or without rescaling) that can be used jointly with the final item parameter estimates; however, the theory is not always clear about the choice. The present study, via simulation, concerns about how well the population and sample ability distributions are recovered from various calibration approaches. Recovery of item parameters are also examined. Simulation factors include variations in IRT model, test length, sample size, and shape of the population distribution.

Assessing students' abilities: an hybrid archetypal analysis and IRT approach

Wednesday, 13th July - 17:05: Estimation of Latent Traits (Room E) - Individual Oral Presentation

Dr. Lucio Palazzo [1], ***Prof. Francesco Palumbo*** [1]

1. *University of Naples Federico II*

Item response theory (IRT, Hambleton & Swaminathan, 1985) measures latent traits from one or more sets of manifest variables, namely items, by defining the relations between the observed variables (e.g., item responses to a test) and the latent variables. Three of the five higher education items that refer to student's abilities, as defined by the Dublin descriptors, are considered in this proposal: knowledge, application, judgment. Moreover, IRT assumes that students belong to homogeneous groups concerning these abilities. The semi-parametric multivariate latent class IRT models (MultiLCIRT, Bartolucci (2007), Bacci et al. (2014)) represent a practical approach to finding groups by aggregating the units with respect to group's average abilities. However, assessors generally want to discover "extreme" groups of students: the most skilled, but especially those profiles that have peculiar deficits for one or more learning abilities, to define a recommendation system helping the student to fill the gaps. Archetypical analysis (AA) represents an effective data partitioning alternative to the clustering approaches around the means. The archetypes are observed or unobserved extreme points lying on the convex-hull, minimizing the sum of the squared distances from all points. The algorithm computes a membership vector for each unit with respect to each archetype. This proposal integrates the LC-IRT model with the probabilistic archetypal analysis (PAA, Seth & Eugster, 2016). It presents a hybrid estimation algorithm for the multidimensional LC-IRT model, which iteratively computes the latent variables and the units' memberships to a set of k archetypes, where k is assumed to be known.

Predicting IRT 3PL parameters via a neural network model

Wednesday, 13th July - 16:20: Neural Networks (Room F) - Individual Oral Presentation

Dr. Dmitry Belov [1], ***Dr. Anna Topczewski*** [1]

1. *Law School Admission Council*

Moving a testing program online may be a necessary risk that an organization needs to take. A major threat in doing so is that content may be easier to harvest. Items suffer from preknowledge because of using the same item repeatedly and exposing items during pretesting. Items suffering from preknowledge must be removed from the bank; thus, increasing item development demands. Creating a more efficient item development process by understanding what item features lead to different IRT 3PL parameters can provide a feedback loop to test developers. And further, if the 3PL item parameters could be known without pretesting, item preknowledge before operational use could be eliminated entirely.

To predict IRT 3PL item parameters 78 features were extracted via natural language processing applied to item's text. Based on a feature's usability analysis, a composite model was developed to predict each 3PL item parameter separately. One neural network (bNN) with more hyperparameters (number of hidden layers, number of nodes in each layer, etc.) was used to predict difficulty and two neural networks with fewer hyperparameters were used to predict discrimination and pseudo-guessing. 1742 items from a high-stakes testing program were used to train and validate the composite model. Our approach was compared with predicting item parameters simultaneously via bNN with three nodes (instead of one) in the output layer. Results of cross validation showed over 15% improvement of the coefficient of determination when using the composite model. Detailed results of validation of the composite model will be presented using pretested items.

Deep learning generalized structured component analysis: A knowledge-based nonlinear multivariate predictive method

Wednesday, 13th July - 16:35: Neural Networks (Room F) - Individual Oral Presentation

Mr. Gyeongcheol Cho [1], ***Dr. Heungsun Hwang*** [1]

1. *McGill University*

Generalized structured component analysis (GSCA) is a multivariate method for specifying and examining interrelationships between observed variables and components. GSCA has been extended to handle various types of data and estimate more complex models. Nonetheless, all the extensions to date assume that a component is always defined as a linear function of observed variables. The linearity assumption can be less optimal when observed variables for a component are nonlinearly related, often reducing the component's predictive power. To address this issue, we combine deep learning and GSCA into a single framework to allow a component to be a nonlinear function of observed variables without specifying the exact functional form in advance. This new method, termed deep learning generalized structured component analysis (DL-GSCA), aims to maximize the predictive power of components while their relationships remain interpretable. Our simulation study shows that DL-GSCA produces components with greater predictive power than those from traditional GSCA in the presence of nonlinear associations between observed variables. We also apply the proposed method to the Human Development Index data to illustrate its empirical usefulness.

Improving measurement models using deep neural networks

Wednesday, 13th July - 16:50: Neural Networks (Room F) - Individual Oral Presentation

Prof. Artur Pokropek [1], ***Dr. Marek Muszyński*** [1], ***Dr. Tomasz Żółtak*** [1]

1. *IFiS Polish Academy of Sciences*

Detecting statistical misspecifications in measurement models is a challenging undertaking. Although some analytical methods exist (e.g., detecting lack of conditional independence or differential item functioning (DIF)), their efficacy is questioned. We present a method based on Deep Neural Networks (DNNs) that overcomes traditional approaches. We present advances in the proposed approach on the example of simulation studies and empirical examples. In this study we focus on: (a) detecting differential item functioning (DIF) in a large number of group settings and (b) detecting misspecifications caused by the existence of response styles (RS). For DIF detection we compared the proposed model with the most popular traditional methods: a) lagrange multiplier (also referred to as modification indices), b) logistic DIF detection, and c) sequential procedure introduced with the IRT alignment approach. Simulation studies showed that the proposed method outperformed traditional methods in almost all scenarios, or that it was at least as accurate as the best one of them. For model misspecification caused by RS in rating scales, we provided a simulation study, where the proposed approach was tested against the multidimensional generalized partial credit model and IRtree models. The obtained results showed that both midpoint and extreme response patterns could be successfully detected by the approach based on DNNs and that the detection rate is generally higher than in the traditional approaches based on model comparisons. Finally, real-life applications of the DNNs approach are presented using ESS and PIAAC data.

PowerGraph: Using neural networks and principal components to determine multivariate statistical power trade-offs

Wednesday, 13th July - 17:05: Neural Networks (Room F) - Individual Oral Presentation

Mr. Ajinkya Mulay [1], ***Dr. Sean Lane*** [1], ***Dr. Erin Hennes*** [1]

1. *Purdue University*

Statistical power estimation for studies with multiple model parameters is a multivariate problem since the power for individual parameters of interest cannot be reliably estimated univariately. For each parameter, sampling variability, correlation with, and variance explained relative to one parameter will impact the power for another parameter, all usual univariate considerations being equal. Explicit solutions in such cases, especially for models with many parameters, are either impractical or impossible to solve, leaving researchers to the prevailing method of simulating power. Furthermore, the point estimates for a vector of model parameters are uncertain, and the impact of parameters' inaccuracy is unknown. In such cases, we use sensitivity analysis wherein we simulate multiple combinations of possible observable parameter vectors to understand power trade-offs. A limitation to this approach is that it is expensive to generate sufficient sensitivity combinations to accurately map the power trade-off function in high-dimensional spaces for the models that social scientists estimate. This paper explores the efficient estimation and graphing of statistical power for a study over varying model parameter combinations. We propose a simple, generalizable machine learning-based solution, POWERNETWORK, to cut the computational cost of power estimation to less than 7% of the baseline. We show that such a model can achieve over 97% testing accuracy, within 1% error for statistical power. Furthermore, we provide a novel sampling technique for studies with high-dimensional models to produce balanced datasets to improve the performance of POWERNETWORK.

A model-assisted approach for distinguishing two nonresponses in achievement test or survey data

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Yu-Wei Chang [1]

1. *National Chengchi University*

Regarding the missing data in survey questionnaire or achievement tests, Weeks, von Davier, and Yamamoto (2016) suggested to distinguish missing data with shorter and longer response time since the later one has high probability to be related to low ability while the former does not. We suggest an IRT tree model with four end nodes of not-all-distinct response categories (TR4) to model nonresponse. The model is capable of distinguishing the two types of nonresponse, based on an proposed index. In the current presentation, we will introduce the index and illustrate how to quantify the uncertainty of the index. Some simulations are conducted to demonstrate the efficiency of distinguishing the two types of nonresponse using the index and the variance quantification. The method is further applied to PISA 2012 data for illustration.

Socioemotional competences and vocational interests: A network analysis

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Nelson Hauck [1], ***Dr. Felipe Valentini*** [1], ***Dr. Ana Carla Crispim*** [2], ***Dr. Ricardo Primi*** [1], ***Dr. Rodolfo Augusto Matteo Ambiel*** [1]

1. *University São Francisco, Brazil,* ***2.*** *Ayrton Senna Institute*

Socioemotional competences (SEC) are empirically connected to a broad array of positive life outcomes, and they also associate with the interests that individuals will manifest in their adult lives, especially vocational interests. However, the available knowledge about the connection between SEC and vocational interests was mainly derived from studies that relied on correlations and confirmatory models. In the current study, we employed network analysis to explore the unique associations between the (Big Five) dimensions of SEC and the Holland's vocational interests. We collected data from 5,184 Brazilian students from 266 elementary and middle schools. Students answered the Senna socioemotional inventory and the 18rest (an instrument designed to assess the Holland's factors). Results revealed that Open-mindedness is a central variable that connects SEC to interests, mainly social, investigative, and artistic interests. The factor Amity was connected to interests for social activities; Engagement with others was connected to entrepreneurship; and Self-Management was connected to all interests. Our findings suggest that, while some SEC might be essential for specific professional activities, open-mindedness appears to be a core competence to all job areas.

Constructing parcels with the continuous response model

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Weldon Smith [1], ***Dr. HyeSun Lee*** [1]

1. *California State University Channel Islands*

Parceling offers higher reliabilities, allows models to be fit with smaller sample sizes, and reduces the likelihood of distributional violations especially with categorical or dichotomous indicators. This study employed the item response theory (IRT) based continuous response model (CRM) to construct parcels. Whereas common approaches to parceling result in a loss of indicator information, the CRM offers all the benefits of IRT as well as additional information about parcel standard errors and item parameters. Thus, using the CRM for parceling lets researchers make more informed decisions on the construction and a greater understanding of the psychometric properties of parcels. Through a simulation, the current study compared the performance of parcels constructed using traditional approaches and the CRM in terms of relationships observed, satisfaction of model assumptions, model fit, and psychometric information provided. The simulation utilized different construction approaches such as combining items with similar loadings, balanced slopes, balanced intercepts, correlated items, and parceling items at random. English language test data from over 1000 test takers were employed to support findings from the simulation study. Also, the procedures to apply the CRM parceling approach to test data were demonstrated based on the empirical data. Finally, common criticisms of parceling were discussed in relation to the use of the CRM approach for parceling, including multidimensionality and measurement invariance issues.

Multi-level reliabilities with missing data

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Minju Hong [1], ***Dr. Zhenqiu Lu*** [1]

1. *University of Georgia*

Reliabilities are widely used in social and behavioral sciences. The most commonly used reliability is the Cronbach alpha (Cronbach, 1951). However, it has been criticized because of its strict assumption (Cortina, 1993). The coefficient omega (McDonald, 1997) has also been proposed. However, they cannot handle multilevel data. Most reliabilities are single-level reliabilities. In 2014, Geldhof et al. (2014) proposed level-specific reliabilities for multi-level data, but they did not investigate the impacts of the missing values. In reality, missing values are very common in educational test due to various reasons.

The main purpose of this study was to investigate the performance of reliabilities for multilevel data with missing values. We examined the accuracy and convergence of multilevel reliabilities with missing values. The single-level reliabilities were also compared. In the simulation study, we considered different conditions, including missing data mechanisms, missing data techniques, missing data proportions, sample sizes at multilevel levels, and intra-class correlations. Results showed that in general multilevel reliabilities performed better than single-level reliabilities. For example, with a high intra-class correlation (0.3), a large number of clusters (100 clusters) and a small (or medium?) cluster sizes (15 samples per cluster), the between-group level reliabilities performed better than single-level reliabilities. Regarding missing data techniques, list-wise deletion method is not recommended. For example, under the condition of non-ignorable missing data with 30% missingness using list-wise deletion method, all reliabilities showed the worst.

Understanding, calculating, and interpreting R-squared effect size measures in multilevel models

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Mairead Shaw [1], ***Dr. Jessica Flake*** [1]

1. *McGill University*

Multilevel models (MLMs) are widely used in the psychological sciences to analyze clustered data structures such as people nested within groups or trials nested within people. Effect sizes are necessary for contextualizing results from statistical models, and are often required by journals and funders. Rights and Sterba (2019) developed a comprehensive approach for R-squared effect size measures in MLMs. Shaw et al. (2020) developed an R package, r2mlm, for calculating them. During this presentation, attendees will learn how to: (1) Define and understand effect sizes for MLMs; (2) Estimate the effect sizes using lme4 and r2mlm in R; and (3) Interpret the R output. Through lecture and a demonstration with an illustrative example, participants will gain an understanding of the R-squared framework and how to apply it in their research.

Evaluating the Rasch tree method for balanced and unbalanced DIF

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Nana Amma Asamoah [1], ***Dr. Ronna Turner*** [1], ***Dr. Wen-Juo Lo*** [1], ***Dr. Brandon Crawford*** [2], ***Dr. Kristen Jozkowski*** [2]

1. *University of Arkansas,* ***2.*** *Indiana University*

The Rasch tree, a differential item functioning (DIF) detection approach, recursively tests all groups that can be defined based on combinations of available grouping variables to identify groups that have different item difficulty parameters. An advantage of the method is that subgroups do not have to be pre-specified. However, as a global DIF detection method, items responsible for DIF are not automatically flagged. A joint Rasch model is fit for all groups and if significant instabilities are detected based on a grouping variable, the sample is split by that variable. A new unpublished study incorporates Mantel-Haenszel effect sizes for identifying DIF items in Rasch tree analyses (Henninger et al., 2022). The simulations compare DIF identification when using statistical testing versus effect sizes as stopping criteria. They further demonstrate the use of effect sizes in identifying DIF magnitude of items.

We build on these prior studies by comparing true and false DIF detection rates using Rasch trees (implementing MH effect size criteria) with previously unstudied conditions. Prior simulations have been restricted to items favoring one group. We are extending conditions to include balanced and unbalanced DIF item proportions that occur in real-world data. Simulation conditions include test length (10, 20 items); sample size (400, 800, 1200); difference in difficulty parameters of selected items (0, 0.21, 0.43, 0.64, 0.85, 1.06); percentage of items with DIF (10%, 20% and 30%); and type of DIF (balanced and unbalanced). The results will provide further guidelines on the use of Rasch trees in empirical DIF studies.

Machine-learning methods for item difficulty prediction using item text features

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Lubomir Stepanek [1], ***Mrs. Jana Dlouhá*** [2], ***Dr. Patricia Martinkova*** [1]

1. *Czech Academy of Sciences,* ***2.*** *Czech Academy of Sciences and Charles University in Prague*

Item difficulty predictions using various text features extracted from items' wordings may help to build a test appropriately, particularly when pre-tests are limited. In this work, we examine and compare different machine-learning methods for prediction of item difficulty using features from text analysis of item wordings. We employ multivariate regression, support vector machine, regression trees, random forests, and back-propagation neural networks in both frameworks, i.e., as supervised regression and classification algorithms, respectively. Furthermore, for item difficulty classification, we also build naïve Bayes classifier, and the multivariate regression designed in multinomial fashion. While the supervised regression algorithms consider the item difficulty as a continuous dependent variable, the supervised classification approaches treat the item difficulty as a variable split into a few disjunctive classes. Methods are illustrated on items of an English language test of the Czech matura exam. Although the regression and classification tasks could not be mutually compared, within the given task, the models differ in their performance. Using k-fold cross validation and several performance metrics, support vector machines and random forests usually outperform others.

Impact of rapid guessing on country rankings in PISA

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Michalis Michaelides [1], ***Ms. Militsa Ivanova*** [1]

1. *University of Cyprus*

Examinee test-taking effort in achievement tests has been found to have significant impact on test performance. Particularly in low-stakes international large-scale assessments, test-takers face few or no consequences and are not motivated to perform at their best. A common criticism for such programs is that suboptimal test-taking effort is a primary reason for low country performance. We study this claim empirically with secondary data from the PISA 2015 Mathematics and Reading tests, and 56 countries that participated in the computerized administration. Using item response times, rapid guessers on each item were identified as those answering faster than the 15% of the item mean response time at the country level. Examinee response behavior across items was summed up to calculate Response Time Effort (RTE), the percentage of items on which they did not respond rapidly.

Analysis examined whether the country mean score and ranking would change after filtering out test-takers with less-than-ideal RTE scores. By removing examinees who rapid guessed on at least one Mathematics item, i.e. RTE<1, the mean change in rank was 1.9 (mode=0). After removing those with RTE<.95, mean rank change was 1.5 (mode=0). For Reading the mean change in country ranking was slightly higher at 3.1 and 2.7 correspondingly (mode=0 and 1). Smaller changes in rankings were observed when a fixed 5-second threshold was applied to identify rapid guessers. Overall, filtering out test-takers who engaged in rapid guessing, led to minor improvements in the country mean scores, but to negligible changes in country rankings, if any.

Bayesian hierarchical stacking in random effects models

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Mingya Huang [1], ***Prof. David Kaplan*** [1]

1. *University of Wisconsin - Madison*

There are several methods to handle model uncertainty in the Bayesian framework such as Bayesian model averaging, which averages over the model space by the posterior probability density. However, BMA seeks a single best model which assumes the true generating model is in the model space (M-closed setting). Unlike BMA, Bayesian stacking can account for the posterior predictive density when the true model is outside the model space (M-open setting). An issue with Bayesian stacking is that it is an optimization technique that uses input-independent model weights and is not a fully Bayesian inference problem. According to Yao et al. 2021, Bayesian hierarchical stacking applies a hyperprior on the stacking weights which allows for uncertainty in the stacking weights. Considering the variation of group-dependent and individual dependent random effects in multilevel models, this project investigates the predictive performance of BMA, Bayesian stacking, and Bayesian hierarchical stacking using PISA 2008 data. Predictive performance is measured by the leave-one-out cross-validation information criterion and the Kullback-Leibler divergence score.

Comparison of equating methods when DIF is present in common items

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Gamze Kartal [1]

1. *University of Illinois, Urbana-Champaign*

Test equating is widely used for educational tests so that test scores on test forms administered at different times to different examinees can be used interchangeably. Extensive research has been conducted on equating, and it has been shown that differential item functioning (DIF) present in anchor items of the test impacts equating (Huggins, 2014). Considering the previous literature, we hypothesized additional information gain by examining the robustness of equating methods if DIF manifested in the common items in a nonequivalent group design. This study compares the performances of three equating methods: Frequency Estimation Equipercentile, Chained Equipercentile, and the Tucker Linear equating methods. The study contains both simulated and real data components. We consider different sample sizes, DIF magnitudes, mean abilities, and types (uniform or nonuniform) of DIF within the simulation portion. We will also examine how well the properties transfer to real-world data. In this case, we expect that investigating the performances of equating methods when DIF is present in common items will aid the selection of the most robust equating method so that equating results will be valid.

Modeling ordinal variables in blavaan

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Edgar Merkle [1], ***Mr. Benjamin Graves*** [1], ***Ms. Ellen Fitzsimmons*** [1], ***Mr. Ronald Flores*** [1], ***Dr. Mauricio Garnier-Villarreal*** [2]

1. *University of Missouri,* ***2.*** *Vrije Universiteit Amsterdam*

The poster will describe recent additions to blavaan, which is an R package for Bayesian estimation of structural equation models using Stan and JAGS. The most notable addition involves the ability to model ordinal observed variables (either exclusively or alongside continuous variables). We will describe our approach to estimating ordinal SEMs in Stan and contrast it with other approaches. We will also describe model summaries and posterior checks that are available for these models, which are difficult to obtain elsewhere. These new features will be illustrated with examples involving code and real data.

Estimating the accuracy of classification into pass/fail conditions of the criterion-referenced chiropractic written clinical competence exam

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Igor Himelfarb [1], ***Dr. Nai-En Tang*** [1], ***Mr. H. Daniel Edi*** [2], ***Mr. Guoliang Fang*** [3]

1. *National Board of Chiropractic Examiners,* ***2.*** *University of Northern Colorado,* ***3.*** *Colorado State University Global*

Validity and reliability of classification decisions made by high-stake exams are of high importance to test developers and psychometricians overseeing operational testing programs. This paper presented a decision consistency (reliability of classification) study into pass/fail conditions for a reduced form of the Chiropractic Written Clinical Competence Exam. The methodology was based on three approaches: Classical Test Theory, Item Response Theory, and Bayesian method. The most important finding is that all analytic methods showed high classification reliability supporting item reduction. The findings also revealed that all three methods yielded almost identical estimates of reliability showing comparable performance of methods based on various theoretical frameworks for dichotomously-scored multiple-choice items.

The classification based on CTT was highly reliable. Approximately 70% of test takes were classified in the mastery condition while the classification into non-mastery ranged between 24 and 26%. The 3PL IRT-based results revealed that the parameter estimates are of the same magnitudes between the full and reduced forms of the exam. The averages of the parameter estimates were comparable. The results based on Bayesian method unveiled comparable classification of the full and reduced forms. The accuracy of classification ranged between 93 and 96%. The kappa values ranging from .84 to .90.

The results of this study contribute to the decision consistency and reliability of classification literature by providing an example of application of previously proposed theoretical methods to evaluate decision consistency to real operational data. The findings revealed that in case of dichotomously scored MC items, all three methods yield comparable results.

Regularized robust confidence interval estimation in Cognitive Diagnostic Models

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Candice Pattisapu [1], ***Dr. Richard M. Golden*** [1]

1. *University of Texas at Dallas*

The goal of this study was to investigate effects of model misspecification on confidence interval coverage probabilities for Cognitive Diagnostic Models (CDMs) in the presence of regularization. Prior work has shown that the Robust covariance matrix estimator (White, 1982) is an effective estimator of standard errors of item parameters in the presence of model misspecification for CDMs (Liu et al., 2019). In the present study, Tatsuoka's (1990) 15 question fraction-subtraction data set (n=536) was fit to a five latent skill DINA CDM using a uniform attribute profile distribution. A sample size dependent prior was introduced to regularize the DINA CDM using MAP estimation (Dai et al., 2019). Next, parametric bootstrap data sets of different sample size sizes were generated from the fitted model and fit to both the original model and a misspecified version of the original model. Consistent with theory (White, 1982), for large sample sizes, we found that the Robust covariance estimator was more effective than both the Hessian and OPG (e.g., Philipp et al. 2018) covariance matrix estimators for estimating confidence interval coverage probabilities in the presence of model misspecification. In addition, we found that when combined with either the Robust or Hessian covariance matrix estimators our regularization term supported good confidence interval coverage probability estimates for relatively small sample sizes (n=240). Implications of these results for estimation and inference in the presence of possible model misspecification in small sample sizes are discussed for not only CDMs but also larger classes of probability models.

Empirical selection of referent variables: Using an iterative MIMIC-interaction modeling

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Cheng-Hsien Li [1], ***Dr. Guo-Wei Sun*** [1]

1. *National Sun Yat-sen University*

Examination of cross-group latent differences in a wide array of research contexts has brought measurement invariance (MI) into the research spotlight in behavioral and social science. One potential limitation related to model identification has been discussed yet received little attention in multiple-group CFA modeling: correctly identifying referent variables. A statistical approach, MIMIC-interaction modeling has been suggested to identify credible referent variables. This study intends to show the superiority of an "iterative" strategy at correctly locating referent variables, compared to a "noniterative" strategy. A Monte Carlo simulation design was used to determine the effects of different percentage of noninvariant variables, magnitude of noninvariance, magnitude of group latent differences, and sample size in a one-dimension measurement model. Data generation and analysis were performed with M*plus* 8. The accuracy rate was used to assess the performance of the two different strategies in correctly identifying credible referent variables from among truly invariant observed variables in the population, along with a benchmark criterion, the probability of randomly selecting truly invariant variables from among all the observed variables. Results showed that most accuracy rates of the iterative strategy were perfect or nearly perfect when the percentage of noninvariance was less than 30%; even when there were more than 30% noninvariance, the iterative strategy still yielded higher accuracy rates than the noniterative strategy, especially for large samples. In addition, the iterative strategy significantly outperformed its counterpart when the percentage of noninvariance increased, and the cross-group latent differences increased, suggesting that the iterative strategy is practically recommendable.

Item difficulty prediction using computational psychometrics and linguistic algorithms

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Jana Dlouhá [1], ***Mr. Lubomir Stepanek*** [2], ***Dr. Patricia Martinkova*** [2]

1. *Czech Academy of Sciences and Charles University in Prague,* ***2.*** *Czech Academy of Sciences*

Item characteristics such as difficulty or discrimination power are typically estimated from data. When little or no data are available at the pre-test, the test developers rely on their experience in how items of different content and wording influence item characteristics. In this work, we explore various item features gathered from text analysis of item wording to predict item difficulty. We illustrate the methods using the English language test of the Czech matura exam.

Investigating differential item functioning via odds ratio in CDM

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Ya-Hui Su [1], ***Ms. Tzu-Ying Chen*** [1]

1. *National Chung Cheng University*

The increasing number of tests being developed has prompted more people to investigate the association between test items and skill attributes and state of knowledge, spurring the development of the cognitive diagnosis model. Studies that detect differential item functioning (DIF) under this model have predominantly adopted the Mantel–Haenszel (MH) method. Jin et al. (2018), assuming that latent traits were continuous, used odds ratio (OR) to examine DIF under the Rasch model and observed that OR outperformed the traditional MH method in terms of type I error rate control and statistical power. However, no studies have applied OR in DIF detection under the cognitive diagnosis model, and none have investigated how DIF detection is affected by the use of different cognitive diagnosis models, proportion of DIF items, and type of DIF. Therefore, this study compared the effectiveness of DIF detection obtained by the MH method, MH method with purification procedure, MH method with attribute mastery profile as matching criterion, OR method, and OR method with purification procedure. According to the results, the effectiveness of DIF detection was affected by sample size and the proportion of DIF items; specifically, a large sample size and a high proportion of DIF items were associated with increased and decreased statistical power, respectively. The purification procedure increased the DIF detection effectiveness and reduced the type I error rate in both the OR and MH methods.

Opportunities and problems of collecting paradata in web-based studies

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Tomasz Żółtak [1], ***Prof. Artur Pokropek*** [2], ***Dr. Marek Muszyński*** [1]

1. *IFiS Polish Academy of Sciences,* ***2.*** *Polish Academy of Science*

Computer-based data collection introduced new opportunities to research response processes in social sciences measurements. A rich set of data, including response time, response editing, number and order of actions, or trajectory of cursor (mouse) movement, can be collected and further used to construct response-process indicators that have proven to be valuable in understanding respondents' actions in cognitive psychology, educational testing or survey methodology.

So far most process data were collected in controlled laboratory or testing service environments, using standardized hardware and software to assure strictly the same conditions of human-interface interaction. Nevertheless, process data may be successfully collected in other settings, such as popular web-survey applications. Although very promising, web-based paradata collection encounters specific problems. Using standardized hardware and software is in general very limited, resulting in technical problems, arising from differences in configurations.

In the presentation we will illustrate the possibility to construct mouse-moves process indicators using data from two web surveys that involved experimental settings. They were conducted using an open-source Lime Survey platform with a rich set of paradata collected using a simple JavaScript applet. We will concentrate on discussing what additional data needed to be collected and what data transformation procedures were needed to be implemented to assure the possibility of further process data standardization to increase its cross-respondents comparability. Finally, we will compare how the aforementioned data standardization improves the validity of mouse-moves measures as careless/insufficient effort responding indicators used as auxiliary variables in IRT scaling of survey items.

Investigating the co-existence of response styles via mixture multidimensional IRTree

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Ömer Emre Can Alagöz [1]***, Prof. Thorsten Meiser*** [1]

1. *University of Mannheim*

IRTree models separate content-related traits from response style (RS) traits by decomposing item responses into multiple binary decision-making processes. These binary processes concern, for instance, 1) whether one agrees with the item content, 2) whether one chooses mid-scale (MRS) or extreme (ERS) categories. This type of modelling involves two constraining assumptions. First, the content trait determines only the direction of a response while extreme or mid-scale category choices are determined only by RS traits. Second, all individuals are assumed to show both types of response styles but differ only in magnitude. In this study, we relax the first assumption by taking into account that not only RS traits but also the content trait affects mid-scale or extreme category choices (Meiser et al., 2019; von Davier & Khorramdel, 2013). For the second assumption, we employ a mixture IRTree approach to investigate whether there are respondents showing only one type of RS. More specifically, the mixture model contains three classes that differ regarding the type of RS that respondents show, while category choices are consistently affected by the content trait in all classes. In a “coexistent-RS” class, distinct RS traits affect both mid-scale and extreme category choices. In an “ERS-only” class, the RS trait influences only the extreme choices, whereas, in an “MRS-only” class, the RS trait determines only the mid-scale category choices. We present simulation results regarding the recovery of the class and IRTree parameters and illustrate our approach with an empirical dataset under realistic conditions.

Measuring patient activation in patients with chronic diseases

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Magdalena Holter [1], ***Mr. Alexander Avian*** [1], ***Prof. Andreas Wedrich*** [1], ***Prof. Andrea Berghold*** [1]

1. *Medical University of Graz*

The number of patients suffering from chronic diseases is increasing. Patient activation plays an important role on their health status (Hibbard et al., 2013), a reliable and valid assessment tool is needed. The Patient Activation Measure (PAM-13) has been proposed for assessing patient activation (Hibbard et al., 2005). It consists of 13 items measuring the skills, knowledge, and confidence of patients in coping with their chronic disease. The main objective of our study is to investigate the psychometric properties of the German PAM-13 in patients with chronic diseases using item response theory (IRT).

This is a questionnaire-based prospective cross-sectional study. The ad-hoc sample investigated is composed of outpatients with macular edema from the Department of Ophthalmology, Medical University of Graz, Austria. Macular edema often occurs in the course of diabetic retinopathy or other retinal vascular diseases like retinal vein occlusion. About 500 patients are included. We are using six questionnaires, including PAM-13. The study started in March 2020 and is finishing recruiting at the end of February 2022.

The results of the IRT analysis will be presented. Several IRT models like the Rasch rating scale model or the partial credit model will be estimated. By comparing fit indices of different models and performing likelihood ratio tests, we will choose the best fitting model for the data. Assumptions like unidimensionality of the construct will be tested. We will discuss the psychometric properties of the German PAM-13 for describing patient activation in patients with chronic diseases exemplified by patients with macular edema.

Vertical scaling of data from a large-scale assessment system

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Prof. Martin Tomasik [1], ***Dr. Charles Driver*** [2], ***Dr. Laura Helbling*** [1], ***Dr. Stéphanie Berger*** [1]

1. *University of Zurich,* ***2.*** *Universtiy of Zurich*

Mindsteps is an online learning and testing platform developed to provide teachers across Switzerland with data that can be used for instruction-related decision making, and to inform students on their current performance and learning gains (Tomasik, Berger & Moser, 2018). Questions in the system are based on the school curriculum, the item bank has tens of thousands of items distributed across eleven competence domains (e.g., “listening comprehension” in English or “form and space” in mathematics). When using the system, teachers and students can set up adaptive or linear assessments that are evaluated using a measurement model based on item response theory. The item parameters of this 1-PL model have so far been estimated for single school grades and subsequently linked to span the whole range between grade 3 and 9 (Berger, Verschoor, Eggen & Moser, 2019). In the past years, more than 100.000 students from Switzerland have been using the system, sometimes irregularly but sometimes also intensively. We present the results of an extensive model testing and the development of new calibration approaches, implemented in a newly developed R package capable of handling the large amounts of raw data at hand. Several models (i.e., from 1-PL to 4-PL, including or not including person-level and item-level covariates such as age, gender, type of item, time of assessment) have been set up and compared with regard to their out-of-sample prediction performance and to grade-to-grade growth. We will discuss the model development and comparison, as well as patterns of results across learning scales.

Evaluating standard error estimators on small clustered samples with heteroscedasticity

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Yichi Zhang [1], ***Dr. Mark Hok Chio Lai*** [1]

1. *University of southern california*

Multilevel modeling (MLM) is commonly used in psychological research to model clustered data. One of the assumptions of MLM is the homogeneity of variance. However, data in experimental research usually come from small samples and have heteroscedastic variances and unbalanced cluster sizes. The fixed-effect estimates produced by the maximum likelihood method remain unbiased, but the standard errors for the fixed-effects are underestimated, resulting in inaccurate inferences and inflated Type I error rates. Small-sample corrections such as the Kenward-Roger (KR) adjustment and the adjusted cluster-robust standard errors (CR-SEs) with the Satterthwaite approximation for t-tests have been used to correct the bias in fixed effects standard errors and provide valid inferences. The current study compares the two standard error estimators with Ordinary Least Squares (OLS), random intercept (RI) and random slope (RS) models with small, clustered data with heteroscedastic variances using a Monte Carlo simulation study. Results show KR-adjusted standard errors with RS models have large biases and inflated Type I error rates for between-cluster effects in the presence of level-two heteroscedasticity. In contrast, the adjusted CR-SEs generally yield results with acceptable biases and maintain Type I error rates close to the nominal level for all examined models. Thus, when the interest is using standard error estimators to make inferences of the between-cluster effect, researchers can choose to use the adjusted CR-SEs with OLS to account for the clustered structure, or RI and RS to guard against unmodeled heterogeneity. The illustrative example demonstrates the use of the adjusted CR-SEs with different models.

Validity study using EFA on the mathematics attitude scale

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Hunwon Choi [1]***, Dr. Youn-Jeng Choi*** [1]

1. *EWHA WOMANS UNIVERSITY*

This article explores the internal structure of the mathematics attitude scale from TIMSS 2019 and examines the similarity and differences in the internal structures among different countries using exploratory factor analysis (EFA). The student background survey related to students' attitudes in mathematics consists of 27 four-point Likert scale items: interest, confidence, and value recognition (nine items per attitude). The International Association for the Evaluation of Educational Achievement (IEA), the institution which conducted this survey, reported a three-factor structure for the survey for decades. But this study assumes that the factor structure can appear differently depending on the country and/or culture. Therefore, we will analyze students' responses from English-speaking countries to prevent language effects. The participants included 25,835 eighth-graders: 6,918 from the United States, 4,755 from Singapore, 3,616 from Ireland, 2,667 from England, and 7,879 from Australia. We extracted a three-factor solution using the inspection of the scree plot and parallel analysis using SPSS version 27 to evaluate the factor structure for this particular population. The initial findings were that several items had low factor loadings for all countries. In addition, three questionnaire items were not suitable for intended attitudes in three or more countries. These results show that the internal structure of the survey might be different among the countries, although the number of factors was the same, and we could propose different sets of subscales for attitudes.

Measurement invariance across age in the Future Events Questionnaire (FEQ)

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Conor Lacey [1], ***Dr. Veronica Cole*** [1]

1. *Wake Forest University*

The current study explores measurement invariance in a scale of work-related expectancies across development in adolescence. Data (N = 663) comes from a longitudinal study which followed young children at risk for maltreatment into adolescence and adulthood. We examine measurement invariance in seven work-related items from the Future Events Questionnaire (FEQ), a measure assessing an adolescents' future expectations in the realm of education, employment, and family. The FEQ was administered twice, once each at ages 14 and 16, and we examine measurement invariance on the basis of age. Separate confirmatory factor analyses run on each age group revealed configural invariance in the work-related items. A unidimensional model fits well among 14-year-olds, $X^2(9) = 13.224$, $p = 0.153$, RMSEA = 0.037, SRMR = 0.035, CFI = 0.999, TLI = 0.998. Though the 16-year-old assessment only contains a subset of items administered to 14-year-olds, a unidimensional model fits here as well, $X^2(5) = 6.693$, $p = 0.245$, RMSEA = 0.032, SRMR = 0.035, CFI = 0.998, and TLI = 0.996. Assuming configural invariance, metric invariance of the common items is assessed by applying moderated nonlinear factor analysis (MNLFA; Bauer, 2017) to a combined dataset including both age groups. Given partial metric non-invariance, we use MNLFA to generate scores which account for the effects of age. The ultimate goal is to use these adjusted scores as indicators in studies of psychosocial functioning in adolescents, tracking how positive attitudes about work relate to mental health outcomes.

Network invariance test: A new way to detect individual heterogeneity

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Ria Hoekstra [1], ***Dr. Sacha Epskamp*** [1], ***Prof. Denny Borsboom*** [1]

1. *University of Amsterdam*

New technological advancements made it possible to collect and analyze data that allow to investigate dynamics on the *individual* level (Hamaker, 2012; Molenaar, 2004). These idiographic research techniques have rapidly gained popularity within psychological research and network analysis in particular. In turn, these applications revealed new challenges specific to the analysis of individual data within network analysis. Identifying and quantifying heterogeneity between individual network models is one of these major challenges. For this purpose, previously, metrics such as visual inspection or centrality measures have been used to quantify the differences between individual network models. However, both measures leave the door wide open to interpret all variability, including supposed differences resulting from ordinary fluctuations in the data such as sampling variation, as heterogeneity. We propose a new way to test individual network model similarity by introducing model equivalence testing techniques to examine heterogeneity between individual network models. By imposing equality (or inequality) constraints between individual networks, one can test, for example, whether all or certain parameters in a network model can be considered equal between individuals. By means of a simulation study, it is shown that imposing constraints on all or several parameters performed well from $t = 100$ onward on both pruned and unpruned estimated network models. Hence providing a more powerful tool to test for individual network differences. The use of this test permits researchers to consider that variability in individual network models can be a result of noise, which otherwise remains overlooked.

Can network analysis help competency modeling in assessment and development centers?

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Molok Khademi [1], ***Mr. Hassan Mahmoudian*** [2], ***Ms. Shirin Rezvanifar*** [2], ***Mr. Meysam Mahmoudian*** [3]

1. *Alzahra University,* ***2.*** *Ph. D candidate, Psychometrics, Allameh Tabataba'i University,* ***3.*** *M. A in General Psychology, Tarbiat Modares University*

Specialists always face many challenges in competency modeling in assessment and development centers. The purpose of this study is to use network analysis in competency modeling in assessment and development centers. For this purpose, 1017 data from the assessment and development center of the National Iranian Oil Company (NIOC) were analyzed in individual interview practice and in six competencies including innovation, planning, decision making, leadership, problem recognition and seeing the big picture. We used network analysis to estimate the network structure of these competencies and calculated the strength centrality. The results showed that all competencies were positively associated in the final network, and the whole network was relatively high-connected. Competency of "innovation" was the most connected node in the network. The present study offers a new perspective on the competency modeling using network analysis for the first time. It carefully explores several links among competencies of the whole network. These results might provide potential targets for the improvement of competency modeling in assessment and development centers.

Stability of “g” loadings in EFA: A safeguard against interpretive hubris

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Ryan McGill [1]***, Dr. Gary Canivez*** [2]

1. *William and Mary School of Education,* ***2.*** *Eastern Illinois University*

Recently, Decker et. al. (2021) argued that bifactor modeling and related variance partitioning procedures, long a staple of structural validity research, are biased in favor of a general factor at the expense of group-specific factors and should thus be eschewed in clinical science. However, this contention was based largely on an inappropriate interpretation of *rotated* EFA loadings (i.e., Varimax) as representing what was regarded as a general factor. Several well-known methods for estimating “g” loadings in EFA (c.f., Jensen, 1998) are available; yet, the method employed by Decker et al. is not one of them. To correct the scientific record, the present study was employed to *correctly* [emphasis added] estimate “g” loadings for several commercial ability measures and illustrate that the interpretations previously rendered by Decker et al. regarding the bifactor model and related procedures are without merit. Across EFA analyses, the first unrotated factor loadings (i.e., correct “g” loadings) remained stable regardless of the number of factors extracted and accounted for substantial portions of explained variance in all measures examined. However, once the rotation method used by Decker et al. was applied, the variance associated with that first factor degraded consistent with the underlying mathematics associated with EFA rotation. Thus, it can be concluded that the claims of “bias” associated with the bifactor model is a result of erroneously interpreting rotated factor loadings, providing illusory meaning.

Evaluating the replicability of network models using oral health data

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Gustavo Hermes Soares [1], ***Dr. Pedro Henrique Ribeiro Santiago*** [1], ***Dr. Fabio Luiz Mialhe*** [2], ***Prof. Lisa Jamieson*** [1]

1. *The University of Adelaide,* ***2.*** *Universidade de Campinas*

Network psychometric models comprise a powerful analytical framework that conceptualises constructs not as a common cause of symptoms/behaviours but rather as a set of mutually reinforcing symptoms/behaviours (Golino & Epskamp, 2017). However, there has been an increased debate in the literature about the replicability of network models across samples. This study aims to use a confirmatory network approach to evaluate the replicability of the Health Literacy Dental Scale (HeLD-14) across two distinct samples. Analyses included data from two oral health surveys conducted in Brazil with 603 and 535 individuals, respectively. Regularized partial correlation networks were estimated with the Graphical LASSO. Model replicability was examined by retrieving the adjacency matrix of the network model estimated in the first sample, which provides information on which edges are present or absent, and fitting the network model to the second sample. Properties compared were: (1) dimensionality and structural stability examined via Exploratory Graph Analysis; (2) global strength; (3) edge weights; and (4) centrality estimates. Model fit was evaluated based on the $\chi 2$ test, Comparative Fit Index (CFI), and Root Mean Squared Error of Approximation (RMSEA). Network models replicated the exact same number and configuration of node communities and model fit was satisfactory ($\chi 2$ (70) = 241.01, $p < 0.001$; CFI = 0.97; RMSEA = 0.064). Strong correlations were observed between edge weights (τ: 0.68; 95% CI: 0.62-0.74) and node strength estimates (τ: 0.63; 95% CI: 0.36-0.89) across samples. Findings indicate that network models of an oral health instrument replicated well across two distinct samples.

Detection of reverse coding effects using a confirmatory factor analysis

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Yelin Gwak [1], ***Dr. Youn-Jeng Choi*** [1]

1. *EWHA WOMANS UNIVERSITY*

The purpose of this study is to examine the effect of insufficient effort responding (IER) on test validity and reliability. Researchers define IER as careless responses to surveys (Ward & Pond, 2015), intentionally random responses (Meade & Craig, 2012), leaving many answers blank, and misinterpreting or misreading items (Johnson, 2005). The data in the study included student responses to 27 survey items asking about interest, confidence, and value recognition in learning science in the eighth grade of TIMSS 2019. The researchers reverse-coded seven of the 27 survey items (26%). Researchers widely use reverse-coding for filtering out students' careless or insufficient effort responses, so we will use it to set IER removal criteria. We will use long string (Meade & Craig, 2012), Mahalanobis distance space (Mahalanobis, 1936), and inter-item standard deviation (Dunn et al., 2018) methods to detect IER. Depending on each method, we reveal the method of IER group identification and how it affects the survey results. We will use the Careless R package to detect IER using the above methods (Yentes & Wilhelm, 2018). We will analyze the effect of IER on factor loading and model fit through the confirmatory factor analysis. We will also examine the test reliability before and after IER removal. When we remove the IER data, we expect to provide more reliable and valid results. This study will contribute to the quality control of education-related data, accurate interpretation, and educational utilization of test results.

Network structure of fear of COVID-19 in Iranian sample

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Shirin Rezvanifar [1], ***Mr. Hassan Mahmoudian*** [1]

1. *Allameh Tabataba'i University*

With the outbreak of the coronavirus disease, national polls indicate sharp increases in fear and worries relating to the virus. As fear may be a central construct in explaining these negative individual and societal consequences of the coronavirus pandemic, it is important to better understand what people are exactly afraid of. The aim of this study was to investigate the structure of the network of fear symptoms in COVID-19. The network of responses of 250 Iranian sample to the fear of COVID-19 scale (Ahorsu et al, 2020) was estimated. This scale includes 7 emotional and somatic symptoms of fear. An EBICglasso network was estimated. This is a regularized partial correlation network suited for ordinal data, given its use of polychoric correlations as input. Results showed that the items of scale have a positive correlation with each other, with the exception of the relation between item 1 (I am most afraid of Corona) with 6 (I cannot sleep because I'm worrying about getting Corona). The severity of the correlation between item 3 (My hands become clammy when I think about Corona) and 6 with a thicker edge is very obvious. Given the strength of the edges, item 3 has the highest degree and the lowest degree, is related to item 5. The use of network analysis demonstrates the interactive importance between symptoms in creating fear of COVID-19 and provides substantive information to researchers and therapists.

Examining structural relationships in multigroup models with small samples

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Emma Somer [1], ***Dr. Milica Miocevic*** [1], ***Dr. Carl Falk*** [1]

1. *McGill University*

SEM is often the method of choice for examining relationships among latent variables. However, small sample sizes and model misspecification can result in convergence issues and bias in the path coefficients. To overcome these limitations, researchers employ Factor Score Regression (FSR), which involves (1) estimating scores on the latent variable using factor analysis and (2) performing linear regression using the resulting factor scores. Simulation studies have found that Croon's (2002) bias corrected FSR method is a viable alternative to maximum likelihood estimation (Devlieger & Rosseel, 2017). While FSR has been extended to the multilevel framework (Devlieger & Rosseel, 2020; Kelcey, Cox, & Dong, 2021), it has yet to be examined in small samples for multigroup models. The aim of the present research is to evaluate the performance of FSR, namely Croon's correction and Skrondal and Laake's (2001) bias avoiding method, for multigroup models and compare the methods to SEM. We conduct several simulation studies to evaluate how the sample size, number of invariant items, and strength of the factor loadings and path coefficient affect conclusions about the relationship between latent variables in multiple group models. Results of our pilot simulation indicate that all methods yield less than 10% relative bias at a sample size of 25 per group and across all percentages of invariant items (0%, 25%, 100%). However, SEM resulted in slightly larger bias and less efficient estimates than the other methods. These results have implications for data synthesis of multiple studies when measurement invariance does not hold.

Assessing the dimensionality of O*NET cognitive ability ratings across job zones

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Prof. Stephen Sireci [1], ***Mr. Brendan Longe*** [1], ***Dr. Javier Suárez*** [1], ***Dr. Maria Elena Oliveri*** [2]

1. *University of Massachusetts Amherst,* ***2.*** *University of Nebraska Lincoln*

In this study we analyzed the mean importance ratings of the cognitive abilities included in the O*NET data base across O*NET job zones to identify the dimensionality of the cognitive ability importance data and to evaluate the consistency of the dimensionality across job zones. Euclidean distances were derived across mean cognitive ability importance ratings and were analyzed using the INDSCAL weighted multidimensional scaling (MDS) model. Using the criteria of fit and interpretability, a three-dimensional MDS solution was selected as the best representation of the data. These dimensions reflected a Processing/Expressing dimension, a Quantitative/Verbal Reasoning dimension, and a Perceptual Speed dimension. Interestingly, the dimensionality was not consistent across job zones. Job zones associated with lower education and training requirements were sufficiently represented by the Quantitative/Verbal Reasoning dimension, and the Processing/Expressing dimension was most relevant to job zones requiring more education and experience. The Perceptual Speed dimension was relevant to only one of the five job zones. The implications of the results for developing assessments for adult learners and employers are discussed, as is the utility of using MDS for understanding the dimensionality of O*NET data.

Computing posterior predictive p-values in Bayesian SEM

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Ellen Fitzsimmons [1], ***Dr. Edgar Merkle*** [1]

1. *University of Missouri - Columbia*

The posterior predictive p-value (ppp-value) is a popular measure of fit for Bayesian SEM and can be calculated in various manners. One way to calculate the ppp-value is to compare an observed likelihood ratio test (LRT) statistic to the posterior distribution of LRT statistics under a fitted model. The LRT statistic involves a marginal likelihood, which integrates out the latent variable(s) within the model. However, it is also possible to calculate an LRT using a conditional likelihood, which conditions on the latent variable(s). In this project, we explore the various ways that ppp-values can be computed, along with the uses of these metrics. One use for ppp-values is to judge absolute model fit via conditional or marginal LRT statistics. Another use for ppp-values is to judge person fit by detecting outliers and influential cases.

Application of the network psychometric framework to measurement burst designs

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Michela Zambelli [1], ***Prof. Semira Tagliabue*** [1], ***Dr. Giulio Costantini*** [2]

1. *Università Cattolica del Sacro Cuore,* ***2.*** *University of Milano-Bicocca*

Network Psychometrics emerged in the last years as an approach that allows investigating how different elements of a system interact and how these interactions change across occasions (e.g., Epskamp, 2020). We applied Network Psychometrics to understand how dynamics of the meaning-making processes in emerging adults changed during different phases of the COVID-19 pandemic. In particular, we involved a sample of Italian participants (N=318; M_{age}=25.4; 30% males), using a measurement burst design (two 14-day bursts). On each of 28 days, participants filled in a six-item self-report measure of situational meaning in life (SMILE; Zambelli & Tagliabue, 2022). The first wave was collected during the first COVID-19 lock-down in Italy, representing a traumatic collective experience. The second wave was ten months later, after the initial health emergency, in a period that was free from restrictions. This work aimed to understand whether the traumatic vs. non-traumatic contextual condition changed the meaning-making process dynamics, by directly comparing the two waves, in terms of both within-subject dynamics (temporal effects over time and contemporaneous effects) and between-subject dynamics (dynamics involving stable individual differences). We discuss how network psychometrics can provide a new perspective on the patterns of stability and change in the meaning-making process.

Sequential analyses for randomized response techniques

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Fabiola Reiber [1]***, Dr. Martin Schnuerch*** [1]***, Prof. Rolf Ulrich*** [2]

1. *University of Mannheim,* ***2.*** *Universität Tübingen*

Randomized response techniques (RRTs) are applied to assess prevalences of sensitive characteristics, such as socially undesirable attitudes or illicit behavior. The key mechanism of RRTs is that they protect the privacy of survey respondents because the responses are masked by a randomization element in the questioning design. Unfortunately, this randomization also increases sampling variance and big samples are required for compensation. To counteract these high sample size requirements, we propose two sequential sampling procedures: First, curtailed sampling is a simple sequential design based on a binomial test, which can be applied in RRT studies to test hypotheses on the size of the prevalence of sensitive characteristics. Second, the sequential maximum likelihood ratio test is an extension of the sequential probability ratio test, which can be applied to test group differences and hypotheses within multi-parameter RRTs. Both these techniques can substantially decrease the sample size of RRT surveys and thereby facilitate their application to study sensitive characteristics.

A comparison of different measures of the proportion of explained variance in multiply imputed datasets

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Joost Van Ginkel [1], ***Dr. Julian Karch*** [1]

1. *Leiden Universty*

Earlier research on nineteen different estimators for the proportion of explained variance in regression analysis showed that the exact Olkin-Pratt estimator produced both unbiased estimates of the proportion of explained variance and the most accurate ones. In the current study, the same nineteen estimators were studied, but now in incomplete data, where the missing data were treated using multiple imputation. In earlier research the estimator R^2_{PS} was shown to be the preferred pooled estimator for regular R^2 in multiply imputed data. For each of the nineteen estimators in the current study two pooled estimators in multiply imputed data were proposed, namely one where the estimator was the average value across imputed datasets, and one where R^2_{PS} was used as the input for the calculation of the specific estimator. Simulation results showed that estimators based on R^2_{PS} performed best regarding bias and accuracy. However, none of the estimators were unbiased, including the exact Olkin-Pratt estimator based on R^2_{PS}.

Differential item functioning in forced-choice response models

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Jacob Plantz [1], ***Dr. Jessica Flake*** [1], ***Dr. Keith Wright*** [2]

1. *McGill University,* ***2.*** *Enrollment Management Association*

The forced-choice response style has the potential to reduce response bias from test-takers when desirable responding is a concern (Brown & Olivares, 2011). The use of high stakes forced-choice assessments is increasing in educational and organizational settings. The Thurstonian-IRT model (T-IRT) (Brown & Olivares, 2011) was proposed as a means of analyzing the ipsative data forced-choice assessment yield, data in which responses are dependent on one another. Only recently has this been extended to evaluate differential item functioning (DIF) (Lee et al, 2021). Few empirical examples of testing DIF for forced-choice models currently exist in the literature, despite that the assessments are used on large and diverse samples and in high stakes settings. We address this by examining DIF of a widely used non-cognitive forced-choice assessment and demonstrate how this can be tested in a T-IRT framework. Steps for conducting DIF testing are reviewed and demonstrated using real testing data, and the implications for measurement practice are discussed.

Detection of cross-loadings in CFA, ESEM and BSEM

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Minying Mo [1], ***Prof. Junhao Pan*** [1]

1. *Sun Yat-sen University*

In traditional confirmatory factor analysis (CFA), cross-loadings on non-target factors are strictly constrained to zero. Under this assumption, CFA with frequentist estimation and Bayesian estimation encountered the problems of poor model fitting or biased estimates when true cross-loadings exist. Exploratory structural equation modeling (ESEM) and Bayesian structural equation modeling (BSEM) are proposed to relax the assumption of zero cross-loadings. In the present research, the performance of traditional CFA using maximum likelihood (ML-CFA) and Bayesian estimation (BCFA), ESEM with Geomin and Target rotation, BSEM with different prior settings for cross-loadings were compared in two simulation studies (96 conditions: 2 model sizes × 3 sample sizes × 2 major-loadings × 4 cross-loadings × 2 factor correlations). Study 1 demonstrated that ML-CFA and BCFA had similar and acceptable performance in general. BCFA was less sensitive in detecting model misspecification and both methods had serious problem in factor-correlation estimations when true cross-loadings were ignored. Study 2 indicated that BSEM with M*plus* default prior settings had severe problems in model fit and parameter estimates. All other four methods generally had acceptable performance. Both ESEMs had higher power to detect cross-loadings under different conditions than BSEMs. But ESEM using target rotation had higher type I error rate in large cross-loadings conditions. Other three methods' type I error rates under different conditions were all acceptable. In summary, we suggest that 1) researchers should use ESEM when the accurate prior information is difficult to obtain and, 2) default prior setting for cross-loadings should be avoided in BSEM modeling.

Designing computer-based assessment: a comparison of linear and adaptive testing

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Luca Bungaro [1], ***Dr. Marta Desimoni*** [2], ***Prof. Mariagiulia Matteucci*** [1], ***Prof. Stefania Mignani*** [1]

1. *University of Bologna,* ***2.*** *INVALSI*

The National Institute for the Evaluation of the Education and Training System (INVALSI) every year administers standardized tests via computer-based testing (CBT) to students attending grades 8, 10, and 13 in Italy. For each subject and grade, multiple parallel test forms are created from a Rasch item bank through automated test assembly (ATA) methods. To date, linear fixed-forms are used. In this study, we investigate the potential of computerized adaptive testing (CAT) over the linear one through a simulation study for grade 10. In particular, we simulated the responses of n = 100, 500, 1000 examinees with different ability levels to each INVALSI-2018 fixed-form and CAT. Both the standard error (SE) and the fixed-length stopping rules are implemented for CAT. Test form assembly constraints are also taken into account. The results are summarized in terms of bias and mean square error (MSE) over 100 replications and in terms of item use. The overall results show that CAT improves the precision of the ability estimates, especially in the tails of the distribution. Also, by using the CAT, it is possible to take full advantage of the CBT administration.

Modeling missing data in factor-analytic investigation of tetrachoric correlations

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Karl Schweizer [1]

1. *Goethe University Frankfurt*

Modeling missing data in factor-analytic investigation means designing an additional latent variable as part of the measurement model in such a way that it accounts for systematic variation associated with what is missing. This approach assumes that missing data bias systematic variation otherwise characterizing the covariance pattern of complete data. The factor loadings on the missing-data latent variable are expected to account for the biased systematic variation. This approach originally proposed for investigating probability-based covariances is transferred to tetrachoric correlations.

The study for investigating the properties of missing-data measurement models with main and missing-data latent variables included applications to structured random data with and without missing data. Incomplete data were generated by eliminating subsets of entries from the generated data matrices. The complete data were investigated by a one-factor model and the incomplete data by one-factor and two-factor models. Furthermore, models with free and fixed factor loadings were employed.

The results showed impairment in model fit because of missing data when investigated by one-factor models. Investigating incomplete data by missing-data measurement models yielded largely recovered fit results. The better degree of correspondence of results for complete and incomplete data was observed when models included constrained factor loadings instead of free factor loadings.

Latent structure model with multilevel groups

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Theren Williams [1]***, Dr. Steven Culpepper*** [1]

1. *University of Illinois, Urbana-Champaign*

Across disciplines in social and educational fields, there is continued growth to further understand the underlying skill and attribute profiles of various classes of respondents. As the implementation of Cognitive Diagnostic Models (CDMs) proliferates to improve the comprehension of such profiles, so does the interest in further advancing their capacity to uncover relationships in their latent structure. Recent advances have aimed at increasingly robust capture of dependence structures present in higher-order records of multivariate polytomous behavioral data. Frequently, these data are recorded in experimental settings where rich group-related influences are present, such as across classrooms, treatment centers, etc. This poster builds upon the current developments in CDMs, proposing new methods for capturing multilevel data configurations within the higher-order elements of the latent structure. We offer a Bayesian framework for inferring the latent structure and a higher-order factor model for attribute and group relations, together with an extension including group effect within the attribute level. Further, we share evidence supporting parameter estimation claims through Monte Carlo (MCMC) simulation results.

Factors affecting item calibration using adaptively administered test data

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Hwanggyu Lim [1], ***Dr. Kyung (Chris) Han*** [1]

1. *Graduate Management Admission Council*

To maintain the integrity of test programs based on computerized adaptive testing (CAT) design, there is a need to periodically recalibrate the operational items in these programs. Unlike the process for calibrating pretest items using randomly administered response data, however, operational items in a CAT-designed program are adaptively administered. Hence, the response data for such operational items usually include examinees from similar performance levels, which could cause biased item parameter estimates (Glas, 2010) due to the violation of the "missing at random" (MAR; Rubin, 1976) assumption.

To avoid such a situation, test practitioners might typically redo the entire pretesting process for the operational items. Such practice takes much time and effort, however, and might increase the test security risks as it would further expose the items. If we can identify the main factors that affect item calibration when the MAR assumption is violated, and if we can control those factors properly, it might be possible to recalibrate the operational items more effectively using the adaptively administered test data.

We conducted a series of simulation studies to evaluate the effect of several factors (e.g., a variance of the proficiency distribution, item difficulty parameter) under different conditions with various CAT item selection methods. For item calibration, we compared the fixed ability parameter calibration (Stocking, 1988) and the fixed item parameter calibration (Kim, 2006). The preliminary results suggest that under certain conditions, item parameters can be accurately estimated in an unbiased way using adaptively administered response data.

Pathway parameter sensitivity across sets of DAGs

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Ronald Flores [1], ***Dr. Edgar Merkle*** [1]

1. *University of Missouri - Columbia*

Directed acyclic graphs (DAGs) are an increasingly common method for researchers to encode their causal assumptions before conducting data analyses. In practice, however, uncertainty about causal relationships motivates the specification of multiple plausible DAGs. In the current study, we develop a test statistic that assesses sensitivity of results across sets of DAGs. The metric specifically assesses whether targeted causal pathways differ across competing DAGs, where the DAGs are operationalized and estimated via SEM. This serves as a sensitivity analysis, where researchers can judge whether the presence of a confounder, collider, or other relationship influences focal paths. For example, if a causal pathway between two key variables does not significantly differ across models, then concerns about the presence of an unwanted source of influence (e.g., confounder, collider, etc.) could be ignored. We will first discuss theoretical details underlying test development. We will then illustrate test performance using both simulated and real data, and we will conclude with future directions.

Modeling the fluctuation of inattention in responding to questionnaires

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Yuki Shimizu [1]

1. *Nagoya University*

Self-reported data reflects not only the trait we want to measure, but also the respondent's habits, called response bias. This contamination is undesirable for item parameter estimation. To eliminate the effect of bias, recently, a measure method which utilizes vignettes has been developed. A vignette contains a short story about a virtual person. Reading the vignette, respondents evaluate the person and themselves. Differences between these data are analyzed and the response bias is detected separately from the trait. One of response bias is inattention. In previous studies, person parameters which reflect inattention were considered, and it was found that accuracy of item parameter estimation was improved. The person parameters were assumed constant. However, the values can be varied when the degree of concentration is changed. Therefore, this study models the fluctuation of inattention by considering linear and nonlinear relationship between inattention and item location. Results from this model are compared with those from models in the previous studies. Data are obtained from simulations and responses to a measure of personality pathology (the Personality Inventory for DSM-5 [PID-5]). In both data sets, we will see that models outperform the fixed model in terms of model fit, accuracy, and parameter recovery.

Reaction time multinomial process trees: comparing parametric and non-parametric procedures.

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Anahí Gutkin [1]***, Prof. Manuel Suero*** [1]***, Prof. Juan Botella*** [1]

1. *Universidad Autónoma de Madrid*

Reaction Time Multinomial Process Tree models (RT-MPT) are recent and lack specific modeling protocol for different data distribution scenarios. Actual procedures differ in the way that RT data is modeled: some include RT histograms and others assume RT distributions, designated here as non-parametric and parametric procedures, respectively. The aim of this study was to investigate, using a Two High Threshold Model (2HTM) simulation, which procedure should be selected considering the effects of extreme process probabilities, the number of trials and the manipulation of latent RT distributions. Preliminary results indicate that for choosing a parametric procedure experimentalist not only must have an approximate idea about the shape of the RT distribution (*e.g.:* RT usually are right skewed with a fat tail) but also need to make distributional assumptions that are compatible with the RT distributions changes between branches (*e.g.:* RT distributions from the 2HTM guessing states have fatter tails than those from detection states). Because the number of observations affected parametric and non-parametric procedures differently, we also provide the number of trials needed for the best procedure in each experimental situation.

Keywords: RT-MPT procedures, 2HTM, RT latent distributions, process with extreme probabilities.

On the relationship between coefficient alpha and closeness between factors and principal components for the multi-factor model

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Kentaro Hayashi [1], ***Dr. Ke-Hai Yuan*** [2]

1. *University of Hawaii at Manoa,* ***2.*** *University of Notre Dame*

Cronbach's alpha remains very important as a measure of internal consistency in the social sciences. The Spearman-Brown formula indicates that as the number of items goes to infinity, the coefficient alpha eventually approaches one. Hayashi, Yuan, and Sato (2021) showed that under the assumption of a one-factor model, the phenomenon of the coefficient alpha approaching one as the number of items increases is closely related to the closeness between factor analysis (FA) loadings and principal component analysis (PCA) loadings, and also the factor score and the PC agreeing with each other. In this work, we extend their partial results to the case with a multi-factor model, with some extra assumptions. These phenomena have an implication to offer another way to characterize the relationship between FA and PCA with respect to the coefficient alpha under more general conditions.

The Concise Health Risk Tracking - Self-Report (CHRT-SR) - A measure of suicidal risk: performance in adolescent outpatients

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Karabi Nandy [1], ***Prof. Augustus Rush*** [2], ***Prof. Thomas Carmody*** [1], ***Ms. Alexandra Kulikova*** [3], ***Mrs. Taryn Mayes*** [1], ***Dr. Graham Emslie*** [1], ***Prof. Madhukar Trivedi*** [1]

1. *University of Texas Southwestern Medical Center,* ***2.*** *Duke-National University of Singapore, and Duke University,* ***3.*** *University of North Texas*

Objectives: The Concise Health Risk Tracking Self Report (CHRT-SR) assesses the risk of suicidal behavior. Its psychometric properties in adolescents seen in primary care practices have not been assessed.
Methods: A sample (n=657) of adolescents (< 18) in primary or psychiatric care completed the 14-item version of the CHRT-SR at both baseline and within three months thereafter. To identify an optimal brief solution for the scale, we evaluated the factor structure of CHRT-SR using multigroup confirmatory factor analysis, which included testing measurement invariance across age and gender categories. We further evaluated item-level parameters based on a graded response model and the overall scale performance using classical test theory analyses. Finally, we assessed the concurrent validity (both cross-sectional and as a change measure over time) of this optimal factor structure by comparing it to responses of an well-known suicide item.
Results: Confirmatory factor analysis and measurement invariance analyses identified the 9-item CHRT-SR ($CHRT\text{-}SR_9$) as the optimal solution. Spearman-Brown coefficient was 0.80 at first visit and 0.86 at second visit. Classical test theory revealed corrected item-total correlations between 0.58 and 0.80. Item response theory analyses revealed excellent item performance and a unifactorial instrument. Cross-sectional and change over time concurrent validity analyses that compared the $CHRT\text{-}SR_9$ with responses to the suicide item of the PHQ-9 revealed that it can measure both the improvement and worsening of suicidality over time.
Conclusion: The $CHRT\text{-}SR_9$ is a brief self-report with excellent psychometric properties for adolescents that is sensitive to changes in suicidality over time.

Discovering trends in high school credit system using topic modeling

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mrs. Eunjeong Jeon [1], ***Dr. Youn-Jeng Choi*** [1]

1. *EWHA WOMANS UNIVERSITY*

The Republic of Korea will implement the high school credit system in 2025. The current standard for high school graduation is the number of units based on the number of attendance days, and the standard for failure is the grade paid, not the subject. Therefore, high schools in the Republic of Korea control subject options. In addition, Korea has standardized high school education for the college entrance examination, and there exists excessive competition without diversity. Thus, the government proposes the high school credit system to solve this problem. This study introduces a high school credit system using topic modeling to explore worldwide trends in the high school system. Therefore, we plan to investigate about 590 highly scholarly articles related to the high school credit system from the Google Scholar website. First, we intend to collect academic journal abstracts using the EndNote program for research. Then, we will discard unnecessary text data through the R program and build a dictionary of terms necessary to extract keywords, create keyword extraction, perform keyword network analysis, word clouding, and heatmaps. After that, we intend to conduct a topic modeling, Latent Dirichlet Allocation Model (LDA) analysis using the R program. We expect to read the clear trends by analyzing previous high school credit system studies and comparing research trends in Korea and international research.

Likelihood ratio test and relative fit indices to evaluate the model fit in typical clinical situations – a simulation study

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Alexander Avian [1], ***Mr. Marko Stijic*** [1]

1. *Medical University of Graz*

Introduction

Medical decision for initializing therapies or adapting already ongoing therapies are often based on patient reported outcomes (PRO). Testing this PRO faces two major challenges: (1) For practical reasons these PRO are often response to an one item questionnaires or calculated out of a small number of items and (2) often only a small number of patients respond to these questionnaire (e.g. questionnaire for rare diseases).

Objectives

The aim of this simulation study is to evaluate the ability of likelihood ratio tests and relative fit indices to identify the appropriate underlying model when analyzing only a few number of items and/or a small number of responders.

Design

Response pattern according to different IRT-models (partial credit model, graded response model, rating scale model) were simulated (1.000 replications). Therefore, responses of 250 to 10.000 respondents (small to very large sample sizes in PRO studies) to 3 to 50 items were generated. Different strategies (likelihood ratio tests and relative fit indices) were used to analyze response.

Results

Depending on the analyzed scenario, likelihood ratio tests and relative fit indices performed different. In general with increasing sample size and increasing number of items the results improved. In typical clinical situations (small sample sizes and only a few items) the analyzed strategies did not perform satisfactorily.

Conclusion

While likelihood ratio tests and relative fit indices are suitable for longer questionnaires and bigger sample sizes, the model fit in sample sizes typical for clinical applications cannot be analyzed in a sufficient way.

Pauci sed moni: An item response theory approach for shortening tests

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Ottavia M. Epifania [1], ***Dr. Pasquale Anselmi*** [1], ***Prof. Egidio Robusto*** [1]

1. *University of Padova*

Item Response Theory (IRT) is the theoretical framework often used for shortening existing tests. This contribution presents new IRT-based item selection procedures for developing short test forms. These procedures are based on the information that each item provides in respect to different trait levels of interest (denoted as targets), which are obtained by segmenting the latent trait continuum in either equal or unequal intervals. In a simulation study, the performances of the new procedures were compared with those of the typical IRT procedure and of a random selection of the items. Different latent trait distributions were considered as well (normal, positively skewed, uniform). The new procedures outperformed the existing ones in recovering central and peripheral regions of the latent trait continuum, particularly when the short test forms consisted of fewer items. Additionally, the tests obtained with the new procedures tended to be more informative than those obtained with the typical procedure.

Scaling properties of pain intensity ratings in adult populations using the Numeric Rating Scale

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Marko Stijic [1], ***Mr. Winfried Meissner*** [2], ***Mr. Alexander Avian*** [1]

1. *Medical University of Graz,* ***2.*** *Jena University Hospital*

Background: Despite having acceptable psychometric features, the interval level of measurement and justification to use parametric statistical methods on the Numeric Rating Scale (NRS) need further examination.
Objective: To evaluate scale properties of the NRS using an item response theory (IRT) approach.
Design: Retrospective analysis of data from an international postoperative pain registry (QUIPS).
Participants: 346.892 adult patients (age range 18-90), 55.7 % are female and 38% had preoperative pain.
Methods: To analyze the scale properties of the NRS three pain items (movement pain, worst pain, minimum pain) were analyzed using three different IRT-models (rating scale model (RSM), graded response model (GRM), partial credit model (PCM)). Subgroup analyses were done for sex and age groups.
Results: After collapsing the highest and the second highest response category, the GRM outperformed the other models (lowest BIC) in all subgroups. Overlapping categories were seen in category boundary curves for worst and minimum pain. This was particularly noticeable for higher pain ratings. Response category widths were wider for categories associated with low pain intensity and smaller for categories associated with high pain intensities. For sex and age groups, similar results were obtained.
Conclusion: According to these results, the response categories after collapsing are ordered but have different widths. Therefore, the interval scale properties of the NRS should be questioned and parametric analysis for single items analysis has to be avoided. In dealing with missing linearity in pain intensity ratings using the NRS, IRT methods may be helpful.

Robustness study of normality-based likelihood ratio tests for testing maximal interaction two-mode clustering and a permutation based alternative

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Zaheer Ahmed [1], ***Dr. Alberto Cassese*** [1], ***Prof. Gerard van Breukelen*** [1], ***Dr. Jan Schepers*** [1]

1. *Maastricht University*

Extended maximal interaction two-mode clustering abbreviated as E-ReMI provides likelihood ratio tests for testing the null hypothesis of no interaction in two-mode data. These likelihood ratio tests are derived under the assumption of Normality and their empirical distributions under the null hypothesis are also based on data sampling from a normal distribution. Unfortunately, in real-life applications, the assumption of Normality is rarely met. A violation of the Normality assumption may have serious implications for the test performance in terms of Type-I error rate and power. This paper is concerned with this problem and has two main goals. Firstly, investigating whether these tests are robust in terms of Type-I error rate against a violation of Normality. Secondly, introducing a new test based on permutations as an alternative, and investigating its robustness. We present simulation studies to assess the performance of the newly proposed test in terms of Type-I error rate and power. Lastly, we show an application to data from a person by situation study.

Revisiting parametrizations for the nominal response model

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Jan Netík [1], ***Dr. Patricia Martinkova*** [1]

1. *Czech Academy of Sciences*

In this work, we revisit the existing parametrizations of the Nominal Response Model (NRM). We consider two parametrizations of NRM directly linked to the baseline-category logit (BL) nature of the model and using the information on correct answer in multiple-choice items: the BL intercept-slope (BL-IS) parametrization, and BL parametrization using a discrimination and difficulty parameters as is usual in the item response theory models (BL-IRT). An advantage of the BL-IRT parametrization is the graphical interpretation of the model parameters while BL-IS is more often encountered in the GLM framework. We hypothesize that the proposed parametrization accounting for the information on correct answer, and the related setup of starting values, may lead to numerically more precise results and less convergence issues. The relationship between the Bock's, Thissen et al.'s, and two newly considered parametrizations is explained and illustrated on practical examples.

Construct meaning in 3-level clustered data

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Mr. Andrea Bazzoli [1]***, Dr. Brian French*** [2]

1. *Washington State University Vancouver,* ***2.*** *Washington State University*

Validity arguments are constructed from several pieces of evidence to inform the extent to which inferences can be supported. A scoring inference (e.g., Kane, 2013) for constructs should use both theoretical arguments and empirical data. A major component of this evidence is understanding an assessment's underlying structure, which is often investigated via confirmatory factor analysis (CFA) at the individual level with item meaning at that level. Often, these individual level scores are aggregated to a different level (organizations) for score interpretation. However, empirical evidence to support such an inference ignores the hierarchical nature of the data which can be reflected in item content. This can be problematic for construct meaning and interpretation at these higher levels. Using the 2021 SOPS Hospital Survey dataset, we explored different CFA models of 3-level data for hospital patient safety climate (Colla et al., 2005) that would be related to different interpretations between constructs at these different levels, as item wording supports such models. We estimated internal consistency reliability via Omega for the constructs. In line with theoretical arguments and the items' wording, a shared construct model fit the data best, implying that those items administered to personnel clustered in units and in hospitals can be used to measure shared patient safety climate. That is, safety climate was specified only at the hospital level, whereas items covariances were specified at the unit and individual level. We highlight different model constraints and assumptions, and how these are related to score interpretation with 3-level data structures.

Investigating equal construct and equity requirements on score transformation precision in true-score equating for test forms with different targeting: A simulation study

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Carolina Fellinghauer [1], ***Dr. Rudolf Debelak*** [1], ***Prof. Carolin Strobl*** [1]

1. *University of Zurich*

The simulation study investigated to what extent departures from construct similarity, but also differences in difficulty and targeting of test forms, impact the score transformation when test forms are equated. The study simulated data for two test forms with ten items, respectively having three response options and a sample size of N = 500. The factor correlation between test forms was used to operationalize construct similarity. Lack of equity of test forms was operationalized through increasing departure from equal difficulty. Targeting was varied through the dispersion of the item and person parameters in each test form. Test forms were equated through concurrent calibration in a common person design using the Partial Credit Model. Anchoring test forms to the difficulty parameters of the common metric provided a transformation rule to link the raw scores of the test forms. Analysis of the RMSE between transformed and truly observed scores showed that departures from construct similarity had an important impact on the score transformation precision. Lack of similarity,between test forms goes along lower transformation precision. With decreasing similarity, score transformation precision benefits from good targeting. This means that when the item difficulties match the person parameters' measurement spectrum, the precision is higher even in scenarios with low similarity. Finally, difficulty shifts up to two logits did not impact the precision of score transformations, indicating the advantage of applying the true-score equating methods over the naive approach of identity equating, that was used as a baseline. The practical implications of these results are discussed.

Natural language processing classifiers application on binary coded twitter messages

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Ting Wang [1]

1. *American Board of Family Medicine*

This presentation demonstrated how to use natural language processing classifiers to differentiate Twitter messages that could be associated with disaster (coded as 1) or not (coded as 0). The investigated algorithms are Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR). The theoretical background of these classifiers was introduced, followed by their empirical performance under different setups. The results showed that the Naive Bayes algorithm outperformed in all scenarios investigated. The possible reason for this phenomenon is discussed in the end.

Exploring the Structure of Speed in Cognitive Diagnostic Models

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Yingshi Huang [1], ***Ms. Tongxin Zhang*** [1], ***Prof. Ping Chen*** [1]

1. *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University*

By incorporating collateral information such as response time (RT), cognitive diagnostic models have the potential to provide a more fine-grained picture of the examinee's latent skill profile. But this naturally creates two questions: (1) whether the mastery of a skill is always coupled with a faster speed compared to the non-mastery state? (2) What patterns of speed are shown in different skill profiles? To answer question one, we analyzed the PISA Math 2012 computer-based dataset via the most widely used deterministic input, noisy 'and' gate model and lognormal model to obtain examinees' pure latent skill profile and speed, respectively. Results revealed that: (1) the examinees who mastered the attribute tended to hold a slower speed than those who did not; (2) for each attribute, the mastered and non-mastered examinees displayed exactly opposite speed patterns. Specifically, the higher the speed of non-mastered examinees, the lower the speed of the examinees who master the skill, and the large differences between these two types of speed were observed in both easy skills with high mastery rates and hard skills with low mastery rates. For question two, we implemented a two-step clustering analysis to reveal natural groupings for speed and latent skills profiles. A 4-cluster solution was found, indicating that the high-ability-level examinees mastering most of the attributes had the slowest but most stable speed while the low-ability-level examinees had the fastest but most fluctuated speed. This study provides food for designing new cognitive diagnostic models incorporating RT.

On online calibration in MCAT with polytomously scored items

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Dr. Lu Yuan [1]***, Prof. Ping Chen*** [1]

1. *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University*

To maintain the vitality and sustainable use of computerized adaptive testing (CAT) item pool, the problematic operational items should be retired periodically and replaced by the new items. Online calibration is a key technology for item replenishment in CAT, which has been widely used in various forms of CAT, including unidimensional CAT (UCAT), multidimensional CAT (MCAT), CAT with polytomously scored items and cognitive diagnostic CAT. However, with multidimensional and polytomous assessment data becoming more common, there have been no published reports on online calibration in MCAT with polytomously scored items (P-MCAT). Therefore, this study proposes new online calibration methods and designs for P-MCAT based on the classical methods used in MCAT [multidimensional “one EM cycle” and “multiple EM cycles” methods (Chen, 2017)] and designs in UCAT [D-VR design (van der Linden & Ren, 2015) and D-c design (He & Chen, 2020)], respectively. Two simulation studies were conducted by manipulating two factors, calibration sample size and correlation between dimensions, to compare a total of four P-MCAT online calibration methods (two new methods and their Bayesian versions) and three designs (two new adaptive designs and random design). The results show that all new methods can accurately recover the item parameters, and the adaptive designs outperform the random design in most cases.

Children's comprehension of part-whole construct of fraction:based on the cognitive diagnosis assessment

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Chuanyue Luo [1], ***Prof. Tao Yang*** [2], ***Dr. Jianqiang Yang*** [1], ***Dr. Yuanting Yang*** [1]

1. *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University,* ***2.*** *Beijing Normal University*

The representation of part-whole relations that are embodied in fractional numbers plays an important role in students' math learning, especially in the primary stages, and continues to be a problematic area. To date, the evaluation of it is mostly based on the grade rating and thus the cognitive structure of the "part-whole" construct for fraction in students' minds is not clear. In addition, it is a relatively lack of studies on the development of children's comprehension of part-whole construct of fractions. Cognitive diagnostic assessment (CDA) , a newly generated evaluation theory, can provide specific and clear feedback information about students subskills, thereby maximizing to promote student learning. This study used a cognitive diagnostic test to investigate elementary students' comprehension about part-whole construct. The assessment was administered to 417 students in grades 4–6 from Beijing, China. The results indicated that (1) 407 students had poor mastery of some cognitive attributes ; (2) there were significant differences in children's probabilities of attribute mastery between grades, and with the growth of grades, probabilities of attribute mastery increase; (3) There was a significant difference between gender in some cognitive attributes.(4) Besides, this study explored the different learning trajectories of students to facilitate individualized remediation of students. Such diagnostic information could be utilized by students, teachers, and administrators of mathematics programs and instruction

An analysis of adaptive learning recommendation based on reinforcement learning

Wednesday, 13th July - 17:20: Poster Session (Belmeloro Building) - Poster Presentation

Ms. Tongxin Zhang [1], ***Ms. Yingshi Huang*** [1], ***Prof. Tao Xin*** [1]

1. *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University*

Adaptive learning can provide a personalized learning trace for each learner. In adaptive learning, recommendation strategy plays an essential role due to its direct influence on learning effects and learners satisfaction. As the reinforcement learning (RL) algorithm can use historical data to make self-improvement and online updating, many researchers see value in implementing recommendation strategies based on RL. However, little message has been left for the practitioners to apply RL-based recommendation strategies under different test scenarios. It is still unclear how the performance of RL-based recommendation strategies vary with different learning times, item parameters of measurement model, test lengths, and sample sizes for the RL algorithm. To address this gap, a Monte Carlo simulation was conducted to investigate the behavior of the recommendation strategy based on Q-learning. Results showed that: first, the recommendation strategy based on Q-learning could gain better rewards than the random strategy across all conditions; second, as the learning time increased, learning rewards grew up sharply in the beginning and then flatted out; third, longer test length or item parameters with lower guessing and slip parameter produced better rewards, but the improvement goes down when the rewards are in a high level. Furthermore, the impact of the sample size was insignificant probably because it had achieved a fulfilled size. Above all, for the practical application of RL in the recommendation of adaptive learning materials, it is suggested that we should give more emphasis on the measurement model to achieve better recommendation in adaptive learning.

Predictor selection for high-dimensional regression via random projection ensembles

Thursday, 14th July - 09:15: Symposium: Statistical methods for the analysis of complex data structures (Room B) - Symposium Presentation

Dr. Matteo Farnè [1], ***Prof. Laura Anderlucci*** [1], ***Prof. Giuliano Galimberti*** [1], ***Prof. Angela Montanari*** [1]

1. *University of Bologna*

Variable selection in high-dimensional settings characterizes many scientific problems. Fan and Lv (2008) introduced the concept of sure screening to reduce the dimensionality, and proposed a procedure essentially based on the magnitude of marginal correlations between the predictors and the response variable. In this work, we propose a variable selection method for multiple linear regression which is based on axis-aligned random projections and accounts for partial correlation between each predictor and the response. In more detail, we standardize the data, and we consider an ensemble of blocks of random models. Each of the models is obtained by choosing d random features out of p potential predictors (with d«p) and by estimating a linear regression via OLS. The optimal model into each block is selected as the one maximizing the linear determination index. Then, we calculate for each potential predictor an importance score based on the optimal models into each block, to account for signal strength and prediction performance simultaneously. Finally, we retain the d predictors with the maximum importance score. Theoretical guarantees are provided about the inclusion of the s true predictors (with s≤d) in the final predictor set, with respect to the strength of predictors and the degree of multi-collinearity. Performances of the proposed method are thoroughly evaluated in an extensive simulation study.

Bayesian inference for mixed models with log-transformed response

Thursday, 14th July - 09:30: Symposium: Statistical methods for the analysis of complex data structures (Room B) - Symposium Presentation

Dr. Aldo Gardini [1]

1. *University of Bologna*

The analysis of variance, and mixed models in general, are popular tools for analyzing experimental data. Bayesian inference for these models is gaining popularity as it allows to easily handle complex experimental designs and data dependence structures. When working on the log of the response variable, the use of standard priors for the variance parameters can create inferential problems and namely the non-existence of posterior moments of parameters and predictive distributions in the original scale of the data. The use of the generalized inverse Gaussian distributions with a careful choice of the hyper-parameters is proposed as a general purpose option for priors on variance parameters. Theoretical and simulation results motivate the proposal.

Detecting latent variable non-normality through the generalized Hausman test

Thursday, 14th July - 09:45: Symposium: Statistical methods for the analysis of complex data structures (Room B) - Symposium Presentation

Ms. Lucia Guastadisegni [1], ***Prof. Irini Moustaki*** [2], ***Prof. Silvia Cagnone*** [1], ***Prof. Vassilis Vasdekis*** [3]

1. *University of Bologna,* ***2.*** *London School of Economics,* ***3.*** *Athens University of Economics and Business*

One of the typical assumptions of Item Response Theory (IRT) models is the normal distribution of the latent variable. However, this assumption is not always appropriate. Assuming normality in the model when the true distribution of the latent variable has a different form, for example skewed or multi-modal, can result in large biases in parameter estimates. In the IRT literature, several models that assume different shapes for the latent variables have been proposed and information criteria have been used for model selection. However, detecting non-normality of the latent variables through a statistical test remains an open issue. In this work, we use the generalized Hausman (GH) test to detect non-normality of the latent variable distribution in unidimensional IRT models for binary data. The GH test compares two different estimators that are both consistent when the model is correctly specified and one also under model misspecification. None of the two estimators need to be fully efficient. We consider the estimator obtained from the two-parameter IRT model that assumes normality of the latent variable with the estimator obtained under a semi-parametric framework. Simulated and real data are used to study the performance of the proposed test and illustrate its use.

Fully symmetric graphical lasso for dependent data

Thursday, 14th July - 10:00: Symposium: Statistical methods for the analysis of complex data structures (Room B) - Symposium Presentation

Dr. Saverio Ranciati [1], ***Prof. Alberto Roverato*** [2], ***Prof. Alessandra Luati*** [1]

1. *University of Bologna,* ***2.*** *University of Padova*

In this work, we propose a method to analyze multivariate data with intrinsic symmetrical structures and, in general, to solve problems belonging to the class of dependent samples inference, such as case-control studies, matched and paired data. To this aim we propose the fully symmetric graphical lasso, a penalized likelihood method with a fused type penalty function that takes into explicit account the natural symmetrical structure within and between symmetrical blocks of the data (or samples). The implementation leverages an alternating directions method of multipliers algorithm to solve the corresponding convex optimization problem. The procedure is applied to different real world dataset, concerning air pollution and brain fMRI scans.

A mixture model for discriminating responses affected by response styles and content-driven preferences

Thursday, 14th July - 09:15: IRT II (Room A) - Individual Oral Presentation

Prof. Sabrina Giordano [1], ***Prof. Roberto Colombi*** [2], ***Prof. Gerhard Tutz*** [3]

1. *University of calabria,* **2.** *University of Bergamo,* **3.** *Ludwig Maximilians Munich*

Questions to investigate perceptions or opinions often collect answers affected by the response style, a tendency to prefer only a few options of the ordinal response scale, such as the extreme or the middle point (EMRS), regardless of the content of the item. It is widely recognised that the response style mechanism introduces heterogeneity in the responses, biased estimates and may impact the validity of the results.

For each item, in our approach, respondents are partitioned into two groups of individuals who have or not a tendency toward extreme or middle categories. The probability function of each item is defined as a mixture of two components, one for the content-driven responses and one for EMRS. Both probability functions of the mixture are modelled by a linear function of covariates in a logit regression model.

The advantages are manifold: for each item, the respondents can be easily classified as content-driven or EMRS respondents; the covariates that affect the response behaviour (content-driven or EMRS) and the intensity of EMRS, when present, can be easily distinguished; the content-related effects of covariates on the observed responses are not confounded with EMRS effects; the EMRS tendency is modelled by an effect that is specific for every response and not by a single effect common to all the items; the effects of EMRS are described in terms of concentration orderings among probability functions.

Our model is applied to capture response styles and content-driven behaviours in the opinion of American citizens about ethnic minorities collected by GSS.

New flexible item response models for dichotomous response with applications

Thursday, 14th July - 09:30: IRT II (Room A) - Individual Oral Presentation

Dr. Jorge Bazán [1], ***Ms. Jessica S. B. Alves*** [1]

1. *University of São Paulo*

Some asymmetric Item Characteristic Curves (ICC)s have already been introduced in the IRT literature introducing a new item parameter associated with the item complexity which explains the asymmetry in the ICC. Although the importance of proposing new models that have asymmetric ICC in IRT is already known, the relation of this with unbalanced numbers of zeros and ones in the testing data in real application has no been explored.

In this work we propose new asymmetric IRT models that have an asymmetric ICC as their main feature. Estimation will be developed using the Bayesian approach. Properties of the proposed models will be discussed and applications in educational data illustrate the benefits of the new ICC for unbalanced data when we compare our IRT models versus the classic IRT models.

Methodological Issues in the IRT Modeling of Recognition Task Data

Thursday, 14th July - 09:45: IRT II (Room A) - Individual Oral Presentation

Ms. Qi (Helen) Huang [1], ***Prof. Daniel Bolt*** [1]

1. *University of Wisconsin*

Traditional item response theory (IRT) models, like the 2PL, are often indiscriminately applied to tests with dichotomously scored items. The results can be misleading without considering the different psychological processes that often underlie such items. In this paper, we investigate the consequences of possible 2PL misspecification in the context of a recognition task (using both North American and Dutch versions of an author recognition test, Brysbaert et al., 2020) which requires identification from a list of names those that are actual authors. Like previous papers, we find a high positive correlation between the item discrimination and difficulty parameter estimates, as well as inconsistent performances across author and foil items, which raise questions about the application of a standard IRT model. These findings motivate the consideration of a one-parameter asymmetric item response model with an additional person parameter (to account for the penalty's effects) as an alternative to the 2PL. We provide a psychological justification for the model, demonstrate its empirical superiority to the 2PL in this context, and illustrate the metric consequences of model choice.

A dynamical framework for the derivation of cumulative response models

Thursday, 14th July - 10:00: IRT II (Room A) - Individual Oral Presentation

Dr. Yvonnick Noel [1]

1. *University of Rennes*

A new generic class of cumulative response models is derived from the dynamical hypothesis that item responses are determined by a set of latent response processes that are chained together in a causal sequence. A parent process is thought to increase in intensity as a function of some latent parameter, while simultaneously increasing the intensity of a child process, that may in turn generate another one, and so on. An arbitrary number of processes may be chained together in this cascading construction, which is defined as a set of differential equations. By varying a delay and a causal weight parameters between two successive processes in the sequence, and also by imposing that process intensities be bounded on [0;1], common IRT models (Rasch, 2PL, 3PL) are recovered as special cases of the generic solution curve, thus providing a new, causal interpretation of a series of items forming an ordered IRT scale. But a new family of non-symmetric response functions also emerges from this framework, with interesting properties: This class of models may help model new forms of cumulation, such as discrimination increasing with difficulty, a phenomenon that has sometimes been reported in ability tests (Lee & Bolt, 2018; Samejima, 2000).

Correcting for extreme response style with IRT: Model choice matters

Thursday, 14th July - 10:15: IRT II (Room A) - Individual Oral Presentation

Mr. Martijn Schoenmakers [1]***, Dr. Jesper Tijmstra*** [1]***, Prof. Jeroen Vermunt*** [1]***, Dr. Maria Bolsinova*** [1]

1. *Tilburg University*

Extreme response style (ERS), the tendency of participants to select extreme item categories regardless of the item content, has frequently been found to decrease the validity of Likert-type questionnaire results (Moors, 2012; Van Vaerenbergh & Thomas, 2013). For this reason, a variety of IRT models have been proposed to model ERS and correct for it. Comparisons of these models are however somewhat rare in the literature. This is especially the case in the context of cross-cultural comparisons, where ERS is even more relevant due to cultural differences between groups. To remedy this issue, the current paper examines two frequently used IRT models that can be estimated using standard software: the multidimensional nominal response model (MNRM) extension by Falk and Cai (2016) and the IRTree model with multidimensional nodes by Meiser (2019). Studying conceptual differences between these models reveals that under the MNRM, the probability of agreeing with a four-category item (scoring a 3 or a 4) is dependent on the extent of ERS. This is not the case for the IRTree model, resulting in different category probabilities between both models. To evaluate the practical impact of these differences in a multigroup context, a simulation study is conducted. Our results show that when the groups differ in their average ERS, the IRTree and MNRM can drastically differ in their conclusions about the size and presence of differences in the substantive trait between these groups. Implications for the future use of both models and the conceptualization of ERS are discussed.

Difference score methods for time-varying covariates in latent transition analysis

Thursday, 14th July - 09:15: Longitudinal Data (Room C) - Individual Oral Presentation

Dr. Paul Scott [1]

1. *University of Pittsburgh*

This study concerns incorporating difference scores into Latent Transition Analysis (LTA) to estimate prediction of transition probabilities from changes in time-varying covariates between successive time points. Methods under consideration are: Single-step approach with either Latent Difference Score (LDS1) or Observed Difference Score (ODS1) incorporated into LTA to simultaneously estimate prediction of transition probabilities with other model parameters, and a three-step approach using either an observed (ODS3) or latent difference score (LDS3) to estimate LTA before assessing prediction of transition probabilities. Distinctions among methods pertain to measurement error, model fitting, parameter estimation and interpretation. The methods are compared in an example examining if changes in BMI predict transitions between symptom clusters of obstructive sleep apnea. LCA established three classes at each time point and measurement invariance across time was imposed. Based on BIC, LDS1 had the poorest fit and ODS3 had the best fit; three-step methods had better fit than single step. LDS3 and ODS3 had similar entropy (0.74), while LDS1 was slightly lower (0.71) and ODS1 had the highest (0.82). LDS1 signaled three significant transition probabilities, while the other models indicated all six transition probabilities were significant. On average, standard errors for the path predicting transition probabilities from difference scores were much larger for LDS relative to ODS (by a factor of 497.70 and 311.45 within single vs. three-step respectively) and slightly larger for single relative to three-step (by a factor of 1.95 and 1.22 between LDS and ODS respectively). Simulation studies are needed for closer investigation.

How and why apply Mokken scaling to longitudinal data?

Thursday, 14th July - 09:30: Longitudinal Data (Room C) - Individual Oral Presentation

Prof. Claus H. Carstensen [1]

1. *University of Bamberg*

With Mokken Scaling, less restrictive measurement models than common parametrical measurement models in Item Response like the Rasch Model, the “PLM or 3PLM for example are defined. The fit of these models to empirical data can be evaluated. They allow to interpret the sum score of a test as an ordinal measure of person ability. Van der Ark and Sijstma (2005) introduced different methods to impute missing values in the item responses in order to offer Mokken Scale based data analyses based on the complete set of cases and items. Analyzing longitudinal data, linking and scaling item responses can be seen as a question of scaling under the condition of missing values by design. For example, in the National Educational Panel Study (NEPS) in Germany, mathematical competence is assessed from the same students in grades five, seven and nine with some items being presented in two of these three assessments according to an anchor item design. If these response data fit to a Mokken Scale, quite simple statistical analyses were justified which can easily be done by less experienced users of the Scientific Use Files of the NEPS.

With this paper, we specify an approach to link the responses over time within a Mokken Scaling using some of the linking procedures presented by van der Ark and Sijstma. The applicability will be evaluated with a simulation study and data from the NEPS mathematics assessments. The limitations and robustness of this approach in comparison to parametrical IRT will be discussed.

Controlling for cohort effects using a latent change score model with moderators

Thursday, 14th July - 09:45: Longitudinal Data (Room C) - Individual Oral Presentation

Mr. Pablo F. Cáncer [1], ***Dr. Emilio Ferrer*** [2], ***Dr. Eduardo Estrada*** [1]

1. *Universidad Autónoma de Madrid,* ***2.*** *University of California, Davis*

Purpose. Accelerated longitudinal designs (ALD) allow studying developmental processes usually spanning many years in a much shorter time frame. The key assumption of ALDs is that individuals from different cohorts (i.e., born in different years) belong to the same population, and thus the populational trajectory can be described by a shared set of parameters. However, participants born in different years may have been exposed to different contextual factors, leading to differences in their developmental patterns. According to previous research, failing to account for such differences will result in unreliable estimates. As a solution to this problem, we propose an extension of the latent change score model in continuous time that captures cohort effects in the context of ALDs. In particular, we focus on cohort differences in the self-feedback parameter.
Method. Through a Monte Carlo study, we examined the performance of the proposed model under different conditions of sample size, sampling schedule, and size of cohort differences.
Results. The proposed model adequately detects and controls for cohort differences in ALDs, regardless of the size of such differences. When the appropriate sampling schedule is selected, the performance of the model is excellent even with sample sizes of 125 individuals.
Discussion. We discuss the most relevant findings, elaborate on the strengths and limitations of our approach, and provide recommendations about the design of longitudinal studies. We encourage researchers to use the proposed model when they expect differences across cohorts in their patterns of change.

Momentary profile similarity measures for multivariate dyadic time series

Thursday, 14th July - 10:00: Longitudinal Data (Room C) - Individual Oral Presentation

Ms. Chiara Carlier [1], ***Dr. Laura Sels*** [2], ***Prof. Peter Kuppens*** [1], ***Prof. Eva Ceulemans*** [1]

1. *KU Leuven,* ***2.*** *Gent University*

People live in continuous interaction with each other and their environments, forming dyads: college roommates, romantic partners, parents, colleagues, … Interactions are of major importance in the course of our lives and impact our well-being both physically and mentally. One important or even essential quality to help us understand interactions is how similar two people are feeling or behaving. Until now, similarity has mostly been examined from a variable-centered approach using bivariate measures and from a cross-sectional approach, using single measurements or aggregate values. However, human interactions take on unique forms and fluctuate with varying contexts. Therefore, a multivariate, dyad-centered, longitudinal approach is better suited for dyadic interactions. In consequence, this requires methods to capture similarity in multivariate dyadic time series. Departing from the existing profile similarity methods, we developed momentary profile similarity measures. These measure takes the profile similarity between the state ratings of two people at a specific moment and can then be looked at over a given time course. During this talk we will guide the audience through several steps of this methodology by applying the momentary measures to an existing longitudinal data set of discrete emotions, rated by romantic couples. We will show how to compute a selection of different momentary profile similarity measures, explore their distribution, give examples of how to relate these measures to other variables, and discuss some strengths and weaknesses of the different measures.

Optimizing multistage adaptive testing designs for international large-scale assessments

Thursday, 14th July - 09:15: Computerized Based Testing (Room D) - Individual Oral Presentation

Dr. Usama Ali [1], ***Dr. Peter van Rijn*** [2]

1. *Educational Testing Service,* ***2.*** *ETS Global*

Multistage adaptive testing (MST) became so popular that several large-scale assessments have switched or are considering switching to enhance measurement efficiency and improve test-taker experience. In large-scale assessments, the target populations (e.g., countries in an international comparative assessment) are of different proficiency levels. For instance, the Programme for the International Assessment of Adult Competencies (PIAAC), as one of the largest and most innovative international assessments focusing on adult populations, is implemented across 38 countries in more than 50 languages (Kirsch & Lennon, 2017). Accordingly, when designing an MST, multiple competing goals need to be considered: a) to better match the proficiency level of a given respondent and the difficulty of assessment; and b) to get sufficient quality data (i.e., enough responses per item across the proficiency levels) to support the estimation of item parameters. In developing the instruments for large-scale assessments, it is critical to ensure content coverage and use the full item pool as well. Therefore, research is required to identify designs best suited for large-scale assessments. Seeking an optimal design for PIAAC, this paper addresses comparing the performance of different designs in terms of various evaluation criteria. The instrument development using optimization methods is also demonstrated. The results and implications are discussed.

Adjusted residuals for evaluating conditional independence in IRT models for multistage adaptive testing

Thursday, 14th July - 09:30: Computerized Based Testing (Room D) - Individual Oral Presentation

Dr. Peter van Rijn [1], ***Dr. Usama Ali*** [2], ***Dr. Hyo Jeong Shin*** [2], ***Dr. Seang-Hwane Joo*** [3]

1. *ETS Global,* ***2.*** *Educational Testing Service,* ***3.*** *University of Kansas*

The key assumption of conditional independence in item response theory (IRT) models is addressed for multistage adaptive testing (MST) designs. Routing decisions in MST designs cause patterns in the data that are not accounted for by the IRT model. This phenomenon relates to quasi-independence in log-linear models for incomplete contingency tables and impacts statistical inference using assumptions on observed and missing data. We demonstrate that generalized residuals for item pair frequencies under IRT models as discussed by Haberman and Sinharay (2013) are inappropriate for MST data without adjustments. The adjustments are dependent on the MST design, and can quickly become nontrivial as the complexity of the routing increases. The performance of the adjusted residuals is demonstrated through simulations and illustrated by an application to real MST data from the Programme for International Student Assessment (PISA). Implications and suggestions for MST design are discussed.

Using unrestricted latent class model to estimate the density of item-score vectors: Towards a flexible computerized adaptive test

Thursday, 14th July - 09:45: Computerized Based Testing (Room D) - Individual Oral Presentation

Mr. Anastasios Psychogyiopoulos [1], ***Dr. Niels Smits*** [1], ***Prof. Andries van der Ark*** [1]

1. *University of Amsterdam*

The construction of computer adaptive tests (CATs) with item response theory (IRT) models is considered standard practice. However, for many tests and questionnaires in psychological research IRT models may be suboptimal due to strict statistical assumptions. Recently, Van der Ark and Smits (2021) proposed replacing the IRT-model in CAT with a latent class model (LCM) and replacing the estimated latent trait by any score convenient for communication (e.g., sum score); they illustrated this approach in a proof-of-principle study. In the construction of the CAT, the LCM estimates the joint density of the item scores (π**)** and the density of the (total) score (π_+). The question at hand is: "Which information criterion (IC) should be used to obtain an adequate estimate of π and π_+?" As the estimation of π typically requires many latent classes, the current literature does not provide an answer. In a large-scale simulation study using an experimental design, we investigated the effects of IC, sample size, item format, number of items, and the number of true latent classes on the bias and accuracy of estimated π, and estimated π_+. On the basis of the outcomes, we will provide guidelines for calibrating sound CATs using LCM.

Useful and proper distractors for multiple choice items in cognitive diagnosis

Thursday, 14th July - 10:00: Computerized Based Testing (Room D) - Individual Oral Presentation

Prof. Hans Friedrich Koehn [1], ***Dr. Chia-Yi Chiu*** [2]

1. *University of Illinois, Urbana-Champaign,* ***2.*** *Graduate School of Education at Rutgers, The State University of New Jersey*

A multiple-choice (MC) item consists of the "stem"—the problem description—and a list of options comprising a correct response known as the "key" and incorrect options, the "distractors." The first cognitively diagnostic model for analyzing MC items with explicit consideration of the distractors was de la Torre's MC-DINA model. It relies on two critical assumptions concerning the q-vectors of distractors: they must be nested within that of the key and within each other. These assumptions are supposed (a) to ensure that examinees do not have to choose between options with identical q-vectors and (b) to impose a strict order on the coded response options to eliminate any ambiguities in determining an examinee's ideal response.

De la Torre's nestedness requirements have been said to be overly restrictive and thus, limit the ability to construct diagnostically meaningful distractors. We propose two criteria labeled "useful" and "proper" that relax the nestedness restriction allowing for greater flexibility in devising distractors without sacrificing de la Torre's original goals. A distractor is said to be useful if it is not redundant. A redundant distractor is one that does not improve the classification of examinees beyond the response options already available for a given item. A distractor is said to be proper if it allows for the nonambiguous identification of an examinee's ideal response. Conditions are defined for the two proposed criteria; proofs of their efficacy are presented and established in simulations.

Toward a quantifiable definition of validity

Thursday, 14th July - 09:15: Validity and Reliability (Room G) - Individual Oral Presentation

Dr. Mijke Rhemtulla [1], ***Ms. Anna Wysocki*** [1], ***Dr. Riet van Bork*** [2]

1. *UCDavis,* ***2.*** *University of Pittsburgh*

While reliability and validity are typically described as differentiable constructs, in practice, evidence for validity and reliability are often highly overlapping (e.g., internal consistency coefficients are interpreted as evidence for both) and empirical "validity coefficients" are often no more than correlations among observed measures. We propose a theoretical definition of validity that is orthogonal of reliability and that can support the development of empirical validity coefficients. Under a modern test theory model in which test scores *X* are composed of attribute (*A*, the attribute measured by the test), specificity (*S*, whatever is reliable but unrelated to *A*), and measurement error (E), we propose that validity with respect to A be defined as:
$\rho^2_{AT_Y}$=var(*A*)/var(*A*+*S*)

We describe the behavior of as a function of test length, amount of specific variance, and shared specific variance among items, relative to that of reliability and internal consistency. From this theoretical definition, we derive empirical estimates of validity under various sets of measurement assumptions. For example, for a scale composed of items that share only attribute variance, scale validity can be estimated as:
$\rho^2_{AT_Y} = (k^2 - \text{cov}(X_i, X_j))/((\text{var}(Y)\text{cor}(Y,Y'))$,

where *Y* and *Y'* represent two repetitions of the same assessment. We introduce model-based estimators for the more plausible scenario in which items share specific variance.

Adjusting scores for systematic bias using additivity analysis

Thursday, 14th July - 09:30: Validity and Reliability (Room G) - Individual Oral Presentation

Dr. Joseph Grochowalski [1]

1. *The College Board*

In this paper, I offer a method to remove bias from G-theory test scores that contain systematic sources of irrelevant variance. Test scores are often contaminated by such bias, which result in correlated errors. Correlated errors inflate reliability estimates, but are detectable using additivity analysis. Using additivity analysis, I propose a method that identifies sources of systematic error, adjusts observed scores to partial out the effects of systematic variance, and estimates reliability for the adjusted scores. I conducted a simulation that illustrated the accuracy of method for recovering true scores and reliability estimates, and applied the method to an applied study of adolescent mental health. The results of the applied study provided validity evidence that the sources of irrelevant variance were predictable using variables external to the scale, so score adjustments were defensible. I argued that the adjusted scores were a more precise and accurate predictors of true scores, and that this method should be applied generally in classical test theory and G-theory applications.

How to estimate ICCs for interrater reliability from incomplete designs

Thursday, 14th July - 09:45: Validity and Reliability (Room G) - Individual Oral Presentation

Ms. Debby ten Hove [1], ***Dr. Terrence Jorgensen*** [1], ***Prof. Andries van der Ark*** [1]

1. *University of Amsterdam*

Many observational studies use an incomplete design in which the observers vary across subjects. Traditional estimation methods of ICCs, which are used to investigate the interrater reliability (IRR) of observations, cannot handle these designs. We therefore conducted a simulation study to compare three novel estimation methods of ICCs for IRR: ICCs based on the variance decomposition of observations using (1) Markov chain Monte Carlo (MCMC) estimation of hierarchical linear models, (2) maximum likelihood estimation (MLE) of latent variable models, and (3) MLE of random-effects models, using the R software packages brms, lavaan, and mle4, respectively. For the MCMC-based method, we computed point estimates and credibility intervals of the ICC, which were readily provided by the software. For the MLE-based methods, we computed point estimates and two types of confidence intervals (CIs) that are specifically useful for coefficients such as ICCs (i.e., functions of parameters whose sampling distributions cannot be expected to be normal): Delta-method-based CIs and Monte-Carlo CIs. Our results showed that—across data-generating conditions (i.e., varying rater variance, numbers of subjects, sizes of the observer pool, and numbers of observers per subject)—MLE estimation of a random-effects model performed best concerning the bias-variance tradeoff, and the Monte-Carlo CIs yielded better CI-coverage rates than Delta-method-based CIs. Also, the estimation method using MLE of random-effects models was computationally least intensive and converged for nearly all replications. We recommend to use MLE of random effects models with Monte-Carlo-based CIs to estimate ICCs from incomplete observational designs, for which we provide user-friendly R scripts.

More on having and eating one's cake: Why modern measurement theory is a poor recipe for preparing predictive tests

Thursday, 14th July - 10:00: Validity and Reliability (Room G) - Individual Oral Presentation

Dr. Niels Smits [1]***, Prof. Andries van der Ark*** [1]

1. *University of Amsterdam*

In many clinical settings, questionnaire-based assessments are used to obtain information about the well-being as experienced by patients. Commonly, measurement models, such as those based on Item Response Theory (IRT), are employed to meaningfully reduce a patient's item scores to a test score (typically denoted θ). Such scores are mostly used to evaluate and monitor the patient, and to provide feedback on her status. For such purposes, measurement precision is key because the scores should accurately represent the patient's attributes. However, tests are also used for predictive purposes, such as forecasting a future health state, or a diagnosis based on the gold standard. Taking a classical test- theoretical approach, it has been shown that in test design a trade-off exists between measurement and prediction, and thus that accurate measurement is no prerequisite for good prediction. This raises several questions concerning the use of test data for prediction purposes. In the present study, three questions are answered: (1) Does the paradox of measurement versus prediction also exist under IRT?, (2) How should the scales from test batteries be combined for predictive purposes?, and (3) How should the reliability of predictions from test batteries be obtained? To answer these questions an illustrative data file is used consisting of 735 patients with scores on a battery of PROMIS-scales assessing health-related quality of life. The patients either did or did not have a clinical condition associated with neuropathic pain; this outcome is considered the gold standard and used as target variable for prediction.

Resolving the test fairness paradox by reconciling predictive and measurement invariance

Thursday, 14th July - 10:15: Validity and Reliability (Room G) - Individual Oral Presentation

Dr. Safir Yousfi [1]

1. *German Federal Employment Agency*

The dominant approach to establish that a psychological or educational test is fair with respect to a demographic characteristic like gender or age is to show that predictive invariance holds (e.g. identical regression of the criterion on the test scores). More recently, psychometricians claim that measurement invariance should be regarded as a major prerequisite for test fairness. Both criteria for test fairness are required by common standards for educational and psychological testing.

However, it has been shown that predictive invariance and measurement invariance are incompatible concepts and cannot hold simultaneously. While psychometricians concluded that a choice between these approaches has to be made, test developers and test users seem to neglect the incompatibility.

A psychometric approach to test fairness is suggested that resolves the incompatibility of predictive and measurement invariance by adopting the key ideas behind both competing concepts of test fairness. Within this framework, predictive invariance and measurement invariance are special cases of psychometric fairness. The integrated psychometric framework will hopefully contribute to a better understanding of statistical aspects of fairness and help to improve the development and use of educational and psychological tests.

Supervised multidimensional scaling for process data

Thursday, 14th July - 09:15: Process Data (Room E) - Individual Oral Presentation

Ms. Ling Chen [1], ***Dr. Xueying Tang*** [2], ***Prof. Jingchen Liu*** [1]

1. *Columbia University,* ***2.*** *University of Arizona*

Computer-based problem-solving items have gained popularity in latent trait assessment. In these items, respondents' interaction with the computer interface is recorded in the computer log files as a sequence of timestamped actions. Such process data contain detailed information about respondents, but the non-standard data format creates difficulty in utilizing the information. Recently, multidimensional scaling (MDS) has been used for extracting informative features from process data. In MDS, the chosen dissimilarity measure plays an essential role in the quality of extracted features. Different dissimilarities may reflect the variability of different traits among individuals, thus capturing different aspects of the information in process data. However, it is unclear how to choose an appropriate dissimilarity measure for understanding a given characteristic of the respondents. In this paper, we extend MDS to a supervised feature extraction method. This method integrates several dissimilarity measures so that the resulting measure produces features that are most relevant to the target variable. We demonstrate the performance of this method through simulation experiments and a case study of PIAAC process data.

Enhancing latent models of rapid-guessing with additional process data indicators

Thursday, 14th July - 09:30: Process Data (Room E) - Individual Oral Presentation

Ms. Jana Welling [1], ***Prof. Claus H. Carstensen*** [2], ***Dr. Timo Gnambs*** [1]

1. *Leibniz-Institute for Educational Trajectories,* ***2.*** *University of Bamberg*

Rapid-guessing poses a threat to the validity of large-scale assessments (Wise, 2017). Most of the existing identification methods, however, rely solely on item responses, item response times or both (e.g. Pokropek, 2016; Wise & Kong, 2005), which raises the risk of misclassifications. Process data provides a rich array of information that is easy to assess and could be thus used to derive additional indicators of rapid-guessing, enhancing existing methods. In the present study we define three new process variables *text reread*, *item revisit* and *answer change* as potential indicators of rapid-guessing in multiple-choice items in a reading task. The predictive power and validity of the new indicators shall be tested in a dependent latent class IRT-model (Nagy & Ulitzsch, 2021; Pokropek, 2016). In this multilevel mixture model, the latent class variable representing rapid-guessing is for each person-item encounter regressed on its indicators. The new indicators shall be added as additional predictors in the logistic regression. Item responses are supposed to follow different distributions, depending on the latent class. In the rapid-guessing class, the probability of a correct response is equated to the guessing parameter, whereas in the solution-behavior class it is defined by an IRT model. The study is based on data of a web-based assessment in wave 12 of starting cohort 5 of the German National Educational Panel Study. The sample comprises 1933 (former) students. All 14 multiple-choice items of the reading test are included, resulting in 27,062 person-item encounters and 71,070 corresponding log events.

Employing process-data indices to account for response biases in questionnaire data

Thursday, 14th July - 09:45: Process Data (Room E) - Individual Oral Presentation

Dr. Marek Muszyński [1], ***Dr. Tomasz Żółtak*** [1], ***Prof. Artur Pokropek*** [2]

1. *IFiS Polish Academy of Sciences,* ***2.*** *Polish Academy of Science*

The presentation aims to broaden knowledge on using process (log) data (Kroehne & Goldhammer, 2018) to improve questionnaire data quality by developing methods based on process-data indices to detect careless responding and response styles - common response biases, threatening self-reports validity. Properly developed process-data indicators can lead to higher detection of aberrant or unmotivated respondents in questionnaire data, allowing for better performance of measurement models.

In this study process data are used to identify careless participants in order to offer additional criteria to careless responding indices (Meade & Craig, 2012) to disentangle aberrant from valid responses and to predict response styles indices based on IRTrees (Boeckenholt, 2012, 2017) and multidimensional generalized partial credit models (Henninger & Meiser, 2019). Finally, log-data indices are validated by comparing response indices collected in a series of large (samples of ca. 3000 participants) web-survey experiments, where task difficulty and participants' motivation are manipulated to simulate satisficing (Krosnick, 1991).

Experiment results show that log-data indices correlated in an expected direction with careless responding indices, e.g. measures of distance traveled and the number of flips of cursor movements were negatively related to straightlining, positively to self-reported diligence and page time, and, although weakly, to the Mahalanobis distance. Log-data indices did not differ much between experimental conditions inducing larger response motivation, although the distance traveled in the horizontal dimension was larger in conditions inducing higher motivation. Furthermore, we compare log-data indices under induced conditions of working memory burden and forced multitasking. Moreover, log-data indices cross-assessment stability is tested.

Bayesian joint modeling of response accuracy and real-time emotions

Thursday, 14th July - 10:00: Process Data (Room E) - Individual Oral Presentation

Prof. JoonHo Lee [1], ***Prof. Yurou Wang*** [1]

1. *The University of Alabama*

Computerized testing has made it possible to collect various types of parallel information to improve our understanding of cognitive processes underlying test performance. Emotions can be a novel source of information to unravel the processes. Activating negative emotions (e.g., test anxiety) can reduce cognitive resources available for information processing while activating positive emotions (e.g., test enjoyment) can help promote task attention.

Using the Contain Intelligent Facial Expression Recognition System (CIFERS), this study measured the real-time emotional status during students' problem-solving process. CIFERS tracked micro-facial expressions as indicators for nine different emotions during the time between the item presentation and the actual response. Then, the item responses and real-time emotions were analyzed jointly within a Bayesian hierarchical modeling framework. At the item level, item responses were modeled by a two-parameter IRT model and emotional status by a log-normal model, with the respective location (item difficulty; emotion intensity) and slope (item discrimination; emotion sensitivity) parameters. At the person level, between-person differences in ability and the latent emotional trait were modeled by the multivariate distribution of crossed random effects.

The positive correlation between item difficulty and emotion intensity indicated that the more difficult items tend to induce higher emotions on average. The discriminating items with respect to person ability also tend to discriminate well with respect to the overall level of emotion. The practical implications of the estimated correlation between person-level ability and latent emotional traits were also discussed.

Introduce confusion matrix to model evaluation in quantitative psychology

Thursday, 14th July - 09:15: Statistical Learning (Room F) - Individual Oral Presentation

Mr. Yongtian Cheng [1], ***Dr. Konstantinos Petrides*** [1]

1. *UCL*

Evaluation of the performance of a proposed model is an important part of both machine learning (ML) and psychological studies. Psychological studies use Monte Carlo simulation (MC) to provide empirical evidence about the performance of the model in various scenarios and conditions. In the model comparison, ML and psychological share some similar criteria like mean square error and root mean square error. The confusion matrix is another criteria mostly used in ML model evaluation. Yet, this criterion and design, especially the condition of null effect, is seldom used in the MC studies in psychology in model evaluation. As proposed by the author, the confusion matrix, both used as a design and a criterion, should be included in MC studies aiming at model evaluation and comparison.

This study has used an influential study done by Trizano-Hermosilla and Alvarado (2016) that proposed Omega should be used rather than Alpha, which are two models to evaluate the internal consistency(IC) of a questionnaire, as a sample to introduce the proposed design and evaluation of confusion matrix.

As we have found, the target study is missing an important part of empirical evidence of the model in the null effect condition. Especially in some conditions with a random number with no IC, over 20 to 40 percent of Omega can reach a satisfying level of .7. This MC simulation with the proposed design provides new and essential empirical evidence to the selection of the IC model, showing the necessity of this design.

A simulation study comparing the use of supervised machine learning variable selection methods in the psychological sciences

Thursday, 14th July - 09:30: Statistical Learning (Room F) - Individual Oral Presentation

Ms. Catherine Bain [1], ***Dr. Jordan Loeffelman*** [1]

1. *University of Oklahoma*

When specifying a predictive model for classification, variable selection (or subset selection) is one of the most important steps for researchers to consider. Reducing the necessary number of variables in a prediction model is vital for many reasons, including reducing the burden of data collection and increasing model efficiency and generalizability. The pool of variable selection methods from which to choose is large, and researchers often struggle to identify which method they should use given the specific features of their data set. Yet, there is a scarcity of available literature to guide researchers in their choice, rather the literature centers on comparing different implementations of a given method rather than comparing different methodologies under vary data features. Through the implementation of a large-scale Monte Carlo simulation and the application to three psychological datasets, we evaluated the prediction error rates, area under the receiver operating curve, number of variables selected, and computation times of five different variable selection methods using R under varying parameterizations (i.e., default vs. grid tuning): the genetic algorithm (*ga*), LASSO (*glmnet*), Elastic Net (*glmnet*), Support Vector Machines (*svmfs*), and random forest (*Boruta*). Preliminary results indicate that the genetic algorithm is the most widely applicable method, exhibiting minimum error rates in hold-out samples with less items when compared to other variable selection methods.

A critical view on interpretation techniques for machine learning methods

Thursday, 14th July - 09:45: Statistical Learning (Room F) - Individual Oral Presentation

Dr. Mirka Henninger [1], ***Dr. Yannick Rothacher*** [1], ***Dr. Rudolf Debelak*** [1], ***Prof. Carolin Strobl*** [1]

1. *University of Zurich*

Machine Learning (ML) methods have become very popular tools in many research areas, including psychology. They have been shown to make good predictions of the outcome variable, in particular when the data contain nonlinear and interaction effects that cannot be presumed a priori. However, many ML methods, such as random forests or neural networks, are so called “black boxes”. This means that one cannot see from the ML method itself which predictor variables are relevant and in what way they influence the outcome variable or interact with each other. A means to this end are techniques from Interpretable Machine Learning (IML): a variety of graphical and numerical tools to support researchers in describing how the black box came to its decision. In this talk, we briefly introduce some of these IML tools, such as permutation importance measures and visualization methods like Partial Dependence (PD) and Accumulated Local Effect (ALE) plots in scenarios that are typical for empirical psychological studies. Furthermore, we illustrate characteristics and potential pitfalls that can occur when these interpretation techniques are applied in psychology. In particular, we show via simulated examples how they might mask effects that are present in the data, or erroneously suggest effects that are not present. We believe that it is important to critically reflect about ML and IML tools, as they will become more and more important in psychological research.

Utilizing machine learning for simulation-based design optimization

Thursday, 14th July - 10:00: Statistical Learning (Room F) - Individual Oral Presentation

Mr. Felix Zimmer [1], ***Dr. Rudolf Debelak*** [1]

1. *University of Zurich*

Planning an adequately powered research design increasingly goes beyond determining an appropriate sample size. More sophisticated scenarios require the simultaneous tuning of various parameters of the design and can only be tackled using Monte Carlo simulation. In addition to the desired statistical power, we want to ensure the optimality of the solutions with respect to a cost metric, such as the financial cost of a study. We introduce a surrogate modeling method to optimize both power and cost with the help of machine learning. Applications include finding design parameters that imply a desired power at minimum cost or, alternatively, maximum power given a cost threshold. As surrogate models, which guide the search process, we use Gaussian process regression and support vector regression. We demonstrate the performance of the method in an extensive simulation study using various hypothesis test scenarios with single and multidimensional design parameters using scenarios from classical statistics, multilevel modeling, and item response theory.

GeneticPower: A genetic algorithm-based framework for learning statistical power manifold

Thursday, 14th July - 10:15: Statistical Learning (Room F) - Individual Oral Presentation

Mr. Abhishek Kumar Umrawal [1], ***Dr. Sean Lane*** [1], ***Dr. Erin Hennes*** [1]

1. *Purdue University*

Statistical power is a measure of the goodness/strength of a hypothesis test. Formally, it is the probability of detecting an effect, if there is a true effect present to detect. Hence, optimizing statistical power as a function of some parameters of a hypothesis test is desirable. However, for most hypothesis tests, the explicit functional form of statistical power as a function of those parameters is unknown but calculating statistical power for a given set of values of those parameters is possible using simulated experiments. These simulated experiments are usually computationally expensive. Hence, developing the entire statistical power manifold using simulations can be very time-consuming. Motivated by this, we propose GeneticPower – a novel genetic algorithm-based framework for learning statistical power manifold. For a multiple linear regression *F*-test, we show that the proposed algorithm/framework learns the statistical power manifold much faster as compared to a brute-force approach as the number of queries to the power oracle is significantly reduced. We also show that the quality of learning the manifold improves as the number of iterations increases for the genetic algorithm.

Structural equation modeling for Errors-in-variables systems

Thursday, 14th July - 10:50: Invited Speaker: Fan Yang Wallentin (Room B) - Individual Oral Presentation

Prof. Fan Yang Wallentin [1]

1. *Uppsala University*

Errors-in-variables (EIV) identification refers to the problem of consistently estimating linear dynamic systems whose output and input variables are affected by additive noise. Several estimation methods have been proposed for identifying linear dynamic systems from noise-corrupted output measurements. This talk introduces Structural Equation Modeling (SEM) setting to EIV identification. Two schemes for how EIV Single-Input Single-Output (SISO) systems can be formulated as SEMs are presented. The proposed formulations allow for quick implementation using standard SEM software. Simulation examples show that compared to existing procedures, such as the covariance matching (CM) approach, SEM-based estimation provides parameter estimates of similar quality.

Factor analysis for multi-way data

Thursday, 14th July - 10:50: Invited Speaker: Paolo Giordani (Room A) - Individual Oral Presentation

Prof. Paolo Giordani [1]

1. *Sapienza University of Rome*

Traditional factor analysis provides an important tool to investigate standard two-way data regarding the measurements of a collection of manifest variables (e.g., various anxiety scales registered by a psychologist) on a set of individuals (e.g., patients). However, it may occur that, for getting more information on the analyzed phenomena, such measurements are repeated in different (time) occasions (e.g., by a number of psychologists, and/or in different stressful conditions, and/or yearly). In these cases, the observed data acquire a multi-way structure, and they can be stored in a multi-way array or tensor (standard matrices can be seen as two-way arrays or tensors). Multi-way data are inherently complex, making conventional factor analysis inadequate. While posing several challenges from a theoretical point of view, such a data complexity has also a positive by-product in the richness of data features that are observed and made available to deepen the knowledge on the phenomena of interest. In this talk, limiting the attention to the three-way framework, I will show how to generalize traditional factor analysis to three-way data with a particular focus on the so-called Tucker3 and PARAllel FACtor (PARAFAC) models. Tucker3 and PARAFAC will be introduced, and their recent advances, developed to fully exploit the rich information given by such a complex and unconventional data structure, will be discussed. These advances include novel penalized estimation methods, as well as new modeling strategies.

Fair algorithms, causality, and measurement

Thursday, 14th July - 11:40: Keynote: Joshua Loftus (Room B) - Individual Oral Presentation

Dr. Joshua Loftus [1]

1. *London School of Economics and Political Science*

I will discuss a line of recent work using causal models to understand algorithmic fairness. Rather than attempting to make minimal assumptions and provide robust inferences, this approach uses strong assumptions for the sake of interpretability, transparency, and falsifiability. Although the application focus is on fairness, causal models can be applied in similar ways toward achieving other values or objectives in responsible machine learning or data-driven decisions more broadly. I will conclude by discussing some of the hard challenges in fair machine learning that connect to measurement issues and problems in psychometrics.

Analyzing clinical scales using full information optimal scoring

Friday, 15th July - 09:15: Applications I (Room B) - Individual Oral Presentation

Dr. Juan Li [1], ***Prof. James Ramsay*** [2], ***Prof. Marie Wiberg*** [3]

1. *Ottawa Hospital Research Institute,* ***2.*** *McGill University,* ***3.*** *Umeå University*

Clinical scales are essential to the health and social sciences, and to the individuals that provide the data. Although statistical models for scale data have been researched for decades, it remains nearly universal that scale scores are sums of weights assigned a priori to question choice options (sum scores). The recently proposed full information optimal scoring method can be used to analyze scale data and provides important improvements in the quality of rating scale scores. Several features of the proposed method are: using percentile rank over [0, 100] as score indices (θ) to place test takers on the information curve; transforming probability into surprisal; fitting ICCs using spline-smoothing; using option weights as functions of θ that better represents option-score interactions. De-identified 'Hospital Anxiety and Depression Scale (HADS)' data of 810 patients with multiple sclerosis (MS) are used for illustration. Standard errors of performance estimates are shown to be as small as a quarter of those of sum scores. The estimation algorithm permits the analysis of data from tens and hundreds of thousands of test takers in a few minutes on consumer level computing equipment. Open access software resources are presented.

How about … no? – Using missing responses and response times to model avoidance behavior

Friday, 15th July - 09:30: Applications I (Room B) - Individual Oral Presentation

Mr. Nico Remmert [1], ***Dr. Robert Krause*** [1], ***Prof. Steffi Pohl*** [1]

1. *Freie Universität Berlin*

Avoidance is one of the central behaviors related to mental health problems, adversely contributes to the maintenance of clinical symptoms, and has an impact on treatment outcomes. So it is hardly surprising that research has brought up several assessment methods being the basis of its investigation. However, there is a considerable demand for a more sophisticated and direct behavioral assessment of different avoidance strategies. In this project, we propose three new behavioral avoidance measures: a) anticipating avoidance tendency, b) anticipating avoidance speed, and c) escaping speed, applying a new psychometric avoidance framework. We herein unify previous assessment approaches (behavioral avoidance tests) with new developments in computerized assessment and psychometric modeling. The proposed measures are based on latent modelling of log-data (i.e., missing responses and response times), which are a valuable source of information accounting for differences in test-taking behavior. We argue that this framework provides a sophisticated and rigorous measurement of directly observed avoidance behavior that can be used for different forms of mental disorders. Our framework is demonstrated through an applied example.

Conditional standard errors of measurement for math modelling assessments

Friday, 15th July - 09:45: Applications I (Room B) - Individual Oral Presentation

Dr. Cigdem Alagoz [1], ***Dr. Celil Ekici*** [2]

1. *Texas Education Agency,* ***2.*** *Texas A and M University-Corpus Christi*

Conditional standard errors of measurement (CSEMs) provide useful information about the performance of the writing assessment. Reporting CSEMs is recommended by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999). Few procedures exist, however, for estimating CSEMs for a polytomously scored single-prompt test and those that have not been well-reported in the context of IRT. The FACETS computer software (Linacre, 2010) provides a separate, unique standard error estimate for each examinee using the Rasch model. Brennan (1998) defined the conditional absolute SEM in the context of generalizability theory. Preliminary research indicates that the CSEMs from these two perspectives do not agree. In this study, we examine the similarities and differences of CSEMs from the perspectives of IRT and generalizability theory.

Calculation of conditional standard errors of measurement for a math modelling assessment are compared within a generalizability theory framework and an IRT Rasch framework. Similarities and differences in patterns and magnitudes of CSEMs will be investigated between the two approaches.

Based on previous research, the IRT approach is expected to yield larger CSEMs at the extremes of the score scale than in the middle. Estimates should also peak between the cut scores. Somewhat different results are reported from generalizability theory, depending on whether raw scores or standard scores are used.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Linacre, J. M. (2010). FACETs Rasch measurement computer program Chicago,IL.:Winsteps.com.

Measuring digital competence internationally: Exploring test dimensionality, position effect and performance differences

Friday, 15th July - 10:00: Applications I (Room B) - Individual Oral Presentation

Dr. Yuan-Ling Liaw [1], ***Dr. Mojca Rozman*** [1], ***Dr. Andrés Christiansen*** [1], ***Dr. Rolf Strietholt*** [1]

1. *International Association for the Evaluation of Educational Achievement (IEA)*

Digital competence is an important capacity for students' learning in today's fast changing world. It is conceived as an overarching concept that includes computer and information literacy (CIL) and computational thinking (CT). Debate exists between the theoretical conceptualizations of CIL and CT as multidimensional constructs and empirical studies reporting unidimensional scores for each domain. Also, little is known about whether item position or module have an undesired effect on test performance. This study uses the assessment from IEA's International Computer and Information Literacy Study (ICILS), which was designed to measure international differences in students' ICT-related skills. The assessment utilized a sequence of tasks contextualized by a real-world theme and required students to use repeat and conditional statements to solve problems. In ICILS 2018, each student completed two of the five CIL modules and students in countries participating in the CT option completed the two CT test modules in randomized order. For example, each CIL test module consisted of a series of smaller tasks and a single large task, whereas each CT module had a unifying theme and a sequence of tasks that related to the theme, but not a large task. Empirical data showed that students' response times and percentages of omissions differed across test modules. The present study applies a multilevel extended item response theory (IRT) framework and uses data from 14 education systems to study test dimensionality and position effect. Additionally, we examine whether country rankings of CIL and CT are robust.

The effects of the Covid 19 pandemics on the mental health of elderly people

Friday, 15th July - 10:15: Applications I (Room B) - Individual Oral Presentation

Prof. Francesco Scalone [1], ***Prof. Rosella Rettaroli*** [1]

1. *University of Bologna*

We aim to assess the effects of the Covid 19 epidemics on the psychological wellbeing among the elderly population in several European countries using the dataset from SHARE (Survey on Health Ageing and Retirement). As previous studies have already shown, most fragile individuals suffered the effects of the severe lockdown, social distancing and separation from their relatives during the recent waves of the Corona Virus. On the one hand, all these measures protected the elderly individuals from the Covid 19 infections; on the other hand, they caused stressful situations and mental health sufferings.

The last SHARE-COVID19 survey has recently provided a special dataset section related to the health issues and socioeconomic impact of COVID-19. Using this data and adopting a structural equations modeling (SEM) approach, we will explore the effects of the epidemics on anxiety, depression and self-reported health status among the elderly population in several European countries.

Generalised additive latent variable models for location, shape, and scale

Friday, 15th July - 09:15: Latent Variable Models and Estimation (Room A) - Individual Oral Presentation

Mr. Camilo Cardenas [1], ***Prof. Irini Moustaki*** [1], ***Prof. Giampiero Marra*** [2]

1. *London School of Economics and Political Science,* ***2.*** *University College London*

Latent variable models are used to analyse multivariate data using a small number of factors. There are well established modelling frameworks for such objective, but often higher order characteristics of the observed items are ignored. In this talk, we extend the Generalized Additive Models for Location, Shape and Scale framework (GAMLSS, Rigby and Stasinopoulos, 2005) to models with latent variables (Bartholomew et al., 2011). The proposed framework allows for linear and nonlinear predictors, as well as heteroscedastic error terms. More specifically, we assume different regression equations for the location, scale, and shape parameters of the univariate (conditional) distributions for each of the observed variables. Modelling the mean, scale, and shape as a function of latent variables and covariates allows for a more flexible and general modelling framework than the classical factor analysis model. A computationally efficient penalised maximum likelihood estimation is proposed. Examples from educational data from large scale surveys are used to demonstrate its applicability.

A Dirichlet response model for the dual-range slider item format

Friday, 15th July - 09:30: Latent Variable Models and Estimation (Room A) - Individual Oral Presentation

Mr. Matthias Kloft [1], ***Dr. Raphael Hartmann*** [1], ***Prof. Andreas Voss*** [2], ***Prof. Daniel W. Heck*** [1]

1. *Philipps-University Marburg,* ***2.*** *Ruprecht-Karls-University Heidelberg*

Single response formats come with the drawback of forcing respondents to condense behavioral distributions into one response value. The variance of these distributions is not considered. A conceivable way of approximating this variance could be the utilization of interval response formats, more specifically the dual-Range Slider (RS2). So far, an item response model for this response format does not exist. Therefore, we develop a measurement method for the RS2 by extending the Beta Response Model (BRM; Noel & Dauvier, 2007) to the Dirichlet Dual Response Model (DDRM).

To evaluate the DDRM's performance, we first assess parameter recovery in a simulation study. The results indicate overall good parameter recovery, although parameters concerning the RS2 interval width perform worse than parameters concerning the RS2 interval location.

Second, we jointly fit the BRM and DDRM to empirical single-Range Slider (RS1) and RS2 responses for two extraversion scales. The DDRM has an overall acceptable fit, but it shows some misfit regarding the RS2 interval widths. Nonetheless, respondents can be differentiated sufficiently. High correlations between person parameters of the BRM and DDRM suggest convergent validity between the RS1 and the RS2 interval location.

Additionally, both the simulation study and the empirical study demonstrate that the RS2's scale inherent interdependence between interval location and interval width is greatly reduced in the latent parameter space of the DDRM compared to mean scores.

In conclusion, the RS2/DDRM could be used in place of single response formats when additional information about the variance is needed.

Modeling measurement error in latent space modeling

Friday, 15th July - 09:45: Latent Variable Models and Estimation (Room A) - Individual Oral Presentation

Ms. Yishan Ding [1], ***Dr. Tracy Sweet*** [1]

1. *University of Maryland, College Park*

In social network analysis, most methods assume that the network is measured without error. Even when network tie data is collected from multiple items, common analysis strategies include either selecting one representative item to measure the network tie or treating each item as a separate type of network and conduct independent analyses (Kitts & Quintane, 2020).

Existing social network measurement error studies have mostly assumed the errors to result from incomplete or inaccurate observation from the manifest variable (Borgatti et al., 2006; Wang et al., 2012). However, many network analyses are based on constructs, such as friendship, that are difficult or even impossible to measure explicitly. As a result, simply equating the latent construct to one of its manifest variables induces measurement error that cannot be addressed by improving any single item, regardless of how complete and accurate the measurement is. Instead, latent variable modeling based on the repeated measurement is needed.

Therefore, we propose a new approach to accommodate network measurement error in social network analysis by integrating an item response theory model (IRT) into a latent space models (LSM). Our resulting model is an IRT-LSM, and we present a simulation study to evaluate the parameter recoveries of IRT-LSM under a variety of practical combinations of sample size and test length. We will also evaluate the IRT-LSM's improvement in recovering network centrality, compared with the naïve method using a single manifest variable.

Model implied instrumental variable approach in structural equation modeling with frequentist model averaging

Friday, 15th July - 10:00: Latent Variable Models and Estimation (Room A) - Individual Oral Presentation

Dr. Shaobo Jin [1]

1. *Uppsala University*

Structural equation models with ordinal data are often estimated by systemwide approaches such as diagonally weighted least squares and unweighted least squares. An alternative approach is the model implied instrumental variable (MIIV) approach. MIIV fits the model in an equation-by-equation manner, which tends to be more robust than systemwide approaches at the potential loss of efficiency. Previous simulation studies have shown that the estimator using all instrumental variables tends to be less viable but more biased than the estimators using a small set of instrumental variables. Hence, approaches such as using the instruments with the highest Shea's R2 have been proposed. In this study, we propose to combine the estimators using different sets of instruments by frequentist model averaging. Instead of choosing one single set of instruments, frequentist model averaging takes the advantages of all sets of instruments. The optimal weights to combine the estimators are derived. Its finite sample performance is investigated in a simulation study.

Fast estimation of generalized linear latent variable models: Thinking out of the box

Friday, 15th July - 10:15: Latent Variable Models and Estimation (Room A) - Individual Oral Presentation

Prof. Maria-Pia Victoria Feser [1], ***Mr. Guillaume Blanc*** [1], ***Dr. Stephane Guerrier*** [1]

1. *University of Geneva*

Generalized Linear Latent Variable Models, that map multivariate data to a lower-dimensional latent space, thereby providing a reduction of dimensionality, are particularly suitable for data sets of high dimensions with correlated mixed type of variables. In practice however, their estimation represent a tremendous challenge for even moderately large dimensions, essentially due to the multiple integrals involved in the likelihood function. Numerous alternative methods have been proposed to estimate GLLVMs, such one based on an approximated likelihood (Laplace approximation, adaptive quadrature, or recently variational approximations), or using an approximation to the EM-algorithm. Of all these likelihood-based methods, the best performing, and indeed current state-of-the art implementation, is the Variational Approach (VA). In this presentation, we propose an alternative route which consists in drastically simplifying the estimating equations and applying numerically efficient bias reduction methods in order to recover a consistent estimator for the GLLVM parameters. The resulting estimator is actually an M-estimator, and can accurately estimate GLLVM models with over 500 (non Gaussian) manifest variables and over 10 latent variables in about one second. Theoretical results as well as simulation studies will be provided.

Selecting interaction effects in additive models using I-priors

Friday, 15th July - 09:15: Model Fit (Room C) - Individual Oral Presentation

Prof. Wicher Bergsma [1], ***Dr. Haziq Jamil*** [2]

1. *London School of Economics and Political Science,* ***2.*** *Universiti Brunei Darussalam*

Additive models with interactions have been considered extensively in the literature, using estimation methods such as splines or Gaussian process regression. We present an alternative empirical-Bayes approach to selecting interaction effects using the I-prior approach introduced by Bergsma (2020). Using a parsimonious formulation of hierarchical interaction spaces, model selection is simplified. Furthermore, we present an efficient EM algorithm for estimating key hyperparameters.

Simulations for linear regressions indicate competitive performance with methods such as the lasso and Bayesian variable selection using spike and slab priors or g-priors. However, our methodology is more general and can also be used with interacting nonlinear regression functions.

Goodness of fit testing of SEM models in cross-validation samples

Friday, 15th July - 09:30: Model Fit (Room C) - Individual Oral Presentation

Prof. Alberto Maydeu-Olivares [1], ***Dr. Dexin Shi*** [1], ***Dr. Raul Ferraz*** [1], ***Dr. Goran Pavlov*** [1]

1. *University of South Carolina*

Model modifications in structural equation modeling (SEM) is almost inevitable, and often desirable, if we wish to approximate well the data generating process. However, in so doing, we risk over-fitting (capturing idiosyncrasies of the data) and confidence intervals for goodness of fit indices do not take into account that multiple testing has been performed, giving an overly optimistic view of the selected model. This problem can be overcome by using a cross-validation sample. We introduce new procedures for cross-validation in SEM in which parameters are held constant at the values estimated in the calibration sample. In this setup, degrees of freedom in the cross-validation sample equal the number of sample statistics and exactly identified models can be tested. We introduce an exact fit test statistic that follows an asymptotic chi-square distribution in this setup, show how to obtain RMSEA based on this statistic, and compare their performance in terms of Type I errors and power with respect to standard cross-validation procedures in which the model is simply re-estimated in the cross-validation sample.

Cross-validation indices for factor model scoring

Friday, 15th July - 09:45: Model Fit (Room C) - Individual Oral Presentation

Mr. Siyuan Marco Chen [1], ***Dr. Daniel Bauer*** [1]

1. *University of North Carolina at Chapel Hill*

Measurement studies in psychology and education commonly apply factor analysis models to assess factor structures of scales and establish scoring rules. The selection of factor models often relies on structural equation modeling (SEM) fit indices. However, while these SEM fit indices examine the discrepancy between the observed and the model-implied covariance matrices, models fitted in scale studies are often used to score existing and future samples. Thus, a model selection criterion better suited for scale assessment may be the generalizable scoring performance of factor models, i.e., how well the factor model performs in scoring individuals across samples. A class of cross-validated fit indices (e.g., Browne and Cudeck, 1989) has been studied in the psychology literature to evaluate the covariance model generalizability across samples, i.e., the appropriateness of a sample-estimated model in representing the population covariance structure, but none has considered how well a factor model calibrated in one sample may produce appropriate scores in another independent sample. This study extends covariance-based cross-validation fit indices to the factor score context by applying the factor score likelihood to constructing these indices. A simulation study is conducted to compare the performance of these proposed cross-validated fit indices relative to other exiting indices in identifying population true factor models and in estimating out-of-sample true scores. An empirical example is presented to show the differential model selection results based on the proposed cross-validated indices and the existing SEM fit indices and their implications.

Beyond Pearson's correlation: General association tests for psychological research

Friday, 15th July - 10:00: Model Fit (Room C) - Individual Oral Presentation

Dr. Julian Karch [1], ***Mr. Andres Felipe Perez Alonso*** [2]

1. *Leiden Universty,* ***2.*** *Tilburg University*

Psychologists traditionally employ traditional tests based on Pearson's, Kendall's, and Spearman's correlation coefficients to investigate whether two continuous variables are associated. These tests are only guaranteed to be consistent (power approaches one as the sample size increases) for linear and monotonic associations, while many associations in Psychology are non-monotonic. Numerous general association tests consistent for any association have been proposed in machine learning and nonparametric statistics to address this problem. Which of these tests has the highest power depends on many factors, such as the association's form. In this study, we propose two new general association tests: a mutual information-based test and a test based on the Heller-Heller-Gorfine distance as well as Pearson's correlation. Contrary to existing general tests, tests were designed to achieve high power in situations that commonly appear in psychological studies. In a simulation study, we compared the new tests with the traditional tests for associations and existing general association tests. As expected, no test was uniformly most powerful. However, the two new tests and distance correlation outperformed the traditional approaches by demonstrating substantially higher average power and being slightly less powerful only in some situations. Consequently, we argue that when testing for an association without knowing the form of the association, one of the new tests or the distance correlation test is preferable. We conclude by discussing those tests' relative strengths and weaknesses.

Testing correlations by using Fisher's z transformation and bootstrapping

Friday, 15th July - 10:15: Model Fit (Room C) - Individual Oral Presentation

Dr. Zhenqiu Lu [1], ***Dr. Ke-Hai Yuan*** [2]

1. *University of Georgia,* ***2.*** *University of Notre Dame*

Pearson product moment correlation (called correlation for short in this paper) is the most widely used statistic for investigating the association of two variables. If we want to test whether the correlation is from a population with mean zero, traditionally, people use the parametric procedure which uses Fisher's z transformation of correlation to compare with a normal distribution. But it required normally distributed data. In reality, data may not be normally distributed or contain outliers and missing values, or with a small sample size. In such cases, bootstrapping method (Efron and Tibshirani, 1993) uses resampling techniques to provide more accurate solutions. To test a correlation, there are bivariate bootstrapping method, univariate bootstrapping method, and bootstrap hypothesis testing method.

In this project, these four methods are studied and compared. First, the distribution of the correlation for each method is derived. Monte Carlo simulations for four testing methods are conducted to verify the conclusions. Also, the theoretical mechanisms in testing correlation are explained. The underlying assumptions of data distribution are discussed. The strengths and weaknesses of each method are discussed. Similarities and differences among them are provided. The connections between them are pointed out. Suggestions are provided under different scenarios.

Performance of covariate-informed factor scores in bivariate latent growth models

Friday, 15th July - 09:15: Growth Curve Models (Room D) - Individual Oral Presentation

Dr. Alexis Georgeson [1]

1. *Arizona State University*

The bivariate latent growth model allows researchers to examine the co-development of constructs, which is indispensable to developmental theory. Despite the advantages of modeling the measurement structure and growth simultaneously in a bivariate second-order latent growth model, the complexity of such models poses a challenge to estimation in practice. Instead, researchers often fit a first-order growth model to scores (e.g., sum scores) of the repeated measures. C*ovariate informed scores* are a novel type of score that are computed from parameters from the moderated nonlinear factor analysis model (MNLFA). By allowing model parameters to vary as a function of age, a single-factor MNLFA model can be used to obtain scores for all timepoints. The present study compared two types of scores: MNLFA scores obtained by fitting the constructs jointly, and MNLFA scores obtained by fitting each construct separately. The effect size of time-adjacent item residual covariances and within-time, between construct residual covariances were varied in the simulation. Using regression factor scores based on MNLFA parameters, the results showed severe relative bias (>10%) for the growth factor variances and covariances for nearly all conditions and score types. However, joint scores showed smaller bias for the between-construct growth parameters. Furthermore, the residuals showed bias-offsetting effects for certain growth parameters. As the first simulation study on this topic, the results suggest that covariate-informed scores from the MNLFA model could be a promising method for growth models, but more work is needed to reduce the bias.

Using bootstrap to test changes in growth curve models with non-normal data

Friday, 15th July - 09:30: Growth Curve Models (Room D) - Individual Oral Presentation

Ms. Stefany Mena [1], ***Dr. Han Du*** [1]

1. *University of California, Los Angeles*

Growth Curve models are widely used by researchers to analyze changes in data over time, such as students' test scores over time or cognition measures across the lifespan. Traditional growth curve models assume that residuals follow a normal distribution. However, non-normally distributed data are not uncommon and can take many forms, including highly-skewed and long-tailed distributions. Currently, researchers with non-normal data usually rely on robust standard errors which require relatively large sample sizes. Thus, the current study proposes a novel bootstrap method to test overall change in growth curve models. The proposed bootstrap method uses estimates of the first four moments from the observed sample to create bootstrapped datasets that are robust to non-normality. This method was tested using Monte Carlo simulations coverage, power and Type I error rates were compared with traditional growth curve models (with and without robust standard error corrections) and another bootstrap algorithm. Conditions were simulated to mimic real dataset situations (e.g., varied number of participants and measurement occasions, slope and error variance, etc.). Results show that the proposed bootstrap method indeed controls Type I error rates while traditional methods even with robust corrections cannot. Power and coverage will also be discussed. Thus, applied researchers can use this novel method when they are faced with non-normal data in growth curve models.

Confidence interval of effect size measures in longitudinal growth models

Friday, 15th July - 09:45: Growth Curve Models (Room D) - Individual Oral Presentation

Ms. Zonggui Li [1], ***Dr. Ehri Ryu*** [1]

1. *Boston College*

When modeling longitudinal data, multilevel modeling (MLM) is the commonly used framework. We previously developed full-model based effect size measures for longitudinal growth models (LGMs) in the MLM framework using R^2 (R-squared) (Li & Ryu, manuscript in preparation). In this study, we extend the previous work to provide confidence interval (CI) computation for the developed effect size measures. Though simulation study has shown that the non-parametric residual based bootstrapping is an appropriate method to compute CI for effect size in MLM (Lai, 2021), two major limitations exist. First, in terms of effect size, previous studies mainly focused on Cohen's d instead of R^2 which has lower and upper bounds. Second, the level-1 residuals are mostly summarized by a single parameter. When doing longitudinal research, a heterogeneity and autocorrelation of within-individuals variances across time is commonly seen. Thereupon, in this study we examine different bootstrapping methods (e.g., observation based bootstrapping, parametric and non-parametric residual based bootstrapping, probability based resampling, cluster bootstrapping) along with different CI computation methods (e.g., studentized CI, percentage CI, and the bias-corrected and accelerated CI) for R^2 based effect size with more complex level-1 residual distribution using simulated data in LGMs. The selected CIs are illustrated using empirical data from the database ECLS-K: 2011 (Early Childhood Longitudinal Studies Kindergarten Class of 2010-11). R functions are provided to implement the proposed CI computation.

Effects of item specific factors on sequential/IRTree model applications

Friday, 15th July - 09:15: IRT III (Room G) - Individual Oral Presentation

Mr. Weicong Lyu [1], ***Prof. Daniel Bolt*** [1], ***Mr. Samuel Westby*** [2]

1. *University of Wisconsin - Madison,* ***2.*** *Northeastern University*

We consider test items whose scores come from sequential or IRTree processes and the strong likelihood for the presence of item specific factors across stages. Through simulation we show why such factors can yield bias in IRT model parameter estimates. This issue is examined for several applications of IRTree models, including ordered categorical items, as well as models for repeated attempts, item change, and item skipping behavior. Through both real data and simulation analyses, we show that item specific factors are commonly seen, and suggest that analyses without awareness of these factors can be misleading. Also, in the presence of item specific factors that are not statistically detectable, item analysis based on the "true" model can even perform much worse than that based on misspecified models as the magnitude of item specific factors grows, suggesting that researchers compare multiple ways of IRT modeling even with known or assumed sequential processes.

A historical perspective on polytomous unfolding Models

Friday, 15th July - 09:30: IRT III (Room G) - Individual Oral Presentation

Ms. Ye Yuan [1], ***Prof. George Engelhard*** [1]

1. *The University of Georgia*

This study provides a review and discussion of unfolding models for unidimensional polytomous data. Unfolding models (ideal point models) have a single-peaked response function. Unfolding models offer an underutilized and alternative approach to cumulative item response theory models for examining measurement data. Engelhard and Yuan (in press) described the basic principles of several key unfolding models for dichotomous responses. This study extends this work to graded responses. First, the study reviews the previous research on unfolding models. Second, we examine in detail the main polytomous unfolding models including the Generalized Hyperbolic Cosine model (Andrich, 1996), Graded Unfolding Model (Roberts & Laughlin, 1996), Generalized Graded Unfolding Model (Roberts, Donoghue, & Laughlin, 2000), and nonparametric unfolding models for multicategory data (van Schuur,1992). One of the major goals of this study is to highlight the underlying principles, formulations, measurement properties, and implementations of selected polytomous unfolding models. Next, the approaches to scaling used by Thurstone and Likert are used to highlight the implications of using cumulative versus unfolding models for attitude measurement. Finally, a summary is presented of our study. The main purpose of this study is to call attention to the use of unfolding models for polytomous responses for modeling measurement data.

Analyzing Spatial Responses: A Comparison of IRT-based Approaches

Friday, 15th July - 09:45: IRT III (Room G) - Individual Oral Presentation

Prof. Amanda Luby [1]

1. *Swarthmore College*

We investigate two approaches for analyzing responses with a spatial component using models inspired by Item Response Theory (IRT). In the first approach, we use a two-stage approach to (1) construct a pseudo-response matrix using the spatial information and then (2) apply standard IRT techniques to estimate proficiency and item parameters. Since the responses are used twice (once to construct the pseudo-response matrix and once to estimate parameters), we must account for the potential "double dipping" problem to avoid artificially small statistical uncertainty. In the second approach, we use an application-informed Bayesian hierarchical model to simultaneously account for the spatial nature of responses and estimate proficiency and item parameters. By using data from a study designed to measure how fingerprint examiners use minutiae (small details in the fingerprint that form the basis for uniqueness) to come to an identification decision, we illustrate the relative strengths and weaknesses of each approach. In this data, participants were asked to mark fingerprint minutiae on a pair of fingerprint images. The outcomes of the study demonstrate substantial participant variability, as different participants tend to focus on different areas of the image. We also find within-participant variability, since participants were sometimes given a repeat image in two different pairs and sometimes change their markup. As a further complication, the images themselves are not available due to privacy concerns. Results, as well as potential downstream impacts in the criminal justice system, will be discussed.

Random effects and extended generalized partial credit models

Friday, 15th July - 10:00: IRT III (Room G) - Individual Oral Presentation

Dr. David Hessen [1]

1. *Universiteit Utrecht*

In this presentation, it is shown that under the random effects generalized partial credit model for the measurement of a single latent variable by a set of polytomously scored items, the joint marginal probability distribution of the item scores has a closed-form expression in terms of item category location parameters, parameters that characterize the distribution of the latent variable in the subpopulation of examinees with a zero score on all items, and item scaling parameters. Due to this closed-form expression, all parameters of the random effects generalized partial credit model can be estimated using marginal maximum likelihood estimation without assuming a particular distribution of the latent variable in the population of examinees and without using numerical integration. In addition, the slightly more general extended generalized partial credit model is presented. Attention is paid to maximum likelihood estimation of the parameters of the extended generalized partial credit model and to testing the goodness of fit of the model using a generalized likelihood ratio test. Attention is also paid to person parameter estimation under the random effects generalized partial credit model. It is shown that expected a posteriori (EAP) estimates can be obtained for all possible score patterns. To show the usefulness of the proposed models, the results of a simulation study are presented.

An algorithm to detect bots in a Likert-type questionnaire

Friday, 15th July - 09:15: Aberrant Test Behaviors (Room E) - Individual Oral Presentation

Mr. Michael John Ilagan [1], ***Dr. Carl Falk*** [1]

1. *McGill University*

Administering Likert-type questionnaires to online samples risks contamination of the data by malicious computer-generated random responses, i.e., bots. Although nonresponsivity indices (NRIs) such as person-total correlation or Mahalanobis distance have shown great promise to detect bots in a target sample, universal cutoff values are elusive. Studies using NRIs typically construct an initial calibration sample of known humans and known bots, either from real data or simulated under a measurement model. A cutoff with high nominal specificity (i.e., low nominal false-positive rate) is then chosen from this calibration sample, for each NRI. However, a high-specificity cutoff is less accurate when the target sample has a high contamination rate. An accurate cutoff requires accounting for the contamination rate, which cannot be estimated in the calibration sample due to its stratified nature. Furthermore, each NRI having its own cutoff amounts to multiple testing, which impacts specificity and assumes that the NRIs are equally important. In this work, we propose the Supervised Components with Unsupervised Mixing Proportions (SCUMP) algorithm, which chooses a multivariate cutoff to maximize accuracy. SCUMP uses a multivariate Gaussian mixture model to estimate, unsupervised, the contamination rate in the sample of interest. A simulation study found that, in the absence of model misspecification on the bots, SCUMP cutoffs maintained accuracy across different contamination rates.

A statistical test for the detection of item compromise based on responses and response times

Friday, 15th July - 09:30: Aberrant Test Behaviors (Room E) - Individual Oral Presentation

Prof. Wim J. van der Linden [1], ***Dr. Dmitry Belov*** [2]

1. *University of Twente,* ***2.*** *Law School Admission Council*

A statistical test of item compromise is presented which combines the simple count of the number of test takers with correct responses with the response times recorded for them on the item. The test can be used to monitor an item in real time during online testing but also as part of post hoc forensic analysis. The null and alternative hypotheses for the test belong to a known family of distributions. Other features of the test are ease of interpretation and simple computation. Empirical examples of the test show extremely high power to detect compromise.

Using item scores and distractors to detect test speededness

Friday, 15th July - 09:45: Aberrant Test Behaviors (Room E) - Individual Oral Presentation

Ms. Kylie Gorney [1], ***Dr. James Wollack*** [1]

1. *University of Wisconsin-Madison*

Test speededness refers to a situation in which examinee performance is inadvertently affected by the time limit of the test. Because speededness has the potential to severely bias both person and item parameter estimates, it is crucial that speeded examinees are identified. In this presentation, we introduce a change point analysis (CPA) procedure for detecting test speededness. Our procedure distinguishes itself from existing CPA procedures by using a multidimensional model to incorporate both item scores and distractor selection. Results show that by attending to distractor selection, the new procedure significantly improves the detection of speeded examinees while also producing less biased estimates of the change point. Therefore, it seems there is a considerable amount of information to be gained from item distractors, which, quite notably are available in all multiple-choice data. A real data example is also included to demonstrate the utility of the new procedure in an operational setting.

Using Information-Theoretic ideas to improve the interpretation of generalized linear and nonlinear exponential family models

Friday, 15th July - 09:15: Regressions Methods and Applications (Room F) - Individual Oral Presentation

Prof. Jay Verkuilen [1]***, Ms. Sydne McCluskey*** [1]***, Prof. Irini Moustaki*** [2]

1. *CUNY Graduate Center,* ***2.*** *London School of Economics and Political Science*

Generalized linear, heteroscedastic, and nonlinear models, with and without latent variables, are very common in application and have proven essential across social and natural sciences for accurate and meaningful data analysis. However, these models can often be extremely difficult to interpret due to the fact that their parameters have no natural interpretation. Even models as widely applied as logistic regression or its generalizations such as the maximum entropy classifier/multinomial logistic regression model suffer from these issues. Recent work in categorical data analysis (e.g., Agresti, Tarantola, & Varialle, in press; Long & Mustillo, 2021) has suggested a number of improvements, focusing on local measures such as partial effects, which are simply derivatives of the expected value function with respect to a variable of interest. These measures are useful but are not standardized and may be difficult to compare across variables in the model. Drawing on insights from Rao (1973), Bjerve and Doksum (1993), and Blyth (1994), among others, and essentially adapting the geometric framework of MIRT, we show how Fisher information and closely related quantities can be used to develop local effect sizes that are standardized and can be visualized readily. In addition, we show how the measures of Agresti et al, are essentially discrete approximations. Using Bayesian estimation, it is possible to perform posterior inference with little additional difficulty past that incurred during model fitting. We provide three empirical examples based on data from a medical experiment, bioassay, and educational measurement.

Monotonic proportional odds cumulative logit regression with ordinal predictors and an ordinal response

Friday, 15th July - 09:30: Regressions Methods and Applications (Room F) - Individual Oral Presentation

Prof. Christian Hennig [1], ***Dr. Javier Espinosa Brito*** [2]

1. *University of Bologna,* ***2.*** *University of Santiago of Chile*

The proportional odds cumulative logit model (POCLM) is a standard regression model for an ordinal response. Ordinality of predictors can be incorporated by monotonicity constraints for the corresponding parameters. It is shown that estimators defined by optimization, such as maximum likelihood estimators, for an unconstrained model and for parameters in the interior set of the parameter space of a constrained model are asymptotically equivalent. This is used in order to derive asymptotic confidence regions and tests for the constrained model, involving simple modifications for finite samples. Tests concern the effect of individual variables, monotonicity, and a specified monotonicity direction. The methodology is applied on real data related to the assessment of school performance.

A global assessment of the predictive capacity of selection tests in partially observed populations

Friday, 15th July - 09:45: Regressions Methods and Applications (Room F) - Individual Oral Presentation

Dr. Eduardo Alarcón-Bustamante [1], ***Dr. Ernesto San Martín*** [1], ***Dr. Jorge González*** [1], ***Dr. David Torres Irribarra*** [1]

1. *Pontificia Universidad Católica de Chile*

The quality of a University admission test is usually assessed by determining how well it predicts the future performance of the applicants. To determine the strength of the relationship between test scores and any variable of interest (e.g., the GPA), researchers usually rely on correlations or regression analyses. Nevertheless, the applicants who take the test do not have the same characteristics (e.g., sex, socioeconomic status). Similarly, the universities that select the applicants have different characteristics (e.g., different undergraduate programs they offer). Thus, the relation between the admission test scores and the (potential) GPA, observed only in selected applicants, is unlikely to be stable across all combinations and interactions of applicant and institution characteristics. Hence, characterising the prediction that test scores have over the GPA only with a single coefficient is an insufficient solution to adequately quantify and understand the potential variation of the predictive quality of the tests.

Based on the law of total probability, we present a new way to analyse the predictive capability of a selection test, which is shown to be a function of the test scores and other covariates, rather than a constant parameter. By using Partial Identification techniques, the potential performances of non-selected applicants are incorporated in the analysis (e.g., the performance of all the non-selected applicants would have been better than the performance of the selected ones). We discuss the interpretation of the new measure and illustrate its use with a real data set involving scores from a university entrance test in Chile.

Handling low quality responses in regression analyses: A simulation study

Friday, 15th July - 10:00: Regressions Methods and Applications (Room F) - Individual Oral Presentation

Dr. Nivedita Bhaktha [1], ***Dr. Clemens Lechner*** [1]

1. *GESIS*

Low quality responses (selection of response options by respondents that do not represent their true position on the trait being measured) are ubiquitous in survey research and their extent is estimated to be around 3.5% - 60% of the collected sample. This forms a major threat to validity as low quality responses have adverse effect on statistical results and inferences. Low quality responses (LQR) can lead to spurious within-group variability, lower reliability, and potential type II errors during hypothesis testing. While there has been a surge in literature for detecting low quality responses, there are very few recommendations on how to deal with such responses. In this study we present some of the methods that can be used in dealing with LQR. We examine three approaches that can potentially help mitigate the issue of bias in regression coefficients due to LQR - exclude LQR, weight LQR, and use LQR indicators as covariates, and compare them to the do-nothing approach in a simulation study. In this simulation study, we are considering three commonly used methods of detecting LQR - maximum longstring, ipsative standard deviation, and Mahalanobis distance for a unidimensional scale with 12 items of 5 response categories, with varying factor loadings and varying levels of LQR.

Bidimensional latent regression Item Response Models for the assessment of financial knowledge in presence of don't know responses

Friday, 15th July - 10:15: Regressions Methods and Applications (Room F) - Individual Oral Presentation

Prof. David Aristei [1], ***Prof. Silvia Bacci*** [2], ***Prof. Manuela Gallo*** [1], ***Prof. Maria Iannario*** [3]

1. *University of Perugia,* ***2.*** *Università degli Studi di Firenze,* ***3.*** *University of Naples Federico II*

Increasing attention has been recently paid to assessing individuals' financial competencies. Financial knowledge appears as a complex and not directly observable phenomenon, whose measurement is usually based on answers to a set of multiple-choice items on key financial concepts, considered as essential for financial decision-making.

The option "Don't Know" (DK) is usually included among the possible answers to capture uncertainty or lack of knowledge. The presence of DK option represents an element of noise that can affect the measurement of financial knowledge. Indeed, DK responses may be due to several unknown reasons, such as poor self-confidence in one's own financial competencies or awareness of lack of knowledge. DK responses are usually considered as incorrect answers or missing values; however, these naïve approaches may lead to biased financial knowledge measures.

In the present study, we address the issue of estimating the latent knowledge construct taking into account DK option, by formulating and estimating a bi-dimensional latent regression two-parameter model. The model at issue relies on the assumption that the response process may be disentangled in two consecutive steps driven by two latent variables: propensity to answer and financial knowledge. At the first step, both latent variables affect the probability of selecting a substantive response versus DK option. At the second step, conditionally on the selection of a substantive response, financial knowledge affects the probability of a correct answer versus an incorrect one. Individual characteristics are also taken into account to explain the two latent traits.

Deriving expected SEM parameters when treating discrete data as continuous

Friday, 15th July - 10:50: SEM (Room A) - Individual Oral Presentation

Dr. Terrence Jorgensen [1], ***Dr. Andrew Johnson*** [2]

1. *University of Amsterdam,* ***2.*** *Curtin University*

I present analytical derivations of univariate and bivariate moments for numerically weighted ordinal variables, implied by the means and covariance matrix of their latent normal responses and the thresholds used to discretize responses. Fitting a SEM to those moments yields population-level SEM parameters when discrete data are treated as continuous, which is more precise and less computationally intensive than Monte Carlo simulation to calculate transformation (discretization) error. A real-data example demonstrates how this method could help inform researchers how best to treat their discrete data, and a simulation replication demonstrates the potential of this method to add value to a Monte Carlo study comparing estimators that make different assumptions about discrete data.

Accounting for uncertainty to remedy two-stage structural fit indices

Friday, 15th July - 11:05: SEM (Room A) - Individual Oral Presentation

Dr. Graham Rifenbark [1], ***Dr. Terrence Jorgensen*** [2]

1. *University of Connecticut,* ***2.*** *University of Amsterdam*

Various Structural Fit Indices (SFIs) have been proposed to evaluate the structural component of a Structural Equation Model (SEM). Decomposed SFIs (Hancock & Mueller, 2011) treat estimated latent (co)variances from an unrestricted Confirmatory Factor Analysis (CFA) as data to fit a path model, from which standard global fit indices are calculated. Conflated SFIs (Lance et al., 2016) fit a SEM with both measurement and structural components, comparing its fit to orthogonal and unrestricted CFAs. Sensitivity of conflated SFIs to the same structural misspecification depends on standardized factor loadings (McNeish & Hancock, 2018), but decomposed SFIs have inflated Type-I error rates (Rifenbark, 2019; Heene et al., 2021) due to treating estimates as data. We use Monte Carlo simulation to explore whether two alternative approaches avoid both shortcomings by separating the measurement and structural model components while accounting for uncertainty of factor-covariance estimates. First, by sampling plausible values of factor scores (Asparouhov & Muthen, 2010), SFIs can be calculated from pooled fit statistics (a numerical solution). Second, the Structural-After-Measurement (SAM; Rosseel & Loh, 2021) approach generalizes Croon's (2002) correction for structural-model estimates (an analytical solution), yielding SFIs from a "pseudo" fit statistic. We hypothesize that plausible values and SAM approaches will maintain Type-1 error rates (for LRT) and 90% confidence-interval coverage (for RMSEA) better than decomposed SFIs, and will show less sensitivity to measurement quality than conflated SFIs. We report simulation results for various magnitudes of factor loadings and numbers of indicators per factor, under populations with simple and complex structure.

Incorporating stability information into cross-sectional estimates

Friday, 15th July - 11:20: SEM (Room A) - Individual Oral Presentation

Ms. Anna Wysocki [1], ***Dr. Mijke Rhemtulla*** [1]

1. *University of California, Davis*

Psychological effects like that of education on income, conscientiousness on work outcomes, and relationship satisfaction on time spent together are inherently longitudinal effects because they involve processes that unfold over time. For a variety of reasons, psychological researchers often rely on cross-sectional data to make inferences about such longitudinal processes even though cross-sectional coefficients are often not unbiased for longitudinal effects.

In this paper, we describe a model that combines cross-sectional data with existing knowledge about a variable's stability—the correlation between instantiations of the same variable at two points in time. This model uses phantom variables within an SEM framework where previous versions of the measured variables are specified as phantom variables. The phantom variable model is identified by imposing three sets of constraints: 1) the auto-regressive paths are calculated based on pre-specified stability values and fixed; 2) stationarity is imposed (i.e., the correlations between the phantom variables are equal to the correlations between the measured variables); and 3) cross-sectional residual correlations are fixed to 0. If these assumptions are met (i.e., the process is stationary and the correct model is specified), then the model allows researchers to estimate longitudinal coefficients with data collected at a single time point. In a series of simulation studies, we explore how small and large violations of the required assumptions produces bias in the stability-informed estimates. We offer recommendations on when to use this model and how to minimize stability misspecification.

Sensitivity analysis of weighted composites in path analysis using partial least squares

Friday, 15th July - 11:35: SEM (Room A) - Individual Oral Presentation

Dr. Ke-Hai Yuan [1], ***Prof. Yong Wen*** [2], ***Prof. Jiashan Tang*** [2]

1. *University of Notre Dame,* ***2.*** *Nanjing University of Posts and Telecommunications*

Structural equation modeling (SEM) and path analysis using composite-scores are distinct classes of methods for modeling the relationship of theoretical constructs. The two classes of methods are integrated in the partial-least-squares approach to structural equation modeling (PLS-SEM), which systematically generates weighted composites and uses them to conduct path analysis of the structural model via the least-squares method. However, the goodness of PLS-SEM depends on the statistical properties of the composites, which are further determined by the formulations of the weights. This article studies how the formulations of PLS-SEM composites are affected by model specification, with focus on the sensitivity of the weights to common specification errors. Results indicate that the weights under PLS-SEM mode A are not affected by within-block error-covariances but those under mode B are. While between-block error-covariances and cross-loadings only affect the weights of the involved items under both PLS-SEM modes A and B, the weights under mode B are much more sensitive than those under mode A. In contrast, the weights under a recently proposed transformed mode (denoted as $\BA$) are a compromise between those of modes A and B. The findings not only advance the understanding of the PLS-SEM methodology but also facilitate model diagnostics. Empirical applications of the results are illustrated via the analysis of a real dataset.

Formalised models of handwriting as a nonverbal instrument and their psychometric properties

Friday, 15th July - 10:50: Applications II (Room C) - Individual Oral Presentation

Dr. Yury Chernov [1]

1. *IHS Institute for Handwriting Sciences*

Under special conditions, experts cannot use traditional psychometric instruments. That relates in particular to forensic and criminal psychology. The person under the expertise can be not available or his/her answers should not be trusted due to the obvious tendency to manipulate them. In these cases, experts apply nonverbal instruments. However, the validity of these instruments is typically weak. That is due to the subjective character of the methods. To make a method more objective and to improve its validity, it should be formalised: a formalised method allows proper statistical researches and validation studies. One of such methods is handwriting analysis.

Handwriting analysis has some advantages over questionnaire-based instruments. The major ones are that it is based on the normal person's activity and practically cannot be manipulated. In the current work, we present formalised models of handwriting and evaluate their psychometric properties. The models are based on statistical regressions and are implemented in the HSDetect computer-aided system. Due to the formalised character, the application is highly objective and transparent. The procedure has been successfully validated both against several well-known psychometric tests and against expert evaluations.

HSDetect was applied in several practical cases. In the current work, two of them are presented. The first one is a suicide investigation with two handwriting samples of the victim. The second one covers the early markers of Alzheimer's disease in handwriting, which could be important for certain forensic investigations. Psychometric properties of handwriting models allow the selection of a proper model in individual cases.

Continuation ratio model for polytomous items with a censored like latent class

Friday, 15th July - 11:05: Applications II (Room C) - Individual Oral Presentation

Dr. Diego Carrasco [1], ***Dr. David Torres Irribarra*** [1], ***Dr. Jorge González*** [1]

1. *Pontificia Universidad Católica de Chile*

Skewed responses are common across polytomous scales of self-reported bullying events within large-scale international assessment (ILSA). Typically, responses to bullying items use ordinal responses that express how often a bullying event has happened to the students (e.g., 'not at all', 'once', '2-4 times', '5 times or more'), in a specific time frame (e.g, '… in the last three months'). The distribution of responses to these types of items presents a high proportion on the least frequent category (e.g., 'not at all'). Previous research has resorted to specifying zero-inflated Poisson models onto sum-scores to account for the high prevalence of zero among participant responses (e.g., Rutkowski & Rutkowski, 2016). Although the zero-inflated models take into account the zero responses, one of the limitations of this approach is that we are left with no information regarding the item side to the instrument. Therefore, though we can produce information regarding students with more significant risks of bullying, we do not know what type of bullying events students are more at risk. In the present study, we use a continuation ratio response model, including a mixture random term to account for the expected higher counts of zero responses across items. We contrast the results with three alternative model specifications: the sum score approximation, an explanatory response model using the continuation ratio model, and the proposed model including a mixture random term. The advantages of the proposed model are layout, in particular for the monitoring of risk events of bullying on students.

The case of the cracked paintings. What cracks in paintings can tell us about their origins

Friday, 15th July - 11:20: Applications II (Room C) - Individual Oral Presentation

Prof. Pieter M. Kroonenberg [1]

1. *Leiden Univer*

Main topics Paintings acquire cracks over time, but the way in which they do varies with the surface on which they were painted, and the painting materials used. Cracks in paintings from different traditions in Europe were characterised by both expert and inexperienced judges.

Data. Small isolated segments of 40 paintings from four different areas in Europe were photographed and used for the judgements. No information about the paintings was provided. The 50cm ´ 50cm parts were judged on seven characteristics by the twenty-seven judges, creating a 40 ´ 7 ´ 27 data block.

Research questions. Can art-historical categories from different countries be distinguished via the judges' subjective scores? Do the different materials on which the paintings were made, influence the cracks in such a way that their differences can be recovered from the judgements? Do experts and inexperienced person judge alike?

Statistical techniques. Three-way data analysis, three-mode principal component analysis, discriminant analysis.

Assessing remote proctors of high-stakes tests: Improving consistency in proctoring decisions

Friday, 15th July - 11:35: Applications II (Room C) - Individual Oral Presentation

Dr. Will Belzak [1]

1. *Duolingo*

High-stakes tests are increasingly being administered online. As a result, remote proctoring is necessary to ensure test security. Remote proctors must evaluate test-taker behaviors and make decisions about rule-breaking, identity verification, and cheating, among other behaviors. Given that proctors vary in disposition and experience, and often make judgements based on imperfect information, it is important to assess how consistent proctors are in their judgements of test-taker behavior. Measuring the consistency of proctoring decisions also helps to identify outliers who are overly stringent or lenient in their application of proctoring procedures. Because remote proctoring is relatively new, there is little-to-no empirical work that evaluates the consistency of proctor judgements. To that end, this work presents a novel application of statistical analyses and quality control approaches for measuring consistency in proctoring decisions on the Duolingo English Test. Specifically, we show how nonlinear mixed effects models and resulting Bayes' estimates can be used to measure proctoring consistency across time while accounting for differences in proctors' workloads and physical locations. We demonstrate how this information is continuously served to proctoring managers and used to intervene on specific proctors through retraining. Our approach shows improved consistency in decision-making across proctors.

Getting the most out of electroretinogram (ERG) data: A methodological review and recommended statistical and machine-learning models

Friday, 15th July - 11:50: Applications II (Room C) - Individual Oral Presentation

Dr. Sunmee Kim [1], ***Dr. Miyoung Suh*** [1]

1. *University of Manitoba*

The electroretinogram (ERG) is an objective measure of electro-physiological function of retina, which is a fundamental research tool to examine the impact of diet and nutrition on retina and vision health. Although an accurate understanding of ERG data is crucial, there is no consensus on which statistical analysis framework is the "best" in the retinal studies using ERG. Few researchers referenced methodological or statistical references to justify their analytical approaches, and what is more worrying is that over-simplified statistical approaches were used in many seminal studies, severely restricting the use of the full range of ERG growth curves. Therefore, the overall aim of the present research is to develop a coherent and effective analytic approach for ERG growth curve data and apply them to real-world data to validate their usefulness. To date, no research has been done to address methodological challenges in the statistical approaches used in the studies using ERG. The present work aims to remedy this by pursuing two interlinking objectives: (1) A systematic review of the literature regarding the impact of nutrition on retinal function to inform efforts toward methodological improvements for the analysis of ERG data. (2) Development of a standardized benchmark statistical method that fully captures the functional attributes of ERG growth curve data.

Dependent latent class models

Friday, 15th July - 10:50: Classification (Room D) - Individual Oral Presentation

Mr. Jesse Bowers [1], ***Dr. Steven Culpepper*** [1]

1. *University of Illinois, Urbana-Champaign*

Latent Class Models (LCMs) are used to cluster multivariate categorical data (e.g. group participants based on survey responses). Traditional LCMs assume a property called conditional independence. This assumption can be restrictive leading to model misspecification and overparameterization. We developed a revised Bayesian model called a Dependent Latent Class Model (DLCM). This revised model permits conditional dependence. We verify identifiability of DLCMs. We also demonstrate the effectiveness of DLCMs in both simulations and real-world applications. Compared to traditional LCMs, DLCMs are effective in applications with time series, overlapping items, and structural zeroes.

MISSION: a MIxture model for Subject by Situation InteractiON in binary data

Friday, 15th July - 11:05: Classification (Room D) - Individual Oral Presentation

Dr. Jan Schepers [1], ***Dr. Alberto Cassese*** [1], ***Dr. Philippe Verduyn*** [1]

1. *Maastricht University*

This talk presents a novel two-mode clustering method for analyzing binary two-mode data. An example of such data is a subject by situation matrix including, for each subject-situation combination, a binary outcome (e.g., presence of anxiety).

The method identifies latent classes of subjects and simultaneously explains differences between situations by a small set of latent features. This talk consists of two parts. Firstly, I discuss how the proposed method is related to other clustering methods for binary two-mode data. A characteristic of the method is that identification of the subject classes is not sensitive to individual differences with respect to the 'average' profile of the log-odds across situations. Instead, the subject classes capture differential treatment sensitivity profiles. Secondly, I present the results of a simulation study as well as an illustrative application on data from an emotion-regulation study.

A new perspective on norming psychological tests.

Friday, 15th July - 11:20: Classification (Room D) - Individual Oral Presentation

Prof. Andries van der Ark [1], ***Mr. Anastasios Psychogyiopoulos*** [1], ***Dr. Niels Smits*** [1]

1. *University of Amsterdam*

In the last two decades, there have been great advancements in norming psychological tests. Regression-based norming has become the standard, and regression models have become more and more flexible, which helps to avoid violations of model assumptions and biased norms. Norming methods usually rely on regression models with a continuous response variable, defined on the real line, whereas test scores are typically discrete with a restricted range. We propose a norming method that provides a discrete and range-preserving estimate of the test-score distribution. Suppose that the test consists of J items with item scores $X_1, \ldots, X_J$. Let $X = X_1 + \ldots + X_J$ denote the test score, let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ denote the predictors for which separate norms should be constructed, and let g denote a density function. We consider the task of norming a psychological test equivalent to estimating $g(X \mid \mathbf{Z})$. From $g(X \mid \mathbf{Z})$, the desired norm may be derived (e.g., percentile ranks, stanines). To achieve a discrete and range-preserving estimate we estimate the joint density $g(X_1, \ldots, X_J, \mathbf{Z})$ using a latent class model or—if $\mathbf{Z}$ contains continuous variables—a general location model, and transform $g(X_1, \ldots, X_J, \mathbf{Z})$ to $g(X \mid \mathbf{Z})$. For a real-data set containing the scores of the SPARTS (a test measuring teacher-student relationships) and several covariates, and for simulated data, we compare the results of the proposed norming method to regression-based norming with GAMLSS, and traditional norming

Mixture multigroup SEM for comparing structural relations among many groups

Friday, 15th July - 11:35: Classification (Room D) - Individual Oral Presentation

Mr. Andres Felipe Perez Alonso [1], ***Prof. Yves Rosseel*** [2], ***Prof. Jeroen Vermunt*** [1], ***Dr. Kim De Roover*** [1]

1. *Tilburg University,* ***2.*** *Ghent University*

Social scientists often examine the relationships between two or more latent variables or constructs, and Structural Equation Modeling (SEM) is the state-of-the-art for doing so. When comparing these structural relations among many groups, they likely differ across the groups. However, it is equally likely that some groups share the same relations, and that clusters of groups emerge in terms of the relations between the latent variables. For validly comparing the latent variables' relations among groups, the measurement of the latent variables should be invariant across the groups (i.e., measurement invariance), whereas often at least some measurement parameters differ across the many groups. Restricting these measurement parameters to be equal across groups, when they are not, causes the structural relations to be estimated incorrectly and thus invalidates their comparison. Therefore, to capture differences and similarities in structural relations while accounting for the reality of measurement non-invariance, we propose mixture multigroup SEM (MixMG-SEM). MixMG-SEM obtains a clustering of groups focused entirely on the structural relations by making them cluster-specific, while allowing for the measurement parameters to be (partially) group-specific. In this way, MixMG-SEM disentangles differences in structural relations from differences in measurement parameters. We present an expectation-maximization estimation procedure, built around the R-package 'lavaan', as well as an evaluation of MixMG-SEM's performance in terms of recovering the group-clustering and the group- and cluster-specific parameters.

When almost all items are endorsed: Extreme responses or substantive classes?

Friday, 15th July - 11:50: Classification (Room D) - Individual Oral Presentation

Ms. Rosario Escribano [1], ***Dr. David Torres Irribarra*** [1], ***Dr. Diego Carrasco*** [1], ***Mr. Fernando Ponce*** [1]

1. *Pontificia Universidad Católica de Chile*

Latent class solutions of survey items often include high and low endorsers classes across all items. However, it is possible to find classes that represent response patterns broadly consistent with an artifact due to the presence of extreme response patterns. If the high (or low) endorsers class comprises persons displaying extreme response styles (ERS), this latent class would likely not have substantive meaning.

In contrast, these unobserved groups deserve substantive interpretation if high (or low) endorsement classes are composed of persons that purposefully respond according to their thoughts, actions, and opinions.

We study this issue using data from the ICCS 2016, which includes responses from students in 24 countries in Europe, Latin America, and Asia. In particular, we inquiry the responses of 8th-grade students to citizenship norms (CN). CN survey the students' endorsement of different civic duties, such as obeying the law, participating in elections, and helping others. Previous literature has used mixture models (Hooghe & Oser, 2015; Hooghe, Oser & Marien, 2016, Torres Irribarra & Carrasco, 2021) to compare CN across countries. The latent class solution from the previous literature includes a class where students endorse almost all norms. We assess, with different strategies, if ERS can account for this latter class. We use generated indexes of ERS, targeting response patterns by rules, and confirmatory mixture models that include extreme response classes. In the present study, we discussed the trade-offs of each strategy, and we addressed if the class with high endorsement should be interpreted as substantially relevant.

Effect of direction of DIF and group ability differences on multidimensional equating

Friday, 15th July - 10:50: Test Equating II (Room G) - Individual Oral Presentation

Dr. Secil Ugurlu [1], ***Dr. Won-Chan Lee*** [2]

1. *Hacettepe University,* ***2.*** *University of Iowa*

Many educational and psychological tests are in multidimensional structure and contain items with Differential Item Functioning (DIF). Not much is known about the relationship between DIF and equating from the multidimensional perspective. Also, the effect of group ability differences on multidimensional equating has not been much studied. The current study is intended to reveal the relationship between bi-directional DIF and equating in the Multidimensional Item Response Theory (MIRT) framework. More specifically, this study aims to investigate, via a simulation study, the performance of the Simple-Structure MIRT Observed-Score (SMO) equating method with respect to the population invariance under various conditions of DIF and group mean ability differences.

Sample size calculation and optimal design for regression-based test norming

Friday, 15th July - 11:05: Test Equating II (Room G) - Individual Oral Presentation

Dr. Francesco Innocenti [1], ***Dr. Frans Tan*** [1], ***Dr. Math Candel*** [1], ***Prof. Gerard van Breukelen*** [1]

1. *Maastricht University*

Normative studies are needed to derive reference values or norms for tests and questionnaires, so that psychologists can use them to assess individuals. Since norms are used to make decisions on individuals, such as the assignment to clinical treatment or remedial teaching, it is important that norms are not strongly affected by sampling error in the sample on which the norms are based. This goal can be attained by drawing a large sample and using regression-based norming, which is more efficient than the traditional approach of splitting the sample into subgroups based on demographic factors and deriving norms per subgroup. Under this norming approach, a procedure for sample size planning to make inference on Z-scores and percentile rank scores is proposed. Sampling variance formulas for these norm statistics are derived and used to obtain the optimal design (i.e. the optimal joint distribution of the predictors in the sample that minimizes the sampling error and thus maximizes the precision of the norms) under the assumed regression model for norming. This is done under five regression models with a quantitative and a categorical predictor, differing in whether they allow for interaction and nonlinearity. To deal with uncertainty about the best model for norming at the design stage, efficient designs that are robust against misspecification of the model are presented. Furthermore, formulas are provided to compute the required sample size, given an optimal design, such that subjects' positions relative to the derived norms can be assessed with prespecified power and precision.

Evolutionary IRT scale maintenance via concurrent calibration designs

Friday, 15th July - 11:20: Test Equating II (Room G) - Individual Oral Presentation

Dr. Richard Luecht [1]

1. *University of North Carolina at Greensboro*

This study proposes an IRT concurrent calibration framework for robust scale maintenance. Rather than focusing on common items for data linking and either running anchored (constrained) calibrations or statistically adjusting locally calibrated item parameter estimates to place them on an item bank metric, this study advocates for allowing the underlying scale to evolve over time to effectively absorb on-going minor changes. A concurrent calibration design: (i) improves the stability of the calibration by maximizing the sample sizes for each item; (ii) allows for optimal fitting of the item parameters for an IRT model of choice to the data—similar to a "local calibration"; and (iii) provides for common-persons linking to move cut scores or trend-related statistics forward or backward. This common-persons linking paradigm can be shown to have minimum error variance of equating and allow for linear or equipercentile equating functions to be applied in an IRT context. A large-scale, multi-year simulation study is used to compare this evolutionary scale maintenance and linking strategy with more traditional non-equivalent groups designs such as anchored calibrations and IRT linear equating using common items to estimate the linking constants.

Does it matter which test scores are used when equating test scores?

Friday, 15th July - 11:35: Test Equating II (Room G) - Individual Oral Presentation

Prof. Marie Wiberg [1], ***Prof. James Ramsay*** [2], ***Dr. Juan Li*** [3]

1. *Umeå University,* ***2.*** *McGill University,* ***3.*** *Ottawa Hospital Research Institute*

An achievement test is often given as different test versions to avoid plagiarism. The results from an achievement test are typically given by a sum score but can also be given as the recently proposed optimal score. Regardless of which score is used, different test score equating methods are used to put the scores on the same scale. The overall aim was to compare different equating methods when both sum scores and optimal scores are used. We focus on the equivalent group design and the nonequivalent groups with anchor test design. The use of different test scores with different methods are illustrated with real test data from a college admission test. Practical implications of the results are discussed.

“Proportions-of-total” ipsative data: Ratio or ordinal?

Friday, 15th July - 10:50: IRT IV (Room E) - Individual Oral Presentation

Dr. Anna Brown [1]

1. *University of Kent*

Ipsative (or relative-to-self) questionnaires ask respondents to compare sets of two or more stimuli from the same domain, such as behaviors, values or interests. Preferences can be expressed as rank orders (e.g. A>C>B), or graded in terms of strength (e.g. ‘prefer a little’ A to C, or ‘prefer a lot’ C to B), or expressed as proportions of a fixed total (e.g. A=50%, B=40%, C=10%), yielding binary, ordinal or ratio outcomes, respectively. Thurstonian IRT (Brown & Maydeu-Olivares, 2011; 2018) and Thurstonian linear factor models (Brown, 2016) were developed to estimate item and person parameters for these different types of ipsative data.

Interesting borderline cases emerge when a relatively small number of points (for example, 10) is distributed between several stimuli. Psychologically, do respondents follow the ratio judgements (such as “I feel A is *twice* more applicable to me than B”) or the graded preference judgements (such as “I feel A is *much more* like me than B”)? And psychometrically, when deciding between using ordinal or ratio models for such ipsative data, what number of points is too ‘coarse’ for comfort? A comparison of two studies using the same Big Five measure consisting of blocks of 3 statements – one with N=326 respondents distributing 15 points within blocks and another with N=256 respondents distributing 7 points – will be used to illustrate the merits of linear and ordinal factor modelling. Qualitative evidence will be used to judge the psychological nature of response process, and recommendations for practice will be made.

Culture-specific faking: Adverse impact on forced-choice personality scores

Friday, 15th July - 11:05: IRT IV (Room E) - Individual Oral Presentation

Dr. HyeSun Lee [1], ***Dr. Weldon Smith*** [1]

1. *California State University Channel Islands*

Forced-choice tests have been suggested as an alternative to Likert-scale measures for personnel selection due to their robustness to faking and response styles. The current study compared degrees of faking occurring in Likert-scale and forced-choice five-factor personality tests between South Korea and the United States. Also, it was examined whether the forced-choice format was effective at reducing faking in both countries. Data were collected from 396 incumbents participating in both honest and applicant conditions. Standardized mean differences (SMDs) between the two conditions were utilized to measure magnitudes of faking occurring in each format and country. In both countries, the degrees of faking occurring in the Likert-scale were larger than those from the forced-choice format, and the magnitudes of faking across five personality traits were larger in South Korea by from 0.07 to 0.12 in SMDs. The forced-choice format appeared to successfully reduce faking for both countries as the average SMDs decreased by 0.06 in both countries. However, the patterns of faking occurring in the forced-choice format varied between the two countries. In South Korea, degrees of faking in Openness and Conscientiousness increased compared to those from the Likert-scale format, whereas those in Extroversion and Agreeableness were substantially decreased. This presentation will discuss potential factors leading to trait-specific faking under the forced-choice format in relation to cultural influence on the perception of personality traits and score estimation in Thurstonian IRT models, especially focused on the loss of information due to response dichotomization in comparative judgement processes.

Peabody quadruplets for forced-choice items

Friday, 15th July - 11:20: IRT IV (Room E) - Individual Oral Presentation

Dr. Felipe Valentini [1], ***Dr. Nelson Hauck*** [1], ***Prof. Rafael Valdece Sousa Bastos*** [1], ***Dr. Ricardo Primi*** [1]

1. *University São Francisco, Brazil*

Assessments using forced-choice items have been receiving increasing popularity, because they avoid the response biases that commonly affect Likert-type items. An important exception might be faking. Even if using forced-choice, examinees can pick a more desirable item within a block, regardless of the factor content. Furthermore, recent research suggests the forced-choice method is not free from social desirability. To overcome this problem, we propose combining the Peabody quadruplets technique with the forced-choice. Peabody quadruplets consist of a set of items that are intentionally balanced for content and value. For instance, a quadruplet for extroversion would include four items: extroversion-desirable, extroversion-undesirable, introversion-desirable, introversion-undesirable. To test the feasibility of this approach, we simulate 200 datasets for each of 12 conditions varying: (a) number of quadruples (6 or 12); (b) variance of social desirability (.1 and .25); (c) model specification (bias control with constrained parameters; control without constraints; no control for bias). Results evidenced consistent estimations for models with parameters constrained (average bias < .01), regardless of the number of quadruplets and the variance of social desirability. Models without constraint were less consistent (average bias between .08 and .19), but still adequate if the social desirability is small (average bias between .02 and .06). However, models lacking control of social desirability always yielded inadequate parameter estimates (average bias from .14 to .48). The approach here studied is promising, albeit future simulations should yet explore different design settings varying the number of factors, balance of items within blocks, and violations of the model's assumptions.

Measuring susceptibility with multidimensional zero-inflated and hurdle graded response models

Friday, 15th July - 11:35: IRT IV (Room E) - Individual Oral Presentation

Dr. Brooke Magnus [1], ***Dr. Mauricio Garnier-Villarreal*** [2]

1. *Boston College,* ***2.*** *Vrije Universiteit Amsterdam*

Much of applied IRT research suggests that clinical instruments are able to measure individual differences only at “pathological” or severe levels of the corresponding trait (Reise & Waller, 2000). When the graded response model (GRM) is fit to data from these types of measures, threshold parameters tend to fall within a high and narrow range along theta, suggesting that many clinical instruments are unable to capture individual differences among milder levels of the (psycho)pathology. Preliminary methodological research suggests that this limited range of measurement may be an artifact of zero inflation within the population, where many respondents are low on susceptibility (Magnus & Garnier-Villarreal, 2021). The multidimensional zero-inflated and hurdle GRMs (MZI-GRM and MH-GRM, respectively) simultaneously model susceptibility to and severity of a construct. In empirical applications, it has been shown that a wider range of individual differences can be accounted for once susceptibility is included in the GRM – more specifically, individual differences at milder levels of the trait. While these models have shown some promise in data applications, their utility needs to be established more rigorously through simulation. Under what data-generating conditions do models that account for zero inflation exhibit superior fit compared to the traditional GRM? Of particular interest is whether the threshold parameters estimated from fitting the GRM to data that are generated from zero-inflated and hurdle models incorrectly suggest that items only capture individual differences at severe levels of the trait, a claim frequently made in the clinical measurement literature.

Application of a new multilevel item response theory model with a latent interaction effect

Friday, 15th July - 11:50: IRT IV (Room E) - Individual Oral Presentation

Dr. Tim Fabian Schaffland [1], ***Prof. Augustin Kelava*** [1], ***Dr. Stefano Noventa*** [1]

1. *University of Tuebingen*

The size and therefore the complexity of collected datasets have been growing over time as computational capacities increase. Therefore, estimation methods that can take this complexity into account are needed.
One possible complication of the datasets is hierarchical structures when data is collected from different backgrounds (e.g., schools, local branches of companies, or countries). Another source of complexity are nonlinear relations of latent variables like interactions, as e.g., in the Expectancy-value theory.
A widely applied method is Item Response Theory models, especially in an educational context (e.g., large scale assessments). Implementations, however, often only include classical models (e.g., 2PL, Partial Credit), while some also include hierarchical structures. Interactions or quadratic effects of latent variables are typically not taken into consideration.
In this talk, a Multilevel IRT model with Nonlinear Latent variable Effects Model (MINoLEM) is presented, which combines hierarchical aspects and nonlinear latent influences. An estimation procedure based on the Expectation-Maximization algorithm is deduced. The accuracy of this estimation approach will be proven in a simulation study and its usefulness will be shown through comparisons to other IRT software with the potential to include multilevel structures or nonlinear latent variable effects.

Improving Psychological Explanations

Friday, 15th July - 10:50: Statistics and Psychometrics (Room F) - Individual Oral Presentation

Dr. Noah van Dongen [1]

1. *University of Amsterdam*

In current practice, psychological explanations typically present a narrative in which a theory renders a putative empirical phenomenon intuitively likely. However, whether the theory actually implies the phenomenon in question is also left to this intuition. We want to remedy this by proposing an account of productive explanation, in which the theory specifies a formal model that produces statistical patterns that reflect empirical phenomena that are purportedly explained by the theory. We propose a workable methodology for establishing empirical implications. This productive explanation methodology involves a) translating a verbal theory into a set of model equations, b) representing empirical phenomena as statistical patterns in putative data, c) assessing whether the formal model actually produces the targeted phenomenon. In addition, we explicate a number of important criteria for evaluating the goodness of this explanatory relation between theory and empirical phenomenon.

Considerations in group differences in missing values

Friday, 15th July - 11:05: Statistics and Psychometrics (Room F) - Individual Oral Presentation

Ms. Ambar Kleinbort [1], ***Dr. Anne Thissen-Roe*** [1], ***Mr. Rohan Chakraborty*** [1], ***Dr. Janelle Szary*** [1]

1. *pymetrics*

In recent years, employers are increasingly using artificial intelligence (AI) systems to summarize multidimensional psychological measurements to support employee selection decisions. It is important to evaluate and maximize the fairness of these AI models with regard to demographic groups such as gender and ethnicity. Different approaches have emerged, including obscuring group labels and maximizing the models' classification parity, neither of which guarantees a reduction in bias in operational settings (Corbett-Davies & Goel). The issue of fairness becomes more complex when missing measurements are imputed in the data used to train a model. The encoding of group differences can vary from the imputed data used for training, to complete real-world data. We tested how this can lead to unexpected observations of bias for the final model in production. To do this, we built imputers that weight the loss for each group equally, and a paired, non-debiased version for each of them. We then built models on data imputed with each pair and tested their fairness with complete datasets labeled by groups (gender and ethnicity). We found that reducing the group differences encoded in imputed training data did not guarantee a more fair AI scoring model, and in some circumstances, it may result in a less fair model. However, we were pleasantly surprised to see that we did not underestimate group differences in scores when we evaluated them on a dataset that was partially imputed using a reduced group difference method.

The Clique Percolation performance to detect cross-loadings across latent factors

Friday, 15th July - 11:20: Statistics and Psychometrics (Room F) - Individual Oral Presentation

Dr. Pedro Henrique Ribeiro Santiago [1], ***Dr. Gustavo Hermes Soares*** [1], ***Dr. Adrian Quintero*** [2], ***Prof. Lisa Jamieson*** [1]

1. *The University of Adelaide,* ***2.*** *Icfes - Instituto Colombiano para la Evaluación de la Educación*

Community detection algorithms (CDA) applied in network/graphical models (e.g.Walktrap algorithm) perform as well as the most traditional factor analytical methods in identifying the correct number of latent factors (Golino et al., 2020). However, one limitation is that these CDAs do not account for cross-loadings and can only assign each node (item) to a single community (latent factor). The Clique Percolation (CP) is a CDA that allows for the identification of nodes belonging to multiple communities and has been recently used in psychological research (Kaiser, Herzog, Voderholzer, & Brakemeier, 2021; Lange & Zickfeld, 2021). However, the performance of the CP in partial correlation networks has not been investigated. This simulation aims to investigate the performance of the CP algorithm to detect items with cross-loadings across latent factors (which cannot be identified based on commonly used CDAs). We investigated 48 conditions (3x2x2x4) across 100 simulated samples: (1) sample sizes of 300, 500 and 1000; (2) 4 or 8 items per factor; (3) 6.25%, 12.5%, and 25% of items with 0.40 cross-loadings; and (4) factor correlations of 0.0, 0.2, 0.5 and 0.7. Performance was evaluated through the percentage of correct items assigned per factor (PCIF), percentage of items with cross-loadings identified (PICI), among other metrics. The simulations indicated high PCIF (M=96.0%) and PICI (M=75.6%) across the simulated conditions. Additionally, the Walktrap algorithm tended to underestimate the number of factors in certain conditions and the CP better estimated the number of factors. This simulation study provides guidelines for the CP use in psychological data.

Incorporating information from historical data for power analysis: A hybrid classical-Bayesian approach for multilevel studies

Friday, 15th July - 11:35: Statistics and Psychometrics (Room F) - Individual Oral Presentation

Ms. Winnie Wing-Yee Tse [1], ***Dr. Mark Hok Chio Lai*** [1]

1. *University of southern california*

Power analysis for planning multilevel studies requires knowledge of the true values of multiple parameters, such as the effect size and intraclass correlation. As the true parameter values are by definition unknown at the study design phase, the conventional approach is to use researchers' best-educated guess based on historical data, such as a meta-analysis study. Due to sampling variability and heterogeneity in the study designs, parameter estimates from historical data entail uncertainty. As shown in the present research, ignoring uncertainty in parameters often result in underpowered future studies and at times leads to an unreasonably large sample size requisite.

We propose a hybrid classical-Bayesian (HCB) approach, which performs power analysis in the Bayesian framework, assuming that classical analyses will follow in the new study. The proposed approach addresses uncertainty from historical data by incorporating the distributions of parameter values, instead of a single best guess, for power analysis. In a simulation study, we showed that the HCB approach adequately adjusts for uncertainty and better controls the expected power and assurance level of achieving at least 80% power than the conventional approach across conditions. We have implemented the HCB approach in an R package *hcbr* and an R Shiny web application and will provide a tutorial on how to utilize our software for addressing uncertainty and designing multilevel studies. We will discuss how to incorporate information from historical studies (e.g., a pilot study or a meta-analysis study) to power analysis and illustrate it in two empirical examples.

Exploratory factor analysis trees: Detecting measurement invariance between multiple covariates

Friday, 15th July - 11:50: Statistics and Psychometrics (Room F) - Individual Oral Presentation

Mr. Philipp Sterner [1], ***Prof. David Goretzko*** [2]

1. *LMU Munich,* ***2.*** *LMU Munich; University of Leipzig*

Measurement invariance (MI) describes the equivalence of a construct across groups. To be able to meaningfully compare latent factor means between groups, it is thus crucial to establish MI. Although methods exist that test for MI, these methods do not perform well when many covariates have to be compared or when there are no prior hypotheses about them. We suggest a method called *Exploratory Factor Analysis Trees* (EFA trees) that are an extension to *SEM trees* introduced by Brandmaier et al. (2013). EFA trees combine EFA with a recursive partitioning algorithm that can uncover subgroups within a dataset with regard to different levels of MI in a data-driven manner. An EFA is fit in each node of a decision tree and then tested for parameter instability on multiple potential split variables (e.g., age, gender, education, etc.). Multiple categorical and continuous split variables (and interactions between them) can be simultaneously tested for violations of MI. Our goal is to suggest a method with which MI can be addressed in the earliest stages of questionnaire development, where changes to the item pool are still easily possible. We show how EFA trees can be implemented in the statistical software *R* using the *R* packages *lavaan* (Rosseel, 2012) and *partykit* (Hothorn & Zeileis, 2015). In a simulation design, we demonstrate the ability of EFA trees to detect violations of MI under various conditions. Limitations and future potential developments of EFA trees are discussed.

Boosting methods for latent variable models

Friday, 15th July - 12:10: Invited Speaker: Michela Battauz (Room B) - Individual Oral Presentation

Prof. Michela Battauz [1]

1. *University of Udine*

Boosting originated as a machine learning procedure developed for classification purposes. In the statistical approach to boosting, it was then viewed as a method for fitting a statistical model by sequentially minimizing an objective loss function. Early stopping criteria are fundamental to obtain regularized estimates and to prevent overfitting. The component-wise version of the algorithm also performs variable selection since, at each step, only the coefficients of the covariate that most improves the fit are updated. However, the non-convexity of the likelihood function of latent variable models poses new challenges. Focusing on factor analysis models for binary data, we propose a new algorithm that exploits the directions of negative curvature. To reduce the computational burden, a pairwise likelihood was chosen as an objective function, and a group lasso penalty was included in order to automatically select the number of latent variables in the procedure. Starting with a model that includes only the thresholds, at each step only two coefficients are updated after selecting either a Newton-type direction or a negative curvature direction. The solution attained tends to be sparse, thus facilitating interpretation without requiring a posterior rotation of the factor loadings.

CDMs that optimize the diagnostic value of multiple-choice data: Real data applications

Friday, 15th July - 12:10: Spotlight Talk: Jimmy de la Torre (Room A) - Individual Oral Presentation

Prof. Jimmy de la Torre [1], ***Dr. Xue-Lan Qiu*** [1], ***Dr. Hartono Tjoe*** [2]

1. *The University of Hong Kong,* ***2.*** *The Pennsylvania State University*

Multiple-choice (MC) items are widely used in educational assessment because of their ease of administration and scoring. To optimize the diagnostic value of MC data, recent research involving cognitive diagnosis models (CDMs) has focused on harnessing diagnostic information that can be found in distractors. One such CDM is the MC deterministic input, noisy "and" gate (MC-DINA) model, where distractors are coded to probe how examinees missing some of the required attributes respond. Although promising, the MC-DINA model has some important limitations—aside from not accommodating misconceptions, it strictly assumes that the knowledge states represented by the distractors are a subset of those of the correct response. To address these limitations, the extended MC-DINA (eMC-DINA) model, which allows for a more flexible coding of the distractors and can simultaneously accommodate skills and misconceptions, has been proposed. Previous studies have shown that the eMC-DINA model yields higher correct classification rates than the MC-DINA model. However, to date, neither the eMC-DINA model nor the MC-DINA model has been applied to real data. This study illustrates the application of the two models using a proportional reasoning (PR) test, where six skills and three misconceptions are represented in the MC options. The data were collected from over 1,400 secondary students in Hong Kong. Results from fitting the eMC-DINA and MC-DINA models to the PR data will be presented, and discrepancies in the model-data fit and attribute classifications will be highlighted. The implications of this study for test design and item writing will be discussed.

Detection of item preknowledge in educational testing: Latent variable models, sequential change detection and compound decision

Friday, 15th July - 14:10: Early Career Award (Room B) - Individual Oral Presentation

Dr. Yunxiao Chen [1]

1. *London School of Economics and Political Science*

In standardised educational testing, items are repeatedly used. Some items may get leaked after exposure in a few test administrations, and some test takers obtain access to the leaked items and gain an advantage in future tests. In this talk, we propose statistical models and methods for detecting item preknowledge in educational tests. We consider two different settings: (1) The detection of leaked items and test-takers with preknowledge based on item responses and response times from a single test, and (2) the online detection of leaked items based on sequentially collected data. We view the first problem as a two-way outlier detection problem for multivariate data and propose a latent variable model and associated compound decision theory to detect the two-way outliers. We view the second problem as a multi-stream sequential change detection problem and propose a compound decision theory to detect changed streams quickly. The proposed methods show superior performance under real and simulated settings.

Estimation methods for simple and complex psychometric models

Friday, 15th July - 15:45: Presidential Address (Room B) - Individual Oral Presentation

Prof. Irini Moustaki [1]

1. *London School of Economics and Political Science*

In this talk, I will review estimation methods and their rationale for simple and complex psychometric models. The complexity of psychometric models depends on the number of observed variables, number of latent variables but also model features reflected in the model parameters (multidimensional 2PL, 3PL, 4PL, non-linear factor analysis, heteroscedastic errors, two-way classification, missing values, etc.). Developments in computing have allowed the evolution of estimation and computational methods. Estimation methods and computational algorithms come with pros and cons. The more recent estimation frameworks such as regularized estimation but also Bayesian estimation have allowed to implicitly test of complex hypothesis of measurement invariance and model selection (e.g. zero factor loadings). Similarly composite likelihood estimation methods have provided a unified framework for estimating and testing models with intractable likelihoods. The talk will draw connections among the various methods and focus on their merits and drawbacks in data analysis.

Authors Index