Abstract Book: Talks

Table of Contents

New science standards, new(ish) psychometrics

Tuesday, 25th July - 10:15: Symposium: New Science Standards, New(ish) Psychometrics (Colony Ballroom) - Symposium Overview

Dr. frank rijmen (Cambium Assessment)

The Next Generation of Science Standards (NGSS) assessments are designed to measure three distinct and equally important dimensions to learning science. Science assessments developed under the NGSS framework rely on multiple interaction items centered around a common scientific phenomenon. The use of these 'item clusters' poses challenges on methods developed for traditional unidimensional assessments. This symposium demonstrates the complications and proposed solutions for NGSS assessments. This session starts with an overview of the NGSS, the innovative item types and how items are scored, and the collaborations among multiple states. Then, four presentations will elaborate on how important challenges in NGSS assessments are addressed from an operational perspective.

The first presentation demonstrates the test design and calibration model. A multigroup Rasch testlet model is used for item calibration. To assess item fit in such a complex assessment, the second presentation discusses the use of the marginal item characteristic curve and the use of a fit index at the item cluster level. The NGSS assessments are adaptive and rely on a calibrated item bank. The third presentation introduces a modified item drift statistic to monitor the adequacy of the item parameters. New person fit z-statistics were developed and used in the NGSS assessments, which will be the focus of the fourth presentation. The fifth presentation proposed a standard setting method that preserves the integrity of the item clusters.

This symposium demonstrates how advanced psychometric methods can be implemented in large-scale assessments.

Test design and calibration model for three-dimensional science assessments

Tuesday, 25th July - 10:20: Symposium: New Science Standards, New(ish) Psychometrics (Colony Ballroom) - Symposium Presentation

Dr. frank rijmen (Cambium Assessment)

Assessments developed under the NGSS framework consist of both traditional stand-alone (SA) items and item clusters, and heavily rely on the latter. In an item cluster, a student is presented with a real-world scenario related to a single performance expectation and is asked to interact with the presented stimulus material in various ways. For each item cluster, a series of explicit assertions can be made about the knowledge and skills that a student has demonstrated based on specific features of the student's responses. Scoring assertions can be supported based on students' responses in one or more interactions within an item cluster. These assertions are binary (TRUE/FALSE) indicator variables and are the basic units of analysis in NGSS assessments.
In unidimensional IRT models, it is assumed that all correlations between item responses are accounted for by a single proficiency variable. The assumption of conditional independence does not hold for NGSS assessments. Our results indicate that cluster effects are too substantial to be ignored. Instead of using a unidimensional IRT model for scale development, we are using a bifactor model. Specifically, the Rasch testlet model is used.
Items are field-tested in several states and item parameters are calibrated concurrently for each grade band. Overall differences across states are considered by specifying a state-specific prior distribution for the latent variable representing overall science proficiency, allowing for a concurrent calibration of all items. This is, to our knowledge, by far the largest application of the bifactor model in an operational context.

Assessing item fit for the Rasch testlet model

Tuesday, 25th July - 10:34: Symposium: New Science Standards, New(ish) Psychometrics (Colony Ballroom) - Symposium Presentation

Dr. Dandan Liao (McKinsey & Company), Dr. frank rijmen (Cambium Assessment)

In this presentation, we present results on different approaches to assessing item fit in the Rasch testlet model. First, we discuss the use of the marginal item characteristic curve (ICC) to assess model fit at the assertion level. The expected marginal item characteristic curve is obtained by computing, for a given value of the overall science proficiency, the marginal probability of a correct response. In addition to visual examination, several statistical tests are generalized to accommodate the structure of the operational calibration model, which is the multigroup Rasch testlet model. The first method examines residual local dependence between assertion pairs (e.g., Chen & Thissen, 1997; Liu & Maydeu-Olivares, 2012). Specifically, we present results on the standardized X^2 to detect assertion pairs with high residual local dependency. The second method generalizes the R_2 statistics proposed by Glas (1988) to assess model fit at the item level (i.e., a group of assertions). This method detects violations of model assumptions regarding the slopes, as well as the residual variances unexplained by the latent variables in the model. By generalizing these methods to the multigroup Rasch testlet model, this study demonstrates how item fit can be numerically assessed and statistically tested in the multidimensional IRT scenario. Additionally, this study shows how these methods can provide information about item fit at different granularity: individual assertion level, assertion pair, and item level. Results obtained from the methods of interest are compared to explore which aspect of item misfit each one can capture.

Evaluation of person fit

Tuesday, 25th July - 10:48: Symposium: New Science Standards, New(ish) Psychometrics (Colony Ballroom) - Symposium Presentation

Dr. Zhongtian Lin (Workera)

This presentation focuses on new statistics we developed to evaluate person fit in the NGSS assessments.
A well-known person fit statistic in the item response theory (IRT) literature is the l_z statistic (Drasgow et al., 1985). Snijders (2001) derived l_z*, which is the asymptotically correct version of l_z, when the ability parameter is estimated. However, both statistics only concern unidimensional IRT models. Considering a maximum likelihood ability estimator, this presentation proposes l_{zg} and l_{zg}*, which are the generalized versions of l_z and l_z*, respectively, and can be used with the Rasch testlet model and beyond. The computation of the proposed statistics relies on several extensions of the Lord-Wingersky algorithm (1984) that are additional contributions of this study. Simulation results show that l_{zg}* has close-to-nominal Type I error rates and satisfactory power for detecting aberrant responses. For unidimensional models, l_{zg} and l_{zg}* reduce to l_z and l_z* , respectively, and therefore allows for the evaluation of person fit with a wider range of IRT models. A real data application is also presented to show how we use the proposed statistics in the NGSS assessments.

Item drift for item clusters

Tuesday, 25th July - 11:02: Symposium: New Science Standards, New(ish) Psychometrics (Colony Ballroom) - Symposium Presentation

Dr. Mengyao Cui (Cambium Assessment)

In this presentation, we present a modified item drift statistic, which is used to monitor the item parameters over time. Item drift occurs when items become relatively more difficult or easier over time. For example, if the curricular emphasis on the relevant subject matter is changing over time. By generalizing methods for detecting item drift to the application of the multigroup Rasch testlet model, this presentation demonstrates how item drift can be examined in the multidimensional IRT scenario.

As assertions serve as the basic unit in the multidimensional IRT calibration, the residual score is computed for each student that took the item as the difference between the observed score and the expected score given a student's ability. Then the assertion-level residuals are summed up within an item cluster. Residuals for each item cluster are averaged across students to flag an item. In computing the residual, the item parameter drift criterion is added to or subtracted from the item difficulty parameters in the item pool. We could assess if an item became harder or easier by examining the upper and lower bounds of the residuals. A simulation study was designed to evaluate the performance of different drift criteria, aiming to search for the optimal drift criterion for item clusters.

The Assertion Mapping Procedure for setting performance standards

Tuesday, 25th July - 11:16: Symposium: New Science Standards, New(ish) Psychometrics (Colony Ballroom) - Symposium Presentation

Dr. Yi-Fang Wu (Test)

The use of item clusters and the way they are scored is a complication for standard setting. Motivated by the Bookmark method (Lewis, Mitzel, & Green, 1996), the single passaged-based method (Skaggs, Hein, & Awuor, 2007), and the Item-Descriptor matching method (Ferrara and Lewis, 2012), we propose the Assertion Mapping Procedure (AMP) to establish performance standards for NGSS assessments. The AMP preserves the integrity of the item clusters and works for multiple performance levels. For each item, standard setting panelists map the ordered scoring assertions to the performance level descriptors which specify the degree to which students demonstrate mastery of the NGSS. A classification method is used to identify the cut points from the set of mapped assertions and their locations on the scale. Cut point are chosen so that the weighted number of misclassifications is minimized.

Mixed integer programming methods can be used to select a set of items that fulfill the test blueprint (e.g., van der Linden, 2005). For standard setting (e.g., the AMP), however, an additional requirement is that the distribution of impact values across all selected items does not show any substantial gaps. Fulfilling this requirement is complicated by the fact that each item cluster has multiple scoring assertions that are selected as a group. We show how this requirement can be formulated as a maximum coverage problem. A maximum coverage problem can also be solved using the mixed-integer programming, and so a comprehensive method is available to select a set of items for standard setting.

An ordinal diagnostic model for inferring item-level and category-specific structure

Tuesday, 25th July - 10:15: Diagnostic Classification Models (Atrium) - Oral

Mr. Auburn Jimenez (University of Illinois Urbana-Champaign), Prof. Steven Culpepper (University of Illinois Urbana-Champaign)

Cognitive diagnostic models (CDMs) are classification-based psychometric models designed to provide diagnostic information about a student's mastery on a set of domain-specific latent attributes, or skills. A fundamental component of CDMs is the Q-matrix which outlines the relationship between the latent attributes and a set of cognitive assessment items. In scenarios where Q is unavailable, exploratory CDM frameworks can be used to estimate the latent structure, providing useful information about the underlying response mechanisms characterizing answer-choice behavior. In various settings, polytomous items are developed and scored with the understanding that an observed response is representative of performance on a set of internal cognitive tasks. Existing CDM frameworks are ill-equipped to model and infer latent structure for such complex item designs. The current study presents a novel diagnostic model framework capable of modeling (potentially) different response mechanisms for different item tasks. More broadly, we extend the concept of latent structure for the class of general CDMs by partitioning item parameters and latent structure into item-level and category-specific components. The proposed framework provides a method for uncovering more detailed structural information in assessment-related domains and can adapt to a range of structural complexities. A Bayesian estimation routine is discussed, and model performance is evaluated under controlled conditions using a Monte Carlo study. Lastly, we discuss an application of the model.

Regularization in cognitive diagnosis models

Tuesday, 25th July - 10:30: Diagnostic Classification Models (Atrium) - Oral

Dr. Yuan Ge (College Board), Dr. Wenchao Ma (The University of Alabama)

Cognitive diagnostic assessment facilitates the fine-grained measurement of students' mastery/ non-mastery of attributes. Previous studies have focused on developing models to conduct diagnostic classifications (e.g., de la Torre, 2009; de la Torre & Minchen, 2014), but some of those cognitive diagnosis models (CDMs) may be too complicated to be useful especially when sample size is small and test is short. To achieve parametrically parsimony while preserving appropriate model fit in CDMs, we hope to examine the performance of regularized CDMs with different penalty terms under varied conditions.

Our research is guided by the following research questions: 1) To what extent regularized CDMs can correctly identify the item response functions? 2) Can regularized CDMs provide better parameter estimation than the unregularized models?

Our preliminary results showed that certain regularized G-DINA models provided better parameter and the regularized models generally had acceptable performance but had different regularization accuracy when looking at the specific parameters about whether the penalties were posed to them when they should be regularized, vice versa.

Balancing the least complexity of the parameters and the most identifiability is important, especially when seeking for sufficient evidence to support acceptable regularization performance in CDMs. We will consider the effect of multiple constraints within L1 regularization approach and with more CDMs in the full paper.

Going deep in diagnostic modeling: Deep Cognitive Diagnostic Models (DeepCDMs)

Tuesday, 25th July - 10:45: Diagnostic Classification Models (Atrium) - Oral

Dr. Yuqi Gu (Columbia University)

Cognitive Diagnostic Models (CDMs) are discrete latent variable models popular in educational and psychological measurement. In this work, motivated by the advantages of deep generative modeling and by identifiability considerations, we propose a new family of DeepCDMs, to hunt for deep discrete diagnostic information. The new class of models enjoys the nice properties of identifiability, parsimony, and interpretability. Mathematically, DeepCDMs are entirely identifiable, including in fully exploratory settings and allowing to uniquely identify the parameters and discrete loading structures (the "Q-matrics") at all different depths in the generative graphical model. Statistically, DeepCDMs are parsimonious, because they can use a relatively small number of parameters to expressively model data thanks to the depth. Practically, DeepCDMs are interpretable, because the shrinking-ladder-shaped deep architecture can capture cognitive concepts and provide multi-granularity diagnostics from coarse- to fine-grained and from high-level to detailed. For identifiability, we establish transparent identifiability conditions for various DeepCDMs. Our conditions impose intuitive constraints on the structures of the multiple Q-matrices, and inspire a generative graph with increasingly smaller latent layers when going deeper. For estimation and computation, we develop Bayesian formulations and efficient Gibbs sampling algorithms. Simulation studies and an application to the TIMSS 2019 math assessment data demonstrate the usefulness of the proposed methodology.

Identifying cognitive diagnostic models for continuous or count responses

Tuesday, 25th July - 11:00: Diagnostic Classification Models (Atrium) - Oral

Mr. Seunghyun Lee (Columbia University), Dr. Yuqi Gu (Columbia University)

Cognitive diagnostic models (CDMs, also known as diagnostic classification models) are a popular family of discrete latent variable models in psychometrics. CDMs introduce binary latent variables to capture students' mastery or deficiency of fine-grained specific skills. Over time, various CDMs have been developed to model dichotomous (binary) item response data, such as the DINA model and log-linear CDM. However, with technological advancements and varying test formats, new response types have emerged in modern educational assessments, including *continuous responses* from response times and *count-valued responses* from assessments with repetitive tasks or eye-tracking sensors.

In the literature, many extended CDMs have been proposed for modeling different response types. However, to the best of our knowledge, the model identifiability of these models has not been theoretically justified, despite their popularity. Also, the proposed models assume a specific parametric family tailored for a specific response type. To address these issues, we propose a flexible CDM framework with minimal assumptions that can model all possible response types. We also prove that our general model is identifiable under conditions similar to those for dichotomous responses. We also propose a universal EM algorithm that can be applied to most parametric CDMs in our general framework. We conduct simulation studies for various response types, which corroborate our identifiability theory and validate our algorithm's superior empirical performance. Additionally, we apply our proposed method to response time and count datasets and discover interesting insights from the educational assessment.

A two-step robust estimation approach for inferring within-person relations in longitudinal design

Tuesday, 25th July - 10:15: Longitudinal Data Analysis (Benjamin Banneker) - Oral

Dr. Satoshi Usami (The University of Tokyo)

Psychological researchers have shown an interest in disaggregating within-person variability from between-person differences, and applications of the random-intercept cross-lagged panel model (RI-CLPM) with stable trait factors has increased rapidly. This presentation provides another recent approach that consists of a two-step procedure: within-person variability scores (WPVS) for each person, which are disaggregated from the stable traits, are calculated using structural equation modeling, and causal parameters are then estimated via a potential outcome approach, such as by using structural nested mean models. This method assumes a data-generating process similar to that in RI-CLPM, and has several advantages: (i) the flexible inclusion of curvilinear and interaction effects for WPVS as latent variables, (ii) more accurate estimates of causal parameters for reciprocal relations can be obtained under certain conditions owing to them being doubly robust, even if unobserved time-varying confounders and model misspecifications exist, and (iii) the risk of obtaining improper solutions is minimized.

Detecting change in dynamics through change-point analysis and time-varying parameters

Tuesday, 25th July - 10:30: Longitudinal Data Analysis (Benjamin Banneker) - Oral

Dr. Meng Chen (University of Oklahoma Health Sciences Center), Prof. Michael D. Hunter (The Pennsylvania State University), Prof. Sy-Miin Chow (The Pennsylvania State University)

Most dynamic modeling practices in psychological science assume that the parameters governing the patterns of change are invariant through time. Having time-varying parameters (TVPs) in dynamic models allows researchers insights into when patterns governing a certain process are also changing over time (e.g., within-person change and within-person variability). In this study, we compare two methods that can incorporate both continuous and abrupt change in TVPs. The first one is a filtering approach based on the extended Kalman filter and the fixed interval smoother (Chow, Hamaker, & Allaire, 2009; You, Hunter, Chen, & Chow, 2020), where a TVP of interest is incorporated as a state variable. Continuous change in the TVP is accounted for with a random walk or another flexible stochastic process, and abrupt change is probed using test statistics for innovative outliers (De Jong and Penzer, 1998) derived from filtering and smoothing byproducts. The second method is a splines approach (Albers and Bringmann, 2020), where continuous change is approximated with regression splines and abrupt change is screened by systematically incorporating potential change points and evaluating model fit improvement. We conduct simulations based on an autoregressive model to compare the performance of these two methods under different types and shapes of dynamic TVPs, as well as both linear and nonlinear models. We also discuss the feasibility of extending these methods to continuous-time dynamic models (i.e., differential equation models).

Repeated measurement analysis for non-linear data in small samples

Tuesday, 25th July - 10:45: Longitudinal Data Analysis (Benjamin Banneker) - Oral

Dr. Sunmee Kim (University of Manitoba)

Longitudinal and experimental studies, in which repeated measurements are collected to track changes in an individual's responses as time or condition changes, play an important role in psychological, health, and biomedical research. However, analyzing this type of data can be challenging as it is often non-linear with respect to a given response function, and can be incomplete, unbalanced, or based on small samples.
This presentation will address methodological issues and best practices in analyzing non-linear repeated measurement data, with a focus on unbalanced and small samples. We will provide a practical guide that compares and discusses three different classes of approaches: generalized additive models, non-parametric bootstrap tests for function differences, and functional ANOVA. Our aim is to strike a balance between technical correctness and ease of use, making it accessible to practitioners with limited statistical background.

Circumplex models with behavioral time series

Tuesday, 25th July - 11:00: Longitudinal Data Analysis (Benjamin Banneker) - Oral

Ms. Dayoung Lee (University of Notre Dame), Dr. Guangjian Zhang (University of Notre Dame)

The circumplex model allows researchers to assess the psychological theory that affect and some personality traits can be placed around the circumference of a circle. Although researchers have successfully used the circumplex model to study stable inter-individual differences with cross-sectional data, there is an increasing need to examine the viability of the circumplex model with multivariate time series data collected on the same individuals due to the rapid developments of new data collection methods like smartphone apps and wearable sensors. Establishing the circumplex model for a particular individual uncovers the relationship between different affect states or interpersonal interactions unique to the individual, and it will contribute to the development of the person-centered treatment or intervention. Estimating the circumplex model with time series data is more complex than with conventional cross-sectional data, because scores at nearby time points tend to be correlated. In this paper, we adapt Browne's (1992) circumplex model to accommodate time series data. We illustrate the proposed method with an empirical data set of daily affect ratings of an individual over a period of 70 days. The affect circumplex of this individual is similar, but not the same, as the one uncovered in the usual between-subject setting. We conduct a simulation study to explore the statistical properties of the proposed method in a variety of situations. The results show that the new method provides satisfactory confidence intervals and test statistics.

Dynamic model with interactions: Insight from the predator-prey model

Tuesday, 25th July - 11:15: Longitudinal Data Analysis (Benjamin Banneker) - Oral

Ms. Minglan Li (Beijing Normal University), Mr. Qingshan Liu (Beijing Normal University), Dr. Yueqin Hu (Beijing Normal University)

In recent years, an increasing number of longitudinal or intensive longitudinal studies have focused on the bidirectional or reciprocal relations between variables. Dynamic models that use the value of one variable at a previous time point to predict the value of another variable at a later time point, such as the dynamic structural equation models (DSEM), have become popular in behavioral science. In this study, we proposed a novel dynamic model that was derived from the discretization of the predator-prey model. Unlike the classic autoregressive and cross-lagged DSEM model, the effect of one variable on another in the predator-prey model is through an interaction term rather than a main effect, representing an influence on the autoregressive effects/inertia rather than on the levels. Simulation studies showed that the new model was able to distinguish between a simulated true interaction effect and a simulated true main effect. A 30-day daily diary study among 359 college students on the bidirectional relation between personality and negative affect supported the practical relevance of the dynamic interaction models. Results indicated that higher levels of neuroticism predicted higher levels of negative affect, whereas higher levels of extroversion predicted lower inertia of negative affect. The dynamic interaction models proposed in this study may have broader applications in revealing the subtle mechanisms in multivariate dynamic systems.

Detecting local item dependency associated with response latency: A test security application

Tuesday, 25th July - 10:15: Response Times (Prince George) - Oral

Dr. Joseph Grochowalski (The College Board), Dr. Amy Hendrickson (The College Board)

Multiple correspondence analysis (MCA)–an unsupervised learning technique that optimally scales categorical data based on multidimensional associations–can be useful for detecting local item dependency (LID) in tests. LID can exist for a number of reasons, including test taking misbehavior. We propose to use MCA to detect LID as a method for cheating detection, and we augment the detection power by using constrained (or canonical) MCA with response time latencies. We illustrate the descriptive and inferential aspects of this method using a simulation analysis, showing that shared answers by test takers can be well detected for varying levels of pre-knowledge and numbers of complicit cheaters. We further demonstrate the improved power of detection when we combine the item response data with response latencies, which shows that any anomalies in response latency–whether they be speed increases, decreases, or patterns of answer speed–can improve LID detection power when it is caused by cheating. We apply the method to a real data set with known test taking misbehavior and discuss the implications of the results and best practices.

Exploring asymmetric relationships between response time and latent traits in noncognitive measurement

Tuesday, 25th July - 10:30: Response Times (Prince George) - Oral

Ms. Tongtong Zou (University of Wisconsin-Madison), Prof. Daniel Bolt (University of Wisconsin-Madison)

In recent years there has been an increasing interest in modeling the relationship between item responses and response time for non-cognitive tests, among which the “inverted-U” approach has been commonly used. Most early models assume that such relationships are symmetric around 0, while Molenaar et al. (2021) proposed a more general version of the “inverted-U” model by incorporating an “asymmetry effect”. The resulting model is shown to often fit better than the traditional symmetric model.

In this study we explore the potential relevance of respondent level factors that may explain this asymmetry, focusing on response style (the frequency with which each category got selected), and the specific item response (which category is chosen) characteristics. The logic of these as predictors follows from the observed tendencies for both respondents and respondent populations to make disproportionate use of some response categories over others, and the previously demonstrated relationships between frequency of category selection and response time, thus potentially explaining the asymmetry phenomenon.

Using three empirical datasets to investigate, we do in fact observe the asymmetry effect to shrink in size across all three when such response style indicators are included in the model. A simulation study with generated extreme response style behaviors and corresponding response time supports the same findings.

Fitting a drift-diffusion IRT model to complex cognitive response times

Tuesday, 25th July - 10:45: Response Times (Prince George) - Oral

Mr. Ritesh Malaiya (University of Texas at Dallas)

Complex cognitive tasks tend to have high variance in the progress of cognitive states while an item is being attempted (Ackerman, 2014). Such a high variance may contribute to the high variance in empirically observed complex cognitive response times (CCRT), with means ranging from several seconds to minutes. The Q-diffusion model (van der Maas et al., 2011) is a joint probability model of response time and accuracy based on the Item Response Theory and Drift-Diffusion Model (DDM). Q-diffusion has been evaluated for response times ranging up to several seconds. The current study further evaluated Q-diffusion for simulated CCRT distributions characterizing empirically observed response times of insight- and math-based tasks (Ackerman, 2014). A random walk simulator was constructed using a probability transition matrix over a set of cognitive states where the simulator can either stay in the current state or move a state closer or further from a decision boundary (Diederich & Busemeyer, 2003). Then, the transition matrix was modified to emulate low to high variability in the probability of state change and six CCRT distributions were simulated. Then using Bayesian methods, the posterior predictive response time distribution (PPD-CCRT) was sampled from Q-diffusion given each simulated CCRT and multiple prior distributions. The mean PPD-CCRT effectively matched the mean CCRT for all scenarios. However, the variance in PPD-CCRT did not match the variance in simulated CCRT, especially when simulated CCRT had a larger variance. Implications of these results on establishing the validity of Q-diffusion for insight- and math-based tasks are further discussed.

Item response theory modeling with response times: Some issues

Tuesday, 25th July - 11:00: Response Times (Prince George) - Oral

Prof. Susan Embretson (Georgia Institute of Technology)

The increased prevalence of item response time (RT) data along with item responses has made applications of several joint item response theory (IRT) models feasible. Molenaar et al (2015) unified several IRT models with joint response accuracy and RT into a common hierarchical framework to possibly increase the measurement precision of trait. Depending on the model, the assumed relationship between response accuracy and response time may be either positive or negative. Item response times also can be included in mixture IRT models (e.g., Ulitzsh et al. 2022) to identify examinee strategy differences. However, a previous study (Embretson,2021) found substantial differences between examinees in the relationship of response time to item difficulty, accuracy and test position on a spatial ability test. Such differences could impact the advantages of the various joint models. In the current study, several different types of tests are examined for examinee differences in within-test response time relationships. Impact on hierarchical and mixture models for these tests is also examined.

Detecting aberrant behaviors of test-takers with Bayesian hierarchical response times models

Tuesday, 25th July - 11:15: Response Times (Prince George) - Oral

Dr. Burhanettin Ozdemir (Prince Sultan University)

The response time data is used to enhance test design, item calibration, detecting aberrant behaviors, and 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 the 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 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 item exposure, item pre-knowledge, and cheating behaviors.

Causal effect sensitivity across sets of competing DAGs

Tuesday, 25th July - 10:15: Causal Inference (Margaret Brent) - Oral

Mr. Ronald Flores (University of Missouri), Dr. Edgar Merkle (University of Missouri)

Directed Acyclic Graphs (DAGs) are meant to explain data. Data, however, could usually be explained by multiple DAGs. In the current study, we develop a test statistic that assesses the sensitivity of results across sets of DAGs, where the DAGs are operationalized and estimated via SEM. Specifically, our metric assesses whether the estimate of a focal path varies across sets of competing models. This serves as a sensitivity analysis wherein researchers can judge whether confounders, colliders, or other model attributes significantly influence a targeted effect of interest. In cases where a path estimate does not vary across competing models, the differing structures of the models in which this effect is situated could be ignored, suggesting robust effect estimation. Conversely, if a path does vary, it could qualify results. We will first discuss the theoretical details underlying our test, followed by illustrations of test performance using both simulated and real data. We will close with some next steps and details on how to automate testing across larger sets of plausible DAGs.

Bounding causal effects of pretest-posttest designs with a control group

Tuesday, 25th July - 10:30: Causal Inference (Margaret Brent) - Oral

Mr. Muwon Kwon (University of Maryland, College Park), Dr. Peter Steiner (University of Maryland, College Park)

Pretest-posttest designs with a control group allow researchers to estimate the causal effect of an intervention either by controlling for the pretest in a regression model or by conducting a gain score analysis (i.e., difference-in-differences, DID). While covariate adjustments rely on the adjustment criterion (Shpitser et al., 2012), DID needs the common trend assumption (Lechner, 2011) to identify a causal effect. Given the uncertainty about the extent to which each of these assumptions might be violated, bounding the causal effect helps researchers in deriving a range of plausible effects without fully relying on either assumption. Building on previous suggestions (Angrist & Pischke, 2009; Ding & Li, 2019; Luedtke & Robitzsch; 2020) and causal graphs (Pearl, 2009), we propose a linear bounding approach that results, in general, in narrower and thus more informative bounds. In a first step, the causal effect is expressed in terms of the pretest's reliability (with respect to the unobserved confounder) and other estimable quantities. Then, to effectively bound the causal effect we derive two estimable restrictions for the pretest's reliability. One uses path tracing rules to bound the reliability from below. The other is based on the ratio of the pretest's and posttest's reliability and allows for further bounding from below and above. Theoretical derivations and simulation results indicate that correspondingly derived bounds are narrower than those of alternative methods and moreover correctly include the true causal effect even when both covariate adjustment and DID either overestimate or underestimate the true causal effect.

Investigating causal relationships between longitudinal treatment patterns and heterogeneous effects

Tuesday, 25th July - 10:45: Causal Inference (Margaret Brent) - Oral

Ms. Hanna Kim (University of Wisconsin-Madison), Prof. Jee-Seon Kim (University of Wisconsin-Madison)

This study focuses on examining treatment effect heterogeneity in longitudinal studies, which yields valuable contextual information about how a treatment works. When treatments are provided for multiple phases, participants may exhibit different treatment participation patterns over time, leading to substantial variability in treatment effects. We propose two frameworks for conceptualizing the impact of longitudinal treatment participation patterns on treatment effect heterogeneity, using the national Head Start program's effect on children's cognitive development as an illustration. Because children in the Head Start Impact Study (HSIS) were randomly assigned to Head Start in the first year but not in the second year of the study, their Head Start attendance was partially observational and resulted in four patterns.

First, we translate the effects of attending Head Start into different causal mediation estimands based on attendance patterns, accounting for confounders and identification assumptions through directed acyclic graphs (DAGs). For example, the benefit of attending Head Start at age three in addition to the regular administration at age four can be defined as the controlled direct effect of an early Head Start. We then use multilevel models combined with generalized propensity score weighting to address selection bias of the four attendance patterns and clustering within program sites. Multiple modeling approaches for propensity score models and outcome models are empirically compared in terms of performance. Our findings present methods for naturally addressing a novel source of treatment effect heterogeneity in longitudinal studies and offer substantive implications for refining social interventions implemented in multiple phases over time.

A practical guide on selecting propensity score matching methods

Tuesday, 25th July - 11:00: Causal Inference (Margaret Brent) - Oral

Ms. Lizzy Wu (University of Illinois Urbana-Champaign), Dr. Ge Jiang (University of Illinois Urbana-Champaign)

As an increasingly frequently used technique in observational studies to estimate the causal treatment effects, propensity-score matching (PSM) has offered certain advantages, like controlling for selection bias (Rosenbaum & Rubin, 1983). However, practitioners have little guidance on selecting the optimal approach to accurately estimate the treatment effect on a binary outcome.

Methods

A Monte Carlo simulation study was conducted to examine the performance of commonly used methods of matching treated with untreated subjects. Specifically, four PSM methods, full, optimal, exact, and greedy matching, were investigated across 54 conditions (two levels of sample size * three levels of treatment model complexity * three levels of outcome model complexity * three levels of true effect size). Moreover, a variety of optimal choices of the matching process were explored, like optimal caliper width, the optimal ratio of the treated matched to the untreated, and the matching order. All simulations were in the context of a binary outcome, a binary treatment, and baseline confounders (categorical and continuous).

Results

Our study showed that the greedy matching method had a superior performance across the four PSM methods. Among 333 algorithms tested, the most recommended one with doubly robust estimation across the 54 simulation conditions was the greedy matching algorithm without a replacement, with a caliper width of .2 and a ratio of 3, and matching from the smallest distance measured to the largest. Meanwhile, PSM methods generally overestimate true treatment effects on binary outcomes. Therefore, it is necessary to assess covariate balance when analyzing propensity-score-matched data.

Moderated treatment effects in nonrandomized partially nested designs

Tuesday, 25th July - 11:15: Causal Inference (Margaret Brent) - Oral

Dr. Xiao Liu (Affiliation: The University of Texas at Austin)

Intervention studies in psychology and education often have a treatment-induced partial nesting design (PND) structure: individuals assigned to the treatment group are subsequently assigned to clusters (e.g., therapy clusters) to receive the treatment, but this clustering does not occur for individuals assigned to the control group. In PNDs, assessing whether and how much the treatment effects are moderated by individuals' characteristics and characteristics of clusters in treatment groups is often of interest. Previous research has focused on assessing moderation in PNDs with randomized treatment assignments, but randomization is not always feasible. For PNDs where treatment assignments may be nonrandomized, methods for investigating treatment effect moderation are underdeveloped. A challenge is that for PNDs, clusters in the treatment group are formed after the treatment assignment, meaning that the observed cluster-level moderators, such as observed cluster-means of baseline covariates, are posttreatment variables (i.e., observed after the treatment assignment); crude conditioning on such posttreatment variables could lead to biased treatment effect estimates. In this study, we employ the expanded potential outcomes framework to define and identify the moderated causal treatment effects in PNDs. For estimation, we develop a multilevel outcome modeling method based on regression with residuals. We also extended a propensity score weighting approach for assessing the moderated treatment effects. The performance of the studied methods is evaluated through simulation studies. An empirical example is provided for illustration. We hope this study can provide useful insights and tools for assessing moderation in PNDs where randomization may be infeasible.

Modeling variance and skewness as functions of latent factors in latent variable models with mixed items

Tuesday, 25th July - 10:15: Item Response Theory (Juan Ramon Jimenez) - Oral

Mr. Camilo Cardenas (London School of Economics and Political Science), Prof. Irini Moustaki (London School of Economics and Political Science), Dr. Yunxiao Chen (London School of Economics and Political Science), Prof. Giampiero Marra (University College London)

We introduce a latent variable model that utilizes linear functions of latent factors to model the location, shape, and scale parameters of the items' conditional distributions. This enables the modeling of mean and higher order moments in terms of the latent factors. We discuss the estimation for the model parameters and present two empirical applications. In our first example, we demonstrate a joint model for items and response times by using data from the PISA 2018 mathematics exam, assuming a Skew-Normal distribution for the (log) response times. In our second example, we use thermometer data from the 2020 American National Election Study and assume a Beta distribution for the items. In both cases, we show how the (conditional) variances and skewness of the items change depending on an individual's position on the latent scale. Our results indicate that modeling the entire distribution of the items, as opposed to just the conditional mean (which is currently the mainstream practice in the field), provides a better model fit and deeper insight into how the items reflect the latent constructs they were designed to measure. This framework provides a powerful tool for researchers in various fields. Our empirical applications demonstrate the effectiveness of this model in both educational and political contexts, and we anticipate that this model will inspire further developments and applications in the future.

Joint modeling of action sequences and action times in problem-solving tasks

Tuesday, 25th July - 10:30: Item Response Theory (Juan Ramon Jimenez) - Oral

Mr. Fu Yanbin (Zhejiang Normal University), Dr. Peida Zhan (Zhejiang Normal University), Mr. Qipeng Chen (Zhejiang Normal University), Prof. Hong Jiao (University of Maryland, College Park)

Process data refer to data recorded in computerized assessments that reflect the problem-solving processes of participants and provide greater insight into how they solve problems and how well they solve them. Action times, namely the amount of time required to complete an action sequence, are also included in such data along with action sequences. In this study, an action-level joint model of action sequences and action times was proposed, in which the sequential response model (SRM) was used as the measurement model for action sequences and a new proposed log-normal action time model was used as the measurement model for action times. The proposed action-level joint model can be regarded as an extension of the SRM by incorporating action times within the joint-hierarchical modeling framework and as an extension of the conventional item-level joint models in terms of process data analysis. Results of the empirical and simulation studies demonstrated that the model setup was justified, the model parameters could be interpreted, the parameter estimates were accurate, and that taking into account participants' action times further was beneficial. Overall, the proposed action-level joint model provides a useful modeling framework for the analysis of process data in technology-enhanced assessments from the perspective of latent variable modeling.

Effect of testlets' difficulty distribution on estimation accuracy and reliability improvement

Tuesday, 25th July - 10:45: Item Response Theory (Juan Ramon Jimenez) - Oral

Dr. Kaili Liang (Collaborative Innovation Center of Assessment for Basic Education Quality,Beijing Normal University), Prof. Hongbo Wen (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University)

There is a special form about tests called testlets which are defined as groups of related items on a test. Researchers mainly focus on its measurement error, that is, testlet effects. Different types of testlets has different measurement error, so the ways to improve reliability were also different. Increasing tests' length is a common path to reduce measurement error and improve reliability.There were two ways to increase testlets' length,one was increasing testlets' number,the other was increasing number of items nested in testlets. Some researchers thought the former was effective, others thought the latter was effective.We think the divergence in above conclusions was due to different testlet effects caused by each measurement facet. We further inferred that maybe testlets' difficulty distribution affects testlet effects by analyzing previous research.Therefore,we investigated effect of testlets' difficulty distribution on testlet effects and reliability' improvement.First,we explored the effect of testlets' difficulty distribution on testlet effects by comparing estimation accuracy of Bi-factor model and 2PL model. Then, we used GENOVA to conduct G and D studies to investigate effective way of reliability' improvement. The main conclusions were as follows:when testlets' difficulty was similar, testlet effects were small, 2PL model was more effective than Bi-factor model; Furthermore, two ways improved reliability were effective which were increasing the number of testlets and items nested in testlets.When testlets' difficulty was inconsistent, the more testlets' difficulty was, the greater testlet effects were, Bi-factor model was more effective than 2PL model;The effective way to improve reliability was to increase testlets' number.

Item response modeling of clinical instruments with filter questions

Tuesday, 25th July - 11:00: Item Response Theory (Juan Ramon Jimenez) - Oral

Dr. Brooke Magnus (Boston College)

Clinical instruments that use a filter/follow-up response format often produce data with excess zeros, especially when administered to nonclinical samples. When the unidimensional graded response model (GRM) is then fit to these data, parameter estimates and scale scores tend to suggest that the instrument measures individual differences only among individuals with severe levels of the psychopathology. In such scenarios, alternative item response models that explicitly account for excess zeros may be more appropriate. The multivariate hurdle graded response model (MH-GRM), which has previously been proposed for handling zero-inflated questionnaire data, includes two latent variables: susceptibility, which underlies responses to the filter question, and severity, which underlies responses to the follow-up question. Using both simulated data and data from the Generalized Anxiety Disorder Scale-7 (GAD-7), the current research shows that compared to unidimensional GRMs, the MH-GRM is better able to capture individual differences across a wider range of psychopathology, and that when unidimensional GRMs are fit to data from questionnaires that include filter questions, item parameter estimates and scale scores can incorrectly suggest that the instrument only measures individual differences among those who are high on the psychopathology; individual differences at the lower end of the severity continuum largely go unmeasured. Practical implications are discussed.

Assessing dimensionality of sparse item response data: Comparison of different data imputation methods

Tuesday, 25th July - 11:15: Item Response Theory (Juan Ramon Jimenez) - Oral

Dr. Fei Zhao (NWEA), Dr. Yong Luo (NWEA)

Assess the dimensionality of sparse item response data collected from computer adaptive testing (CAT) or multi-stage testing (MST) is psychometrically challenging. Traditionally, the most common method for dimensionality assessment of sparse item response data is to conduct principal component analysis (PCA) of the standardized residual (Linacre, 1998) and impute the standardized residual as 0 for the missing item responses, which is implemented in the popular Rasch software program Winsteps. Recently, Bulut & Kim (2021) explored the performances of different data imputation techniques (two-way and multivariate) when investigating dimensionality of sparse item response data collected from CAT simulation and found that data imputation with multivariate imputation by chain equations with classification and regression trees (MICE-CART) performed promisingly. While both the traditional method and MICE-CART are data imputation methods that can be applied in conducting PCA of the standardized residual , they differ noticeably in the sense that the former simply imputes the standardized residuals of missing item responses as 0 while the latter use model-based predictions to impute the raw item responses before computing the standardized residuals, and it is not clear which method performs better. The current simulation study will compare the performances of those two methods in assessing dimensionality of CAT-based sparse item response data to provide methodological guidelines to researchers and practitioners who need to conduct dimensionality assessment of sparse data.

Longitudinal measurement invariance of Christian scales

Tuesday, 25th July - 10:15: Measurement Applications (Thurgood Marshall) - Oral

Mr. Hiroki Matsuo (Baylor University)

When researchers are interested in individual changes in constructs over time, such constructs are repeatedly measured for the same individuals at multiple time points. One of the assumptions is that the measured variables such as responses to survey items represent the same construct at each time point. Violating this assumption could lead to misinterpretation of proceeding analyses (i.e., latent growth models). Different levels of invariant models have been suggested including configural, metric, and scale invariant models where with longitudinal data, the same constructs at different time points are allowed to covary. In this study, two widely used measures of Christian beliefs, the Christian Orthodoxy Scale (Hunsberger, 1989) and the Faith Maturity Scale (Benson et al., 1993), were administered to a group of undergraduate students at a regional faith-based institution. Participants answered the surveys during their first, third, and fourth years. Measurement invariance tests were conducted for each survey separately to confirm its factor structure. Examinations of latent means using robust maximum likelihood estimation suggested that the structures of these two measures were relatively stable across the time points. Our recommendations include that researchers who are interested in individual changes/growth in Christian beliefs should carefully choose their measurements, and further investigation is still needed to test other psychometric properties of these measures including multigroup invariance between genders.

A study on the influence of family moral education on college students' moral outlook

Tuesday, 25th July - 10:30: Measurement Applications (Thurgood Marshall) - Oral

Ms. Qile Liu (University of Macau), Dr. Biao Zeng (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University)

Moral values are a key part of the social development of college students, and family education is an important factor affecting the development of moral values. However, due to the lack of standardized measurement instruments, the impact of family education on the moral values of college students is unclear.

Based on this, this study developed the "Family Moral Education Scale" and "College Student Morality Scale" to assess the development status and relationship between family moral education and moral values among Chinese college students. A total of 376 college students responded to these two scales. We then conducted an empirical analysis of the questionnaire and analyzed the relationship between family moral education and moral values.

The results showed that the reliability and validity of the "Family Moral Education Scale" and "College Student Morality Scale" were sound; parents' moral literacy, moral education philosophy, and moral education methods can significantly and positively predict the parents' moral literacy, moral education philosophy, and moral education methods can significantly and positively predict the moral values scores of college students, but the parent–child relationship cannot significantly predict the moral values scores of college students; gender, family structure, and family socioeconomic status cannot moderate the impact of family moral education on college students' moral values.

Therefore, it is necessary to focus on the major influence of family education and adopt appropriate moral education concepts and methods while improving the development of a family education system that is based on the principles of the family.

Assessing the consistency of affective responses

Tuesday, 25th July - 10:45: Measurement Applications (Thurgood Marshall) - Oral

Mr. Niels Vanhasbroeck (KU Leuven)

Experimental studies have increased our understanding of fluctuations in affect, relating them to monetary outcomes, prediction errors, and learning. However, few of these studies account for measurement error in the affective responses. Adding to this the issue of using single-item measures of affect in such studies, one can wonder the extent of this error on our inferences. In this study, we address this question for several different affective measure formats (*format*) that were either discrete or continuous (*continuity*). A total of 1398 participants completed a probabilistic reward task in which they were exposed to multiple iterations of the same sequence of monetary wins and losses. After each trial outcome, participants had to report their affective state using the measure assigned to them (between-subject manipulation), ultimately allowing us to examine the correspondence of one's responses across iterations. Results indicate that there is little difference in consistency across levels of format and continuity. As to the consistencies themselves, we found that data-based consistencies were relatively low. However, there was a high agreement between the parameter estimates of linear models across iterations. This suggests that patterns in the data can still be adequately picked up by statistical models, despite the presence of measurement error in the data.

Psychometric analysis of patient reported outcomes

Tuesday, 25th July - 14:30: Symposium: Psychometric Analysis of Patient Reported Outcomes (Colony Ballroom) - Symposium Overview

Prof. Jeff Douglas (University of Illinois Urbana-Champaign)

Patient reported outcomes (PROs) are playing an increasingly important role in both clinical trials and medical observational studies. With new technologies available to collect PROs more easily and more frequently that load directly into databases there is a growing abundance of these data that give a more complete picture of a patient's well-being. This creates a need for advanced psychometric methods whether it is for assessing treatment effects on multivariate and multidimensional endpoints or simply monitoring the daily health of patients. Item response models and classical test theory have received much attention in the analysis of PROs and the construction of valid and reliable measurement instruments. However, PROs for health and quality of life measurement tend to be longitudinal and multidimensional and provide great opportunities for a wide variety of latent variable modeling and psychometric research. Models for a variety of mixed data types are needed for PRO data that can be a mixture of binary, nominal, ordinal, or even continuous measures of time or points on visual analogue scale. There are also opportunities for development of methods to jointly model PROs with other study endpoints such as time-to-event or biomarker data. This symposium includes presentations ranging from an overview of the general area and use of PROs in medicine and health and the psychometric research opportunities that they afford, to very specific examples of statistical models especially tailored to the analysis of PROs.

Opportunities and challenges in the field of patient reported outcomes

Tuesday, 25th July - 14:33: Symposium: Psychometric Analysis of Patient Reported Outcomes (Colony Ballroom) - Symposium Presentation

Dr. Charlie Iaconangelo (Janssen)

Recent trends in the field of patient reported outcomes (PROs) present opportunities for the psychometric community to make high-impact contributions. Regulatory bodies such as the US Food and Drug Administration (FDA) value the incorporation of patients' perspectives in medical/pharmaceutical product development. FDA draft guidance released in June 2022 emphasized advanced psychometric methods for analyzing data from clinical studies to assess outcomes, such as impact of treatments on symptoms, functioning, and health-related quality of life (HRQoL). This offers opportunities for psychometric researchers. For example, new data collection techniques, such as personal device data, yield an unprecedented level of detail about the patient condition while on therapy. Another opportunity exists as a result of stakeholders (such as regulatory agencies) updating standards for defining meaningful change, which is used to interpret the treatment effect. Similarly, the field of PROs seeks updated approaches for establishing responder definitions, which are used to help quantify therapeutic benefits. These definitions are used to make decisions about drug approval and treatment decisions. The field of psychometrics already implements high-stakes validity arguments for assessments used in, for example, licensure exams. To address these new challenges in PRO research, psychometric models should be expanded in ways that best accommodate clinical trial data. Existing psychometric research may be translated to the clinical field to address the new, higher validity standards. In doing so, the field of psychometrics can take advantage of new opportunities to provide solutions that address the challenges faced by drug developers and clinical researchers more broadly.

Some methodologic challenges in analyzing patient-reported outcomes in health sciences

Tuesday, 25th July - 14:51: Symposium: Psychometric Analysis of Patient Reported Outcomes (Colony Ballroom) - Symposium Presentation

Dr. Edward Ip (Wake Forest University School of Medicine)

This presentation discusses methodological challenges related to the analysis of patient-reported outcomes (PROs) within the health sciences. Specifically, it poses the following questions: (1) How can analytical methods make PROs more clinically actionable? (2) How can psychometrics improve the interpretation of PROs, given that summing item responses may present an incomplete picture of a patient because of the compensatory property of sum score? (3) Can measurement science be applied to PROs to demonstrate their added value beyond clinician-reported outcomes? Drawing on his experience analyzing PROs from multiple projects, the presenter will highlight the relevance of psychometric methods such as latent variable modeling for addressing these challenges. The presentation will cover predictive modeling methods for improving the clinical actionability of PROs (Challenge 1), hidden Markov modeling to illustrate the value of a non-compensatory approach for PROs (Challenge 2), and a new PRO system for mapping clinician-assessed adverse events in oncology trials. The presenter will also discuss potential factor analytic models that can be used to explore the potential added value of this new measurement system (Challenge 3).

A latent variable mixed-effects location scale model for patient reported longitudinal data

Tuesday, 25th July - 15:09: Symposium: Psychometric Analysis of Patient Reported Outcomes (Colony Ballroom) - Symposium Presentation

Prof. Shelley Blozis (University of California, Davis)

Patient reported outcomes that are measured over extended periods of time are typically marked by individual differences, both in the extent of change and the degree of within-person variation over time. Thus, characterizing these sources of variation and understanding the role of time-varying and person-level characteristics on these sources are critical in understanding patient experiences. For patients living with multiple sclerosis (MS), patient reported measures can exhibit a general trend at the individual level, such as generally increasing or decreasing, with individuals differing in these trends. In addition, there may be year-to-year variation about each patient's trend, with the degree of variability differing from one patient to the next. Indeed, understanding general trends, as well as within-person variation, can improve understanding of the disease process. A mixed-effects location scale model allows for the study of within- and between-person variation in measures over time. If multiple indicators of the outcome are available, then a latent variable version of the model may be applied. These models are applied to patient reported outcomes of individuals living with MS to gauze the utility of the analytic approach in this domain of study.

A hidden Markov modelling approach for understanding patient health over time

Tuesday, 25th July - 15:27: Symposium: Psychometric Analysis of Patient Reported Outcomes (Colony Ballroom) - Symposium Presentation

Mr. Eric Wayman (University of Illinois Urbana-Champaign), Prof. Jeff Douglas (University of Illinois Urbana-Champaign), Prof. Steven Culpepper (University of Illinois Urbana-Champaign)

In this paper we present a hidden Markov-type model for longitudinal patient-reported outcomes (PRO) data that is polytomous and where questions can have different numbers of levels. We treat these response items as reflecting the evolution of a latent, multi-attribute state, where the attributes are ordinal and polytomous and where each attribute can take potentially a different number of levels. In addition to a description of the model, we describe a multiple-block Metropolis-Hastings algorithm for estimating its parameters, display simulation results for multiple scenarios, and describe how the model could potentially be used to perform classification of individuals suffering from conditions that are best measured using PRO data.

Developing CD-CAT algorithms for college gate-way STEM courses

Tuesday, 25th July - 14:30: Computerized Adaptive Testing (Atrium) - Oral

Ms. Xiuxiu Tang (Purdue University, West Lafayette), Mr. Yuxiao Zhang (Purdue University, West Lafayette), Prof. Hua Hua Chang (Purdue University, West Lafayette)

The ideal learning environment in STEM education is one where instruction can be customized to best fit each student, so instructors can focus on each individual's specific needs. To deliver individualized instruction at scale, instructors need assessments that efficiently identify course proficiency for students from diverse backgrounds and abilities. However, instructors must also have a complete picture of their students' skill mastery to effectively implement targeted interventions. Cognitive diagnostic computerized adaptive testing (CD-CAT) combines the strength of both computerized adaptive testing (CAT) and cognitive diagnosis (CD), which can not only provide more precise information about student achievement levels, but also diagnose student mastery of the skills tested by an assessment, thereby allowing for effective and tailored instruction and targeted individualized intervention.

This study proposes an item selection algorithm for the CD-CAT design in the setting of college gate-way STEM courses. The algorithm is a dual information-based method, which combines information from both θ (overall proficiency) and α (skill mastery). In the beginning of the test, the selection of items will use more information from the estimation of θ. Once a desired estimation accuracy for is achieved, the item selection algorithm will focus more on the assessment of the examinee's skill mastery. Content balancing will also be considered by using the maximum priority index (MPI) method. The efficiency of the algorithm will be demonstrated using mastery pattern correct classification rate (PCCR) for α and root mean square error (RMSE) for θ .

A dynamic balancing attribute coverage method for CD-CAT

Tuesday, 25th July - 14:45: Computerized Adaptive Testing (Atrium) - Oral

Dr. Chia-Ling Hsu (Hong Kong Examinations and Assessment Authority), Prof. Shu-Ying Chen (National Chung Cheng University), Dr. Yi-Hsin Chen (University of South Florida)

Cognitive diagnosis computerized adaptive testing (CD-CAT) can offer each test-taker with tailored and detailed feedback. Balancing attribute coverage is crucial in the design and implementation of CD-CAT. This study presents a balancing attribute coverage method–dynamic attribute balancing index (DABI)–to dynamically utilize multiple balancing attribute coverage methods at various phases of CD-CAT. Based on simulation results, DABI demonstrated more item utilization than previous balancing attribute coverage methods, better coverage for each attribute to promote test validity, and produced equivalent or higher classification accuracy for estimating test-takers' mastery statuses. Overall, these findings support the feasibility of employing DABI in CD-CAT to enhance measurement accuracy, test validity, and item consumption efficiency.

Designing variable-length multidimensional multistage computerized adaptive testing

Tuesday, 25th July - 15:00: Computerized Adaptive Testing (Atrium) - Oral

Dr. Yi-Ling WU (National Taiwan Normal University), Dr. Chia-Ling Hsu (Hong Kong Examinations and Assessment Authority)

Multistage adaptive testing (MST) selects items adaptively at the module level rather than the item level as in computerized adaptive testing, resulting in fewer adaptive processes, more efficient test assembly, and better content balancing. In practice, assessments/tests are often evalute multiple proficiencies or literacy levels, such as the National Assessment of Educational Progress, Progress in International Reading Literacy Study, and Trends in Mathematics and Science Study. This study proposes an approach for designing variable-length multidimensional MST by considering the correlation between proficiencies. The simulation results showed that, as compared to fixed-length uni- and multi-dimensional MSTs, the new approach increased measurement precision for estimating individuals' latent proficiencies and decreased test length for terminating the MST.Multistage adaptive testing (MST) selects items adaptively at the module level rather than the item level as in computerized adaptive testing, resulting in fewer adaptive processes, more efficient test assembly, and better content balancing. In practice, assessments/tests are often evalute multiple proficiencies or literacy levels, such as the National Assessment of Educational Progress, Progress in International Reading Literacy Study, and Trends in Mathematics and Science Study. This study proposes an approach for designing variable-length multidimensional MST by considering the correlation between proficiencies. The simulation results showed that, as compared to fixed-length uni- and multi-dimensional MSTs, the new approach increased measurement precision for estimating individuals' latent proficiencies and decreased test length for terminating the MST.

A new approach to evaluating item parameter drift in computerized adaptive testing

Tuesday, 25th July - 15:15: Computerized Adaptive Testing (Atrium) - Oral

Mr. Hwanggyu Lim (Graduate Management Admission Council), Dr. Kyung T. Han (Graduate Management Admission Council)

In computerized adaptive testing (CAT), once new pretest items are calibrated based on the IRT model using response data from non-adaptive administrations, they are added to an item bank for operational uses in adaptive administrations. In IRT, it is assumed that the probability of answering the same item correctly for test takers with the same proficiency level is invariant across test occasions. In practice, however, test taker's behavior may change over time for various reasons. Consequently, response patterns could differ considerably from what was expected based on the initially calibrated items, leading to changes in item characteristics known as item parameter drift (IPD). Several IPD detection methods have been developed for the fixed test forms, but many of them are infeasible or impractical for CAT. For example, most IRT-based IPD detection methods require reestimating item parameters, which is complicated (if not infeasible) for items that are adaptively administered.

We propose a new IPD detection approach effective in CAT by modifying the residual-based differential item functioning (RDIF) detection framework (Lim et al., 2022). The RDIF framework is appealing for its effectiveness and practicality for DIF analysis in CAT because it does not require recalibration of items. In the preliminary simulations, the new method demonstrated its comparable performance to other existing measures in detecting IPD items in CAT, while significantly reducing the required analysis time. The study will offer comprehensive insights regarding the IPD in CAT and provide a new tool for streamlining the effort of continuously monitoring IPD for operational items.

Reducing measurement errors for PISA's computerized multistage testing by using an on-the-fly multistage design that incorporates response time.

Tuesday, 25th July - 15:30: Computerized Adaptive Testing (Atrium) - Oral

Prof. Hua Hua Chang (Purdue University, West Lafayette)

The 2018 PISA Reading Test uses a multi-stage adaptive test (MST) designed to improve measurement efficiency, especially in extreme cases of proficiency levels. However, the results reported (OECD 2019a and 2019b) did not show significant improvement compared to the pen-and-paper version as most people would expect. PISA-MST was designed by Educational Testing Services and is similar to NAEP-MST. Since PISA focuses on country performance, one might argue that reducing measurement error on each test taker may not improve accuracy at the country level. This study proposes an on-the-fly multistage adaptive testing (OMST) that can balance the advantages and limitations of Computerized Adaptive Testing (CAT) and MST. The new design also uses students' response time collected during the course of the testing. The simulation studies demonstrated that the proposed OMST design resulted in higher measurement efficiency, less violated constraints, and higher test security compared to the simulated MST design mimicking PISA 2018 reading assessment. In particular, for both low-end and high-end proficiency levels, the measurement efficiency has been greatly improved.

Priors in Bayesian estimation under the graded response model

Tuesday, 25th July - 14:30: Computational Methods for Latent Variable Models (Benjamin Banneker) - Oral

Prof. Seock-Ho Kim (University of Georgia)

A review of various priors used in Bayesian estimation under the graded response model is presented together with clear mathematical definitions of the hierarchical prior distributions. A Bayesian estimation method, Gibbs sampling, was compared with other methods including marginal Bayesian estimation and marginal maximum likelihood estimation using rating data from a performance assessment instrument under the graded response model. In addition, simulated item response data from a manual of a commercially available computer program were analyzed with the marginal maximum likelihood estimation method and Gibbs sampling under the graded response model. The data represent responses of 1000 examinees for a test containing 20 items with 4 ordered categories. All methods yielded nearly identical item parameter estimates. The shrinkage effect was observed in the ability estimates from Gibbs sampling. The computer program WinBUGS that implemented the rejection sampling method of Gibbs sampling was the main program employed in the study.

A doubly stochastic gradient algorithm for high-dimensional latent variable models

Tuesday, 25th July - 14:45: Computational Methods for Latent Variable Models (Benjamin Banneker) - Oral

Mr. Motonori Oka (London School of Economics and Political Science), Dr. Yunxiao Chen (London School of Economics and Political Science), Prof. Irini Moustaki (London School of Economics and Political Science)

High-dimensional latent variable models are becoming increasingly popular for analyzing large-scale data, with many observations and observed variables. These models contain a large number of latent variables which measure unobserved traits and provide insight into the relationships among the observable variables. A popular approach for estimating a latent variable model is by maximizing the marginal likelihood. However, when the dimension of the latent space is high, the expectation-maximization algorithm becomes computationally infeasible, and stochastic approximation algorithms—also known as stochastic gradient algorithms—have become the standard for solving the optimization problem. These algorithms, however, do not scale well with the dimension of the latent space. To address this computational bottleneck, a stochastic gradient MCMC sampler—a recent advance in Bayesian computation—has been incorporated into the sampling step. This sampler uses information from the gradient of the unnormalized posterior densities to efficiently sample the latent variables. This algorithm is named the doubly stochastic gradient algorithm, as both the latent variable samples and the fixed parameters are updated by some stochastic gradients. The algorithm has been extended to the problem of maximizing regularized marginal likelihoods, and its convergence properties have been established. To assess its performance, the proposed algorithm has been compared with existing stochastic gradient algorithms via simulation studies. It has also been applied to analyzing response data from large-scale psychological and educational assessments.

Assessing fitting propensities of item response models using limited-information methods

Tuesday, 25th July - 15:00: Computational Methods for Latent Variable Models (Benjamin Banneker) - Oral

Dr. Yon Soo Suh (NWEA), Dr. Li Cai (University of California Los Angeles)

Fitting Propensity (FP; Preacher, 2006) has been suggested as a method to better study and understand the utility of a model by being able to account for multiple types of model complexity and, thus, more adequately consider the principle of parsimony or Occam's Razor in model evaluation. An intuitive method for studying FP involves repeatedly fitting models to datasets sampled randomly and uniformly across a data space of interest and examining summaries of model fit indexes. However, computational issues in the data generation and estimation process due to the high-dimensional discrete space involved have impeded investigating the FP of Item Response Models (IRMs) when applying conventional full-information (FI) methods. We propose that limited-information (LI) methods can overcome these issues by focusing on the lower-order marginal moments up to only pairs of items instead of the full multinomial probabilities to reduce computational burden significantly. We present an efficient data-generating algorithm and corresponding estimation methods inspired by classical literature on sampling and estimating contingency tables with fixed margins. The feasibility of examining FP via the suggested LI approach is shown in comparison with the previously used FI approach with focus on exploratory factor analytic models, bifactor models, and two diagnostic classification models following the set up in Bonifay and Cai (2017). Furthermore, using the same IRMs, we demonstrate the flexibility of the proposed LI method to assess the FP of IRMs under various combinations of data sampling and estimation methods and discuss their suggested implications on factors influencing model complexity.

Fast M-estimation of GLLVM in high dimensions

Tuesday, 25th July - 15:15: Computational Methods for Latent Variable Models (Benjamin Banneker) - Oral

Prof. Maria-Pia Victoria-Feser (University of Geneva), Dr. Guillaume Blanc (University of Geneva), Prof. Silvia Cagnone (University of Bologna), Prof. Stephane Guerrier (University of Geneva)

Dimension reduction for high dimensional data is an important and challenging task, relevant to both machine learning and statistical applications. Generalized Linear Latent Variable Models (GLLVM) provide a probabilistic alternative to matrix factorization when the data are of mixed types, whether discrete, continuous, or a mixture of both. However, estimation of GLLVM parameters represents a tremendous challenge for even moderately large dimensions, essentially due to the multiple integrals involved in the likelihood function. Methods based on approximations of this latter, such as Laplace approximation, adaptive quadrature, or variational approximation, do not scale well to high dimensions. Alternatively, estimators based on penalized joint-likelihood functions, allow to fit GLLVM with p>n, but need the specification of the penalty scale and their consistency property is achieved when p diverges. Consequently, even for relatively large values of p, the estimators exhibit a relatively large finite sample bias. We propose instead an M-estimator, which has a negligible efficiency loss compared to the (exact) MLE. It can provide (bias reduced) estimates for GLLVMs with p>n, and with up to p=500 (non Gaussian) manifest variables and over q=10 latent variables it can be computed in mere seconds on commodity hardware. To compute the M-estimator, we propose an extended EM-algorithm, combined with a stochastic approximation algorithm, leading to a computational burden that is linear in npq . We also derive its statistical properties, establish the convergence of the associated algorithm, and provide an accompanying R package.

Using cross-validation for parameter selection in cubic-spline postsmoothing

Tuesday, 25th July - 15:30: Computational Methods for Latent Variable Models (Benjamin Banneker) - Oral

Dr. Stella Kim (University of North Carolina at Charlotte), Dr. Hwanggyu Lim (Graduate Management Admission Council), Ms. Yeonwho Kim (Seoul National University), Prof. Won-Chan Lee (University of Iowa)

Equating is a statistical process used to adjust scores from one test form to make them comparable to scores from another form. Conducting equating does not necessarily require the use of smoothing processes. However, previous research has indicated the improved accuracy of equating results with the aid of smoothing (Kolen & Brennan, 2014). A smoothing technique is used to reduce the amount of random error in equating, with the aim that the reduction of random error is greater than the potential systematic error it may introduce. Of many smoothing procedures, cubic-spline postsmoothing is one of the most frequently used methods in practice[LWC1] .

One challenge in using cubic-spline postsmoothing is the lack of statistics that can be used to determine a smoothing parameter, leading investigators to rely primarily on subjective judgment based on the examination of fitted plots. This subjective approach raises concerns about the accuracy of parameter selection, as different investigators may make different choices. To address this issue, the current study proposes the use of cross-validation methods to assist in the selection of the smoothing parameter. The leave-one-out cross-validation error (LOOCV) is used to evaluate each level of smoothing parameters, and the model with the lowest LOOCV will be selected as a final solution. A set of simulation studies will be carried out to examine the effectiveness of the proposed method.

A novel framework of diagnostic classification model for multiple-choice items and a simulation study.

Tuesday, 25th July - 14:30: Diagnostic Classification Models (Prince George) - Oral

Mr. Kentaro Fukushima (The University of Tokyo), Dr. Kensuke Okada (The University of Tokyo)

Polytomous responses provide rich information on the levels of learners. Some recent studies on Diagnostic Classification Models (DCMs) aim to take advantage of their characteristics for efficient diagnosis. For example, selected distractors in multiple-choice items may reflect insufficient learning of the examinees. Several models have been psed to diagnose more accurate learning proficiency from the information. This study proposes a novel framework of DCMs for multiple-choice items, Multiple-Choice Log-linear Cognitive Diagnostic Model (MC-LCDM). This framework expresses response probabilities as the main effects and interactions of attribute mastery profile and Q-vectors representing each option's character. Existing DCMs for multiple-choice items can be described as sub-models of the proposed framework, which lets us compare the models from a unified perspective. After a theoretical discussion of the nature of the sub-models, we examine adequate models for some situations through a simulation study.

Restricted HMM for latent class attribute transitions

Tuesday, 25th July - 14:45: Diagnostic Classification Models (Prince George) - Oral

Mr. Theren Williams (University of Illinois Urbana-Champaign), Prof. Steven Culpepper (University of Illinois Urbana-Champaign), Dr. Yuguo Chen (University of Illinois Urbana-Champaign)

throughout the multitude of fields with a vested interest in underlying attribute and skill profiles, a wide variety of formulations of cognitive diagnostic models (CDMs) exist for analyzing and understanding these profiles. one popular application across fields, such as education and the social sciences, has respondents record information over time, for example, pre/post-testing scenarios. in these cases of longitudinal polytomous data, constructing CDMs requires special care to account for potential transitions between attribute profiles. in some cases, shifts may be characterized freely or constrained if some profiles cannot transition between one another; for example, an expert cannot shift to novice in most scenarios. recent efforts have investigated the generic identifiability of hidden markov model (HMM) constructions for the deterministic inputs, noisy “and” gate (DINA) model. this project builds upon these advances, examining the HMM structure within a more general restricted latent class model. further, we discuss identifiability given restrictions on the attribute transition matrix, constraining which attributes may transition to which others. additionally, via markov chain monte carlo (MCMC) simulations, we provide evidence supporting claims of adequate parameter estimation.

Higher order personalized slip tendency model for cognitive diagnosis

Tuesday, 25th July - 15:00: Diagnostic Classification Models (Prince George) - Oral

Ms. Yunting Liu (University of California, Berkeley), Dr. Hongyun Liu (Beijing Normal University)

Several ways are proposed to identify the “noisy” input in Psychometrics model – asymptotes value in Item Response Theory, slip and guess parameters in Cognitive diagnosis model. However, neither of them was able to elucidate the reason how it occurs in the cognitive process.

Considering the two postulated components in reasoning from cognitive psychology, the improvements in terms of goodness of fit, the diagnostic value of personalized information, a new model was put forward by setting the item “slip” parameter on individual level, changing fixed effect on item into random effect term, named Higher Order Deterministic Input Personalized Noisy Output “And” gate model (HO-DIPNA). In this way the slip tendency can be measured for diagnostic purpose.

The Metropolis-Hastings algorithm with Gibbs approach was applied for parameter estimation. The simulation result indicated that when the data were generated from Higher Order the deterministic inputs, noisy “and” gate model (HODINA, de la Torre and Douglas, 2004), using both models to fit the data result in similar parameter recovery. However, when the data were generated from HO-DIPNA, using HO-DIPNA resulted in unbiased estimates, whereas using the HODINA result in biased estimates. Two empirical example (fraction subtraction data and cognitive test data) was illustrated in which three models were compared: the deterministic inputs, noisy “and” gate (DINA) model, the HODINA and the HO-DIPNA. Real data shows that when item quality is high, the need for a random effects formulation becomes clear, more diagnostic information can also be gained through using the HO-DIPNA model.

Q-matrix identification using text classification: TF-IDF and word embedding

Tuesday, 25th July - 15:15: Diagnostic Classification Models (Prince George) - Oral

Mr. Yuxiao Zhang (Purdue University, West Lafayette), Mr. David Arthur (Purdue University, West Lafayette), Ms. Xiyu Wang (Purdue University, West Lafayette), Prof. Hua Hua Chang (Purdue University, West Lafayette)

The Cognitive Diagnosis Models (CDMs) are highly useful in educational assessment, providing fine-grained information on students' skill mastery. However, the lack of a Q-matrix which defines the relationships between test items and the latent attributes limited the practical application of the CDMs. Constructing a Q-matrix can be time-consuming and resource-intensive, as it requires expertise in the subject matter and an understanding of the cognitive process involved.

This study proposes the use of Text Classification (TC) in machine learning to automatically identify the Q-matrix. The approach involves extracting two features from questions for classifying them regarding the attributes being measured: Term Frequency – Inverse Document Frequency (TF-IDF) and word embeddings. TF-IDF is a statistical measure evaluating the importance of a word in a question, while word embeddings are numerical representations of words in high-dimensional vector space. The two features are combined to produce a single vector representation for each question, which serves as input for machine learning classification algorithms. Three classifiers (K-nearest neighbours, logistic regression, and support vector machine) are used to implement the classification for each attribute separately. The proposed approach is evaluated on a sample of 300 college physics items, and the recall, precision, and F1-measure are calculated to assess its effectiveness. TC has achieved high level of accuracy in classifying questions for other purposes in previous studies (e.g., Mohammed & Omar, 2020). The proposed approach has the potential to significantly reduce the resources required for Q-matrix construction and can improve the application of CDMs in educational assessment

Omitted response treatment using a modified Laplace smoothing for approximate Bayesian inference in item response theory

Tuesday, 25th July - 14:30: Item Response Theory (Margaret Brent) - Oral

Dr. Matthias von Davier (Boston College)

This presentation applies the approach of adding artificially created data to observations to stabilize estimates to the problem of treating missing responses for cases in which students choose to omit answers to questionnaire items or achievement test items.

This addition of manufactured data is known in the literature as Laplace smoothing or the method of data augmentation priors. It can be understood as a penalty added to a parameter's likelihood function. This approach is used to stabilize results in the National Assessment of Educational Progress (NAEP) analysis and implemented in the MGROUP software program, which is important in generating results in the form of plausible values (PVs) for NAEP.

The modified data augmentation approach presented here aims to replace common missing data treatments used in IRT so it can be understood as special deterministic cases of data augmentation priors that add fixed information to the observed data, either by conceptualizing these as adding a fixed form to the likelihood function to constant represent prior information or by understanding the augmentation as a conjugate prior that 'emulates' non-random observations.

Asymptotic standard errors of model-based oral reading fluency score equating

Tuesday, 25th July - 14:45: Item Response Theory (Margaret Brent) - Oral

Dr. Xin Qiao (Southern Methodist University), Dr. Akihito Kamata (Southern Methodist University), Dr. Cornelis Potgieter (Texas Christian University)

Oral reading fluency (ORF) assessments are commonly used to screen at-risk readers and to evaluate the effectiveness of interventions as curriculum-based measurements. Same as other assessments, equating ORF scores becomes necessary when we want to compare ORF scores from different test forms. Recently, Kara et al. (2023) proposed a model-based equating method for ORF scores. However, they did not provide closed-form asymptotic standard errors (SEs) of the equated ORF scores while it is advocated to report SEs of equating in practice. Therefore, we aim to address this remaining question in this study. Specifically, we adopt the delta method to derive the asymptotic SEs of equated ORF scores. Delta method is a general method to calculate SEs for transformed parameters that are functions of some other parameters with known asymptotic variances. In the current study, ORF score is a function of passage parameters and latent variables. Then, we conduct a simulation study to evaluate the recovery of derived equating SEs in various simulated conditions. The empirical standard deviation of SE estimates across replications is used as the criterion SE in each simulation condition. Results suggest both accuracy and precision of the derived asymptotic SEs of equated ORF scores.

Factors impacting high dimensional graded response model calibration

Tuesday, 25th July - 15:00: Item Response Theory (Margaret Brent) - Oral

Mr. Kenneth McClure (University of Notre Dame), Dr. Ross Jacobucci (University of Notre Dame)

Psychological researchers are fitting larger factor models, particularly in clinical psychology (e.g., HiTOP). Instruments assessing these constructs are frequently measured using Likert type scale for which the multidimensional graded response model (MGRM; Muraki & Carlson, 1993) is appropriate. Estimation of item parameters in multidimensional models is computationally challenging using marginal likelihood methods, particularly for large numbers of latent variables; however, the Metropolis-Hastings Robbins-Monro (MH-RM) estimation (Cai, 2010) provides a promising alternative for high-dimensional models.

The current study examines the impact of sample size, test length, and characteristics of the underlying latent traits on item parameter recovery in high dimensional (5, 7, and 9 dimensions) MGRMs with MH-RM estimation. In addition to sample size, test length, and trait dimensionality, the influence of non-normal trait distributions are examined by varying skewness and excess kurtosis of marginal trait distributions using non-gaussian copula methods. Correlation between test dimensions is also varied.

Results suggest that samples of 1000 – 2500 are often sufficient to obtain accurate estimates; however, large tests with highly correlated dimensions exhibit poor recovery across sample sizes. Large deviations from normality also impede recovery; increases in kurtosis impact discrimination and skewness influences boundary estimates. This is the first study to examine MH-RM estimation for item parameter calibration in the MGRM and to explicitly study the role of trait kurtosis and skewness in item calibration. Findings provide vital information for researchers developing item banks or adaptive testing procedures for correlated latent trait models. An R tool will be developed to recommend sample size.

Asymptotic standard errors of equating coefficients for non-parametric ability distribution

Tuesday, 25th July - 15:15: Item Response Theory (Margaret Brent) - Oral

Dr. Ikko Kawahashi (Meiji Gakuin University)

Over the past two decades, the derivation of asymptotic standard errors of equating coefficients has been a key topic in the fundamental study of the Item Response Theory and is still actively discussed. Therefore, the current study derived asymptotic standard errors of equating coefficients for the common-examinee design. Unlike Ogasawara (2002), the ability distribution is denoted by a nonparametric function. We assumed that all item parameters in the two-parameter logistic model are known and that the equating coefficients are derived from the expectation and variances of the ability distributions of common-examinee in the two anchor tests. First, we applied Mislevy's (1984) approach of calculating the asymptotic covariance matrix corresponding to the expectation and variance of the nonparametric ability distribution to the item response matrix collected in the common-examinee design. Using the matrix, then, we used the delta approach to calculate the asymptotic standard errors of the equating coefficients. Moreover, a parametric approach based on a normal distribution was developed for comparison. According to the simulation study, when the ability distribution is highly skewed and the difference between the two scales is significant (A = 1.2 and B = −0.5), the parametric method estimates of the equating coefficients are biased and the asymptotic standard errors are underestimated, whereas the nonparametric methods have small biases in the estimates but need a large sample size (>3000) for the asymptotic standard error to approximate the true value.

IRTree models of co-occurring dominance and ideal point response processes

Tuesday, 25th July - 15:30: Item Response Theory (Margaret Brent) - Oral

Ms. Viola Merhof (University of Mannheim), Prof. Thorsten Meiser (University of Mannheim)

Item responding is a multidimensional process, since not only the substantive trait being measured, but also additional personal characteristics can affect how respondents select rating scale categories.A flexible model class for analyzing such multidimensional responses are IRTree models, in which rating responses are decomposed into a sequence of sub-decisions. Different response processes can be involved in item responding both sequentially across those sub-decisions and as co-occurring processes within sub-decisions. In the previous literature, modeling co-occurring processes has been exclusively limited to dominance models, in which higher trait levels are associated with higher expected scores. However, some response processes may rather follow an ideal point rationale, in which the expected score depends on the proximity of a person's trait level (i.e., his/her ideal point) and the item's location. Therefore, we propose a general multidimensional IRT model, in which the co-occurrence of multiple dominance processes, multiple ideal point processes, as well as a combination of both is modeled in a consistent way. IRTree models parameterized by this new approach allow multiple dominance and ideal point processes to be involved in the response selection sequentially across sub-decisions and simultaneously within sub-decisions. Simulation analyses revealed good parameter recovery and a clear advantage of IRTree models with the new parameterization compared to traditional ones. Two application examples from the field of response style analysis demonstrated the benefits of such IRTree models under real-world conditions.

Bayesian stacking in multilevel models

Tuesday, 25th July - 14:30: Mixed-Effects and Multilevel Models (Juan Ramon Jimenez) - Oral

Ms. Mingya Huang (University of Wisconsin-Madison), Prof. David Kaplan (University of Wisconsin-Madison)

The issue of model uncertainty has been gaining interest in education and the social sciences community over the years and the dominant methods for handling model uncertainty are based on Bayesian inference by each model's posterior model probability. However, Bayesian model averaging seeks a single best model which assumes that the true generating model is in the model space. Unlike Bayesian model averaging, the method of Bayesian stacking can account for model uncertainty without assuming that a true model exists. An issue with Bayesian stacking, however, is that it is an optimization technique that uses input-independent model weights and is, therefore, not fully Bayesian. Bayesian hierarchical stacking, proposed by Yao, Pirš, Vehtari, and Gelman (2021), incorporates uncertainty by applying a hyperprior to the stacking weights. Considering the importance of multilevel models commonly applied in educational settings, this paper investigates the predictive performance of Bayesian stacking and Bayesian hierarchical along with two other readily available methods based on Bayesian model averaging weighting via a simulation study and a real data example using data from PISA 2018. Predictive performance is measured by the log predictive density and the Kullback-Leibler divergence scores. Results suggest that Bayesian hierarchical stacking performs consistently better than original Bayesian stacking and other BMA-based weighting systems in terms of predictive performance in multilevel models.

On generating plausible values for multilevel modeling in large-scale assessments

Tuesday, 25th July - 14:45: Mixed-Effects and Multilevel Models (Juan Ramon Jimenez) - Oral

Dr. Xiaying Zheng (American Institutes for Research)

Large-scale assessments (LSAs) routinely employ latent regressions to generate plausible values (PVs; Mislevy, 1991) for unbiased estimation of the relationships between examinees' background variables and performance. LSA data often have a nested structure (e.g., students within schools) due to the complex sample design. To handle the clustering effect, there has been an increased interest in using PVs for multilevel modeling. However, most LSAs use a single-level latent regression model, the PVs from which may not support unbiased multilevel modeling. It is operationally impractical to conduct a true multilevel latent regression in LSAs. Instead, two single-level methods have been used in LSAs to create PVs for multilevel modeling. The first is to include dummy-coded cluster identifiers in the single-level regression to estimate a fixed effect (FE) for each cluster (e.g., in PISA). The second is to include the cluster-level interim means (IM) aggregated from the unconditioned examinee EAP scores (e.g., in TIMSS). Via stimulations, this research evaluates these two single-level methods that aim to address the multilevel modeling concerns. Multilevel response data and covariates were generated using a random-intercept and -slope model. Analyses of the yielded PVs showed that both the FE and IM methods could produce PVs suitable for random-intercept models only, but not for random-slope models. In light of the results, we proposed an extension to the IM method to include cluster-specific interim slopes in the latent regression. We demonstrate in a second simulation study that the proposed method could provide PVs suitable for a random-intercept and random-slope model.

Efficient additive Gaussian process models for large-scale balanced multi-level data

Tuesday, 25th July - 15:00: Mixed-Effects and Multilevel Models (Juan Ramon Jimenez) - Oral

Ms. Sahoko Ishida (London School of Economics and Political Science), Prof. Wicher Bergsma (London School of Economics and Political Science)

Gaussian process (GP) regression is a statistical learning method that models the underlying regression function as a GP. The covariance function of a GP, known as the kernel, plays a crucial role in determining the relationship between the predictors and the response variable, and different kernel choices allow for flexibility in modelling. While there is extensive literature on GP models in machine learning communities, the aspect of statistical modelling, such as the inclusion and interpretation of interaction effects, has received much less attention. Our approach to these issues builds on recent research in additive GP models. More specifically we use a general class of ANOVA decomposition kernel for constructing the covariance function. One of the primary hurdles to the practical application of such models is their computational complexity. Large-scale datasets encountered in social, behavioural, and medical science applications are often deemed prohibitive in GP models. In this talk, we focus on the cases where the datasets have a balanced multi-level structure, and we use Kronecker methods to enable efficient implementation. The use of this scalable method in the previous literature has been limited to when the covariance function has a so-called tensor product structure, which implies a model with only the highest-order interaction term. We extend the method to a more general class of additive GP models. Our approach yields significant improvement in computational time, enabling the analysis of large-scale real-world datasets. We also discuss some extensions of the methods, including the handling of missing values and multivariate responses.

Evaluation of factors impacting predictor importance results in multilevel models

Tuesday, 25th July - 15:15: Mixed-Effects and Multilevel Models (Juan Ramon Jimenez) - Oral

Ms. Soonhwa Paek (University of Wisconsin - Milwaukee), Prof. Razia Azen (University of Wisconsin - Milwaukee)

Dominance Analysis (DA) was originally proposed to determine the relative importance of predictor variables in OLS regression models by comparing the change in model fit (i.e., R^2) resulting from adding each predictor to each possible subset model (Azen & Budescu, 2003; Budescu, 1993). The DA procedure has not been studied extensively with Multilevel Linear Models (MLMs), which are commonly used to analyze nested data structures, and the focus of this study is to examine how DA can be extended to determine predictor importance in MLMs. A Monte Carlo simulation study will be conducted with various MLMs to compare and evaluate DA results across several simulation conditions. Conditions will include various sample sizes and predictor effects, the way in which the level-1 predictors are centered (i.e., grand- or cluster-mean centering), the multilevel measure of fit used to determine the change in model fit (e.g., the variance component explained, as discussed by Rights and Sterba, 2019), and the type of MLM used (e.g., its random effects).

The DA rank ordering of predictors will be evaluated using the Kendall rank correlation (Kendall, 1955). Simple random sampling and multilevel bootstrapping will be used to evaluate inference for the quantitative general dominance measure (e.g., bias, RMSE, coverage of 95% CIs, Type I error and power rates) and make recommendations for applied research.

This study will contribute useful information for applying DA in MLM studies, and will also provide a demonstration of how to apply the procedure using an empirical dataset with R code.

Best practices for centering categorical predictors in multilevel models

Tuesday, 25th July - 15:30: Mixed-Effects and Multilevel Models (Juan Ramon Jimenez) - Oral

Ms. Haley Yaremych (Vanderbilt University), Dr. Kristopher Preacher (Vanderbilt University), Dr. Donald Hedeker (University of Chicago)

The topic of centering in multilevel modeling (MLM) has received substantial attention from methodologists, as different centering choices for lower-level predictors present important ramifications for the estimation and interpretation of model parameters. However, the centering literature has focused almost exclusively on continuous predictors, with little attention paid to whether and how categorical predictors should be centered, despite their ubiquity across applied fields. Algebraically and statistically, continuous and categorical predictors behave the same, but researchers using them do not, and for many, interpreting the effects of categorical predictors is not intuitive. Thus, the goals of this presentation are twofold: to clarify why and how categorical predictors should be centered in MLM, and to explain how multilevel regression coefficients resulting from centered categorical predictors should be interpreted. I will first provide algebraic support showing that uncentered coding variables used in isolation yield a conflated blend of within- and between-cluster effects of a multicategorical predictor, whereas appropriate centering techniques yield level-specific effects. Next, I will review algebraic derivations that illuminate precisely how the within- and between-cluster effects of a multicategorical predictor should be interpreted under dummy, contrast, and effect coding schemes. All conclusions will be demonstrated with an empirical example. Implications for practice and opportunities for future research, including relevance of these findings for multilevel structural equation models, will be discussed.

Validation of the household food security survey module using confirmatory factor analysis and Rasch modeling

Tuesday, 25th July - 14:30: Measurement Applications (Thurgood Marshall) - Oral

Ms. Jing Li (Universality of Georgia), Prof. Seock-Ho Kim (University of Georgia), Prof. George engelhard (university of Georgia)

The Household Food Security Survey Module (HFSSM), as part of the Current Population Survey Food Security Supplement (CPS-FSS), is the main tool used in the U.S. to measure food security conditions (Bickel et al., 2000). The purpose of this study is to apply Confirmatory Factor Analysis and Rasch Modeling to HFSSM with the goal of achieving four objectives: (a) to verify its construct validity by examining the factor structure; (b) to evaluate the items' effectiveness in measuring the continuum of household food insecurity; (c) to use Differential Item Functioning analysis (DIF) to identify any variations in how certain subgroups of households respond to the items; (d) to compare the outcomes of both models to establish their consistency. The research is based on data collected from 7,324 low-income households with children who completed the HFSSM between 2012 and 2014. The study's findings will reveal whether the 18 items of HFSSM significantly contribute to assessing household food security as a unidimensional construct. The methods used in this study can be applied to validate and assess scales in various other areas.

Examining the measurement invariance of the Chinese Short Grit Scale

Tuesday, 25th July - 14:45: Measurement Applications (Thurgood Marshall) - Oral

Ms. Roti Chakraborty (Georgia State University)

This study examined the measurement invariance of the Chinese version of the Short Grit Scale (Grit-S) Questionnaire to determine if it accurately measured the same latent variables across groups. Grit-S is a self-report questionnaire of an 8-item scale in which four positively-worded items construct the CI (Consistency of interests) sub-scale, and four negatively-worded items are the PE (Perseverance of effort) sub-scale. We collected data from 2749 students taking Chinese language (2444 middle and 305 high school students) and 1260 students taking Math (817 middle and 443 high school students) online classes at 12 secondary schools in China during COVID-19, in June 2020. Measurement invariance was tested by Confirmatory Factor Analysis using the R Lavaan package. CI subscale was considered as Factor 1, PE subscale as Factor 2, middle school students as Group 1, and high school students as Group 2. A good configural invariance model fit was found for the Chinese language (CFI = .98, RMSEA = .063, SRMR = .048) and Math (CFI = .96, RMSEA = .064, SRMR = .067) classes. However, the metric invariance model showed that the chi-square test was significant for both the Chinese language (χ^2 = 276.90, *df* = 44, *p* = 0.016) and Math (χ^2 = 227.53, *df* = 44, *p* = 0.024) classes. The factor loadings of the items were not constrained equally and the items did not have the same meaning to students across middle and high schools. Next, the partial invariance will be tested to crosscheck the initial findings.

A shortened Positive and Negative Symptom Scale (PANSS): Harmonizing classical item response theory with the perspectives from network approach

Tuesday, 25th July - 15:00: Measurement Applications (Thurgood Marshall) - Oral

Dr. Jinyuan Liu (Vanderbilt University), Dr. Lénie Torregrossa (Vanderbilt University), Dr. Kristan Armstrong (Vanderbilt University), Dr. Brandee Feola (Vanderbilt University), Dr. Alexandra Moussa-Tooks (Vanderbilt University), Dr. Stephan Heckers (Vanderbilt University)

Schizophrenia is a complex, heterogeneous behavioral and cognitive syndrome with profound impacts on society. As a “gold standard,” the Positive and Negative Syndrome Scale (PANSS) is a 30-item rating scale to assess the dimensions of schizophrenia symptoms. Despite being proposed to constitute three subscales measuring positive and negative syndromes, and general psychopathology, growing studies report an unsatisfactory fit of this original three-factor model. Moreover, recent studies have confirmed the feasibility of a shortened version of PANSS as an alternative regulatory endpoint under the randomized clinical trial (RCT) setting. Here we use a sample of 362 schizophrenia patients aged 13 to 65 years from the Psychiatric Genotype/Phenotype Project (PGPP) longitudinal study to 1) reevaluate the factor structure of the PANSS and 2) harmonize the item response theory (IRT) and recent network approach to identify the best-performing items and derive a shortened PANSS scale. Compared with previous studies, this new scale is established beyond the strict inclusion and exclusion criteria in RCTs, hence providing better representations for patients in clinical practice. PANSS items were assessed in detail from two perspectives to assist the decision-making in selecting functional items. Specifically, besides the classical IRT, the newly developed centrality measures in the network model were accommodated into the selection criteria. Therefore, balancing considerations of internal factor structure, content validity, and expert judgment, this newly proposed shortened scale has the potential to reduce the assessment burden while being validated to be functional, internally consistent, and parsimonious.

The impact of teachers' instructional methods on the reading achievement of resilient students

Tuesday, 25th July - 15:15: Measurement Applications (Thurgood Marshall) - Oral

Ms. Qile Liu (Faculty of Education, University of Macau), Prof. Fu Chen (Faculty of Education, University of Macau), Dr. Biao Zeng (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University)

Teachers are known to play a significant role in promoting resilience among students. Enhancing teachers' instructional methods can lead to improved reading proficiency among students. However, the link between teachers' instruction and the reading proficiency of resilient students remains unclear, and few studies have investigated the mechanism of varying teacher instruction affecting resilient students' reading achievement. The distinctive characteristics of resilient students during the learning process warrant further investigation.
Based on the PISA 2018 data of 625 resilient students and 3,150 non-resilient students in Macao, China, this study explored the varying effects of teachers' instructional methods on the reading achievement of resilient students and non-resilient students. Results show that reading interest completely mediates the relations of teacher instruction, students' perceived teacher interest, and teacher stimulation for reading engagement to the reading achievement of resilient students. In addition, mastery goal orientation completely mediates the relationships between the reading achievement of resilient students and predictors of teachers' instruction and students' perceived teacher interest. However, the mediation effect of mastery goal orientation was not observed for non-resilient students. Moreover, students perceived cooperative atmosphere completely mediates the relation of students' perceived teacher interest to the reading achievement of non-resilient students.
The findings suggest that teachers' instructional strategies in reading should be adapted in accordance with students' resilience levels. This study highlights the importance of recognizing the unique needs of resilient students and tailoring instructional methods accordingly.

Assessing raters rating quality under holistic and analytic scoring schemes in writing assessment. An empirical example.

Tuesday, 25th July - 15:30: Measurement Applications (Thurgood Marshall) - Oral

Dr. Diego Carrasco (Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile), Dr. Natalia Ávila (Facultad de Educación, Pontificia Universidad Católica de Chile), Ms. Carolina Castillo (Facultad de Educación, Pontificia Universidad Católica de Chile), Dr. Rosario Escribano (Facultad de Educación, Pontificia Universidad Católica de Chile), Dr. María Jesús Espinosa Aguirre (Universidad Diego Portales), Dr. Javiera Figueroa Millares (Universidad Alberto Hurtado)

Holistic rubrics are favored in writing assessment when writing quality is to be judge (White, 1984). Analytic scoring is often dismissed in the holistic tradition, due to its over-reliance on superficial characteristics of written samples, including indicators such as word complexity, or spelling (e.g., Hamp-Lyons, 2016), instead of writing communicative attributes. However, "analytic rubrics" in its more descriptive meaning refers to scoring methods using two or more indicators (Frey, 2018). These different views on rubrics, regarding what is holistic or analytic, confound scoring methods with writing assessment approaches. In the present work, we compare holistic and analytic rubrics designed with a common writing assessment approach. We designed two rubrics where writing quality is judged in terms of the fulfillment of the communicative purpose. We recruit 88 raters to rate 30 written samples of 5th graders using the two designed rubrics. Raters were trained in a rubric, and then rate written samples. The order was randomly assigned (holistic-analytic, vs analytic-holistic). We assess raters rating quality using variance partition and rater accuracy models (Engelhard et al, 2018). We fit cross-classified models, with students and raters as random terms of total scores for each rubric and found no substantial difference between the two scoring methods. We use generalized cross-classified models to compare raters' accuracy in contrast to expert raters' responses. We found that observed raters presented higher accuracy under the analytic (five ordinal indicator) rubric, than with the holistic rubric. We discussed the importance of distinguishing methods and traditions in writing assessments development.

A latent hidden Markov model for process data

Wednesday, 26th July - 09:00: Process Data (Atrium) - Oral

Dr. Xueying Tang (University of Arizona)

Response process data from computer-based problem-solving items describe respondents' problem-solving processes as sequences of actions. Such data provide a valuable source for understanding respondents' problem-solving behaviors. Recently, feature extraction methods have been developed to compress the information in unstructured process data into relatively low-dimensional features. Although the extracted features can be used as surrogates for quantifying the variations of response processes, the results are often not easily interpretable as each extracted feature is usually a complicated function of the original response process. In this paper, we propose a statistical model for describing response processes and how they vary across respondents. The proposed model assumes a response process follows a hidden Markov model given the respondent's latent traits. The structure of hidden Markov models resembles problem-solving processes, with the hidden states interpreted as problem-solving subtasks or stages. Incorporating the latent traits in hidden Markov models enables us to characterize the heterogeneity of response processes across respondents in a parsimonious and interpretable way. We demonstrate the performance of the proposed model through a case study of PISA process data and simulation studies.

Leveraging process data and variable selection for TIMSS achievement estimation

Wednesday, 26th July - 09:15: Process Data (Atrium) - Oral

Ms. Dihao Leng (Boston College), Dr. Ummugul Bezirhan (Boston College), Dr. Matthias von Davier (Boston College)

The transition from paper-based to computer-based assessments in prominent international large-scale assessments (ILSAs) has resulted in the availability of process data, which are derived from raw log data capturing student-computer interactions. However, these data have been underutilized in operational scaling.

The conventional scaling method in ILSAs typically involves the use of item response theory and latent regression models. Prior to implementing latent regression modeling, principal component analyses (PCA) are usually performed to reduce the number of contextual variables and retain only principal components (PCs) that account for a significant proportion (e.g., 80% or 90%) of the variance to avoid overparameterization. Nonetheless, the number of remaining PCs is still often larger than desired and can lead to overfitting, thereby jeopardizing numerical stability.

This study investigated the potential improvement in estimating group-level achievement in TIMSS 2019 by incorporating process variables (i.e. speed estimated from lognormal models and revisit variables) into the latent regression models. Moreover, this study explored whether alternative variable selection approaches including LASSO, random forests, and gradient boosting can lead to measurement precision similar to that of the traditional PCA approach but with fewer covariates.

The results show that using variable selection with process data before latent regression yielded the highest measurement precision. Specifically, the latent regression models with the 150 most important predictors (including process variables) outperformed the one with 330 PCs in country A and 270 PCs in country B. The utility of process data and variable selection in operational ILSAs scaling is further discussed.

Supporting the process evaluation of assessing others' skills: A content analysis using deep learning

Wednesday, 26th July - 09:30: Process Data (Atrium) - Oral

Dr. Xue Wang (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University), Prof. Sheng Zhang (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University)

The importance of learning as a peer assessor is receiving increasing focus in peer assessment, the desire to evaluate and track the development of students' skills of assessing others has led researchers to establish analysis framework based on process text. However, with the growing number of peer assessment, it becomes a challenge for instructors to provide good quality feedback timely. In this context, this paper proposed a content analysis of feedback text from cognition, emotion, standard and metacognition dimensions. ALBERT and ALBERT-Seq2Seq-Attention classifiers were trained and evaluated at single label and multiple labels dimensions. Through comparison with several classical and baseline models, such as CNN、LSTM、FastText、RCNN、ALBERT-CNN , the results achieved a higher accuracy ranged from 78.88% to 97.73% in each dimension. The prediction consistency in some dimensions is higher than the manual double-coding consistency. The paper has promoted the cultivation of students' skills of assessing others and the development of unstructured process data analysis.

Process data enhanced diagnostic measurement using topic modeling on constructed responses

Wednesday, 26th July - 09:45: Process Data (Atrium) - Oral

Ms. Constanza Mardones-Segovia (University of Georgia), Dr. Jiawei Xiong (Pearson), Dr. Allan Cohen (University of Georgia)

Interactive computer-based assessments provide a wealth of data collected during the examinees' assessment process that can be used to improve estimates of student ability in item response theory (He & von Davier, 2016). One potential source of data is the textual responses generated by examinees in response to constructed-response (CR) items, which can be analyzed using topic modeling (Blei et al., 2003) techniques (Wheeler et al., 2022). Traditional diagnostic classification models (DCMs; Templin & Bradshaw, 2013) have mainly focused on item scores and have also included process data to estimate the attribute mastery probabilities (Zhan & Qiao, 2022). However, DCMs have not yet used the richness found in students' answers to CR items to improve the parameter estimates. To address this gap and investigate whether information from CR answers can help understand the diagnostic measurement of students' responses, this study proposes a process-based regression method that integrates the DCMs and process data from the CR answers to analyze students' attribute mastery status. Specifically, this method models students' attribute mastery probabilities using a DCM and incorporates information from topic modeling to extract topic proportion from the textual responses. The resulting probabilities are then regressed on the responses and topic proportion, yielding process-based attribute mastery probabilities. Empirical results indicate that the proposed approach is comparable in accuracy to traditional DCMs and provides a meaningful interpretation of students' mastery status.

The nonparametric item selection method for multiple-choice items in CD-CAT

Wednesday, 26th July - 09:00: CD-CAT (Benjamin Banneker) - Oral

Ms. Yu Wang (University of Minnesota - Twin Cities), Prof. Chia-Yi Chiu (University of Minnesota - Twin Cities)

The multiple-choice (MC) item format has been used extensively in education and allows for collecting rich diagnostic information. This format has also been adopted to the cognitive diagnosis (CD) framework. However, early approaches simply dichotomized responses and analyzed them with a CD model developed for binary responses, which limited the use of the additional diagnostic information provided by MC items. De la Torre's (2009) Multiple-Choice Deterministic Inputs, Noisy "And" Gate (MC-DINA) model was the first for the explicit analysis of MC items. However, the model has limitations such as restricted attribute vectors of the distractors and computational difficulties with small sample sizes. In response to these obstacles, the nonparametric classification method for MC items (MC-NPC; Wang, Chiu, & Köhn, 2022) was proposed and shown to outperform the MC-DINA method and other methods for binary responses. In this study, we go one step further to develop a nonparametric item selection method for MC items (MC-NPS) by combining the MC-NPC and the general nonparametric item selection (GNPS; Chiu, & Chang, 2021) methods. Additionally, the Q-optimal criterion (Xu, Wang, & Shang, 2016) for binary responses was extended to the MC items, and the new Q-optimal criterion for MC items was implemented in the MC-NPS method to determine the initial items. The preliminary study shows that the proposed method outperforms the GNPS method for dichotomous items and achieves high agreement rates quickly.

CD□CAT for the extended multiple-choice DINA model

Wednesday, 26th July - 09:15: CD-CAT (Benjamin Banneker) - Oral

Prof. Jimmy de la Torre (The University of Hong Kong), Mr. Zechu Feng (The University of Hong Kong)

In addition to its flexibility to accommodate misconceptions, the extended multiple-choice deterministic inputs, noisy “and” gate (eMC-DINA) model has been shown to outperform cognitive diagnosis models (CDMs) that partially (i.e., MC-DINA model) or wholly (i.e., generalized DINA model) ignore distractors designed to measure partial knowledge or incongruous skills when non-adaptive tests are involved. This paper investigates the cognitive diagnosis computerized adaptive testing (CD-CAT) implementation of the eMC-DINA model using the Jensen–Shannon divergence (JSD) index as the item selection criterion. A simulation study is designed to compare the CD-CAT performance of the eMC-DINA model against those of CDMs that do not or only partially utilize distractors designed to be diagnostically informative. Moreover, the efficiencies of the adaptive and non-adaptive implementations of the eMC-DINA models are compared. In this study, a number of factors, which include item quality, distractor quality, and stopping criterion are manipulated. Preliminary results show that, when the test is short, the eMC-DINA model is 20% and 10% more efficient than the MC-DINA model for attribute profile and individual attribute classification accuracies, respectively; under the same context, the eMC-DINA model is 42% and 17% more accurate than the G-DINA model. As expected, the classification accuracies become more similar with longer tests. Finally, the adaptive implementation of the eMC-DINA model is close to 60% more efficient than its non-adaptive counterpart.

Termination rules for hierarchical cognitive diagnosis computerized adaptive testing

Wednesday, 26th July - 09:30: CD-CAT (Benjamin Banneker) - Oral

Dr. Ya-Hui Su (Department of Psychology, National Chung Cheng University), Mr. Yen-Ting Chen (Department of Psychology, National Taiwan University)

Computerized Adaptive Testing based on Cognitive Diagnostic Model (CD-CAT; Cheng, 2009; Huebner, 2010) is a popular test format which includes advantages on both cognitive diagnostic models and computerized adaptive testing. The CD-CAT approach not only obtains useful cognitive diagnostic information measured in psychological or educational assessments, but also has great efficiency brought by CAT. The CD-CAT algorithms can be adopted for terminating the fixed-length or fixed-precision CD-CAT (Cheng, 2008; Guo & Zheng, 2019; Hsu et al., 2013; Hsu & Wang, 2015; Tatsuoka, 2002; Wang et al., 2012), but most of these studies have been conducted when the fixed-length termination rule is considered (e.g., Cheng, 2009; Wang et al., 2012), which yields different degrees of measurement precision for different examinees. To achieve nearly the same degrees of measurement precision for examinees, the fixed-precision termination rule can be considered (Tatsuoka, 2002; Hsu et al., 2013; Guo & Zheng, 2019). These previous CD-CAT studies assumed that attributes had a non-hierarchical relationship. In practice, attributes might have a hierarchical relationship, meaning some attributes are a prerequisite for the presence of others (Kuo et al., 2016). However, the fixed-precision termination rules have not been investigated in hierarchical CD-CAT. Therefore, the purpose of this study was to investigate the efficiency of termination rules in the hierarchical CD-CAT when the fixed-precision was considered.

Handling missing data in ecological momentary assessments via later retrieval

Wednesday, 26th July - 09:00: Missing Data (Prince George) - Oral

Dr. Manshu Yang (University of Rhode Island)

The past decade has seen a rapid growth in research using ecological momentary assessments (EMA) to examine how human behavior and experience unfold and interact in real time and in natural context. Intensive longitudinal measurements via EMA maximize ecological validity, mitigate recall bias, and allow the examination of acute momentary factors. However, missing responses in EMAs are almost inevitable, often occur simultaneously in outcomes and covariates, and could be missing not at random, thereby posing a major challenge in data analysis. On the other hand, EMA brings a unique opportunity to address this challenge, by re-prompting participants shortly after they missed an EMA survey to *retrieve* their data, which provides *direct information* on missing responses. The current study compares three methods for handling missing EMA data, including (1) maximum likelihood estimation (MLE), (2) multiple imputation (MI), and (3) fully Bayesian estimation (FBE). All methods were first carried out based on both initially observed and later-retrieved data, and an additional MI was conducted based on later-retrieved data alone. Monte Carlo simulations were conducted to compare the methods given varying sample sizes, missing data patterns, score variability, as well as proportions of initially observed and later retrieved data. Preliminary results suggested that simply combining initially observed and later-retrieved data to conduct MLE or MI could lead to biased estimates, and careful consideration should be given when incorporating later-retrieved information into analysis. Complete findings and implications will be discussed in the presentation.

Evaluating model fit with missing nonnormal data in SEM

Wednesday, 26th July - 09:15: Missing Data (Prince George) - Oral

Dr. Fan Jia (University of California, Merced)

Missing data and nonnormality are two common challenges researchers face in structural equation modeling (SEM). When multiple imputation is used for handling missing data, a pooling approach needs to be considered for model fit evaluation. Under the normality assumption, Li et al. (1991) and Meng and Rubin (1992) proposed two most popular pooling approaches for the test statistic. Research has found that the pooled test statistics and the fit indices based on them were comparable to those from FIML (full information maximum likelihood) when normality holds (e.g., Enders, 2010; Enders & Mansolf, 2018). It is not clear, however, whether these approaches work for missing nonnormal data when robust estimators are employed. In the current simulation study, I examined the overall performance of these approaches in pooling robust test statistics and fit indices across multiply imputed data sets, and their sensitivities to different design conditions. The goal of this study is to provide researchers with guidelines on dealing with missing data and nonnormality at the same time in SEM.

Mixture of missing-data mechanisms in multigroup invariance testing

Wednesday, 26th July - 09:30: Missing Data (Prince George) - Oral

Dr. Young Min Kim (Ohio State University), Dr. Brenna Gomer (Utah State University)

Missing data commonly arise in the social sciences, and, in the context of multigroup comparisons, missing data may occur unequally between the groups. Furthermore, the reasons for missingness itself – the so-called missing-data mechanisms – may also differ between the groups. For example, survey items are not always cross-culturally equivalent with respect to how comfortable participants feel about providing a response. This scenario has not been studied in the literature and thus there are no guidelines on how to handle missing data in such cases. The current study aims to identify the impact of differences of missing-data mechanisms on groups in multigroup equivalence testing when missing data is handled according to two different approaches: 1) the standard practice of handling missingness at the dataset-level; and 2) handling missingness separately within the groups. To address this purpose, we evaluate measurement invariance - weak and strong invariance -in a Monte Carlo simulation study when partial invariance is expected between groups and the missingness in each group can be attributed to a different mechanism. In this talk, we share our results and provide recommendations to researchers regarding the best approach (handling missing data at the dataset vs. group level).

Regularized estimation of the Gaussian graphical model under planned missing data

Wednesday, 26th July - 09:45: Missing Data (Prince George) - Oral

Dr. Carl Falk (McGill University), Mr. Joshua Starr (McGill University)

Many applications of network modeling involve cross-sectional data of symptom prevalence for one or more psychological disorders with analyses conducted using a regularized Gaussian graphical model (GGM). Despite the close relationship between the GGM and covariance structure analysis (CSA), methodology for appropriate handling of missing data is often not applied for the regularized GGM. Many applied researchers are apt to apply listwise deletion which precludes the possibility of a planned missing data design in which all participants may have some missing data. In this research, we compare two approaches to handling missing data for the regularized GGM. The first resembles a two-stage estimation approach whereby a saturated covariance matrix among the items is estimated (e.g., using the expectation-maximization algorithm) prior to estimating the desired regularized GGM. The second estimates a regularized GGM using the expectation-maximization algorithm in a single stage. This latter strategy is thus equivalent to direct maximum likelihood (or full-information maximum likelihood in CSA) using regularization. We compared these approaches in a simulation study that presumes a planned missing data design with a variety of sample sizes, proportions of missing data, and network saturation. Results indicate that the EM algorithm tended to experience fewer estimation problems and had better recovery of true partial correlations at smaller sample sizes and with a higher proportion of missing data. However, both methods tended to yield similar results as sample size increased, supporting the viability of either approach at large sample sizes.

Performance of item-fit test statistics in cognitive diagnosis modeling based on imputed missing data

Wednesday, 26th July - 10:00: Missing Data (Prince George) - Oral

Dr. Kevin Carl Santos (University of the Philippines College of Education)

In low-stakes tests such as cognitively diagnostic assessments, some examinees might not be motivated enough to answer all the questions resulting in missing response data. With incomplete assessment data, the validity of the inferences based on statistical tests for item fit evaluation can be affected. Using the generalized deterministic inputs, noisy “and” gate model framework, this study investigates the impact of the presence of missing data on the performance of four inferential item-fit statistics: the S-X^2test, the likelihood ratio test, the Wald test, and the Lagrange Multiplier test (Sorrel et al., 2017). To minimize their impact, we examine the viability of using single and multiple imputation methods (Zhang & Wang, 2022) first to the incomplete assessment data and apply the four item-fit tests to them. A simulation study manipulating the sample size, attribute correlation structure, test length, item quality and generating cognitive diagnosis models is conducted. Different sources of misfit (e.g., model misspecification, Q-matrix misspecification), misfit proportion, and missing proportion are also examined in the simulation. To evaluate their performance, the Type I error and power are calculated and compared under different conditions.

Diagnosing skills and misconceptions with Bayesian Networks applied to diagnostic MC tests

Wednesday, 26th July - 09:00: Diagnostic Classification Models (Margaret Brent) - Oral

Prof. James Corter (Teachers College Columbia University), Prof. Jihyun Lee (University of New South Wales)

Cognitive scientists investigating STEM achievement have long been interested in the nature of misconceptions and “bugs” in procedural skills. Assessing misconceptions seems educationally important because such assessments might be used to guide individualized instruction, but it has long been recognized as a difficult problem, because misconceptions are not exhibited consistently even by a single examinee. Lee (2003; Lee & Corter, 2003, 2011) demonstrated that in order to assess misconceptions effectively using MC tests, it is necessary to leverage information from incorrect alternative in MC tests designed for that purpose, and that diagnosis of misconceptions or bugs is most stable when bugs ang skills are assessed simultaneously. Lee and Corter proposed using Bayesian Networks as the inference engine to learn from the tests and diagnose individuals. More recently, these ideas have been rediscovered in the context of traditional CDM models. Specifically, models have been proposed that can use distractor information in MC items for diagnostic purposes (de la Torre, 2009), and that enable simultaneous assessment of misconceptions and skills (e.g., Kuo et al., 2018). Finally, two very recent CDM models have been described (Elbulok, 2021; de la Torre, Qiu, & Tjoe, 2022) that can use MC-diagnostic information for simultaneous diagnosis of bugs and subskills. In this paper we describe the approach taken by Lee and Corter in some detail, and discuss practical advantages and disadvantages of the BN and CDM approaches to the problem of diagnosing misconceptions.

A sparse latent class model incorporating response time

Wednesday, 26th July - 09:15: Diagnostic Classification Models (Margaret Brent) - Oral

Ms. Siqi He (University of Illinois Urbana-Champaign), Prof. Steven Culpepper (University of Illinois Urbana-Champaign), Prof. Jeff Douglas (University of Illinois Urbana-Champaign)

The sparse latent class models (SLCMs) have been developed to provide fine-grained classification regarding examinees' latent skills in cognitive ability tests. The inclusion of response times (RTs) allows researchers to evaluate the cognitive theories underlying the test design and better understand examinees' problem-solving behaviors. To date, no study has looked specifically at how response times (RTs) can be incorporated into the exploratory diagnostic models (DMs). This study proposed an SLCM-RT framework with an interrelationship assumed between the RTs and examinees' attribute profiles. A novel Bayesian formulation with variable selection techniques was adopted for parameter estimations. A Monte Carlo simulation study was conducted to evaluate recovery accuracy. In the end, a real data example was reported to demonstrate the method.

Testing the whole testlet: An application of the Mantel-Haenszel statistic

Wednesday, 26th July - 09:30: Diagnostic Classification Models (Margaret Brent) - Oral

Prof. Youn Seon Lim (University of Cincinnati)

Testlets—item bundles sharing a common theme—call into question one of the key statistical assumptions underlying modern psychometric frameworks in educational measurement: local independence of the item responses. (Violations of local independence result in biased parameter estimates and erroneous ability diagnostics.) For cognitive diagnosis assessments, a new tool is proposed for testing an entire testlet at once for the presence of a testlet effect, which relies on a sophisticated procedure based on the Mantel-Haenszel $\chi 2$ statistic. (Existing testing routines can either check only an entire test or pairs of items for the presence of a testlet effect, but not a subset of items all at once. The performance of the proposed diagnostic statistic is evaluated in large-scale simulations, with a focus on the Type I error rate and the power of tests relying on the proposed statistic. An application to a real data set illustrates its usage in practice.

Identifiability of estimated Q-Matrices: Implications for estimation algorithms

Wednesday, 26th July - 09:45: Diagnostic Classification Models (Margaret Brent) - Oral

Ms. Hyunjoo Kim (University of Illinois Urbana-Champaign), Prof. Hans Friedrich Koehn (Dep. of Psychology, University of Illinois, Urbana-Champaign), Prof. Chia-Yi Chiu (University of Minnesota - Twin Cities)

The true Q-matrix underlying a cognitively diagnostic assessment is never known; and estimating the Q-matrix is a challenging task. In practice, the Q-matrix is typically estimated by content experts, which, however, can result in its misspecification causing examinees to be misclassified.

In response to these difficulties, algorithms have been developed for estimating the entire Q-matrix based on item responses collected with the test in question.

Chen and collaborators defined sufficiency conditions for the identifiability of the true, but unknown Q-matrix. They proved that if the true Q-matrix satisfies these identifiability conditions, then a consistent estimator of the true Q-matrix exists.

Extant algorithms for estimating the Q-matrix either impose these identifiability conditions on the estimator, or they don't. The debate on which is "right" is ongoing—especially, as these sufficient but not necessary conditions are rather restrictive so that viable alternatives may be ignored.

Results from a large scale simulation are reported for four Q-matrix estimation algorithms (MCMC, MMLE-EM, EFA-based, neural network) that do not impose the identifiability conditions on the Q-matrix estimator. Item responses were generated under the DINA and GDINA model using an identifiable, "true" Q-matrix and eight experimental conditions varying sample size, test length, number of attributes, and amount of error perturbation added to the data.

Estimated Q-matrices were evaluated (i) whether they met the identifiability conditions and (ii) in their capacity to enable the correct classification of examinees—also in comparison with estimated Q-matrices obtained in imposing the identifiability conditions.

DIF detection in a response time measure: An LRT method

Wednesday, 26th July - 09:00: Differential Item Functioning (Juan Ramon Jimenez) - Oral

Dr. Anne Thissen-Roe (Harver)

For assessments used in hiring decisions, it is essential to address discrepancies in measurement across subgroups from which job candidates may be compared for the same position, so that scores and subsequent decisions are fair and valid. Computer-administered employment tests today commonly use objective items that capture a response reflecting a job-relevant construct, and a response time. Response times may be used to assess individual speed, to calibrate the difficulty of a speeded test, or for other purposes. In order to improve the fairness and validity of speed measures and time limits, we want to check for parametric DIF not only in item responses, but in response times.

While many joint models of item responses and response times fall into the category of complex multidimensional systems in which DIF is challenging to interpret, the hierarchical framework proposed by van der Linden (2006) isolates response and response time models sufficiently that, to the extent that the framework holds, the possibility of DIF in responses and response times may be evaluated separately. Using this framework, likelihood ratio tests of parametric DIF are applied to a three-parameter form of the lognormal response time model. The lognormal response time model can be fit and interpreted as a transformation of a factor model (Finger & Chuah, 2008). Through this relationship, the accepted procedures of factorial invariance testing suggest an order for nested parameter constraint tests in a DIF sweep, corresponding to sequential tests of loadings, intercepts and unique variances. This facilitates interpretation of DIF findings.

Empirical evaluations of DIF detection methods

Wednesday, 26th July - 09:15: Differential Item Functioning (Juan Ramon Jimenez) - Oral

Dr. Yevgeniy Ptukhin (Western Illinois University), Dr. Yanyan Sheng (University of Chicago)

Differential item functioning (DIF) occurs when test items function differently between subpopulations and therefore has been an important consideration for test validity. The fact that DIF items can be potentially biased has called for its increasing attention in measurement practices over the last three decades. Extensive research has been conducted to develop statistical methods for detecting DIF, which can be classified into three broad categories: 1) approaches based on classical test theory, such as Mantel-Haenszel statistic, Angoff's Delta method, Breslow-Day statistic, standardization, logistic regression, simultaneous item bias test (SIBTEST), 2) approaches based on item response theory, such as Lord's chi-squared test, Raju's area method, likelihood-ratio test, and 3) multiple indicator multiple cause (MIMIC) via the use of structural equation models. Much work has been conducted in the DIF literature to evaluate and compare performances of two or three of these methods in different test situations (e.g. Swaminathan & Rogers, 1990; Cohen & Kim, 1993; Raju et al., 1993; Herrera & Gomez, 2007; Aguerri et al., 2009; DeMars, 2009; Woods, 2009; Magis & Facon, 2012; Apinyapibal et al., 2015; Diaz et al., 2021) and can be limited in providing overall recommendations on performances of the available methods. The purpose of this study is to conduct a comprehensive comparison of these methods by evaluating their Type I error and power rates. Results of the study provide a set of guidelines for researchers and practitioners on the use of each DIF detection method in different test situations where uniform or non-uniform DIF presents.

Asymmetry-induced model misspecification and the observation of cross-national DIF

Wednesday, 26th July - 09:30: Differential Item Functioning (Juan Ramon Jimenez) - Oral

Ms. Qi (Helen) Huang (University of Wisconsin-Madison), Prof. Daniel Bolt (University of Wisconsin-Madison), Mr. Weicong Lyu (University of Wisconsin-Madison)

Large scale international assessments depend on invariance of measurement across countries. An important consideration when observing cross-national differential item functioning (DIF) is whether the DIF actually reflects a source of bias, or might instead be a methodological artifact due to IRT model misspecification. Certain forms of misspecification may be accommodated by allowing psychologically plausible generalizations of traditional IRT models on an item-by-item basis (e.g., allowing a small number of items to be asymmetric while others are modeled as 2PL). In this paper, we focus on a generalization of the 2PL which accommodates ICC asymmetry as an example illustration. The model is easily applied using the mirt routine in R (Chalmers, 2017). We show through a simulation study how model misspecification induced by ICC asymmetry produces artifactual cross-national DIF, and how such DIF quantifications can often be substantially reduced when asymmetry is allowed. These results are further supported using real data examples from TIMSS Math and Science.

Unveiling gender bias in SET: A text mining approach

Wednesday, 26th July - 09:45: Differential Item Functioning (Juan Ramon Jimenez) - Oral

Dr. Wen Qu (Fudan University), Dr. Zhiyong Zhang (University of Notre Dame)

Student evaluation of teaching (SET) is a crucial tool for teachers to enhance their pedagogical methods and for academic institutions to assess faculty performance, thereby aiding student achievement. However, gender bias is an issue that often arises, leading to debates on the validity of SET data. Empirical evidence reveals that female instructors face bias in student evaluations, and such prejudice is not due to their teaching proficiency but their gender. SET data is widely used for teaching quality for hiring and promotion, so the form of gender inequality necessitates a thorough investigation. This study goes beyond traditional analysis of only numerical rating scores and focuses on the textual comments provided in SET data. It utilizes domain-specific aspect-based sentiment analysis, deep learning models, and other machine-learning methods to detect potential gender bias in teaching evaluations. By analyzing a large-scale, long-term SET dataset with multiple methods, our results demonstrate that different analyses of gender bias from student feedback reveal distinct patterns. As such, the bias can be mitigated with a different analysis method and focus, making the student evaluation more reliable and valuable.

Bayesian location-scale model for assessing reliability differences with ordinal ratings

Wednesday, 26th July - 09:00: Bayesian Methods (Thurgood Marshall) - Oral

Mr. František Bartoš (Department of Statistical Modelling, Institute of Computer Science, Czech Academy of Sciences), Dr. Patricia Martinkova (Department of Statistical Modelling, Institute of Computer Science, Czech Academy of Sciences), Dr. Marek Brabec (Department of Statistical Modelling, Institute of Computer Science, Czech Academy of Sciences)

The quality of ratings and quantitative assessments depends on the reliability of the rating instrument. Especially important is the measurement error – a high measurement error results in high uncertainty of the resulting scores. Detected systematic differences in measurement error due to applicant/raters-related characteristics might provide guidance on which groups to focus on in interventions designed to lower the measurement error. A flexible approach for detecting differences in measurement error was proposed in Martinková et al., 2023) for cases when scores are assumed to be continuous. In this work, we build on this approach by focusing on ordinal ratings. We highlight cases where treating ordinal rating as continuous might result in biased estimates and outline a Bayesian cumulative probit multi-level location-scale model to mitigate the issue. We use spike-and-slab prior distributions to obtain inclusion Bayes factors of individual predictors and model-averaged posterior distributions within a single model fit. We demonstrate the superiority of the proposed ordinal approach with a simulation study.

Further examination of fully Bayesian information criteria for mixture IRT models

Wednesday, 26th July - 09:15: Bayesian Methods (Thurgood Marshall) - Oral

Dr. Rehab AlHakmani (Emirates College for Advanced Education), Dr. Yanyan Sheng (University of Chicago)

Mixture item response theory (MixIRT) allows the presence of several latent classes that are qualitatively different but within which a conventional IRT model holds. The increase in the popularity of MixIRT models calls for an efficient estimation approach under the fully Bayesian framework via the use of Markov chain Monte Carlo (MCMC) techniques. Recent attention focuses on the no-U-turn sampler (NUTS; Hoffman & Gelman, 2011), a non-random walk MCMC algorithm that converges to high dimensional target distributions more quickly than conventional random walk MCMC algorithms. In an effort of applying NUTS to a MixIRT model, previous research (AlHakmani & Sheng, 2022) evaluated the performance of fully Bayesian measures of information criteria under limited test situations. The purpose of this study is to extend the previous work to further evaluate the performance of the widely applicable information criterion (WAIC; Watanabe, 2010) and the leave-one-out cross validation (PSIS-LOO; Vehtari et al., 2017) in terms of the accuracy in determining the number of latent classes of a two-parameter MixIRT model via implementing NUTS under a wider test situations. Monte Carlo simulations were carried out for tests with two latent classes while manipulating four factors including sample sizes (500 and 1000), test lengths (15, 20, and 30), class sizes (with ratios being 1/4, 1/3, 1/2, and 1), and distances between locations of the latent classes (ranging from 1 to 5 standard deviations). Results of the study provide guidelines on the fit and hence use of MixIRT under various situations using fully Bayesian information criteria.

Bayesian nonparametric latent class analysis for different item types

Wednesday, 26th July - 09:30: Bayesian Methods (Thurgood Marshall) - Oral

Mr. Meng Qiu (University of Notre Dame), Dr. Sally Paganin (Harvard University), Prof. Ilsang Ohn (Inha University), Prof. Lizhen Lin (University of Notre Dame)

The conventional latent class analysis (LCA) requires the number of classes to be predetermined, and model selection criteria are utilized to select the optimal number of classes through the class enumeration process. However, this process neglects the variability of the number of classes, leading to overfitting or underfitting. Moreover, different information criteria can result in different optimal models, making it challenging to reach a consensus on the best criterion. By integrating the Dirichlet process mixture model (DPMM), Bayesian nonparametric LCA can avoid class enumeration by assuming an infinite number of classes and deducing the number of classes from the data. This study aims to introduce an extended DPMM-LCA approach that can cluster individuals by observed variables measured on various metrics. A simulation study was conducted to evaluate the performance of DPMM-LCA in model selection, parameter estimation, and classification accuracy in comparison with the conventional approach. A step-by-step tutorial with the R package NIMBLE was provided for easy implementation. Three real-data examples were presented for illustration of use. The extended DPMM-LCA approach offers a promising expansion of previous DPMM methods for LCA.

Bayesian approaches to quantifying the practical impact of measurement non-invariance: Extending dMACS

Wednesday, 26th July - 09:45: Bayesian Methods (Thurgood Marshall) - Oral

Mr. Conor Lacey (Wake Forest Universti), Dr. Veronica Cole (Wake Forest University)

Measurement invariance assesses the psychometric equivalence of a latent construct across groups. Whenever a measurement construct is said to be non-invariant this implies the overall construct is different or has a different meaning across groups and therefore groups cannot be meaningfully compared under the same measure. Different methods exist to assess measurement invariance. However, it has become clear with advancing research that complete measurement invariance isn't an absolute necessity for group comparison (e.g., partial invariance) and there is no clear threshold for when non-invariance becomes a problem. As a result, the field has begun to develop effect size measures for answers. One of the limitations with effect size research, however, is that it does not currently consider the plausibility of an effect size of zero in the calculation of an estimate. In this study we take the commonly used measurement non-invariance effect size estimator and modify it using a spike-and-slab approach created by Bergh et al., (2021) that allows the probability of an effect of zero to be considered in its calculation. We evaluate the implications of this method as well as demonstrate its use in an empirical example.

How does prior distribution affect model fit indices of BSEM

Wednesday, 26th July - 10:00: Bayesian Methods (Thurgood Marshall) - Oral

Ms. Yonglin Feng (Department of Psychology, Sun Yat-sen University), Prof. Junhao Pan (Department of Psychology, Sun Yat-sen University)

Bayesian structural equation model (BSEM) integrates the advantages of Bayesian method into the framework of structural equation modeling and ensures the identification by assigning priors with small variances. Previous studies have revealed that prior specifications in BSEM influence model parameter estimation, but the impact on model fit indices remains unclear and needs to be explored. Therefore, two simulation studies were conducted. Normal distribution priors were specified for factor loadings, while inverse Wishart distribution priors and separation strategy priors were applied for the variance-covariance matrix of latent factors. Conditions included five sample sizes and 24 prior distribution settings. Simulation study 1 compared the model fitting performance of BCFI, BTLI, and BRMSEA proposed by Garnier-Villarreal and Jorgensen (2020), and PP*p* value. Simulation study 2 compared the performance of BCFI, BTLI, BRMSEA, and DIC in model selection between three data generation models and three fitting models. Results showed that prior settings would affect Bayesian model fit indices in evaluating model fitting and selecting models, especially in small sample sizes. Even under large sample size, the highly improper factor loading prior led to poor performance of the Bayesian model fit indices. BCFI and BTLI were less likely to reject the correct model than BRMSEA and PP*p* values under different prior specifications. For model selection, different prior settings would affect DIC on selecting the wrong model, and BRMSEA preferred the parsimonious model. Our findings recommended that the Bayesian approximate fit indices may be better for evaluating model fitting and selecting models under the BSEM framework.

Response time modeling: Inference, evaluation, and new modeling approaches

Wednesday, 26th July - 14:40: Symposium: Response time modeling: Inference, evaluation, and new modeling approaches. (Colony Ballroom) - Symposium Overview

Dr. Hyeon-Ah Kang (The University of Texas at Austin)

As response times have been used as an important auxiliary variable to traditional item response data, methodological approaches to modeling response times have become of great interest in the measurement community. In this symposium, we share recent work on the response time models that straddle over statistical inference and new modeling approaches. The first study, presented by Dr. Debelak, addresses an estimation method in the current development and proposes a new procedure that uses deep learning technique for small and large-scale timing data. The second study, presented by Ms. Sung, evaluates fit of a response-time model based on the generalized residuals so that items' fitness can be evaluated using moment estimators. The following two studies present new modeling approaches for real-life data. Dr. Kang, a third presenter, presents an exploratory graphical modeling approach to modeling local item dependency within the latent factor response-time model. Mr. Mutak, a fourth presenter, proposes a modeling framework that relaxes conditional independence of items and probes into patterns in the missing responses and omission behavior. Our selected presentations span across estimation, evaluation, and modeling, and yet, bring new interesting insights into the current applications of response time models. Our studies, as a whole, seek to improve upon the current development and provide methodological alternatives that better serve the need of the field in utilizing the response times.

Deep learning approaches for factor analysis of responses and response times

Wednesday, 26th July - 14:43: Symposium: Response time modeling: Inference, evaluation, and new modeling approaches. (Colony Ballroom) - Symposium Presentation

Dr. Rudolf Debelak (University of Zurich), Mr. Christopher J. Urban (University of North Carolina at Chapel Hill)

An important problem in the application of psychometric models is the selection of suitable algorithms for parameter estimation. In a recent publication, Urban and Bauer (Psychometrika 86:1-29, 2021) proposed an estimation algorithm based on deep learning for item parameter estimation for large sample sizes. We first give an overview on the principal ideas of this approach, and how it related to classical estimation methods for models of item response theory. Second, we will evaluate the accuracy of this approach for two types of factor models: a) a log-normal factor model for response times, b) a hybrid factor model for responses and response times, which is related to previously proposed methods for responses and response times. The results of the simulation studies indicate that the deep learning algorithm can be used for an accurate item parameter estimation for these models even in relatively small datasets and is computationally fast in large datasets. The evaluated algorithm is freely available in a Python package.

Assessing the fit of response time factor models by generalized residuals

Wednesday, 26th July - 15:01: Symposium: Response time modeling: Inference, evaluation, and new modeling approaches. (Colony Ballroom) - Symposium Presentation

Ms. Youjin Sung (University of Maryland, College Park), Mr. Youngjin Han (University of Maryland, College Park), Dr. Yang Liu (University of Maryland, College Park)

Individual differences in processing speed are reflected in item-level response time (RT) data and are often investigated using factor analysis. Commonly used RT factor models, such as the log-normal RT model (van der Linden, 2006), tend to rely on restrictive parametric assumptions, which can be violated in practice. In cases where violation of assumptions results in a poor fit to the observed data, it is often necessary to relax untenable assumptions and fit a more complex model. In contrast, if the amount of misfit is negligible, one may stick to the restrictive parametric model, taking benefits from parsimony. In the present study, we propose a method that can be used to assess the fit of RT factor models using empirical moment functions, which estimate the conditional moments of item-level RT given the latent speed. To obtain formal statistical tests, we extend the theory of generalized residuals (Haberman, Sinharay, & Chon, 2013), originally developed for assessing item fit in IRT models, to factor models for continuous data. The empirical Type I error rate and power of the proposed tests are evaluated in a Monte Carlo experiment.

Gaussian graphical model for evaluating local item dependency in response times

Wednesday, 26th July - 15:19: Symposium: Response time modeling: Inference, evaluation, and new modeling approaches. (Colony Ballroom) - Symposium Presentation

Dr. Hyeon-Ah Kang (The University of Texas at Austin)

The study proposes an exploratory modeling framework for evaluating local item dependency in response times. Current approaches to modeling local dependency require pre-specification of dependent items and have limited utility when such information is not available. In this study, we present an exploratory graphical modeling approach to evaluating local item dependency within the latent factor response-time models. The model integrates a Gaussian graphical model to the log-normal response-time model so that items' own interplay can be modeled via an undirected spatial network. An inferential framework is proposed based on the regularized pseudo-likelihood and is implemented by an EM algorithm applying soft-thresholding and gradient-based Newton iteration. Numerical experimentation from Monte Carlo simulation suggests that the estimation adequately recovers generating parameters and yields reliable standard error estimates. Comparison with a regular response-time model suggested that the graph-integrated model yields substantially lower bias in parameter estimation when items are locally dependent. The study demonstrates application of the model using two secondary data (e.g., a national licensure assessment and an international educational assessment) and discusses practical benefits of modeling local item dependence.

Modeling omissions in tests as dependent on previous test behavior

Wednesday, 26th July - 15:37: Symposium: Response time modeling: Inference, evaluation, and new modeling approaches. (Colony Ballroom) - Symposium Presentation

Mr. Augustin Mutak (Freie Universität Berlin), Dr. Esther Ulitzsch (IPN - Leibniz Institute for Science and Mathematics Education), Mr. Sören Much (Martin-Luther-Universität Halle-Wittenberg), Dr. Jochen Ranger (Martin-Luther-Universität Halle-Wittenberg), Dr. Steffi Pohl (Freie Universität Berlin)

In order to adequately account for missing values in tests, it is essential to have a good understanding of missing data mechanisms. Most of the current approaches place missing responses into the context of low ability, disengagement, or general test-wiseness. However, the mechanism of emergence of missing values in psychological tests is still not fully known, as it likely stems from different sources. There are little findings in the literature which reveal how a person's behavior on previous items can be relevant for omissions in subsequent items. However, previous behavior in the test, such as taking too much time or scoring low on items may impact test-takers strategy. To explore this, we develop a new model, which includes responses, response times and omissions on the manifest level and their respective latent constructs. In the model, we relax the assumption of conditional independence between responses or response times on an item and omissions in the subsequent item. By allowing for these residual correlations to exist in the model, we explore if spending relatively too much time on one item can lead examinees to become more hesitant into investing time to solve the next item. We also investigate whether comparably worse performance on a previous item impacts the occurrence of missing values on later items. We conduct a simulation study to test the performance of our model. To illustrate its use, we apply it to an empirical dataset exploring whether behavior in responding to previous items may explain omissions in later items.

How are item difficulties and item discriminations related? Does that matter?

Wednesday, 26th July - 14:40: Validity and Reliability (Atrium) - Oral

Dr. Sandip Sinharay (Educational Testing Service), Dr. Matthew Johnson (Educational Testing Service), Dr. Sandra Sweeney (Cognia Inc.), Dr. Eric Steinhauer (ETS)

The focus of this presentation will be on the empirical relationship between item difficulty and item discrimination. Two studies—an empirical investigation and a simulation study—will be used to shed light on the association between item difficulty and item discrimination under classical test theory and item response theory (IRT), and the effects of the association on various quantities of interest. Results from the empirical investigation will be used to argue that item difficulty and item discrimination are negatively correlated under classical test theory, mostly negatively correlated under the two-parameter logistic model, and mostly positively correlated under the three-parameter logistic model; the magnitude of the correlation varies over the different data sets. Results from the simulation study will be used to demonstrate that a failure to incorporate the correlation between item difficulty and item discrimination in IRT simulations may provide the investigator with inaccurate values of important quantities of interest, and may lead to incorrect operational decisions. Implications to practice and future directions will be discussed.

Exploring attenuation of reliability in categorical subscore reporting

Wednesday, 26th July - 14:55: Validity and Reliability (Atrium) - Oral

Dr. Richard Feinberg (National Board of Medical Examiners)

Research on subscores has consistently advocated discontinuing reporting when they lack sufficient psychometric properties (Feinberg & Jurich, 2017; Wainer & Thissen, 2001), yet many operational testing programs have not changed (Sinharay, 2010). This may be due to several real-world complications such as user demand, competitors, or contractual obligations. Given these challenges, some test publishers have continued to provide subscores, but in a categorical format to mitigate misinterpretation of small differences likely due to error. However, there also exists robust literature on how continuous scores grouped into categories can be less reliable than the scores from which they were constructed (Ramsay, 1973) and could even be used fraudulently to skew the outcome (Wainer, Gessaroli, & Verdi, 2006).

In this study, we will compare different methods of discretizing subscore information into categories, for different criteria, and the impact on subscore reliability. For instance, Feinberg & von Davier (2020) presented a method where each examinee's observed subscore is compared against a discrete probability distribution of subscores conditional on the examinee's overall ability. This method could also be generalized to instead use a discrete probability distribution of subscores conditional on the average ability of the group (norm-referenced) or an external standard (criterion-referenced). Implications for practice, operational utility, and the extent to which categorical subscore reporting represents an appropriate compromise will be discussed.

Exploring response biases in rating scales data with interaction map

Wednesday, 26th July - 15:10: Validity and Reliability (Atrium) - Oral

Mr. Jinwen Luo (University of California Los Angeles), Prof. Minjeong Jeon (University of California Los Angeles)

Rating scales are widely used to measure personality, attitudes, and beliefs in psychological, health, and social research. Using a rating scale assumes that the observed scores can be attributed solely to the latent traits being measured and that other factors are not responsible for individual differences. However, it is also well-established in the literature that many factors, including social desirability, response styles, or other abnormal response behaviors, will violate the conditional independence assumption and introduce response bias. (Podsakoff et al., 2003; Steenkamp et al., 2010; Bockenholt, 2017; van der Linden & Guo, 2008). In this paper, we extended the interaction map approach (Jeon et al., 2021) and explored its usability in detecting various types of response bias in rating scale response data. Dependencies between respondents and item response categories, unexplained by the person and item parameters, are visualized as respondent-category distances in the map, reflecting the presence, pattern, and size of potential response bias. We estimate the proposed approach using a fully Bayesian approach with MCMC implemented in the R nimble package (de Valpine et al., 2017). We illustrate the potential utility of the proposed approach in simulation and empirical data settings. Our results have demonstrated that the interaction map can be utilized without making assumptions about the sources or nature of response biases. The interaction map will show potential response biases, while the size and uncertainty are quantifiable. This paper provides evidence for using interaction map for screening and monitoring purposes.

Impact of ignoring rater effects in objective structured clinical examinations

Wednesday, 26th July - 15:25: Validity and Reliability (Atrium) - Oral

Dr. Daniel Edi (Pearson), Dr. Igor Himelfarb (National Board of Chiropractic Examiners)

Objective Structured Clinical Examination is (OSCE) is the performance-based clinical competence assessment methods used in healthcare (Harden et al., 1975). The objective of the OSCE is to evaluate the skills of future health practitioners with a special focus on the objectivity of the process (Harden et al., 1975). Bartfay et al. (2004) argued that OSCE has become a benchmark of health professional assessments. Nevertheless, many researchers have argued that in performance-based assessments such as OSCEs, examinees' performance ratings are seldom free from examiner bias, even after extensive examiner training sessions. The purpose of this study was to explore the impacts of discounting rater effects on examinee's ability estimates in operational OSCEs. The data were obtained from operational administrations of chiropractic OSCE exam in the U.S. (NBCE, 2022). Two models were used to estimate examinee's ability: (1) the partial credit model (PCM; Masters, 1982) that ignores potential rater effects, and (2) a PCM version of the many-facet Rasch model (MFRM; Linacre, 1989). Three indices were used to evaluate the impact of ignoring rater effects on examinee's ability estimates: (1) bias values, (2) root means square errors, and (3) mean absolute differences. Model-based confusion matrix was used to evaluate the impact of ignoring rater effects on classification accuracy. The findings emphasized the need for practitioners to employ the MFRM as an additional tool that controls for persistent rater effects. Additionally, testing companies should continue to offer extensive training to their raters with the goal of achieving appropriate levels of intra-rater reliability.

Extending reliability to intensive longitudinal data with the Kalman filter

Wednesday, 26th July - 15:40: Validity and Reliability (Atrium) - Oral

Dr. Michael Hunter (The Pennsylvania State University)

Reliability is at the core of how researchers approach measurement in standard, group-bases analyses of single-timepoint data, yet this critical aspect is often overlooked in the analysis of repeated observations. In this presentation, we relate notions of reliability in group-based designs to previously established results from the Kalman filter on single-subject designs. We then extend these results to multisubject designs with multiple timepoints, and relate our findings to results from the multilevel modeling literature. Using a combination of analytic results and simulation studies, we explore how classical measurement concepts like reliability can be extended to modern designs and analysis techniques like dynamic structural equation modeling (DSEM) that allow for person-specific and time-varying parameters. We note that the concept of reliability itself might need to be updated to account for modern research designs and methods.

Classifying normal and aberrant behaviors through machine learning

Wednesday, 26th July - 14:40: Machine Learning (Benjamin Banneker) - Oral

Ms. Suhwa Han (The University of Texas at Austin), Dr. Hyeon-Ah Kang (The University of Texas at Austin)

One of the problems of the existing aberrant behavior detection methods in educational measurement is that they are ignorant of distinct patterns of aberrant behaviors across examinees. For example, in real test settings, examinees can differ in their location, duration, and intensity in exhibiting abnormal behaviors. However, a majority of existent detection methods overlook such individual differences and examine the entire vector of observations as a whole when evaluating the aberrancy. In this study, we suggest a novel application of machine learning classification algorithms for identifying aberrant test-taking behaviors. We employ unsupervised classifiers to automatically identify the training and evaluation items, and conduct multivariate hypothesis tests to determine the significance of abnormality. The proposed scheme allows for fully automated anomaly detection without making artificial assumptions on the reference and testing samples. All procedures are applied to the cognitive and behavioral indicators (e.g., item response, response time, behavioral frequency) obtained from large-scale standardized testing. Through extensive simulation studies and empirical data analysis, we evaluate the adequacy of the classifiers and suggest some adjustments for use in testing scenarios.

Neural networks approach to estimate IRT models in small samples

Wednesday, 26th July - 14:55: Machine Learning (Benjamin Banneker) - Oral

Dr. Dmitry Belov (Law School Admission Council), Dr. Esther Ulitzsch (IPN - Leibniz Institute for Science and Mathematics Education), Dr. Alexander Robitzsch (IPN - Leibniz Institute for Science and Mathematics Education), Dr. Oliver Lüdtke (IPN - Leibniz Institute for Science and Mathematics Education)

Given an assessment context with a large item pool aiming to reduce calibration samples (or to address sparsity in calibration samples), the general problem of estimating the parameters of IRT models having item characteristic curves (ICC) is considered. This class covers popular models including the 1PL, 2PL, 3PL, and nominal response models. We have built a neural network (NN) predicting item parameters using the following features generated from item pool response data: (1) the discrete ICC computed for five bins using responses and percentage scores and (2) the sample mean and standard deviation of the percentage score. Assuming normally distributed populations, we have developed a special training procedure for the NN to accommodate differences in score distributions. Our approach was evaluated on 1742 3PLM items from the Law School Admission Test (LSAT). In cross-validation studies with just 20 (50) simulees, the NN achieved an average coefficient of determination (R2) of 0.57 (0.70). In contrast, parameters obtained by gradient descent running for the same sample sizes provided an average R2 of as low as 0.33 (0.46). One important practical application of our approach is to resolve the issue of item compromise. With recent migration of high stakes testing programs from offline to online this issue became prominent due to pretesting new items in settings that are hard to reliably proctor. Using the presented approach, testing organizations can pretest new items only in test centers with strict proctoring, thus, completely eradicating item compromise.

Online calibration for P-MCAT: A neural network based approach

Wednesday, 26th July - 15:10: Machine Learning (Benjamin Banneker) - Oral

Dr. Lu Yuan (Collaborative Innovation Center of Assessment for Basic Education Quality,Beijing Normal University), Ms. Yingshi Huang (Collaborative Innovation Center of Assessment for Basic Education Quality,Beijing Normal University), Prof. Ping Chen (Collaborative Innovation Center of Assessment for Basic Education Quality,Beijing Normal University)

Online calibration is a key technology for calibrating new items in computerized adaptive testing (CAT). As the multidimensional polytomous data become popular, the marginal maximum likelihood estimation with an expectation maximization (MMLE/EM)-based online calibration methods applicable to multidimensional CAT with polytomously scored items (P-MCAT) have been proposed (Yuan et al., 2022). However, the existing methods are mainly based on the MMLE/EM algorithm, which suffer from convergence problem when faced with high dimensionality and inaccessible proper initial values. To conquer these challenges, a neural network (NN)-based online calibration framework was put forward in this study. The new method differs profoundly from the traditional ones in that the parameter estimates of new items are obtained by learning the patterns between input and output data instead of finding solutions to the log-marginal likelihood function. This change in perspective gains critical benefits: (1) sidestepping the high-dimensional integration and high-dimensional optimization procedures on the ability posterior calculation and item parameter estimation, respectively; (2) reducing the impact of initial values by approaching item parameters indirectly via the optimization of network parameters. In this study, full-scale simulation studies were conducted to evaluate the performance of the NN-based method and MMLE/EM-based methods under varying scenarios. Furthermore, an alternative solution was presented for traditional methods to obtain appropriate initial values. Results showed that both the NN-based method and the alternative solution found their strengths in recovering the item parameters of new items while the MMLE/EM methods struggled to converge when more than three dimensions were involved in the test.

Validating automated methods for measuring psychological constructs in text

Wednesday, 26th July - 15:25: Machine Learning (Benjamin Banneker) - Oral

Mr. Daniel Low (Harvard Medical School & Massachusetts Institute of Technology), Dr. Patrick Mair (Harvard University), Prof. Matthew Nock (Harvard University), Dr. Satrajit Ghosh (Massachusetts Institute of Technology & Harvard Medical School)

Rating scales for mental health are often retrospective (e.g., ask about symptoms in the last two weeks), rely on an individual accurately interpreting the wording of items, and can miss key risk factors if not included in the questionnaire. When humans describe their symptoms using their own natural language in therapy sessions or on social media, they can potentially describe many more risk factors and perceived causes in an ecological way. The challenge is reliably measuring symptoms from text as has been done from rating scales. Current approaches consist of manually building lexicons or coding thousands of text documents for supervised learning models. Here we present a series of methods that measure constructs in text in more automated or unsupervised ways without the need of a labeled training set. These include semi-automated ways of generating a lexicon for a given construct as well as deep learning methods such as few-shot learning that capture aspects of meaning to overcome lexicons' exact-match approach. We demonstrate that these methods can capture 40 known risk factors for suicidal thoughts and behaviors. Using over 200k crisis counseling sessions from Crisis Text Line, we then validate these measurements of risk factors by using them to predict the severity of a crisis counseling session using a held-out test set. We also test how well each method is able to quantify the presence of one of 15 types of crisis (e.g., suicide, sexual assault, anxiety). We compare these methods to standard approaches including existing lexicons and supervised learning.

Application of MCMC algorithm with Davidian curve in multidimensional IRT models

Wednesday, 26th July - 14:40: Multidimensional IRT (Prince George) - Oral

Dr. Xue Zhang (Northeast Normal University), Prof. Chun Wang (University of washington), Prof. David Weiss (University of Minnesota - Twin Cities)

Under item response theory (IRT) framework, misspecified latent distribution may lead to biased parameter estimations, such as normality assumption for a non-normal distribution. For unidimensional IRT models with non-normal latent trait distributions, the MCMC algorithm with Davidian curve (MCMC-DC) mostly outperforms the MML method, especially when sample size is small (Zhang, Wang, Weiss & Tao, 2021). The purpose of this study was to extend the MCMC-DC to handle the multidimensional IRT models with flexible latent trait distributions (i.e., normal, skewed, and bimodal), and to propose and explore an adaptive selection method of the best fit DC order during estimating parameters. The performance of the proposed method was illustrated via simulation studies and a real data example. Preliminary results indicated that the MCMC-DC method with the adaptive selection method could fit normal and bimodal distributions well and skewed distributions reasonably well, and the method provided good estimates of item parameters.

Accommodating curvilinear unidimensional approximations to multidimensionality: IRT modeling implications

Wednesday, 26th July - 14:55: Multidimensional IRT (Prince George) - Oral

Ms. Xiangyi Liao (University of Wisconsin-Madison), Prof. Daniel Bolt (University of Wisconsin-Madison), Prof. Jee-Seon Kim (University of Wisconsin-Madison)

Item difficulty and dimensionality frequently correlate in measurement settings, implying that unidimensional approximations (i.e. reference composites) will often take a curvilinear form through the multidimensional space (Carlson, 2017). Although this issue is most commonly discussed in the context of vertical scaling applications, such phenomena can also occur in cross-sectional within-grade applications. Measurement of reading proficiency, for example, frequently entails the use of different forms of measurement (e.g., letter recognition versus decoding tasks) having items that uniquely discriminate at different locations along the reading proficiency continuum. We demonstrate through simulations and application of a latent regression procedure that the multidimensionality associated with the different forms of measurement yields a curvilinear dimensional composite where the weight of each dimension changes along the unidimensional continuum. Effectively, different dimensions come to dominate different regions of the unidimensional continuum according to where the item difficulties of a particular form of measurement are concentrated.

We then show how this form of curvilinearity can and should produce systematic forms of misspecification in the models (e.g., 2PL) traditionally applied in unidimensional measurement. Failure to attend to such misspecification may produce subsequent distortions in the latent metric. A real data example using item level data from the ECLS-K reading proficiency measurement is provided for demonstration. Using a semiparametric unidimensional IRT model, we show how ICCs demonstrating the anticipated forms of asymmetry due to multidimensionality provide a closer fit to the data than a traditional 2PL representation.

Multidimensional beta factor analysis for bounded and asymmetric item response data

Wednesday, 26th July - 15:10: Multidimensional IRT (Prince George) - Oral

Mr. Alfonso J. Martinez (University of Iowa)

In this presentation, we propose a multidimensional beta factor-analytic model (beta IFA) for modeling bounded and skewed item response data. The beta IFA model is based on the beta distribution, a flexible distribution that can accommodate a variety of response distributions (e.g., bimodal, heavily skewed). In addition, the beta IFA respects the bounds of the response range and will not provide out-of-range predictions. We present the model as a compromise between normal-theory factor analysis (NTFA; which assume that response variables are continuous, unbounded, and symmetric) and ordinal factor analysis (which assumes underlying latent category thresholds). We derive an expectation-maximization algorithm for full-information maximum likelihood estimation and an expected a posteriori procedure for estimating factor scores. We investigate the performance of the algorithm/model via three simulation studies. In Simulation Study I, we examine the performance of the algorithm in a one-factor setting by varying sample size and test length. Simulation Study II focuses on performance in the multi-dimensional case with correlated latent factors. Simulation Study III compares the beta IFA model to NTFA under different data-generating model mechanisms. Results from Simulation Studies I and II provide evidence that the algorithm performs well in small- and large-sample settings with as few as ten items. Simulation study III demonstrates that the beta IFA model performs comparable to NTFA with respect to model fit when the latter is the data-generating model but outperforms NTFA when item response data are skewed. We demonstrate the utility of the model with real data via three empirical applications.

Multidimensional item response tree model and its application to investigating response styles

Wednesday, 26th July - 15:25: Multidimensional IRT (Prince George) - Oral

Dr. Biao Zeng (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University), Ms. Yanmei Li (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University), Prof. Hongbo Wen (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University)

This study developed a Multidimensional Item Response Tree Model and used this model to investigate the response style in three Chinese versions of the Undergraduate Learning Burnout (ULB) scale.

We found that Model 5 which contained three learning burnout factors, two trait RES factors, and one acquiescence factor fits all three versions the best, indicating that participants showed an acquiescence response style and extreme responses style at either high or low ends of the trait being measured. Participants had a weak acquiescence response style and a strong tendency to avoid extreme responses. Compared to participants with high levels of learning burnout, those with lower levels of learning burnout showed a stronger tendency to avoid extreme responses, as they were less likely to choose extreme options, indicating that a significant interaction between learning burnout and response styles, which means different levels of learning burnout correspond to different intensities of response styles.

The present study has demonstrated that Multidimensional Item Response Tree models can be used to control response styles by recoding the participants' responses to multiple node responses, separating different decision-making processes, and effectively detecting response styles. In addition, researchers should prioritize examining the more extreme response styles due to the interaction between the traits of interest and response styles, and finally, estimate the role of potential response styles more accurately.

Method effects of item wording: MIRT estimation based on equivalence method

Wednesday, 26th July - 15:40: Multidimensional IRT (Prince George) - Oral

Dr. Biao Zeng (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University), Ms. Yanmei Li (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University), Prof. Hongbo Wen (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University)

This study combined Multidimensional Item Response Theory and Linear Equivalence Method to estimate the effect of item wording directions.

We formed the original-reverse and positive version of the scales based on the Undergraduate Learning Burnout (ULB) scale by changing the item wording directions through common item design. 871 undergraduate and graduate students completed one of the three versions of the ULB scale. The Graded Rating Scale model and the quasi-Monte Carlo EM algorithm were used to estimate item parameters. Then, based on the positive version of the scale, the item parameters are converted through using Linear Equivalence in the other two versions of the scale.

The results showed that: before equivalence, the average difficulty of the original version of the scale was lower than that of the original-reverse and positive versions of the scale, indicating that the participants had different responses in different versions of the scale, and the participants tended to choose options which represent higher levels of learning burnout in the original version of scale; after the equivalence, neither the original nor the original-reverse version of the scale had significant changes in the item difficulty parameters of negatively worded items, indicating that the participants had no significant tendency when answering positively and negatively worded items.

Together, this study indicated that the change of item wording directions will indirectly bring a significant method effect through the combination of items with different wording directions, thereby threatening the validity of the scale and decreasing the effectiveness of the results.

Clusterwise Joint-ICA for studying heterogeneity between subjects in multi-modal components

Wednesday, 26th July - 14:40: Classification and Clustering (Margaret Brent) - Oral

Dr. Tom Wilderjans (Leiden University), Mr. Jeffrey Durieux (Erasmus University Rotterdam), Prof. Serge Rombouts (Leiden University)

In this digital era, more and more interesting research questions in different fields of science (e.g., psychology, neuroscience, genetics) call for the collection and joint analysis of (big) data from different sources or modalities. In neuroscience, for example, brain scans yield simultaneous information on brain functioning (i.e., connectivity as measured by fMRI) and structure (i.e., MRI scan) for a set of subjects, implying a two-modal data set. To extract the mechanisms underlying each modality and investigate the covariation of mechanisms across modalities, Joint-Independent Component Analysis (Joint-ICA) was proposed. Often, however, sample heterogeneity in these underlying mechanisms exists. For example, brain diseases (and subtypes thereof), like dementia and depression, are known to be very heterogeneous across patients (groups). To capture this heterogeneity, subjects should be clustered based on similarities and differences in the multi-modal components underlying their coupled data. To obtain this, in this presentation, Clusterwise Joint-ICA, which combines (K-means type of) clustering with Joint-ICA, is proposed. In this model, subjects are clustered and the multi-modal components characterizing each subject group are estimated simultaneously, using an Alternating Least Squares (ALS) type of algorithm. The performance is evaluated by means of an extensive simulation study and by an illustrative application to coupled brain data (combining functional and structural information). Finally, it is also investigated whether clusterwise J-ICA performs better than single-modal analysis strategies and alternative clustering procedures (i.e., K-means on the raw data and an sequential Joint-ICA with K-means approach).

Bootstrap standard errors for LDA, NCA, and transformation-matrix methods with multiple solutions

Wednesday, 26th July - 14:55: Classification and Clustering (Margaret Brent) - Oral

Mr. Yikai Lu (University of Notre Dame), Prof. Ying Cheng (University of Notre Dame)

Bootstrap is a nonparametric method for estimating sampling distributions of various statistics, without assuming any asymptotic distribution. It is commonly used in psychological research for statistics generated from statistical models such as structural equation modeling, factor analysis, and mediation. However, the application of bootstrap to dimension reduction methods for classification, such as linear discriminant analysis (LDA) and neighborhood components analysis (NCA), has not been studied extensively. These methods generate a transformation matrix which has a column reflection problem and possible multiple solutions. The paper proposes a clustering method which allows the use of the bootstrap for LDA and NCA by addressing both issues. It enables identifying multiple solutions in bootstrap samples and statistical testing of the weights of the transformation matrix which separates groups or classes of data. In this study, we conducted a simulation study and demonstrated that we can obtain reliable standard error estimates with a large sample size. In addition, we will illustrate a multiple-solutions example with real data using NCA.

MixML-SEM: A parsimonious approach for finding clusters of groups with equivalent structural relations in presence of measurement non-invariance

Wednesday, 26th July - 15:10: Classification and Clustering (Margaret Brent) - Oral

Ms. Hongwei Zhao (KU Leuven), Prof. Jeroen Vermunt (Tilburg University), Prof. Kim De Roover (KU Leuven)

Structural equation modeling (SEM) is commonly used to explore relationships between latent variables, such as beliefs and attitudes. However, comparing structural relations across a large number of groups, such as countries, can be challenging. Existing SEM approaches may fall short, especially when measurement non-invariance is present. In this project, we propose Mixture Multilevel SEM (MixML-SEM), a novel approach to comparing relationships between latent variables across many groups that gathers groups with the same structural relations in a cluster, while accounting for measurement non-invariance in a parsimonious way. Specifically, MixML-SEM captures measurement non-invariance using multilevel CFA and then estimates the structural relations and mixture clustering of the groups by means of the structural-after-measurement (SAM) approach. In this way, MixML-SEM ensures that the clustering is focused on structural relations and unaffected by differences in measurement. MixML-SEM is particularly useful when sample sizes per group are too small to estimate partially group-specific measurement models (e.g., by multigroup CFA). In this case, accounting for measurement non-invariance with random parameters is more accurate and efficient. We demonstrate the effectiveness of MixML-SEM through simulations and a real data example, showing that it outperforms existing mixture SEM approaches.

Model-agnostic unsupervised detection of bots in Likert-type survey data

Wednesday, 26th July - 15:25: Classification and Clustering (Margaret Brent) - Oral

Mr. Michael John Ilagan (McGill University), Dr. Carl Falk (McGill University)

To detect bots in online survey data, there is a wealth of literature on statistical detection using only responses to Likert-type items. There are two traditions in the literature. One tradition requires labeled data, forgoing strong model assumptions. The other tradition requires a measurement model, forgoing collection of labeled data. In this work, we consider the problem where neither requirement is available, for an inventory that has the same number of Likert-type categories for all items. We propose a bot detection algorithm that is both model-agnostic and unsupervised. Our proposed algorithm involves a permutation test with leave-one-out calculations of outlier statistics. For each respondent, it outputs a p-value for the null hypothesis that the respondent is a bot. Such an algorithm offers nominal sensitivity calibration that is robust to the bot response distribution. In a simulation study, we found our proposed algorithm to improve upon naive alternatives in terms of 95% sensitivity calibration and, in many scenarios, in terms of classification accuracy.

Power analysis for correspondence measures of replication success

Wednesday, 26th July - 15:40: Classification and Clustering (Margaret Brent) - Oral

Mr. Patrick Sheehan (University of Maryland, College Park), Dr. Peter Steiner (University of Maryland, College Park)

Although replicability is an important foundation of scientific research and knowledge generation, there is little guidance for planning replication studies or assessing their results. This presentation discusses the importance of power analysis for replication studies comprised of two studies, with a focus on two correspondence measures: Correspondence in Significance Pattern (CSP), which concludes replication success if both effects are (non-) significant and replication failure otherwise; and the Correspondence Test (CT), which uses a difference test and an equivalence test of the two effects of interest to assess whether or not the two effects correspond (Steiner & Wong, 2018; Tryon & Lewis, 2008). While CSP examines both studies independently and only concludes replication success or failure, CT considers both studies jointly and has four possible results: *Equivalence* of effects, *Difference* of effects, a potentially negligible *Trivial Difference* between effects, or *Indeterminacy*– insufficient evidence to reach a conclusion. The measure used to assess replication success directly impacts how large a sample is necessary to be likely to show effect correspondence. While CSP and CT have differing sample size requirements, with either measure the necessary sample size to have a reasonable chance to show replication success is larger than that needed to show the presence of an effect in a single study. Ultimately, power analysis is necessary to ensure that the constituent studies of a replication effort are large enough to be likely to show replication success.

Contributions of equating error and measurement error to score variability

Wednesday, 26th July - 14:40: Test Theory (Juan Ramon Jimenez) - Oral

Dr. Dongmei Li (ACT, Inc.)

In educational measurement, error is assumed to affect all test scores. Likewise, equating error is an inevitability in longstanding testing programs, where equating is routinely conducted to make scores from different administrations comparable. Theoretically, equating error can be treated as a component of measurement error, since equating error contributes to observed score variability. In practice, however, common methodologies for estimating the standard error of measurement (SEM) do not incorporate equating error. This inconsistency between theory and practice and the lack of understanding of the relationships between SEM and the standard error of equating (SEE) may cause confusion for researchers and practitioners, leading to inappropriate or inadequate uses of these estimates even when both are available.

The purpose of this study is to (1) demonstrate how equating error contributes to measurement error for individual scores and group means and (2) provide guidance on appropriate uses of commonly used SEM and SEE statistics in score interpretation. The relative contributions of equating error and other sources of measurement error will be demonstrated through the Generalizability theory framework (Brennan, 2001) using empirical data, where equating is treated as one facet in the universe of admissible observations so its contribution to measurement error can be estimated.

This investigation will inform best practices in score reporting and score interpretation, especially for reporting and interpretation of changes in group means, since the relative contributions of equating error to group mean variability can be larger than other sources of measurement error (Brockmann, 2011; Li, 2022).

Test item instructional sensitivity indices incorporating instructional content and quality

Wednesday, 26th July - 14:55: Test Theory (Juan Ramon Jimenez) - Oral

Prof. Anne Traynor (Purdue University), Dr. Cheng-Hsien Li (National Sun Yat-sen University)

For achievement test scores to provide information about student learning progress, test-takers' item responses must be affected by differences in the content of instruction (i.e., the implemented curriculum) over time. Test item scores are "instructionally sensitive" if they are affected by variation in the quality or content of instruction received by test-takers. Many statistical indices have been proposed to evaluate the instructional sensitivity of test items that are scored dichotomously (e.g., Naumann et al., 2017), and a few instructional sensitivity indices exist for polytomous items, also. A longstanding problem is how to represent instructional quality in these indices (Ing, 2018). Using data from the TIMSS Grade 8 Science teacher questionnaire and student achievement test, we compare test item instructional sensitivity indices that incorporate measures of content coverage only, or both content coverage and instructional quality. A multiple-group MIMIC model with a continuous covariate will be used to generate the instructional sensitivity index values. We will use differential item functioning effect size criteria from Roussos et al. (1999) to classify items' instructional sensitivity values. The results will allow us to determine if the more complex instructional sensitivity index is useful, and to characterize features of science test items that have high instructional sensitivity.

Latent equating: the case of the NEAT design

Wednesday, 26th July - 15:10: Test Theory (Juan Ramon Jimenez) - Oral

Dr. Inés Varas (Pontificia Universidad Católica de Chile)

In the equating literature, the proposed approaches for the estimation of the equipercentile function are based on the continuization step, i.e. continuous approximations of the test scores distributions. Considering scores as ordinal random variables Varas et al (2019) and Varas et al (2020) proposed the Latent equating method to tackle the discreteness of scores in test equating procedures. Considering a Bayesian nonparametric model for the latent representation of the
ordinal scores distributions authors defined a procedure to obtain discrete equated scores.
Several data collection designs have been described in the equating literature to control test takers ability differences (Von Davier et al, 2004). The nonequivalent groups with anchor test (NEAT) design is widely used in test equating. In this project we extend the Latent equating method to several equating designs with emphasis on the NEAT design. Several methods to evaluate the performance of the method are discussed and applied in both a simulation study and in a real dataset.

A comparison of equating method to detect longitudinal trends.

Wednesday, 26th July - 15:25: Test Theory (Juan Ramon Jimenez) - Oral

Dr. Haruhiko Mitsunaga (Nagoya University), Dr. Yuri Uesaka (The University of Tokyo)

An assessment program which measures lerners' ability often requires a set of longitudinal dichotomous data in schools. In order to obtain a common vertical scale, it is necessary that 'anchor test design' be applied to monitor groups with small sample sizes. It is also required that the data with anchor test design should be equated using item response theory (IRT), however, the equating method to obtain proper solution when the sample sizes of monitor examinee groups are smaller than that of focal examinee groups. This paper presents two studies. In study 1, an anchor test was constructed from existing test forms. The existing data and anchor test data were combined and analyzed using item response theory. Scales from test forms for grade 2 to 6 were equated by two different methods: concurrent and separate calibration. Results of a comparison showed that the concurrent calibration method returned theoretically accurate estimates. In study 2, test data from different academic years between grade 2 and 6 were organized according to student. Different types of equating procedure were adopted to find which procedure is feasible in order to explain longitudinal trends of the students' latent ability estimation. It revealed that the posterior distribution estimates of multigroup IRT model should be adopted when non-hierarchical cluster analysis was used.

Self-normalized, score-based tests of models with dependent observations

Wednesday, 26th July - 15:40: Test Theory (Juan Ramon Jimenez) - Oral

Dr. Ting Wang (American Board of Family Medicine), Dr. Edgar Merkle (University of Missouri), Dr. Thomas O'Neill (American Board of Family Medicine)

Score-based tests have been used to study parameter heterogeneity across many types of statistical models. This study describes a new self-normalization approach for score-based tests of mixed models, which addresses situations where there is dependence between scores. This differs from the traditional score-based tests, which require independence of scores. We first review traditional score-based tests and then propose a new, self-normalized statistic that is related to previous work by Shao and Zhang (2010) and Zhang, Shao, Hayhoe, and Wuebbles (2011). We then provide simulation studies that demonstrate how traditional score-based tests can fail when scores are dependent, and that also demonstrate the good performance of the self-normalized tests. Next, we illustrate how the statistics can be used with real data. Finally, we discuss the potential broad application of self-normalized, score-based tests in mixed models and other models with dependent observations.

Decision-making when representing survey data: Is one dimension enough?

Wednesday, 26th July - 14:40: Measurement Applications (Thurgood Marshall) - Oral

Mx. Linda Galib (Loyola University Chicago), Dr. Ken Fujimoto (Loyola University Chicago), Ms. Naomi Brown (George Mason University), Dr. Elizabeth Levine Brown (George Mason University), Dr. Kate Phillippo (Loyola University Chicago)

Many decisions must be considered when investigating the association among variables (latent and observed), often balancing simplicity and accounting for the full data structure. For example, composite scores are used to represent latent variables in multiple regression, thereby ignoring each item's unique contribution to the underlying latent construct and assuming no measurement error. Structural Equation Modeling (SEM) addresses these limitations but introduces new decisions about the dimensionality of the latent variables. For this presentation, we investigated the impact of decisions in SEM, specifically: 1) the trade-offs of representing latent constructs with a unidimensional structure or a bifactor structure when the latter is theoretically more appropriate, and 2) how conclusions are impacted.

We used data from a survey of K-12 teachers in the United States (N=1,442), representing students' wellness, teacher competency to support student wellness, and teachers' feelings of burnout during the beginning of the COVID-19 pandemic. We examined the relationships among the variables using SEM (with unidimensional and bifactor representations for appropriate latent variables) and composite scores representing the latent variables (i.e., multiple regression models).

We found that these modeling approaches led to different results. Representing the latent variables as bifactor models in SEM led to the largest effect sizes and explained more variance in the outcome than the other approaches. In addition to demonstrating how the decisions impact findings, we discuss why attenuated effects are observed when neglecting to represent the latent variables with a bifactor structure when such a structure is theoretically more appropriate.

Global validity of assessments: Location and currency effects

Wednesday, 26th July - 14:55: Measurement Applications (Thurgood Marshall) - Oral

Ms. Ambar Kleinbort (Harver), Ms. Amy Li (Harver), Dr. Janelle Szary (Harver), Dr. Anne Thissen-Roe (Harver)

As assessments are used in an increasingly multicultural and connected world, there is a growing need to verify that they are equally valid across different populations. More specifically, when using hiring assessments to select people for jobs, it is important to corroborate that direct comparisons of individuals from different populations are valid, leading to fair and accurate hires. Populations differ in many interesting ways, but in this study, we examined how cultural group differences affect assessment behavior. Thus, we set out to disentangle the effects of location and currency, as elements of cultural behavior, on constructs used in hiring assessments: fairness, altruism, and decision-making speed. These constructs are measured in our gamified implementation of the Trust Game (Berg, Dickhaut, & McCabe, 1995) and Dictator Game (Forsythe et al., 1994). We had data from job candidates in many world regions, who responded in various languages and game money currencies. Using this, we tested the factorial invariance of the measures from the two games. We compared large groups across different regions (controlling for language and currency), such as the US and China, and across different currencies (controlling for language and region), such as Euros and Reales. While the general factor structure held across all groups, we found differences in the observed variables, with specific cultural differences by country.

Are we playing the same game? Translating fairness content

Wednesday, 26th July - 15:10: Measurement Applications (Thurgood Marshall) - Oral

Ms. Amy Li (Harver), Ms. Ambar Kleinbort (Harver), Dr. Anne Thissen-Roe (Harver), Dr. Janelle Szary (Harver)

"Traduttore, traditore." An act of translation is always an act of betrayal. For years, language has been shown to impact individual perception and decision making. Some have argued that preferences and evaluations may be modulated as a result of linguistic contexts (Vidal et al., 2021). Therefore, when translating assessments across languages and cultures, it is important to verify that the validity of these tests is preserved. Harver (pymetrics) offers hiring solutions in the form of various assessments. As international hiring increases and the candidate population diversifies, more emphasis is being placed on ensuring that the same constructs apply across populations and that the meanings are preserved. Our gamified implementations of the Trust Game (Berg, Dickhaut, & McCabe, 1995) and the Dictator Game (Forsythe et al., 1994) have been found to follow a three-factor model. However, the model is built on English-language response patterns and must be verified for fair comparisons and inferences about individuals who test in different languages. This research aims to disentangle the effects of language in observed behaviors of fairness, altruism, and decision-making speed. Factorial invariance testing reveals overall equal factor loadings with small intercept differences. Distribution analysis provides insights into the different behaviors in languages. In this talk, we will show and discuss these differences.

Development and comparison of regular and reversed items measuring the need for intimacy in the workplace

Wednesday, 26th July - 15:25: Measurement Applications (Thurgood Marshall) - Oral

Mr. Seito Nakamura (University of Tsukuba, R & D Center for Working Persons' Psychological Support), Dr. Junichi Maruyama (University of Tsukuba, R & D Center for Working Persons' Psychological Support), Dr. Soichi Nagano (University of Tsukuba, R & D Center for Working Persons' Psychological Support), Prof. Naoya Todo (Tokyo Metropolitan University), Prof. Hiroko Endo (Saitama Gakuen University), Prof. Kei Fuji (University of Tsukuba)

In Japan, loneliness in the workplace is regarded as a growing problem. To solve this problem, based on Peplau & Perlman (1982), we attempted to develop a scale to measure both the need for intimacy and the situation of isolation within the workplace. For scale development, it has been suggested to include reversed items. However, recent studies showed that reversed items have not often worked as designed by the researcher (e.g., Kam et al., 2021), even for concepts considered to be similar to loneliness in the workplace (Ferris et al., 2008). Whether this phenomenon is also observed in scales related to the need for intimacy has not yet been studied, and few studies compare how the item parameters vary between regular and reversed items based on IRT. Using data from 2000 Japanese working people, we developed the scale measuring the need for intimacy and examined how the parameters of the graded response model (Samejima, 1969) differed between regular and reversed items. The results showed that, although most of the items with high discrimination in the regular expressions had similar properties in the reversed them, reversed items had more difficulty relating to each category than the regular items. In other words, the responses to higher-scoring options were more likely to be suppressed when the textual expression was negative, even if the items had the same content. We discussed the possibility that the phenomenon of reverse item setting does not function as intended due to differences in degrees of difficulty.

Characterizing individual differences in medical certification testing

Wednesday, 26th July - 15:40: Measurement Applications (Thurgood Marshall) - Oral

Ms. Haeju Lee (University of North Carolina Greensboro), Dr. Drew Dallas (NCCPA)

One goal of medical certifications is often to provide test-enhanced learning experiences to maintain the depth of core medical knowledge and skills. The certification also measures how learners' knowledge has changed by retaking portions of the examinations where a test-taker performed poorly. Previous research has used multi-level approaches to characterize the different patterns of groups over time. For example, latent class analysis (LCA) and latent profile analysis (LPA) has been used to examine potential changes in groups based on their ability levels and external factors. While both techniques have the same goal, LCAs are based on categorical observed variables, while LPAs are based on continuous observed variables. Although both techniques have been used to characterize class membership, there is little research on comparing LCA and LPA in the context of medical certification. In addition, further research is needed to explore the relationship between the groups' performance and multiple external variables in the certification. Therefore, this current study investigates class membership, changes in group performance over time, and illustrates what external variables interact with group performance.

Emerging trends in psychometrics: Opportunities and challenges

Wednesday, 26th July - 16:15: Symposium: Emerging Trends in Psychometrics: Opportunities and Challenges (Colony Ballroom) - Symposium Overview

Dr. Kadriye Ercikan (Educational Testing Service)

This symposium will explore emerging trends in psychometrics, particularly those related to emerging technologies and personalization in assessment, and will examine innovative solutions for meeting the challenges presented by the new paradigms of assessment. The symposium will address key topics around the psychometric issues related to socioculturally responsive assessments, automatically generated content, innovative digital assessments designed for hard-to-measure constructs, and issues of fairness and validity of AI- and AI-assisted scoring algorithms.

Statistical and psychometric models for culturally responsive assessments

Wednesday, 26th July - 16:33: Symposium: Emerging Trends in Psychometrics: Opportunities and Challenges (Colony Ballroom) - Symposium Presentation

Dr. Sandip Sinharay (Educational Testing Service)

- Culturally responsive assessments have been proposed as potential tools to ensure equity for examinees from all backgrounds. This presentation discusses some psychometric issues that may arise while analyzing data that originate from some proposed forms of culturally responsive assessments. Some psychometric models are discussed, along with suggestions on the conditions under which those models may be used to analyze data from such assessments.

Predicting the psychometric properties of automatically generated items

Wednesday, 26th July - 16:36: Symposium: Emerging Trends in Psychometrics: Opportunities and Challenges (Colony Ballroom) - Symposium Presentation

Dr. Jiyun Zu (Educational Testing Service)

- Recent advances in large language models has simplified automated content creation for many types of assessment. However, the ability to generate many new items may not be that useful if we still need large sample sizes to calibrate the psychometric properties of theses items. This presentation will explore different deep learning methods to predict the psychometric properties of automatically generated items.

Assessing bias in AI-powered scoring models and developing strategies for reducing the risk

Wednesday, 26th July - 16:54: Symposium: Emerging Trends in Psychometrics: Opportunities and Challenges (Colony Ballroom) - Symposium Presentation

Dr. Matthew Johnson (Educational Testing Service)

- The widespread availability of pre-trained language models has allowed for new methods of automated scoring for short constructed response items. However, the algorithmic complexity of these models makes them particularly susceptible to algorithmic bias, thus producing scores that might disadvantage individuals based on their backgrounds. This presentation will discuss methods for investigating these types of algorithmic bias and strategies for mitigating such issues of fairness.

Assessing collaborative problem solving: Psychometric challenges and strategies

Wednesday, 26th July - 17:12: Symposium: Emerging Trends in Psychometrics: Opportunities and Challenges (Colony Ballroom) - Symposium Presentation

Dr. Jiangang Hao (Educational Testing Service)

- Developing psychometrically sound assessments for collaborative problem solving (CPS) can be very challenging. In this presentation, we summarize the main psychometric challenges drawing upon years of empirical research and introduce a set of strategies around assessment design, task design, evidence identification, scoring, and modeling to address the challenges. We also introduce how advanced AI, such as ChatGPT, can help to enable the scalable assessment of CPS.

Reconceptualizing idiographic and nomothetic relations in dynamic models powered by machine learning methods: Implications on causality inference

Wednesday, 26th July - 16:15: Machine Learning (Atrium) - Oral

Dr. Sy-Miin Chow (The Pennsylvania State University), Dr. Linying Ji (The Pennsylvania State University), Mr. Jyotirmoy Nirupam Das (The Pennsylvania State University), Dr. Orfeu Buxton (The Pennsylvania State University), Dr. Soundar Kumara (The Pennsylvania State University)

Much of the developments in continuous- and discrete-time dynamic modeling in psychometrics in the past decades have focused on confirmatory approaches for fitting differential, difference, and related dynamic models. Some of these approaches leverage fit indices and inferential approaches from related frameworks, such as structural equation modeling (SEM), to facilitate understanding and testing of causal hypotheses. In parallel, methodological developments from machine learning and data mining frameworks have made groundbreaking strides at forecasting and classification with nonparametric (e.g., deep learning) models in the absence of strong theoretical knowledge. In this talk, we use simulated examples generated using the van der Pol model, a well-known nonlinear differential equation model, to illustrate the strengths and limitations of deep learning models in approximating derivative information in the presence of mixed effects, and without precise knowledge of the true processes of change. We use a similar framework to illustrate how idiographic and nomothetic change processes can be integrated when using machine learning approaches to perform inferences with intensive longitudinal data from multiple subjects. Other possible extensions, challenges, and implications on causality inference are discussed.

Multilevel causal machine learning methods for CATE with confidence bands

Wednesday, 26th July - 16:30: Machine Learning (Atrium) - Oral

Prof. Jee-Seon Kim (University of Wisconsin-Madison), Ms. Xiangyi Liao (University of Wisconsin-Madison), Prof. Wen Wei Loh (Emory University)

Treatment effect heterogeneity is critical yet often overlooked in the social and behavioral sciences. A one-size-fits-all approach to causal inference is inadequate and can even be detrimental for certain treatments and interventions whose benefits are limited to certain subgroups. While recent studies have explored various methods for estimating conditional average treatment effects (CATEs), most have concentrated solely on obtaining point estimates of the CATE and paid little attention to its variance or precision for statistical inference. However, the difficulty of correctly quantifying the degree of sampling uncertainty is compounded by using data-adaptive nonparametric algorithms and is thus critical for drawing correct inferences and making effective decisions.
This study comprehensively examines several multilevel causal machine learning methods that account for the nested structure of data in observational studies. The pool of methods under investigation includes multilevel versions of widely-used causal forests (CF), Bayesian additive regression trees (BART), and meta-learners. We investigate the accuracy and sensitivity of CATE estimates generated by these methods, as well as the error bars and confidence bands for varying values of CATE. Confidence bands are constructed based on asymptotic normality of the CATE estimator when theoretical justifications are available, or using cluster bootstrapping when a distributional assumption is not warranted. We conduct extensive simulation studies and empirical comparisons of these methods with large-scale assessment data. We clarify important factors for the empirical consistency and efficiency of the CATE estimation methods. The paper concludes by summarizing the strengths and limitations of the compared methods for researchers in practice.

MxML: Exploring the relationship between measurement and machine learning in recent history

Wednesday, 26th July - 16:45: Machine Learning (Atrium) - Oral

Prof. Yi Zheng (Arizona State University), Dr. Steven Nydick (Duolingo), Prof. Sijia Huang (Indiana University Bloomington), Prof. Susu susu.zhang1992@gmail.com (University of Illinois Urbana-Champaign)

The explosive growth of machine learning (ML) over the past decades has impacted many disciplines, including educational and psychological measurement. The measurement literature has seen a growing volume of studies exploring ML methods to solve measurement problems. However, the typical ML paradigm conflicts with several fundamental principles of measurement. Looking to the future, our MxML project aims to understand how the Zeitgeist of ML may change the measurement landscape and what role ML will play in future measurement practices. To conjecture about the future, we first need to examine the past. As Phase I of the MxML project, this study systematically examines the latest 10 years' literature to explore the role ML has played in measurement. Specifically, we have (1) systematically searched databases (i.e., Scopus, ProQuest, Wiley) to identify an initial collection of articles (N=1314) potentially relevant for the intersection between ML and measurement, (2) screened all articles to determine whether they exemplified the intersection between ML and measurement, and we are currently (3) reviewing the full-texts of the included articles (N=170 for the last 10 years) to extract themes and form a theory based on the grounded theory methodology. We will present results in terms of (1) the different perspectives the measurement community exhibited towards ML; (2) the areas of measurement that have been discussed in light of ML; (3) the categories of ML methods that have been explored; and (4) how the gaps between the ML and measurement principles have been addressed.

Data preprocessing techniques using machine learning algorithms in large-scale assessment

Wednesday, 26th July - 17:00: Machine Learning (Atrium) - Oral

Mrs. Mingying Zheng (University of Iowa)

Different online assessments have generated a huge volume of “big data” from large-scale digitally based learning management and assessment platforms. Such big data with complex structures that are so readily available will soon influence the way we define, operationalize, and derive evidence of target constructs in measurement. Today, with the expansion of rich data and rapid advances in computer technologies, psychometrics should incorporate innovative and advanced techniques from machine learning that can contribute to the advancement of behavioral and social sciences.

The appropriate utilization of innovative machine learning methods to efficiently explore such rich data and to maintain validity and fairness of interpretation and the proper use of scores should be the primary interests in educational research.

The current study proposes different data preprocessing techniques including standardization, normalization, discretization, and non-linear transformation in addition to some basic data cleaning techniques (e.g., data formatting and dropping missing values) to handle raw data with complex structure for the purpose of improving estimators’ accuracy scores and to boost the model performance by using a dataset from a large-scale assessment. Different machine learning algorithms are used to examine the performance of proposed data preprocessing techniques. Accuracy scores, precision, recall and area under the ROC curve (AUC) are compared and rank ordered on different machine learning models. Recommendations are given based on the results of different data preprocessing techniques using machine learning algorithms.

A mixed effects model in machine learning

Wednesday, 26th July - 17:15: Machine Learning (Atrium) - Oral

Dr. Pascal Kilian (University of Tübingen), Dr. Sangbeak Ye (University of Tübingen), Prof. Augustin Kelava (University of Tübingen)

Clustered data can frequently be found not only in social and behavioral sciences (e.g., multiple measurements of individuals) but also in typical machine learning problems (e.g., weather forecast in different cities, house prices in different regions). This implies dependencies for observations within one cluster, leading to violations of independent and identically distributed assumptions, biased estimates, and false inference. A typical approach to address this issue is to include random effects instead of fixed effects. We introduce the general mixed effects machine learning framework (mixedML), which includes random effects in supervised regression machine learning models, and present different estimation procedures. A segmentation of the problem allows to include random effects as an additional correction to the standard machine learning regression problem. Thus, the framework can be applied on top of the machine learning task, without the need to change the model or architecture, which distinguishes mixedML from other models in this field. With a simulation study and empirical data sets, we show that the framework produces comparable estimates to typical mixed effects frameworks in the linear case and increases the prediction quality and the gained information of the standard machine learning models in both the linear and non-linear case. Furthermore, the presented estimation procedures significantly decrease estimation time. Compared to other approaches in this area, the framework does not restrict the choice of machine learning algorithms and still includes random effects.

Exploring the effect of item calibration and scoring methods on growth mixture model results

Wednesday, 26th July - 16:30: Growth Mixture Models and Longitudinal Analysis (Benjamin Banneker) - Oral

Dr. James Soland (University of Virginia), Dr. Veronica Cole (Wake Forest University), Mr. Stephen Tavares (University of Virginia)

Growth mixture models (GMMs) are an extremely popular approach to modeling human development. However, a large body of methodological research has shown that the results of GMMs are highly sensitive to a number of common model specification choices. We know less about how measurement issues impact GMMs. In particular, when scores resulting from multiple items are used in GMMs, how do choices about item calibration and scoring impact the number and nature of classes GMMs find? We first address this question with an empirical data analysis, using two datasets (N=9212 and N=4337) measuring social-emotional learning outcomes in children. We generated scores in a variety of ways, including sum scores, as well as EAP and MLE scores resulting from unidimensional, multidimensional, and multiple-group IRT models. We then applied GMMs to all of the resulting scores. For each of the two datasets, the number of classes favored by fit indices in GMM (including information criteria and likelihood ratio tests) varied widely over the scoring methods, ranging from 3-7 classes in the first dataset and 2-7 classes in the second. However, this variation did not resolve according to a clear pattern; contrary to our predictions, more heavily parameterized scoring models did not result in GMMs with fewer classes. Moreover, this being an analysis of real datasets, it was unclear whether any of the chosen scoring models was the correct one. Thus, we followed up with a simulation study, varying parameters both in the mixture and measurement components of the data-generating model.

Interpretable machine learning vs. linear mixed models for longitudinal data

Wednesday, 26th July - 16:45: Growth Mixture Models and Longitudinal Analysis (Benjamin Banneker) - Oral

Ms. YOUNG WON CHO (The Pennsylvania State University), Prof. Sy-Miin Chow (The Pennsylvania State University)

This study compares two approaches to analyzing longitudinal data: mixed-effect random forest (MERF) models with SHapley Additive exPlanations (SHAP), and conventional linear mixed-effect models (LMM). We demonstrate how SHAP can help us interpret the effects of variables in machine learning models and discuss the advantages and disadvantages of using SHAP compared to LMM with empirical data collected from couples over 22 days.

First, multiple MERF models, ranging from simple random intercept to complex random slope models, were fitted. The best model was selected based on validation loss and mean squared error, and an LMM that best matched the selected MERF model was fitted. A SHAP summary plot was used to interpret the importance and effects of variables on model prediction.

Our results show the top three influential variables identified by SHAP were also significant predictors in LMM, indicating that both methods can provide similar insights. However, SHAP was able to reveal non-linear interaction effects and between-individual heterogeneity that were not captured by LMM. This suggests that using SHAP in combination with machine learning models enables extremely efficient model explorations and synthesis of results, and provide novel, interpretable insights that complement conventional hypothesis testing approaches. However, SHAP's quantification of predictor importance also requires some further considerations in cross-validating results across studies that utilize different measurement scales as it is based on changes in the units of the dependent variables. Overall, our study provides insights into the potential benefits and limitations of using SHAP with machine learning models compared to conventional LMM.

An investigation of missing data analytical methods in longitudinal research: Traditional and machine learning approaches

Wednesday, 26th July - 17:00: Growth Mixture Models and Longitudinal Analysis (Benjamin Banneker) - Oral

Ms. Dandan Tang (University of Virginia), Prof. Xin Tong (University of Virginia)

Missing data are inevitable in longitudinal studies. Traditional methods, such as the full information maximum likelihood (FIML) and the 2-stage robust (TSR) methods, are commonly used and perform well for handling ignorable missing data. However, they may lead to biased model estimation for non-ignorable missing data. Recently, machine learning methods, such as tree-based (TB) and K-nearest neighbors (KNN) imputation methods, have been proposed to cope with missing values. Although the machine learning imputation methods have been gaining popularity, few studies have investigated the tenability and utility of these methods in longitudinal research. Through Monte Carlo simulations, this study systematically evaluates and compares the performance of traditional and machine learning approaches (FIML, TSR, TB, and KNN) in growth curve modeling with ignorable and non-ignorable missing values and potential outliers. The effects of sample size, missing data mechanism, missing data rate, and outliers rate on four outcome measures are investigated: convergence rate, parameter estimate bias, standard error, and model goodness of fit. Results indicate that the performance of these methods varies across different conditions. The study provides practical advice for longitudinal data analysts and uses a real-data example to illustrate the application of different methods.

A two-stage approach to a latent variable mixed-effects location scale model

Wednesday, 26th July - 17:15: Growth Mixture Models and Longitudinal Analysis (Benjamin Banneker) - Oral

Prof. Shelley Blozis (University of California, Davis), Dr. Mark H. C. Lai (University of Southern California)

Understanding the within- and between-person variation in repeated measures is central to behavioral investigations. Mixed-effects location scale models include distinct variance models to permit study of heterogeneity of within- and between-person variation. Recent developments have extended the model to address measurement error in the longitudinal response. Accounting for variation in the response that is due to measurement error is especially important in behavioral studies that focus efforts to understand the within-person variation. Relative to a mixed-effects location scale model for a variable assumed to be measured without error, the latent variable version of the model is more complicated and this complexity can be carried over to increased computational demands. This paper proposes a two-stage approach in which factor scores and their corresponding standard errors of measurement are estimated and then incorporated into a mixed-effects location scale model. This paper considers the approach in the context of daily diary data from a large sample of adults in the U.S.

A Bayesian hierarchical item response theory model for estimating attributes of regular exams and students' knowledge levels

Wednesday, 26th July - 16:15: Bayesian IRT (Prince George) - Oral

Mr. Jiatong Li (School of Data Science, University of Science and Technology of China), Dr. Mengxiao Zhu (Department of Communication of Science and Technology, University of Science and Technology of China), Dr. Xiang Liu (Educational Testing Service)

Abstract: Several latent growth item response theory (IRT) models (e.g. Embretson, 1991) have been developed in previous research. These models typically require anchor items across multiple time points of measurement. However, in most K-12 classroom assessment scenarios, anchor items are often not available. In this talk, we will introduce a Bayesian hierarchical IRT model that can fit longitudinal measurement data of the same population of students without anchor items. As a result, inferences about students' relative proficiency levels, as well as relative test difficulty levels at each time point, can be drawn. In addition, we will discuss methods of checking the goodness-of-fit of the model and comparing alternative model specifications. To demonstrate the utilities of the proposed model, we will analyze a real dataset consisting of 854 high school students being measured 6 times during three years on the English subject from a non-native English-speaking country.

Sparse Bayesian joint modal estimation for item factor analysis

Wednesday, 26th July - 16:30: Bayesian IRT (Prince George) - Oral

Mr. Keiichiro Hijikata (The University of Tokyo), Mr. Motonori Oka (London School of Economics and Political Science), Dr. Kensuke Okada (The University of Tokyo)

In this study, we develop a sparse estimation method for item factor analysis (IFA) models. The proposed method exploratorily reveals the sparse structure of factor loadings with a Bayesian joint modal estimation (BJME) approach. In the Bayesian estimation framework, Markov chain Monte Carlo (MCMC) methods are most frequently used; however, they are often time-consuming or infeasible for parameter estimation, especially when the sample size and the number of latent/observed variables are large. In contrast, a BJME method estimates parameters of interest by directly maximizing joint posterior probabilities, allowing for computationally efficient parameter estimation in such large-scale settings. In the IFA framework, Bergner et al. (2022) developed a BJME approach to a shrinkage estimation with L2 penalty for dichotomous responses. However, the method does not consider the sparsity of latent variables and is only applied to binary responses. In psychological research, the sparse structure of factor loadings is preferred because it facilitates interpretability in understanding the relationship between latent and observed variables. Furthermore, instead of binary responses, polytomous response data are commonly used in psychological research. Hence, we propose a BJME method for estimating the sparse structure of factor loadings with L1 penalty that can be applied to polytomous response data. We carried out simulation and empirical studies to evaluate our algorithm's estimation accuracy and scalability compared to an MCMC method. The results will be presented at the talk.

WAIC and PSIS-LOO for Bayesian diagnostic classification model selection

Wednesday, 26th July - 16:45: Bayesian IRT (Prince George) - Oral

Ms. Ae Kyong Jung (University of Iowa), Prof. Jonathan Templin (University of Iowa)

Bayesian diagnostic classification models (Bayesian DCMs) are becoming increasingly popular for diagnosing students' skills. However, selecting the appropriate Bayesian DCM can be challenging, and the necessity of relative model fit indices has emerged. This study aims to investigate the performance of Bayesian relative model fit indices, including the widely applicable information criterion (WAIC) and leave-one-out cross-validation using Pareto Smoothed importance sampling (PSIS-LOO), in comparison to simpler and more widely used model fit indices, such as AIC and DIC.

The study would conduct a simulation to evaluate the performance of WAIC and PSIS-LOO by detecting the true Bayesian log-linear cognitive diagnosis model (Bayesian LCDM), which is a generally parameterized DCM. The simulation study includes variants of the Bayesian LCDM with varying sample sizes and test lengths. The R package called 'blatent' will be used for computation and demonstration.

The results of the study would indicate that WAIC and PSIS-LOO outperform AIC and DIC in detecting the true Bayesian LCDM, especially for smaller sample sizes and shorter test lengths. Therefore, the use of WAIC and PSIS-LOO as model fit indices for Bayesian DCMs can improve the accuracy of the information provided to students and aid researchers and practitioners in selecting the appropriate model for diagnosing student skills.

In conclusion, the findings of this study suggest that Bayesian relative model fit indices, specifically WAIC and PSIS-LOO, should be used in conjunction with Bayesian DCMs to select the appropriate model for diagnosing student skills accurately.

A comparative study of Bayesian samplers in MCMC estimation for joint modeling of response, response time, and item revisit count

Wednesday, 26th July - 17:00: Bayesian IRT (Prince George) - Oral

Ms. Jinglei Ren (University of Maryland, College Park), Prof. Hong Jiao (University of Maryland, College Park)

Bayesian joint modeling of response, response time, and count data provides a new perspective in analyzing item-level product and process data from psychological and educational assessments. This proposed study aims to compare the accuracy and efficiency of different Bayesian sampling algorithms in Markov Chain Monte Carlo (MCMC) estimation of parameters in the joint models using MultiBUGS, Stan, NIMBLE, and the MCMC R package. MultiBUGS is the latest version of BUGS which is under development. It uses the default Gibbs sampling algorithm, but automatically selects other algorithms such as adaptive rejection sampling, slice, sampling, and Metropolis-Hasting sampling if needed. NIMBLE builds on BUGS and extends it by allowing manual selection of sampling methods. Stan is a popular and efficient MCMC software package that implements the Hamiltonian Monte Carlo (HMC) algorithm. The R package “MCMC” is a flexible MCMC software package that provides several MCMC algorithms and allows for customization of MCMC sampling strategies. This study compares different samplers in these four software programs in terms of their model parameter estimation accuracy, computational efficiency, and convergence properties. data simulation study will be conducted to evaluate the performance of the different samplers under different study conditions due to varying sample sizes, model misspecification, and correlations among the latent parameters. The results of this study will provide guidance on the selection of the most appropriate Bayesian sampler for MCMC estimation of joint models for response, response time, and item revisit count, which has important implications for applications in psychology, education, and other fields.

The effect of the Projective IRT model on DIF detection

Wednesday, 26th July - 16:15: Differential Item Functioning (Margaret Brent) - Oral

Dr. Ye Ma (aws), Dr. Terry Ackerman (the University of Iowa), Dr. Edward Ip (Wake Forest University School of Medicine)

Ip (2010) developed the Projective IRT (PIRT) model to eliminate unwanted dimensionality in test response data. Several researchers have shown that when items are measuring irrelevant dimensions and groups of examinees differ in their latent ability distributions on these invalid dimensions, there are possibilities that DIF exists. The purpose of this study is to examine the potential of the Projective IRT Model to control DIF in dichotomously scored tests.

In this study we first simulated several MIRT data sets to determine if the PIRT approach would eliminate or reduce the effect of DIF. It was assumed that θ_1 was the valid dimension and θ_2 was the nuisance dimension. Four factors were examined: correlation, ($r_{\theta 1 \theta 2}$ =.0, .3, .6 and .8), angular composite of the potential DIF item with the θ_1-axis (i.e., the amount the invalid skill is measured, $\alpha = 30^o, 45^o, 60^o$), sample size (500, 1500, 3000), and mean difference of the Reference and Focal groups on the invalid dimension ($\mu Ref_{\theta 2} - \mu Foc_{\theta 2} = 0, .5, 1.0$). The number of items was fixed at 33. The 30 non-DIF items had angular composites from 0^o to $5^{o.}$ The difficulties of the non-bias items ranged from -2 to +2 whereas DIF item difficulty was fixed at 1.0. There will be 100 iterations for each cell in the totally crossed design.

Preliminary results show that differences in latent ability distributions on the invalid dimension can result in DIF when items measure primarily the invalid dimension. More results about DIF detection will be presented.

Comparing Frequentist and Bayesian approaches for detecting differential item functioning

Wednesday, 26th July - 16:30: Differential Item Functioning (Margaret Brent) - Oral

Dr. Eric Schuler (American University), Ms. Huan KUANG (University of Florida)

There is an increase in using Bayesian methods for latent variable modeling. Bayesian methods for detecting measurement invariance (MI), namely Bayesian approximate MI (Muthén & Asparouhov, 2018) and Bayesian interval MI (Shi et al., 2019) have been developed, and compared with Frequentist (Authors, 2023). Relatedly, researchers applied the Bayesian approximate method (Sideridis, Tsaousis & Alamri, 2020) for detecting differential item functioning (DIF) under the item response theory. However, other Bayesian methods were less discussed in the literature, and there has not been a comparison of different Bayesian and Frequentist DIF methods.

To address this gap, we propose and investigate the application of the Bayesian region of practical equivalence (ROPE) for detecting DIF and compare the proposed method with existing Bayesian DIF methods and common Frequentist DIF tests. In this study, we will first walk through how the Bayesian ROPE that can be applied to detecting DIF and then, conduct a Monte Carlo simulation study to compare Bayesian ROPE, Bayesian approximate method, and three Frequentist DIF tests.

The dichotomous item response data were simulated based on three conditions: (a) group sample size (150, 300, 450, 600, 750, 900), (b) categorization of DIF (negligible, medium, large) based on Lin and Lin's (2014) modification of the Educational Testing Service's DIF categorizations, and (c) type of DIF (uniform or non-uniform). Results and interpretations will be discussed along with recommendations based on the simulation results. We hypothesize that both Bayesian methods will outperform the Frequentist methods and provide more information for decision making purposes.

The deconstruction of measurement invariance (and DIF)

Wednesday, 26th July - 16:45: Differential Item Functioning (Margaret Brent) - Oral

Dr. Safir Yousfi (German Federal Employment Agency)

Measurement invariance is violated if some item parameters of a measurement model differ across groups (DIF: differential item functioning). However, due to identification issues it is often hard to tell which item parameters actually differ. Moreover, DIF describes a situation where a set of items measures a latent variable in each group, but the same trait is only measured across groups if the DIF items are deleted. However, deleting the DIF items does not change the latent variable in each group. Given these considerations, the question arises whether DIF and measurement invariance (as the absence of DIF) have a viable conceptual basis. Indeed, it can be shown that each DIF model (derived from an one-dimensional IRT model or a compensatory latent variable model) is equivalent to higher-dimensional latent variable model with a degenerate distribution of person parameters in each group. Parameter vectors of individuals of different groups are almost surely (P=1) unequal which renders DIF models implausible. More realistic options of modelling data that seems to violate measurement invariance will be proposed.

An examination of the interaction between reliability and DIF detection

Wednesday, 26th July - 17:00: Differential Item Functioning (Margaret Brent) - Oral

Dr. Terry Ackerman (the University of Iowa), Dr. Ye Ma (Amazon Web Services), Dr. Jinmin Chung (the University of Iowa)

This study examines how reliability, sample size, and test length affect the performance of four DIF methods. A fully crossed four-factor Monte Carlo research design was used. The factors examined were: **DIF detection methods** (Mantel-Haenszel, SIBTEST, Lord's χ^2, and Raju's DFIT), **reliability levels (**low .2-.4, medium .4-.7, high, .7-.9**)**, **types of DIF** (uniform/non-uniform), **sample size** (200/500/1000) and **test length** (20/40). DIF was created by adjusting the generated 2PL IRT b and a parameters, to simulate non-uniform and uniform DIF respectively for the reference and focal groups. Reliability (α, IRT) was adjusted by using different levels of discrimination when simulating the response data. In all, each of the 144 conditions were replicated 100 times.
Results indicated in 40-item test:
▪ DIF detection rate increased as reliability increased in Non-uniform/ Uniform DIF conditions. DIF detection rate increased as sample size increased in Non-uniform/Uniform DIF conditions. It appears that the IRT methods had better detection rates in most conditions than nonparametric methods. The MH procedure had much lower detection rates when the Non-uniform DIF was symmetric.
In 20 items test:
• The DIF detection rates increased as reliability increased as sample size increased in non-uniform/ Uniform DIF conditions with similar pattern in 40 items test. MH method showed highest detection rates in Uniform DIF condition, but low detection rates for non-uniform DIF items regardless of the test length. The SIBTEST method's detection rates are higher in 20-item test compared to its performance in 40- item test when sample size is 200.

Pervasive DIF and DIF detection bias

Wednesday, 26th July - 17:15: Differential Item Functioning (Margaret Brent) - Oral

Dr. Paul De Boeck (Ohio State University)

The stream of methodological studies of differential item functioning (DIF) is unstoppable. One reason is that DIF detection is very important. Another reason is that DIF cannot be identified without implicit or explicit assumptions. We believe that most approaches minimize DIF detection (and detection of measurement invariance violations), including the regularization methods and factor model alignment methods. In the presence of DIF that pervades a large part of the test, minimizing DIF detection can have far reaching consequences for group mean differences. A possible cause of pervasive DIF is that one or more item covariates (item properties) affect the item difficulties (intercepts) in a way that depends on group membership. This can be the cause of group mean differences. The principle is explained by De Boeck and Cho (2021).

We will present:

1. A simulation study showing that pervasive DIF can create group mean differences while no substantial DIF is detected. Traditional DIF methods and measurement invariance methods fail to detect pervasive DIF that causes group mean differences.
2. An outline for the detection of pervasive DIF.
3. Results from a pervasive DIF analysis using data from the Boston Naming Test (BNT; Kaplan, et al., 1983). The BNT is the most commonly used language test among neuropsychologists and a critical diagnostic step in detecting and characterizing neurocognitive disorders. A common finding is that there are group mean differences between Caucasians and African Americans. The research question is whether not pervasive DIF can explain these group mean differences.

Prediction-based selection of individual predictors in generalized structured component analysis

Wednesday, 26th July - 16:15: Structural Equation Models (Juan Ramon Jimenez) - Oral

Ms. Bellete Lu (McGill University), Mr. Gyeongcheol Cho (McGill University), Dr. Heungsun Hwang (McGill University)

Generalized structural component analysis (GSCA) is a component-based approach to structural equation modeling (SEM). It is a constituent of GSCA-SEM, which focuses on specifying and estimating models with components only. GSCA can be useful for predictive modeling as it can produce the deterministic scores of components for new observations. An overall cross-validation index and prediction-oriented specification algorithms have been developed to search for the best GSCA model in terms of predictive generalizability. However, both the index and algorithms are geared toward evaluating the model's overall predictive performance. There exists no index or algorithm available for evaluating the contribution of an individual predictor to the prediction of outcome variables in the model. To address this issue, we propose a new prediction-based index for GSCA, called variable importance (VIMP) cut-off, which informs whether each predictor contributes to reducing prediction error for an outcome variable in the model. The calculation of this index is well-coupled with GSCA's estimation procedure, which utilizes the bootstrap method to estimate the standard errors or confidence intervals of parameter estimates. We conduct a simulation study to test the performance of the VIMP cut-off and apply the index to real data to demonstrate its practical usefulness.

Examining SEM trees for investigating measurement invariance concerning multiple violators

Wednesday, 26th July - 16:30: Structural Equation Models (Juan Ramon Jimenez) - Oral

Dr. Yuanfang Liu (University of Cincinnati), Dr. Mark H. C. Lai (University of Southern California)

Measurement invariance is needed for accuracy and meaningful interpretations of statistical results. In addition, measurement invariance status can be associated with multiple covariates. This study explored a novel use of the structural equation model (SEM) tree (Brandmaier et al., 2013) to detect measurement noninvariance concerning multiple violators (i.e., covariates contributing to noninvariance) under a Monte Carlo simulation. We evaluated measurement invariance under the SEM tree by comparing three models: one pre-split model assuming full invariance and two post-split models with different invariance status assumptions. Specifically, the pre-split model assumed that no covariates were associated with data split due to noninvariant measurement parameters. One post-split model assumed a targeted invariance level (e.g., intercept invariance only), and one post-split model did not assume that invariance. Preliminary results showed that likelihood comparisons under the SEM tree had Type I error rates ≤ .052 when $n \leq 1000$ and statistical power rates of .964–1.00 in detecting both linear and quadratic nonlinear intercept noninvariance when $n = 1000$. The SEM tree distinguished true violators from noise covariates well related to intercept noninvariance in terms of split rates ≥ .928, $n = 1000$, especially a dichotomous violator. The SEM tree can identify covariates associated with parameter estimate heterogeneity. The invariance testing can be applied to datasets with many covariates to uncover relations between items, covariates, and constructs and thereby facilitate instruments and theory development.

Keywords: measurement invariance, SEM trees, multiple violators

Accommodating multiple time metrics using the latent change score model

Wednesday, 26th July - 16:45: Structural Equation Models (Juan Ramon Jimenez) - Oral

Dr. Sarfaraz Serang (University of South Carolina), Dr. Shawn Whiteman (Utah State University)

Longitudinal models can model change as a function of time, but time can be operationalized in many ways. Longitudinal studies often collect data in waves, which serve as a natural time metric. However, waves of data collection can be wide enough such that there is substantial variability in how much chronological time has passed between observations for different individuals. Researchers with multiple time metrics are often forced to choose a single time metric for their models, for example wave or age, while treating the other as a covariate instead of a time metric or simply not involving the other at all. We propose a novel methodological approach for accommodating multiple time variables as time metrics in the same model simultaneously using the latent change score model framework. We discuss benefits and limitations of this approach, including necessary constraints and possible applications. We also demonstrate the practical utility of the approach using data on adolescent expectations involving marijuana use collected over the course of the COVID-19 pandemic.

Two-stage asymptotically distribution free method in structural equation modeling

Wednesday, 26th July - 17:00: Structural Equation Models (Juan Ramon Jimenez) - Oral

Mr. You-Lin Chen (National Taiwan University), Dr. Li-Jen Weng (National Taiwan University)

Asymptotically distribution free (ADF) method (Browne, 1984) that requires no specific distribution form for data in structural equation modeling (SEM) with covariance matrices is a legitimate approach to analyze non-normal data commonly observed. Yet ADF test statistics tend to overreject the correct models unless with very large samples. In the context of SEM for correlation matrices, Steiger and Hakstian (1982) suggested a two-stage ADF (TADF) method to improve the performance of ADF statistics. The first stage of TADF estimates the structured weight matrix by model-reproduced correlations and the second stage applies this weight matrix with the ADF method. TADF was also introduced by Yuan and Bentler (1997) for covariance structure analysis. This method has been evaluated only in SEM for correlation matrices under restricted conditions (Bentler & Savalei, 2010; Mels, 2000). In this study, six levels of sample size (150, 250, 500, 1,000, and 5,000) were crossed factorially with four levels of data distribution (normal, elliptical, skewed factor, and skewed error) to evaluate the performance of TADF in SEM with covariance and correlation matrices. Preliminary results indicate that, under the true model, mean TADF statistics were closer to the expected values than ADF across all the conditions. The empirical Type I error rates of ADF statistics were much higher than TADF, yet TADF rejected the true model less than expected in most conditions. TADF also produced more accurate parameter estimates and standard errors than ADF, though with underestimated standard errors at small sample sizes.

An empirical study of testing nonlinear latent relationships

Wednesday, 26th July - 17:15: Structural Equation Models (Juan Ramon Jimenez) - Oral

Prof. Fan Yang-Wallentin (Department of Statisitcs, Uppsala University)

Nonlinear latent relationships exist in many empirical settings. In this presentation, we demonstrate two different estimation approaches, parametric and semiparametric, on an organizational behavior dataset. Within the organizational context, we expect nonlinear latent relationships. The results show that, when compared with imposing a quadratic functional form, the semiparametric approach has greater flexibility in estimating the true functional form of the nonlinear latent relationships.

Psychometric perspectives on modeling and measuring synergistic risk factors

Wednesday, 26th July - 16:15: Measurement Applications (Thurgood Marshall) - Oral

Dr. Wilco Emons (Tilburg University)

Personality traits may be associated with unfavorable health prospects. However, research has suggested that oftentimes the negative effects on health are only substantial if multiple risk factors are simultaneously present. For example, distressed personality (Type D) theory hypothesizes that experiencing negative emotions only leads to an increased risk of a cardiac arrest if it is accompanied by a tendency to inhibit these emotions. Such constellations of contingent risk factors are referred to as synergies. For dichotomous risk factors, the conceptualization of a synergy is rather straightforward. However, for continuous risk factors, such as personality traits, the concept of synergy is much more more involved. A complicating issue is the fact that personality measures cannot be observed directly and have to be assessed using questionnaires. The resulting scores have measurement errors and, moreover, clinically oriented personality questionnaires are often only informative for trait differences over a limited range. In this presentation, I will take a psychometric (latent variable) perspective on defining and studying synergies between continuous psychological risk factors. In particular, I will present different (psychometric) models providing different conceptualizations of a synergy and discuss different analytic methods to unravel synergies in empirical data, while taking into account measurement properties. Results will be presented from simulations and from empirical examples. Implications will be discussed both at the group level (i.e., for use in scientific studies) and at the individual level (i.e., for individual-level clinical predictions).

Improving preference analysis: Joint models for ordinal and cardinal data

Wednesday, 26th July - 16:30: Measurement Applications (Thurgood Marshall) - Oral

Mr. Michael Pearce (University of Washington), Dr. Elena Erosheva (University of Washington)

Preference data is used to make important decisions in many real-world settings, including peer review, elections, and policy surveys. Traditionally, preferences are elicited from individuals in the form of either ordinal (e.g., rankings or pairwise comparisons) or cardinal (e.g., ratings or scores) data. Psychological and psychometric literatures point out several distinct and complementary properties between ordinal and cardinal data. Notably, ordinal data allow for scale-free but coarse comparisons, while cardinal data provide granular assessments with only implicit comparisons that may be highly subjective. Despite some suggestions that joint models may improve our ability to understand preferences and make accurate decisions, few principled methods exist. In this work, we propose a statistical modeling framework to jointly analyze ordinal and cardinal data. Our framework flexibly allows for the incorporation of different types of ordinal and cardinal preference data, estimates heterogeneity in preferences among the individuals, measures the strength of consensus, and quantifies the inherent uncertainty in estimated preferences via a fully Bayesian statistical approach. We demonstrate our models' ability to accurately learn preferences and make decisions using large-scale conference and small-scale grant peer review data.

Investigating interactive patterns in simulation-based inquiry tasks using sequential analysis

Wednesday, 26th July - 16:45: Measurement Applications (Thurgood Marshall) - Oral

Ms. Shuang Wang (Beijing Normal University), Dr. An Hu (Peking University), Dr. Wei Tian (Beijing Normal University), Dr. Tao Xin (Beijing Normal University)

Scientific inquiry is one of the most crucial competencies in the 21st century. Conducting inquiry activities via interactive simulations has gained more and more popularity in recent years. Mining students' behaviors within these environments using process data approaches can identify interactive patterns which may unveil inquiry processes and applied strategies. While a body of previous research has paid attention to students' behaviors when conducting scientific inquiry simulations, only a few studies have focused on such interaction processes from the temporal perspective. In the current study, a total of 334 fourth-graders were engaged in simulation-based scientific inquiry tasks and their interactions with the system were recorded and analyzed. This study integrated two sequential analysis methods: lag sequential analysis revealed the significant adjacent behaviors, and sequential pattern mining identified the most frequent sequential patterns. Overall results demonstrated that during the inquiry task, students' behaviors were involved mostly in two-way transitions between collecting evidence and drawing conclusions. The successful group showed more active and strategic patterns in collecting evidence, for instance, the intentional use of the control of variables strategy. The unsuccessful group exhibited more reckless and less task-oriented patterns such as rushing to conclusions with insufficient evidence or conducting simulated experiments without aforehand planning. The findings can shed some light on both the task design for interactive inquiry tasks and the scaffolding design for inquiry-based learning.

Prediction of cognitive impairment using machine learning models

Wednesday, 26th July - 17:00: Measurement Applications (Thurgood Marshall) - Oral

Dr. Lihua Yao (Northwestern University Feinberg School of Medicine Department of Medical Social Sciences), Dr. Yusuke Shono (Claremont Graduate University), Dr. Elizabeth McManus Dworak (Northwestern University), Dr. Cindy Nowinski (Northwestern University), Dr. Marie Curtis (Northwestern University), Dr. Aaron Kaat (Northwestern University), Dr. Emily Ho (Northwestern University), Ms. Zahra Hosseinan (Northwestern University), Dr. Michael Wolf (Northwestern University), Dr. Richard Gershon (Northwestern University)

Early detection of Cognitive impairment (MCI) is very important for aged

adults. MyCog is a self-administered cognitive screening assessment designed for use in different clinical settings. It is a standardized iPad-based assessment for older adults or any patient with a concern of impairment. This paper examined 105 Clinical patients who were administered MyCog two assessments: Picture Sequence Memory (PSM) and Dimensional Change Card Sorting (DCCS). Different models such as varying combinations of the number of correct scores, unidimensional and multidimensional item response theory
models, and different machine learning models were examined for the purpose for a better prediction of MCI patients. Supervised machine learning models both traditional machine-learning classifiers and neural-network-based approaches were applied.
The highest recall/sensitivity value of 0.875 was found for model ANN, a neural network, with a AUC, Precision , Accuracy, and F1 values of 0.906, 0.875, 0.9, and 0.875, respectively. The features used contains the composite scores derived from multidimensional 2 parameter model, item response times, age, education, race and income. The impairment is a complicated issue and any linear or transformed linear prediction using only the assessment scores is not sufficient for predicting. Incorporating information such as item response time, age, education, and income into the summarized scores for the assessment is a powerful tool in detecting impairment.

Generating reading assessment passages using a large language model

Wednesday, 26th July - 17:15: Measurement Applications (Thurgood Marshall) - Oral

Dr. Ummugul Bezirhan (Boston College), Dr. Matthias von Davier (Boston College)

The increasing popularity of computer-based assessments resulted in a greater demand for high-quality items. Unfortunately, the process of creating items is typically expensive and labor-intensive due to its heavy reliance on human content specialists. While automated item generation has been used for some time, the use of machine learning algorithms can significantly improve the efficiency and effectiveness of the process.

In this research, we describe an approach that uses OpenAI's latest Transformer-based language model, GPT-3, to generate reading passages for the Progress in International Reading Literacy Study (PIRLS). Creating high-quality reading passages is challenging since they must be written at an appropriate complexity level, engaging, and relevant to the intended audience. This process requires significant time, effort, and proficiency from subject matter experts, editors, and other professionals involved in the development process.

We used carefully engineered prompts to ensure that the AI-generated text had similar content and structure to 4th-grade reading passages used in previous PIRLS assessments. Multiple passages were generated for each prompt, and the final passage was chosen based on its Lexile score agreement with the original passage. Lastly, human editors reviewed the selected passage to ensure it was free of grammatical and factual errors. All passages, including original ones, were evaluated by judges based on their coherence, readability, and appropriateness. Initial results showed that the AI-generated passages were as appropriate for 4th graders as the previously used PIRLS passages. We will present further results for both literary and informational passages generated using GPT-3.

Innovative methods for variable selection in different scenarios of model building

Thursday, 27th July - 09:00: Symposium: Innovative Methods for Variable Selection in Different Scenarios of Model Building (Colony Ballroom) - Symposium Overview

Prof. Hairong Song (University of Oklahoma)

Variable selection helps researchers build a stronger, more parsimonious predictive model by identifying the best subset of predictors that are importantly associated with the given outcomes. This symposium introduces three innovative methods of variable selection and their applications in different scenarios of model building, including (1) a method that incorporates random regularized penalized quasi-likelihood estimation for jointly selecting both random and fixed effects in building generalized linear mixed models, (2) a machine learning technique (i.e., the genetic algorithm) for variable selection in the presence of missing data as well as when the missing data are handled by multiple imputation and random forest imputation, and (3) a Lasso estimator for variable selection when a VAR-based method is used to cluster high-dimensional intensive longitudinal data. Each method will be demonstrated through a simulation and/or an empirical study. Suggestions and recommendations will be provided for empirical uses of these methods.

A new algorithm for variable selection in building GLMMs

Thursday, 27th July - 09:03: Symposium: Innovative Methods for Variable Selection in Different Scenarios of Model Building (Colony Ballroom) - Symposium Presentation

Dr. Yutian Thompson (University of Oklahoma Health Sciences Center), Ms. Yaqi Li (the University of Oklahoma), Prof. Hairong Song (University of Oklahoma), Dr. David Bard (University of Oklahoma Health Sciences Center)

Variable selection presents a pathway to preventing generalized linear mixed models (GLMMs) from encountering model overfitting, nonconvergence, low external validity and estimation biases. Among various selection approaches, the method of regularized parameter estimation has shown superiority on jointly selecting both fixed and random effects. However, challenges, such as high computational cost, the outnumbered predictors problem, and multicollinearity have created barriers to its wider usage in practice. This study proposes a new algorithm, called random *rPQL* that incorporates a regularized penalized quasi-likelihood estimation with a stability selection process to overcome the aforementioned challenges. Simulation results indicate that the *rPQL* algorithm is capable of selecting fixed and random effects with high accuracy and efficiency, even when the number of candidate variables exceeds within-group observations or when severe multicollinearity exists. A new R package is presented that allows researchers to appropriately and pragmatically select variables in GLMMs.

Investigating variable selection techniques under missing data: A simulation study

Thursday, 27th July - 09:27: Symposium: Innovative Methods for Variable Selection in Different Scenarios of Model Building (Colony Ballroom) - Symposium Presentation

Ms. Catherine Bain (the University of Oklahoma), Dr. Dingjing Shi (the University of Oklahoma)

Variable selection is an essential step researchers must take when constructing a predictive model because it allows for a) the identification of variables that are importantly associated with the given outcome, and b) the formation of a stronger, more parsimonious predictive model. Variable selection techniques such as LASSO, Elastic Net, and more recently machine learning techniques like the genetic algorithm have been a strong point of interest for researchers in the behavioral sciences. While there is a growing interest in understanding these techniques and their applications, most of them have focused on complete datasets. Missing data, however, is ubiquitous in the behavioral sciences. Therefore, there is a gap in the existing literature that needs to be addressed regarding the performance of these methods in the presence of missing data. This study incorporates missing data handling techniques, including multiple imputation and random forest imputation, into the above-mentioned variable selection techniques and evaluates the performance of the techniques through a Monte Carlo simulation study. Preliminary results showed that the genetic algorithm provides the smallest bias in the presence of ignorable missing data. Theory indicates that no variable selection techniques will perform well under non-ignorable missing data; results from this study will provide quantitative evidence regarding this. Future directions on strategies to handle non-ignorable missing data in this context will be discussed.

Clustering intensive longitudinal data using VAR models with Lasso estimator

Thursday, 27th July - 09:51: Symposium: Innovative Methods for Variable Selection in Different Scenarios of Model Building (Colony Ballroom) - Symposium Presentation

Ms. Yaqi Li (the University of Oklahoma), Prof. Hairong Song (University of Oklahoma)

Model-based methods for clustering intensive longitudinal data is a powerful technique for quantifying between-individual differences in within-individual dynamics of psychological and behavioral processes. However, the performance of such clustering techniques depends heavily on the accuracy of parameter estimation of the chosen time-series models. A satisfactory performance often requires a large number of time-ordered observations ($t > 100$), which imposes a huge challenge in clustering such data, particularly when high-dimensional, multivariate observations are involved. To address this challenge and improve the performance of model-based clustering techniques, we proposed a clustering method that incorporates the Lasso estimator in vector autoregressive (VAR) model parameter estimation. The Lasso estimator is well-suited for estimating VAR models with high-dimensional data collected from a limited number of measurement occasions, thereby enhancing the performance of the VAR-based clustering technique. Results from the extensive simulations, where we implemented a two-step clustering procedure based on VAR models with the Lasso estimator using the Gaussian mixture model (GMM) and k-means clustering algorithm, exhibited that the application of the Lasso estimator in VAR-based clustering was associated with improved accuracy on recovering the number clusters and cluster membership, which was particularly true with the GMM clustering algorithm.

DIF analysis with unknown groups and anchor items

Thursday, 27th July - 09:00: Differential Item Functioning (Benjamin Banneker) - Oral

Dr. Gabriel Wallin (London School of Economics and Political Science), Dr. Yunxiao Chen (London School of Economics and Political Science), Prof. Irini Moustaki (London School of Economics and Political Science)

Measurement invariance across items is key to the validity of instruments like a survey questionnaire or an educational test. Differential item functioning (DIF) analysis is typically conducted to assess measurement invariance at the item level. Traditional DIF analysis methods require knowing the comparison groups (reference and focal groups) and anchor items (a subset of DIF-free items). Such prior knowledge may not always be available, and psychometric methods have been proposed for DIF analysis when one piece of information is unknown. More specifically, when the comparison groups are unknown while anchor items are known, latent DIF analysis methods have been proposed that estimates the unknown groups by latent classes. When anchor items are unknown while comparison groups are known, methods have also been proposed, typically under a sparsity assumption - the number of DIF items is not too large. However, there does not exist a method for DIF analysis when both pieces of information are unknown. In this talk, a framework for DIF analysis is presented that fills the gap. In the proposed method, we model the unknown groups by latent classes and introduce item-specific DIF parameters to capture the DIF effects. Assuming the number of DIF items is relatively small, a regularised estimator is proposed to simultaneously identify the latent classes and the DIF items. A computationally efficient Expectation-Maximisation algorithm is developed to solve the non-smooth optimisation problem.
The performance of the proposed method is evaluated by simulation studies and an application to response data from a real-world educational test.

Integrated approach for detecting Differential Item Functioning (DIF) in survey adapted Montreal Cognitive Assessment (MoCA-SA)

Thursday, 27th July - 09:15: Differential Item Functioning (Benjamin Banneker) - Oral

Dr. Ji Eun Park (NORC), Mr. Brian Geistwhite (NORC), Dr. Vi-Nhuan Le (NORC)

Early detection of mild cognitive impairment or early signs of dementia among the older population is crucial for timely diagnosis and treatment. However, developing a robust screening tool for a diverse population can be a challenge. This study used an integrated approach for detecting education-based DIF among any of the 18 dichotomously scored items in the survey-adapted version of the Montreal Cognitive Assessment (MoCA-SA), a multidomain clinical tool developed to detect mild cognitive impairment (MCI) and early dementia. Several approaches were used to evaluate uniform and non-uniform DIF in the survey-adapted version of MoCA included in the National Social Life, Health and Aging Project (NSHAP). NSHAP is a population-based study of health and social factors to understand the well-being of older, community-dwelling Americans. We used Mantel-Haenszel (MH), logistic regression, and Item Response Theory-Log-likelihood Ratio Test (IRT-LRT) to detect item-level DIF in the MoCA-SA. We also examined the effect sizes and Item Characteristics Curves to assess the magnitude and directionality of DIF. Both MH and logistic regression detected evidence for differential item functioning on 11 items. The effect size estimation using the Mantel-Haenszel method classified nine items for negligible DIF and two items for moderate DIF, while the logistic regression method classified all 11 items for negligible DIF. More items were detected of DIF using Item Response Theory-Log-likelihood Ratio Test (IRT-LRT). None of the items were detected for severe DIF. The study provides valuable insight into understanding an integrated approach for detecting DIF and its methodological challenges.

A comparison of methods for adjusting samples and test score distributions for group differences

Thursday, 27th July - 09:30: Differential Item Functioning (Benjamin Banneker) - Oral

Dr. Tim Moses (The College Board), Dr. YoungKoung Kim (College Board)

A longstanding question in psychometrics, equating and linking activities is how to account for examinee group differences when analyzing test score distribution differences. Methods include post stratification/frequency estimation procedures, minimum discriminant function adjustment, and propensity weighting. Because some of these methods are more established than others, it is helpful to periodically review the more recently described ones and especially to compare their results with those of the established procedures.

The focus of this presentation/paper is on comparing weighted test score distributions resulting from the use of several methods to account for group differences. Test score distributions reflecting ranges of examinee differences will be compared after adjusting for the differences using post stratification, minimum discriminant function adjustments, and propensity weighting based on logistic regression. Results show that similar weighted distributions can be achieved when using each of the considered methods to account for group differences in the same way, but that some methods are better in terms of their flexibility and their potential applications to a range of datasets, sample sizes, and group differences. The presentation will conclude by addressing how the flexibility of the propensity weighting procedures can be used to expand evaluations of test score distributions while accounting for group differences in a range of increasingly complex models that can provide detailed explanations for how group differences can affect test score distributions and also how these group differences can be addressed in equating and linking procedures.

Understanding differential item functioning using process data

Thursday, 27th July - 09:45: Differential Item Functioning (Benjamin Banneker) - Oral

Ms. Ling Chen (Columbia University), Prof. Jingchen Liu (Columbia University)

Differential item functioning (DIF) is an important concept in testing fairness. It occurs when items function differently among different subgroups. Previous research on DIF has mainly focused on statistical detection, yet understanding why DIF occurs remains a challenge. Process data, obtained from respondents interacting with a computer-based assessment item, provides a unique opportunity to understand DIF as it contains rich information about the respondents' progress and strategies towards solving the problems. Using features extracted from process data, we are able to construct a variable that mediates the relationship between the grouping variable and the response variable, which helps in detecting behavioral patterns that could lead to DIF and thus provides a deeper understanding of the underlying mechanism of DIF.

Scalable explanatory IRT modeling with sparse data structures

Thursday, 27th July - 09:00: Item Response Theory (Prince George) - Oral

Dr. J.R. Lockwood (Duolingo), Dr. Steven Nydick (Duolingo)

Sparse response data structures, in which there are large numbers of persons and items but relatively few person-item pairings, commonly arise in applied psychometrics. These structures often occur in assessment or pedagogical contexts in which tasks from large banks of potentially diverse item types are adaptively presented to persons. Data with large degrees of sparsity may overwhelm estimation procedures that assume dense data representations and methods, hampering the use of the data to make inferences about persons and/or items. To address this issue, we present a Bayesian explanatory IRT model that allows for multidimensional person and item traits, all of which can be modeled with latent regressions on features. We then describe a Markov Chain Monte Carlo estimation procedure, using the Metropolis-Hastings algorithm within a Gibbs sampling framework, that exploits the sparseness of the observed person-item pairings to efficiently sample the posterior distribution of model parameters. The model is designed to support activities such as estimating person-level and item-level traits, analyzing dimensional structures of these traits in populations, relating traits to observed features of persons or items for research purposes, and constructing shrinkage estimators of traits that synthesize feature information with response data. In our presentation, we will summarize the model and estimation algorithm, and will discuss corresponding numerical methods that we implemented to scale the procedure to large datasets that increasingly arise in computational psychometrics contexts. Finally, we will demonstrate performance for several canonical problems using data from the Duolingo English Test.

Mitigating bias in ability estimates during routing in multistage testing

Thursday, 27th July - 09:15: Item Response Theory (Prince George) - Oral

Ms. Merve Sarac (University of Wisconsin-Madison), Prof. James Wollack (University of Wisconsin-Madison)

Mitigation strategies used for anomalous response behavior are critical following detecting such behavior in real time. We utilize the robust ability estimation between stages in multistage testing (MST) by modifying the weights during MST in the presence of preknowledge. We propose a new preknowledge-adjusted weighting method that downweights observations on potentially compromised items. The weighted ability estimation is an intervention to mitigate the bias in interim ability estimates in MST. This approach is expected to route examinees to more accurate modules of appropriate difficulty adaptive to their (purified) weighted ability estimates.Most importantly, more accurate routing of examinees based on weighted ability estimates improves the unweighted ability estimates after completing the adaptive test. We compared four methods: MLE, Hubert weighting, preknowledge-adjusted weighting, and MLE based only on potential secure items. We evaluate the performance of different weighting methods using RMSE and bias for the ability parameter estimates, which are calculated for examinees with preknowledge and honest examinees separately for each discrete ability value. Preliminary results showed that preknowledge-adjusted weighting produced lower bias and RMSE than Hubert weighting and MLE.

Using item response theory to investigate whether rater assessments measure rater quality: Is there such a thing as a “correct” rating?

Thursday, 27th July - 09:30: Item Response Theory (Prince George) - Oral

Dr. William Belzak (Duolingo), Dr. Yigal Attali (Duolingo), Ms. Danielle Mann (Duolingo)

Rating tasks are common in the behavioral and social sciences, and across many industries. For example, the development of ChatGPT relied on humans to evaluate the quality of output text and assign scores according to a standard rubric; these scores were then used to fine-tune model output. As part of large-scale rating tasks, assessments are often used to evaluate rater quality and ensure that raters are making appropriate ratings. However, evaluating rater quality through assessment implies that there are “correct” answers to rating tasks, where “correctness” is typically derived from an “expert” (or sometimes from consensus). In this talk, we investigate whether certain types of rating tasks have “correct” answers and whether rater assessments measure what they intend to measure (i.e., rater quality). Our application involves remote proctors (raters) of a high-stakes English proficiency test who make decisions about whether test takers have violated rules during test sessions (rating tasks). We use item response theory (IRT) to demonstrate that defining “correctness” is not always defensible for certain types of rating tasks. Namely, divergent estimates of item discrimination and low estimates of test-retest reliability suggest that some rater assessments fail to measure rater quality. Alternative standards with which to judge rater quality, such as rating “severity” (determined by raters), rating “reasonableness” (determined by experts), or rating in “agreement with policy” (determined by policy-makers), may prove more tractable as measurement goals.

The Gumbel-Reverse Gumbel (GRG) model for binary data: A new asymmetric IRT model

Thursday, 27th July - 09:45: Item Response Theory (Prince George) - Oral

Prof. Jay Verkuilen (CUNY Graduate Center), Mr. Peter Johnson (CUNY Graduate Center)

Recent work by Daniel Bolt and colleagues (e.g., Bolt & Li, 2021) have suggested that asymmetric IRT models, such as the Logistic Positive Exponent (LPE; Samejima, 2000), the Heteroscedastic Residuals (HR; Molenaar, 2014), or the complementary loglog model (Goldstein, 1980; Shim, Bonifay, & Wiederman, 2023) can address known issues, such as correlation between discrimination and difficulty or poor identification of the 3PL and 4PL models while still allowing for guessing or slipping. Unfortunately, asymmetric models have potentially problematic issues of their own. The LPE model is known to be very difficult to estimate due to confounding between the exponent and the discrimination parameters. The HR model's heteroscedasticity parameter is difficult to interpret. Finally, the cloglog or loglog models are more readily estimated, but they require that researchers know the proper direction of the link function in advance. We propose the Gumbel-Reverse Gumbel (GRG) model that is based on a weighted average of the loglog and cloglog links, with the weight being an estimated parameter of the model. The loglog and cloglog links are based on the CDFs of the Gumbel and Reverse Gumbel distributions, respectively, which are the asymptotic distributions of the sample maximum and minimum, respectively, of properly standardized random variables with exponential tails. As such, they represent disjunctive and conjunctive processes, respectively. While the 2PL is not a special case of the GRG model, it is very closely approximated when the estimated weight parameter is equal to 0.5. We illustrate this model using simulated and real data.

Quantile multilevel item response theory model with a change point

Thursday, 27th July - 10:00: Item Response Theory (Prince George) - Oral

Dr. Hongyue Zhu (Academy for Research in Teacher Education, Northeast Normal University)

Multilevel item response theory models are widely used in educational and psychometric research problems. This model takes the latent trait (ability) as the output variable and aims to analyze the influence of the factors of interest (explanatory variable) on ability. However, most of the current studies are limited to linear regression analysis and also do not consider whether the relationship between ability and explanatory variables changes significantly. Based on the quantile multilevel item response theory model (Zhu et al., 2021), this study intends to extend it to polytomous response data, and incorporate change point detection into the model, so as to establish a quantile multilevel item response theory model with a change point and broaden its theoretical and application value. Meanwhile, this study intends to develop the corresponding Bayesian estimation methods for estimating model parameters, including the identification of the change point. Finally, this model is applied to the research on the relationship between teachers' professional happiness and work intensity of primary and secondary school teachers in China, aiming to conduct a more comprehensive and in-depth analysis of the relationship between the two variables, and determine whether there is a certain threshold of work intensity that makes the relationship between them change significantly after this.

Using Gaussian process ordinal regression with mixture errors to understand student well-being

Thursday, 27th July - 09:00: Measurement Applications (Margaret Brent) - Oral

Mrs. Elizabeth Gibbs (University of Connecticut), Dr. Xiaojing Wang (University of Connecticut)

With the ubiquity of sensing devices comes the potential to use sensing data to understand human behavior. Models used to fit sensing data must be flexible enough to account for the diversity in individuals' behavior and efficient enough to handle the large amount of information in the data. In this paper, we analyze sensing and repeated survey data collected in an educational setting. We use Bayesian ordinal Gaussian process regression (GPR) to model the relationship between students' moods and their behavior. Using a Gaussian process (GP) prior gives us the flexibility to capture the nonlinear structures of dynamic changes of students' behavior for data collected in an irregularly spaced time series.

Fitting ordinal GPR models is very challenging because sampling from the joint posterior distribution often requires sampling from high-dimensional truncated distributions. Further, there is strong correlation between unknown parameters. This correlation can make it difficult to adequately explore the sample space. We propose an efficient Markov chain Monte Carlo (MCMC) algorithm by introducing a mixture distribution on the error term, which gives us additional flexibility to sample from the parameters' joint posterior distribution and make Bayesian inference.

We find there is not a statistically significant relationship between students' mood and sensed covariates, including conversation time and percent of the day spent in the dark. While the proposed model is applied in an educational setting, it has great potential for use in applications where we wish to understand dynamic changes for other types of ordinal phenomena of interest.

Psychometric properties of the Dual-Range Slider response format

Thursday, 27th July - 09:15: Measurement Applications (Margaret Brent) - Oral

Mr. Matthias Kloft (Philipps-University Marburg), Dr. Jean-Paul Snijder (Heidelberg University), Prof. Daniel W. Heck (Philipps-University Marburg)

The measurement of variability in persons' behaviors using ecological momentary assessment is time-intensive and costly. We propose to use interval responses (i.e., the Dual-Range Slider response format, DRS) as a simple and efficient alternative. Respondents indicate variability in their behavior in a retrospective rating by providing a lower and an upper bound on a continuous, bounded scale. We investigate the psychometric properties of this response format as a prerequisite for further validation. First, we assess the test-retest reliability of factor-score estimates for the *width of DRS intervals*. Second, we test whether factor-score estimates of the Visual Analog Scale (VAS) and the *location of DRS intervals* show high convergence. Third, we investigate whether factor-score estimates for the DRS are uncorrelated between different personality scales. We present a longitudinal multi-trait multi-method study using two personality scales (Extraversion, Conscientiousness) and two response formats (VAS, DRS) at two measurement occasions for which we estimate factor-score correlations in a joint IRT model. The test-retest reliability of the width of DRS intervals is high ($r > .74$). Also, convergence between the location scores of VAS and DRS is high ($r > .89$). Conversely, the discrimination of the width of DRS intervals between Extraversion and Conscientiousness is poor ($r > .94$). In conclusion, the DRS seems to be a reliable response format which might not be perfectly suited for measuring variability in personality. Further, we present preliminary results of a second study demonstrating that the width of DRS intervals shows higher discrimination for more dissimilar tasks.

Mapping of the data science skillset of Dutch master study graduates

Thursday, 27th July - 09:30: Measurement Applications (Margaret Brent) - Oral

Dr. zsuzsa BAKK (Leiden University), Mr. Mathijs Mol (Leiden University)

Despite the growing popularity of the field, no clear general definition of the skillsets of data science and artificial intelligence professionals are readily available. We analyze 41 data science and AI Master of Science programs from 7 Dutch universities to derive the skillsets of young graduates entering the job market. Correlated topic modeling is used to extract the general topics from various data science or artificial intelligence related program and course descriptions. Afterwards, an analysis of posterior classification of the topics per university was performed to explore the differences and similarities between the universities on their orientation of data science and artificial intelligence programs. General and specific skill sets are uncovered and differences between the universities are described in this paper. The results of this paper are aimed both for researchers but also for universities that aim to develop relevant educational programs and companies that have no clear view whether their vacancies might be fit for data science or artificial intelligence graduates.

Research on adaptive learning system based on learners' academic emotions

Thursday, 27th July - 09:45: Measurement Applications (Margaret Brent) - Oral

Ms. Chang Nie (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University), Ms. Tao Xie (Statistics Bureau of DongGuan Municipality), Dr. Tao Xin (Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University)

With the development of information technology, adaptive learning systems based on reinforcement learning algorithms have emerged, which can recommend the most appropriate learning materials to learners, enabling them to personalize their learning and achieve maximum knowledge growth throughout the learning process. It is important to note that when recommending learning materials to learners, we should also focus on whether learners are able to learn happily, in addition to improving their knowledge. However, to date, related researches have focused on improving recommendation algorithms but have not considered the negative academic emotions that learners experience when faced with inappropriate recommendations. This study aims to model learners' negative academic emotions in adaptive learning systems and improve recommendation strategies to help learners learn happily.

This study consists of two sub-studies. Study 1 uses the A3C algorithm to construct an adaptive learning system with a continuous level of knowledge, and model the negative academic emotions generated during the recommendation process based on the "Control-Value" theory. Study 2 proposes two improvement methods for the recommendation strategy: Method 1, which reduces the pool of materials before recommending them to exclude materials that cause negative academic emotions; Method 2, which adds a penalty term for negative academic emotions to the reinforcement learning algorithm to reduce the generation of negative academic emotions. The simulation experiments revealed that both types of methods can effectively reduce the occurrence of negative academic emotions.

Differential item functioning treatment in computerized adaptive testing item pools

Thursday, 27th July - 09:00: Item Response Theory (Juan Ramon Jimenez) - Oral

Mr. Juyoung Jung (University of Iowa), Ms. Ae Kyong Jung (University of Iowa)

Computerized adaptive testing (CAT) is utilized in various contexts, but differential item functioning (DIF) in the item pool can create systematic differences in item difficulty. To address this issue, test developers should detect and treat DIF to improve item selection and accurately estimate respondents' abilities. In this study, we focus on the performance of DIF treatment when designing the item pool through simulation studies. This study examines various patterns of DIF that may impact item response theory (IRT) parameter estimates, including DIF magnitude, contamination, and type. After detecting DIF, we apply a confirmatory multi-group multidimensional item response model to treat DIF of item parameters. We compare the accuracy of ability parameter estimates using the selection of items with or without considering multiple group DIF in the item pool.

Linking method for writing tests using item response theory and automated essay scoring

Thursday, 27th July - 09:15: Item Response Theory (Juan Ramon Jimenez) - Oral

Mr. Kota Aramaki (The University of Electro-Communications), Dr. Masaki Uto (The University of Electro-Communications)

For essay writing tests, a difficulty is that the scores given for essays depend on rater characteristics, such as severity and consistency. To resolve this problem, the generalized many-facet Rasch model (GMFRM) has been proposed, which is an item response theory (IRT) model that can estimate examinees' ability while considering rater characteristics. When applying such an IRT model to multiple writing tests with different examinees and raters, test linking is necessary to unify the scale of model parameters estimated from individual test results. In test linking, test administrators generally need to design multiple tests such that examinees or raters are partially shared, but this is often difficult in actual testing environments. Therefore, in this study, we propose a novel method to link the GMFRM parameters estimated from multiple writing tests with different examinees and raters using neural automated essay scoring technology, which has recently been intensively studied in artificial intelligence and natural language processing research. Experimental results show that our method successfully realizes test linking without common examinees or raters.

Towards advancing precision environmental health: Developing a customized exposure burden score to PFAS chemicals using mixture item response theory

Thursday, 27th July - 09:45: Item Response Theory (Juan Ramon Jimenez) - Oral

Dr. Shelley Liu (Icahn School of Medicine at Mount Sinai), Dr. Leah Feuerstahler (Fordham University), Ms. Yitong Chen (Icahn School of Medicine at Mount Sinai), Dr. Joseph Braun (Brown University School of Public Health), Dr. Jessie Buckley (Johns Hopkins Bloomberg School of Public Health)

In environmental health, researchers are interested in quantifying a person's total exposure to a chemical class, using an exposure burden score. For example, in the case of PFAS (per- and polyfluoroalkyl substances), clinical recommendations suggest to measure a small set of PFAS chemicals in a person's blood and use a simple summation of the concentrations. We previously showed IRT as an alternative scoring approach, with each chemical considered an "item". IRT can account for the fact that different studies or laboratories measure different sets of PFAS chemicals, and IRT-derived scores are more sensitive to associations with health outcomes. Here, we extend that work by developed a customized PFAS burden scoring algorithm using mixture item response theory (mixIRT). This accommodates the fact that individuals are exposed to different sets of PFAS chemicals, due to potentially complex, unknown combinations of individual factors (e.g. differing drinking water sources, diet, behaviors). We apply this to data from the U.S. National Health and Nutrition Examination Survey (NHANES). We estimate mixIRT PFAS burden scores using weighted combination of subpopulation-specific scoring algorithms. Using the mixIRT burden score, we find that non-Hispanic Asians have significantly higher PFAS burden compared with non-Hispanic Whites. However, this disparity is masked when using a single IRT algorithm for all participants or summed concentrations. MixIRT burden scores also have more precise associations with health outcomes. Mixture IRT may allow us to develop personalized exposure burden metrics to ensure that the quantification of PFAS exposure burden is equitable and informative for all people.

Identifiability of polychoric models with latent elliptical distributions

Thursday, 27th July - 09:00: Statistical Methods and Theory (Thurgood Marshall) - Oral

Mr. Che Cheng (National Taiwan University), Mr. Hau-Hung Yang (National Taiwan University), Prof. Yung-Fong Hsu (National Taiwan University)

The family of polychoric models (PM) considers ordinal data as categorization of latent multivariate normal variables. Such framework is commonly used to study the association between ordinal variables, often leading to the polychoric correlation model (PCM). Moreover, PM subsumes several psychometric models, such as the graded response model (Samejima, 1968; 1997). However, the property of identifiability of PM has not been addressed in the literature. To make the issue more complicated, the normality assumption underlying PM has been challenged recently; researchers have suggested that the latent variables underlying PM could be generalized to elliptical distributions. Two unsolved questions can be posited: (a) Is PM and/or PCM with latent elliptical distributions identifiable? (b) If not, can we find the identifiability constraints of it?

In this research, we investigate the identifiability issue of PM and PCM with latent elliptical distributions by generalizing Rodriguez and Mouchart's (2003) argument. We first prove the identification of PCM based on the copula representation. We then proceed to find the set of identifiability constraints of PM through the equivalence-classes approach introduced by Tsai (2000). Our results show that the PM of Likert scales (LS) can be identified if we set the first cut-off of items to be zero and the final cut-off of items to be one. Future research should investigate how to equate the scales constructed by LS and by comparative judgment items.

A familiy of discrete kernels for presmoothing

Thursday, 27th July - 09:15: Statistical Methods and Theory (Thurgood Marshall) - Oral

Dr. Jorge González (Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI), Chile Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile Interdisciplinary Laboratory of Social Statistics (LIES), Chile)

Although the sample relative frequency distribution can be used as an estimate of a score distribution, presmoothing is often used to estimate and correct irregularities of score distributions. Both, log-linear and mixture models (i.e., Beta4) have been used for presmoothing in equating. In this paper, we introduce a class of discrete kernels that can be used to estimate the probability mass function of scores, serving thus for the purpose of presmoothing. An empirical illustration shows that the proposed discrete kernel estimates work either equally well or better than current methods for presmoothing score distributions.

Nonparametric estimation of the risk or rate ratio in rare events meta-analysis with arm-based and contrast-based approaches

Thursday, 27th July - 09:30: Statistical Methods and Theory (Thurgood Marshall) - Oral

Prof. Heinz Holling (University of Muenster), Ms. Katrin Jansen (University of Muenster)

Rare events are events which occur with low frequencies. These often arise in studies where the data are arranged in binary contingency tables. In this contribution, the estimation of effect heterogeneity for the risk-ratio parameter in meta-analysis of rare events studies is investigated through two likelihood-based nonparametric mixture approaches: an arm-based and a contrast-based model. Maximum likelihood estimation for both approaches is derived by means of the EM algorithm. The likelihoods under the contrast-based approach and the arm-based approach are compared and differences are highlighted. First, the methodologies are illustrated using two typical meta-analytic data sets. Then, a simulation study is reported which assesses the performance of these two methods. Under the design of sampling studies with nested treatment groups the results show that the nonparametric mixture model based on the contrast-based approach is more appropriate in terms of model selection criteria such as AIC and BIC. Comparisons of the estimators are provided in terms of bias and mean squared error. Estimating effect heterogeneity in the case of the contrast-based method appears to behave better than the compared method although differences become negligible for large within-study sample sizes.

Equivalence test and sample size procedures for ANCOVA designs

Thursday, 27th July - 09:45: Statistical Methods and Theory (Thurgood Marshall) - Oral

Dr. Gwowen Shieh (National Yang Ming Chiao Tung University), Dr. Show-Li Jan (Chung Yuan Christian University)

Traditional procedures mainly aim to establish the existence of substantial differences in comparative studies. There has been an increasing concern on the equivalence tests to assess an observed effect size that is practically negligible for psychological research. This study presents equivalence procedure for appraising standardized mean difference in ANCOVA designs. The general formulation of flexible equivalence bounds permits a wide range of research questions to be tested. The associated power and sample size formulas are also derived under the random model framework. Accordingly, the stochastic features of both the response and covariate variables are simultaneously taken into account. Numerical example and simulation study are utilized to explicate the usefulness and accuracy of the proposed procedures for equivalence evaluation in ANCOVA.

Maximum likelihood estimation using a possibly misspecified parameter redundant model

Thursday, 27th July - 10:00: Statistical Methods and Theory (Thurgood Marshall) - Oral

Dr. Richard M. Golden (University of Texas at Dallas)

Maximum Likelihood (ML) estimation methods for smooth finite-dimensional probability models are widely used to support parameter estimation in many modeling applications. When the probability model is possibly misspecified, it is well known that the ML estimates are asymptotically Gaussian centered at a locally identifiable optimal solution with an ML covariance matrix derived from the first and second derivatives of the log-likelihood function. Such asymptotic results, in turn, support the derivation of confidence intervals, hypothesis testing, and model selection methods such as the Akaike Information Criterion (AIC). All these results, however, assume the probability model is not parameter redundant near the maximum likelihood estimates and the maximum likelihood estimates are locally identifiable. Furthermore, in the presence of model misspecification, it is possible to have a model that is parameter redundant near a locally identifiable optimal solution.

This paper explores a method for eliminating the effects of parameter redundancy near the maximum likelihood estimates by assuming the maximum likelihood estimates are generated by a special linear projection from a lower-dimensional subspace. This setup eliminates the parameter redundancy while preserving the original parameter space. It is helpful for applications such as cognitive diagnostic models (CDMs) whose parameter values have important semantic interpretations. Next, proofs of theoretical results are provided based upon this projection method which establish the consistency and asymptotic distribution of the maximum likelihood estimates when parameter redundancy is present. Some simulation studies are then briefly presented which are designed to investigate the relevance of this theoretical framework in practice.

Capturing sample heterogeneity in dynamic psychological processes

Thursday, 27th July - 13:45: Symposium: Capturing sample heterogeneity in dynamic psychological processes (Colony Ballroom) - Symposium Overview

Ms. Di Jody Zhou (University of California, Davis)

Modeling intensive longitudinal data (ILD) enhances our comprehension of the complexity of psychological dynamics, which are commonly known to vary greatly across individuals. Data-driven subgrouping methods provide an impetus for investigating quantitative and qualitative individual differences that are often obscured when aggregating information across individuals. This symposium will present six data-driven subgrouping methods to capture sample heterogeneity in dynamic processes, including five model-based (i.e., VAR model) and one graph-based methods. Specifically, five presenters will discuss: (1) Bayesian mixture multilevel VAR to account for latent between-individual heterogeneity (Xiao); (2) multilevel VAR Tree (ml-VARTree) to discover covariate association with subgroup differences using decision-tree (Zhou); (3) the impact of temporal order selection on a two-stage VAR-based clustering method integrated with Gaussian mixture model and K-means technique (Song); (4) the graph-based community detection technique Louvain clustering to uncover homogeneous patterns across time (Aragones); and (5) a comparison between the Subgrouped Chain Graphical VAR (scgVAR) and the Subgroup Group Iterative Multiple Model Estimation (S-GIMME) in exploring subgroups of continuous-time processes (Park). The utility and performance of each method will be demonstrated through either a Monte Carlo simulation or an empirical study.

A Bayesian mixture multilevel vector autoregressive model

Thursday, 27th July - 13:50: Symposium: Capturing sample heterogeneity in dynamic psychological processes (Colony Ballroom) - Symposium Presentation

Ms. Xingyao Xiao (University of California, Berkeley), Dr. Anja Ernst (University of Groningen), Dr. Feng Ji (University of Toronto)

The growing availability of intensive longitudinal data has opened up new opportunities of modeling dynamics to comprehend complex relations and population heterogeneity in psychological and behavioral sciences. In this study, we introduce a novel Bayesian mixture multilevel vector autoregressive (BMMVAR) model that accommodates latent between-individual heterogeneity. Unique challenges in Bayesian implementation such as label-switching will also be discussed. Using real-world data, we showcase the potential of our Bayesian approach to discover similarities and differences between individuals and enhance our comprehension of psychological dynamics.

Identifying and explaining sample heterogeneity in dynamic psychological processes using ml-VARTree

Thursday, 27th July - 14:04: Symposium: Capturing sample heterogeneity in dynamic psychological processes (Colony Ballroom) - Symposium Presentation

Ms. Di Jody Zhou (University of California, Davis), Dr. Emilio Ferrer (University of California, Davis), Prof. Siwei Liu (University of California, Davis)

With the increasing popularity of intensive longitudinal data in psychological studies, it is now well-recognized that individuals vary meaningfully in their associations among behaviors, emotions, and other psychological states. While data-driven subgrouping methods exist to explore potential qualitative, as well as quantitative, differences in these processes, they often fall short in explaining these subgroup differences unless further post-hoc analyses are conducted. Our study aims to introduce ml-VARTree, a novel method that uses a decision-tree-based subgrouping algorithm to identify subgroups of individuals characterized by different vector autoregressive (VAR) models. Importantly, the algorithm automatically searches for covariates and their interactions that predict subgroup memberships, providing insights into why individuals are different or similar in their dynamic processes. At the same time, this method estimates group-level parametric means and within-group variances and covariances based on the multilevel VAR model. We examine the subgrouping accuracy of ml-VARTree in a Monte Carlo simulation study including data conditions varying in effect size, time series length, sample size, and number of covariates. We also demonstrate the utility of this novel method through an empirical data analysis.

Impact of temporal order selection on VAR-based clustering of intensive longitudinal data

Thursday, 27th July - 14:18: Symposium: Capturing sample heterogeneity in dynamic psychological processes (Colony Ballroom) - Symposium Presentation

Ms. Yaqi Li (the University of Oklahoma), Prof. Hairong Song (University of Oklahoma)

Model-based methods for clustering intensive longitudinal data is a powerful tool to study similarity and dissimilarity in dynamic characteristics among multiple individuals. When implementing such clustering techniques, researchers need to set the temporal orders of the within-individual processes to be identical for all individuals. Two methods have been proposed here. One applies the most complex structure or highest order (HO) for all individual processes, while the other chooses the most parsimonious structure or the lowest order (LO) for all individuals. How each method, HO vs. LO, would perform in clustering multivariate dynamic processes under various data conditions? In this study, we introduced a two-step VAR-based clustering procedure and investigated the performance of HO and LO using this procedure with Gaussian mixture model (GMM) and K-means clustering algorithms. Our simulation study showed that LO generally performed better than HO on recovering the number of clusters and cluster membership. In addition, the GMM algorithm exhibited an improved accuracy over the K-Means on recovering numbers of clusters and cluster membership; however, the GMM algorithm demonstrated a higher sensitivity to the selection of temporal orders and showed worse performance on clustering dynamic processes with high-order VAR models.

Clustering analysis of time series of affect in dyadic interactions

Thursday, 27th July - 14:32: Symposium: Capturing sample heterogeneity in dynamic psychological processes (Colony Ballroom) - Symposium Presentation

Mr. Samuel Aragones (University of California, Davis), Dr. Emilio Ferrer (University of California, Davis)

Longitudinal data, by nature, is a series of repeated measurements that can capture dynamics of individuals over time. These dynamics, such as those found in romantic relationships, can have effects on outcomes such as relationship satisfaction. In analyzing multivariate time series data, a problem that analysis techniques focus on is characterizing the heterogeneity within the data. While longstanding analysis techniques exist, such as ANOVA and linear regression, these are limited to patterns that describe the population as a whole. Even through developments in longitudinal data analysis, there are still difficulties with addressing variability within individuals. Clustering analysis is an approach taken to capture homogeneity in various subspaces and may be promising as a method of analyzing longitudinal time series data in a manner that can detect patterns within the repeated measures of individuals.

A popular clustering method, Louvain (Blondel et. al, 2008), is used to assess the viability of such techniques in longitudinal data analysis. Concerns about the technique are addressed, including: [1] how is heterogeneity described through our clusters, [2] what are the interpretations of said clusters, and [3], are we able to determine how informative these clusters of partial time series are? We assess Louvain through a variety of metrics aimed at answering these questions which include measures from information theory.

Implications of clustering continuous-time processes using discrete-time methods

Thursday, 27th July - 14:46: Symposium: Capturing sample heterogeneity in dynamic psychological processes (Colony Ballroom) - Symposium Presentation

Mr. Jonathan Park (The Pennsylvania State University), Dr. Zachary Fisher (The Pennsylvania State University), Prof. Sy-Miin Chow (The Pennsylvania State University), Dr. Peter Molenaar (The Pennsylvania State University)

Several methodologies have been developed in recent years for identifying heterogeneous subgroups in intraindividual dynamics, particularly dynamic network models such as vector autoregression (VAR) models and their extensions. Many of these developments are from the discrete-time framework, including the Subgroup Group Iterative Multiple Model Estimation (S-GIMME; Gates et al., 2017) procedure and the Subgrouped Chain Graphical VAR (scgVAR; Park et al., *Under Review*).Discrete-time approaches bear some inherent limitations. For instance, potential inferential confounds may arise when time intervals vary within or across studies. These challenges may be ameliorated by fitting models in the continuous-time framework; however, it is unclear how subgrouping algorithms formulated in discrete-time perform at identifying meaningful subgroup differences when applied to continuously unfolding processes. The current work examined how discrete-time clustering approaches such as S-GIMME and the scgVAR perform at identifying meaningful subgroup differences when applied to continuously unfolding processes. We assessed the subgrouping accuracy and the quality of point-estimates from these two methods when applied to continuous-time processes under various sampling designs and time intervals in a Monte Carlo simulation study. Insights on design considerations to facilitate identification of meaningful individual and subgroup differences in dynamics are provided.

Assessment Engineering meets generative AI: Unlocking new opportunities for digital assessment

Thursday, 27th July - 13:45: Generative AI (Atrium) - Oral

Prof. Jaehwa Choi (George Washington University)

Digital assessment has become increasingly popular in recent years due to its many benefits. However, there are challenges associated with designing and implementing digital assessments that meet the needs of diverse participants.

To address these challenges, Assessment Engineering (AE) has emerged as a promising solution. AE has undergone various evolutions in the last decades with the advent of computer technology, which has made the process of assessment more efficient, and more effective. However, the rise of generative AI, such as ChatGPT, has brought a new dimension to AE research, making it imperative to explore the potential benefits of integrating generative AI into AE. The recent emergence of generative AI has created a dire need for research on how to integrate it into AE due to the lack of existing research on this topic.

This article proposes the use of generative AI techniques in AE to unlock new opportunities for digital assessment. I present a framework for incorporating generative AI into assessment engineering, including the use of generative models to automatically create items, construct assessments, and evaluate student responses. I discuss the potential benefits of this approach, including reduced development time and increased flexibility in assessment design and development. I also identify challenges , such as the need for high-quality data and the potential for biases in generative models. Overall, I argue that the integration of AE and generative AI has the potential to transform digital assessment for testing/learning. Computer Adaptive Formative Assessment (CAFA) platform was used in the illustrations.

Generative distractor modeling with generative AI

Thursday, 27th July - 14:00: Generative AI (Atrium) - Oral

Mrs. Sunhyoung Lee (University of Nebraska-Lincoln), Prof. Kyongil Yoon (Notre Dame of Maryland University), Prof. Jaehwa Choi (George Washington University)

Generative artificial intelligence (AI) has had a significant impact on assessment engineering (AE), particularly in developing generative distractor models for multiple-choice questions (MCQs). By mimicking human thought processes, generative AI analyzes vast amounts of data to differentiate between students who possess the knowledge and those who do not. This approach provides valuable insights into common misconceptions and misunderstandings. To create plausible distractors, AI can be further trained on a large dataset of problems and corresponding answers. Then, the model generates higher-quality distractors for given problems, which are evaluated based on their plausibility and their ability to mislead students who lack a deep understanding of the underlying concept.

Once the AI model generates plausible distractors and rationales for them, they can be used to develop a generative distractor model that generates new assessment items and corresponding distractors within the AE framework. In this paper, we provide concrete illustrations of how generative AI techniques can aid in developing high-quality distractor models in AE. By utilizing AI-generated distractors and their underlying logic, it is possible to reduce subject-matter experts' efforts and the possibility of bias. The resulting generative distractor models can provide concrete examples of how AI can further strengthen the power of the explicit rule-based model approach in efficient and effective assessment development. In summary, this article demonstrates how leveraging generative AI techniques in AE can enhance the quality of assessment development practices for researchers and practitioners who are interested in developing generative distractor models.

Features for detecting essays produced by generative AI

Thursday, 27th July - 14:15: Generative AI (Atrium) - Oral

Prof. Hong Jiao (University of Maryland, College Park), Dr. Chandramani Yadav (University of Maryland, College Park), Mr. Haowei Hua (Culver Academies), Dr. Lei Wan (College Board)

Generative AI, especially ChatGPT drew much attention recently in education. Some educators are worried about the misuse of the latest AI technology in assessing learning outcomes in both low-stakes classroom assessments and un-proctored high-stakes assessments such as prompt-based essay assessment. Though some generic detectors are being developed in different stages of maturity for ChatGPT, few detectors are on essays with high-stakes nature. This study intends to develop automated detectors of AI-generated essays based on essay prompts and identify effective features in detecting AI-generated essays. The identified effective features intend to be used to train human raters to catch such essays in their routine rating activities. To develop automated detectors of AI-generated essays, three prompts from the 2012 Hewlett Foundation Automated Essay Scoring challenge, namely prompts 1, 2, and 7 representing essay assessments for grades 8, 10, and 7 respectively. The genre of these three prompts are all persuasive/narrative/expository. ChatGPT is used to generate essays for each prompt. Student essays are combined with the AI-generated essays in developing the automated detectors. Both handcrafted features and BERT model extracted features are explored and compared. Handcrafted features include length-based features, syntactic features, word-based features, readability features, semantic features, argumentation features, and prompt-relevant features. To facilitate the interpretability of the developed detectors, non-ensemble and ensemble supervised learning algorithms are compared including support vector machine, decision tree, naïve Bayes, random forest, gradient boosting, and stacking learning algorithms. Critical features contributing to the detection accuracy are highlighted.

Reassess the item response theory simulation with generative adversarial networks

Thursday, 27th July - 14:30: Generative AI (Atrium) - Oral

Dr. Jiawei Xiong (Pearson), Dr. Bowen Wang (University of Florida)

Simulation studies are a common approach in Item Response Theory (IRT) research. These simulations often generate data based on specific IRT models, such as simulating item and person parameters from a normal distribution of N(0,1) and simulating student responses using a selected IRT model. However, the source or validity of those specific assertions is often conspicuously lacking and real data could be more complex as real data can exhibit non-normal ability distributions (Micceri, 1989; Zhang et al., 2021). Generative Adversarial Networks (GANs; Goodfellow et al., 2014), served as unsupervised networks, are used for generatively modeling data with deep learning methods. GANs can automatically discover and learn the regularities or patterns in input data, and then generate or output new samples that plausibly could have been drawn from the original dataset and resemble the original samples with high efficiency and accuracy. To address these limitations, we explore the use of GANs as an alternative simulation method. The main objective of this research is to study and compare the effectiveness of GANs and the traditional simulation method in terms of their ability to generate realistic item response data. In particular, the parameter-level (e.g., person and item parameters) comparison and model-level (e.g., model fit and response distribution) comparison are both included. Empirical examples are also employed to demonstrate the effectiveness of GANs in terms of learning data distribution.

Automatic generation of cognitive test items using large language models

Thursday, 27th July - 14:45: Generative AI (Atrium) - Oral

Mr. Antonio Laverghetta Jr. (University of South Florida), Dr. John Licato (University of South Florida)

Writing high-quality items is critical to building psychometric measures but has traditionally also been a time-consuming process. One promising avenue for alleviating this is automated item generation (AIG), whereby methods from artificial intelligence are used to generate new items with little to no human intervention. Researchers have explored using large language models (LLMs) from natural language processing (NLP) to generate new items and have demonstrated promising results. However, much of this work has examined item generation for non-cognitive assessments, especially those related to personality. Important questions remain as to the generalizability of LLM AIG to other measurement domains, and especially whether it will perform just as well for *cognitive* assessments, which have thus far received limited attention. We investigate using the GPT-3 LLM to generate items for a popular language assessment used in NLP research: the GLUE benchmark. Our method is based on *prompting* GPT-3 to generate new items using existing, human-written items, where the human items are chosen based on their psychometric properties (e.g., those possessing the highest discrimination). We use this approach to generate new GLUE items and evaluate their quality through a series of human studies. We find that the AIG items have good content validity (as assessed using content experts) and have comparable or even superior item discrimination and internal consistency reliability. We discuss the limitations of our approach and scope for improvements, as well as our recommendations for using GPT-3 to generate cognitive items, to aid those wishing to use LLMs to streamline test development.

State space models for ordinal measurements: Towards a generalized dynamic IRT

Thursday, 27th July - 13:45: Longitudinal and Dynamic Models (Benjamin Banneker) - Oral

Dr. Teague Henry (University of Virginia), Ms. Lindley Slipetz (University of Virginia), Mr. Ami Falk (University of Virginia)

State space modeling is a method for representing the dynamics of unobserved processes by way of their observed measurements, and can be seen as a form of generalized dynamic structural equation modeling. However, the methodological development of the framework has focused on continuous measurements (i.e. location data during flight), and applications of state space modeling to psychological data typically applies continuous measurement models to ordinal measurements, such as Likert scales. In this talk, I will discuss a general estimating framework for state-space models, multiple iterative filtering, that allows for arbitrary measurement models to be used, and show how it can be applied to estimate state space models with graded response model measurements, thus allowing for a dynamic IRT model. In a simulation study, I show that we can obtain consistent model estimation using this method, and that using continuous approximations to ordinal measurements results in severe bias in parameter estimates of interest. Finally, I will discuss the use of multiple iterative filtering as a general approach for state space model estimation, and next steps to improve the framework.

Power considerations in dynamic structural equation models

Thursday, 27th July - 14:00: Longitudinal and Dynamic Models (Benjamin Banneker) - Oral

Mr. Hyungeun Oh (The Pennsylvania State University), Prof. Michael D. Hunter (The Pennsylvania State University), Prof. Sy-Miin Chow (The Pennsylvania State University)

Dynamic structural equation models (DSEMs), a special case of which is the multilevel dynamic factor models, are a powerful tool for analyzing intensive longitudinal data (ILD). DSEMs integrate techniques from multilevel modeling, time series analyses, and structural equation modeling to examine the interrelations among past and current values of latent factors identified with manifest indicators. Despite the increasing popularity of Bayesian DSEMs due to their versatility (in software such as Mplus), many applications of DSEMs utilize composite scores without accounting for measurement errors and differences in indicator quality. The goal of this study is to provide a framework for performing power and sample size planning involving DSEMs as related to the number of indicators, reliability of the indicators as defined using several well-known indices, effect sizes of the dynamics, and choice of priors. We provide results from a Monte Carlo simulation study and discuss unresolved challenges in conducting power analysis for DSEMs.

Dynamic structural equation modeling with missing data

Thursday, 27th July - 14:15: Longitudinal and Dynamic Models (Benjamin Banneker) - Oral

Ms. Yuan Fang (University of Notre Dame), Dr. Lijuan Wang (University of Notre Dame)

Dynamic structural equation modeling (DSEM) is a useful technique for analyzing intensive longitudinal data. A challenge of applying DSEM is the missing data problem. The impact of missing data on DSEM, especially on widely applied DSEM such as the two-level vector autoregressive cross-lagged (VAR) models, however, is understudied. To fill the research gap, we evaluated how well the fixed effects and variance parameters in two-level bivariate VAR models are recovered under different missingness percentages *m*, sample sizes *N*, the number of time points *T*, within-person (WP) missingness patterns, and between-person (BP) missingness distributions through three simulation studies. Simulation 1 results provided guidelines for the data requirements (*N* and *T*) of using the two-level bivariate DSEM with missing data. In simulation 2, the random WP missingness pattern (missingness scattered randomly) was found to perform better than the drop-out pattern (missingness gathered towards the end of a study) when the missingness percentage is large. This suggests that when data collection is challenging (e.g., 80% missingness expected), researchers may plan for fewer measurement occasions and minimize the missingness proportions. In simulation 3, the BP homogeneous scenario (individuals have the same *m*) outperformed the BP heterogeneous scenario (individuals have different *m*) in estimating between-person variances. Finally, to facilitate evaluations of DSEM under customized data and model scenarios (different from those in our simulations), we provided illustrative examples of how to conduct Monte Carlo simulations in M*plus* to determine whether a data configuration is sufficient for fitting a specific DSEM.

Challenges for psychometric evaluations in intensive longitudinal data

Thursday, 27th July - 14:30: Longitudinal and Dynamic Models (Benjamin Banneker) - Oral

Prof. Holger Brandt (University of Tübingen), Dr. Patrick Schmidt (University of Zurich)

Collecting intensive longitudinal data has become very popular in many applied fields including clinical and health psychological research due to the feasible implementation via smartphones. Researchers use such data to gain insight in the underlying processes of psychotherapies, for example, using ecological momentary assessments. Newly developed frameworks such as dynamic latent class structural equation modeling (DLC-SEM) help to evaluate psychometric test properties in innovative ways because they account for dynamic changes and inter-individual heterogeneity in these changes. In this talk, I will highlight and discuss important challenges that may not have been of relevance in traditional psychometric evaluations but that become more important when using the same test (very) often for the same persons. I will provide details on how DLC-SEM can be used to flexibly model a temporal person-specific evolution of underlying test dimensions in a clinical study of cognitive behavioral therapy for generalized anxiety disorder. Results from a simulation study indicate that some Bayesian model fit indices may favor overparameterized models but that the model estimates themselves correctly reflect the underlying population model. Generalizations of DLC-SEM to include time- and person-specific differential item functioning via shrinkage priors will be discussed.

Zero inflation in longitudinal data: Why is it important and how should we deal with it?

Thursday, 27th July - 14:45: Longitudinal and Dynamic Models (Benjamin Banneker) - Oral

Ms. Sijing (SJ) Shao (University of Notre Dame), Ms. Ziqian Xu (University of Notre Dame), Mr. Kenneth McClure (University of Notre Dame), Dr. Zhiyong Zhang (University of Notre Dame)

Behavioral research increasingly investigates changes over time to understand the dynamics of process-oriented mechanisms. Intensive time sampling methods, such as ecological momentary assessment (EMA) in repeated measures designs, are often utilized for this purpose. These designs result in longitudinal data with correlated observations over time, requiring the use of methods such as growth modeling in multilevel models (MLMs) to account for within-person changes. However, the assumption of normality in MLMs is often violated in behavioral research, particularly with frequency or number of behaviors (e.g., number of alcoholic drinks) that result in excess zeros generated from different processes. To account for zero-inflated count outcomes, zero-inflated models, such as zero-inflated Poisson (ZIP) and hurdle models, were developed and are accessible to researchers, with the choice depending on the theoretical basis for generating zeros. However, MLMs with autoregressive processes are often used for data collected from EMA studies to capture rapid changes. In this framework, the zero-inflated variable from the previous time point is controlled as a predictor in the model, but the consequences of entering the zero-inflated variable as a predictor in multilevel autoregressive models have not been examined. In this talk, we discuss the results from simulation studies that evaluate these consequences. Additionally, we introduce estimation method in both ZIP and hurdle frameworks with JAGS (Plummer, 2003). We also compare ZIP and hurdle models, emphasizing interpretation and model selection when analyzing zero-inflated count data in EMA studies. Finally, we provide an illustration of an empirical study to further clarify these concepts.

The assessment collaboration skill using multilevel multidimensional partial credit model

Thursday, 27th July - 13:45: Item Response Theory (Prince George) - Oral

Ms. Guiyu Li (East China Normal University)

Researchers have proposed numerous measurement methods to assess collaboration skill from response time and response data perspective, but the group variables are still ignored. However, Collaboration skill goes through the entire process of collaborative tasks and involves group cooperation. Therefore, the process data of collaborative skill and group variables must be considered in the study.

This study aims to develop a model to measure the collaboration skill with the entire data of collaboration process including response process, individual response data, and group response data. It is a multilevel nested data that the process data nested within person data and person data nested within group data. As the data is a nested data structure, a multilevel psychometric model is required in this study. Level 1 is response process level that describes the relationship between response time and response behaviors. Level 2 is person level that curves the relationships between the student's ability and the item difficulty. Level 3 is group level which refers to the ability of students impacted by group variables.

The collaboration skill usually contains multidimensional elements and polytomous response data. In the study, a multilevel multidimensional partial credit model will be used to analyze data from the entire process of collaborative tasks.

In data analyzed, the rstan package and mirt package in R and the Hamiltonian Monte Carlo of No-U-Turn Sampler was used to estimate model parameters.

Finally, the Rhat, RMSE, outfit and infit index were used to evaluate the parameters recovery and model fit.

Computational aspects of modelling item responses

Thursday, 27th July - 14:00: Item Response Theory (Prince George) - Oral

Dr. Patricia Martinkova (Institute of Computer Science of the Czech Academy of Sciences), Dr. Adela Hladka (Institute of Computer Science of the Czech Academy of Sciences)

Item response theory models can be derived in factor analytic framework as well as in the framework of generalized linear and nonlinear mixed-effect models. In this work, we focus on the latter one. We first describe step-by-step development of IRT models via empirical characteristic curves and generalized linear and nonlinear models (GLNM) with emphasis on didactic value of such approach. In addition to that, we discuss wide usage possibilities of GLNM in terms of criterion-related item validity and we demonstrate these aspects with real data examples. Finally, we present some novel approaches to parameter estimation in this framework together with their challenges in practical implementation.
Martinková, P., and Hladká, A. (2023) Computational Aspects of Psychometric Methods. With R. Chapman and Hall/CRC (In press). ISBN 9780367515386

Fisher information-based difficulty and discrimination measures in binary IRT

Thursday, 27th July - 14:15: Item Response Theory (Prince George) - Oral

Prof. Jay Verkuilen (CUNY Graduate Center), Mr. Peter Johnson (CUNY Graduate Center)

While difficulty and discrimination parameters have appealing and intuitive meanings in the 2PL model, the parameters in IRT models outside of the 2PL are much harder to interpret. For example, even adding the pseudo-guessing parameter with the 3PL model makes the difficulty and discrimination parameters no longer have the meaning they have in the 2PL, and they are not directly comparable when the pseudo-guessing parameter differs. Increasingly, however, models even more complicated than the 3PL, such as the 4PL (e.g., Loken & Rulison, 2010) or various asymmetric IRF models, such as the logistic positive exponent (LPE) or heteroscedastic residuals (HR) (e.g., Bolt & Liao, 2021), etc. have been considered. These models help resolve some issues encountered in IRT, such as discrimination-difficulty confounding that is often noted in practice or the substantial potential bias for examinee scores when guessing and/or slipping are not taken into account. Unfortunately, they sacrifice the interpretable nature of difficulty and discrimination that the 2PL provides. We propose to use two properties of the Fisher information function – the maximizer and a transformation of the information at the maximum – in analogy to the 2PL model, for which the model parameters and the Fisher information function are in close correspondence, as measures of effective difficulty and discrimination, respectively. For most models, it is not feasible to find these values analytically, but they can be approximated numerically from IRT program output when the Fisher information function is generated. We illustrate using data from TIMSS.

A mixture IRTree approach to deal with heterogeneity in response strategies

Thursday, 27th July - 14:30: Item Response Theory (Prince George) - Oral

Mr. Ömer Emre Can Alagöz (University of Mannheim), Prof. Thorsten Meiser (University of Mannheim)

IRTree models can improve the validity of our inferences from questionnaire data by controlling response style (RS) effects. In traditional IRTree models, responses are modelled as a product of several distinct decision processes, namely informative response (whether to choose mid-category), direction (whether to choose dis/agreement categories), and intensity (whether to choose extreme categories). The decision about the direction is affected by the substantive trait, whereas the decisions about the informative response and intensity are largely determined by midpoint RS (MRS) and extreme RS (ERS), respectively. Two limitations of traditional IRTrees are: 1) effects of the substantive trait on informative response and response intensity decisions are not accommodated, 2) all respondents are assumed to utilize ERS and MRS in their response strategy. However, the types of RS being used and the strength of their effects on responses can differ between respondents. We address these limitations by proposing a mixture multidimensional IRTree (MM-IRTree) model that detects heterogeneity in response strategies, and it consists of four latent classes. In all latent classes, all decisions are affected by a common substantive trait, but the classes differ in the set of RS that are used in response strategies. More specifically, class-specific response strategies include following RS: 1) ERS, 2) MRS, 3) both ERS and MRS, 4) neither ERS nor MRS. A simulation study showed excellent recovery of latent classes and reliable estimation of item/person parameters. An empirical analysis unravelled distinct classes of noticeable sizes suggesting that respondents indeed utilize different combinations of response styles.

Model averaging of (nonlinearly related) item response models

Thursday, 27th July - 14:45: Item Response Theory (Prince George) - Oral

Dr. Leah Feuerstahler (Fordham University)

Model uncertainty is endemic to applied statistics and psychometrics, and it is often advisable to evaluate several candidate models in applied settings. Model averaging avoids the problem of having to select a single most appropriate model by taking weighted averages of model predictions where weights are based on theory, model fit, or other considerations. To date, relatively little research has applied model averaging to item response theory (IRT), despite much research on IRT model comparison methods (e.g., Kang & Cohen, 2007). One of the significant challenges in averaging IRT models is that item and person parameters potentially have different meanings across different models. Rights et al. (2018) addressed this problem by linearly scaling all candidate models to have the same latent mean and variance. However, this strategy may not lead to latent trait metrics that are comparable across models, as different models can lead to nonlinearly related latent traits (Bolt et al., 2014; Lord, 1975).

This presentation will present the problem of IRT model averaging in the context of averaging both item response functions and person scores. A Bayesian model averaging (BMA) approach will be proposed that uses quantiles of the trait distribution to account for potential nonlinear relationships between latent traits. Simulation study results will be presented that compare the accuracy of item and person estimates averaged using the proposed method, the method of Rights et al. (2018) and its Bayesian implementation, and model selection. In addition, different weighting strategies, including pseudo-BMA weights and model stacking, will be compared.

Nominal category models and the independence of irrelevant alternatives assumption

Thursday, 27th July - 13:45: Categorical Data Analysis (Margaret Brent) - Oral

Mr. Weicong Lyu (University of Wisconsin-Madison), Prof. Daniel Bolt (University of Wisconsin-Madison)

Nominal categories models (NCMs) are widely used in psychometric research for polytomous responses due to their flexibility and generality. As a latent variable generalization of the multinomial logit model, the practical use of the NCM relies on an important assumption, namely the independence of irrelevant alternatives (IIA), which requires that given the true latent traits, the error terms of all the categories are independent. This assumption can be easily violated in practice, such as when the model is applied to outcomes of a sequential process. This problem seems rarely discussed in the psychometrics literature. In this study, we discuss possible scenarios where IIA is violated, and in turn show a potential for bias due to the violation of IIA through simulation. We consider a hypothetical scenario involving on-demand requests for hints on test items. Sensitivity analysis may be helpful in understanding the degree to which IIA is a problem in practice.

Examining different approaches to treating zero-frequency cells in polychoric correlation estimation

Thursday, 27th July - 14:00: Categorical Data Analysis (Margaret Brent) - Oral

Ms. Jeongwon Choi (Vanderbilt University), Dr. Hao Wu (Vanderbilt University)

When examining the correlation among ordered categorical variables, polychoric correlations are in fact more appropriate than Pearson correlations. When the contingency tables of data include zero-frequency cells, the polychoric correlation approaches ±1. This is problematic as it yields a singular or indefinite correlation matrix. Zero-frequency cells emerge when the sample size is insufficient for sampling the cell with a low probability or when a specific combination of variables cannot exist. Although the most commonly employed solution is to add a small value to zero-frequency cells before the estimation, this method is implemented inconsistently among researchers. There exist disagreements on determining the value to be added, to which cell these values will be added, and whether to maintain the marginal distribution. Therefore, to further the previous simulation study (Savalei, 2011), this study comprehensively investigated the consequences of these manipulation methods in different dimensions under a variety of conditions by using the Monte Carlo simulation method. Simulation results indicate that zero-frequency cell corrections are helpful for reducing the error of estimations when correlations are not extremely strong but the signs of thresholds are dissimilar and far apart, or when thresholds are extreme. This study makes a significant contribution by providing guidance to better estimate polychoric correlations in the presence of zero-frequency cells and presenting ways to improve computational efficiency for simulation studies by detecting conditions that yield equivalent estimates to avoid redundant computation.

Using extreme threshold constraints for partially-known latent class models

Thursday, 27th July - 14:15: Categorical Data Analysis (Margaret Brent) - Oral

Dr. Paul Scott (University of Pittsburgh)

We consider fitting a latent class model (LCM) where some indicators provide information about known class membership, warranting a partially-known latent class modeling approach. Important known class indicators can result from “gold standard” measures, case-control assignments, or other personal characteristics considered key to account for when findings classes. The approach suggested here involves setting extreme threshold constraints to known class indicators to approximate perfect probabilities forcing individuals’ membership into certain classes. This approach is exemplified through a study aiming to find classes of sleep symptom constellations for individuals with and without obstructive sleep apnea (OSA). The presence of OSA, based on “gold standard” diagnostic criteria, is considered an important known class indicator. When estimating the LCM, we set extreme threshold constraints to approximate the probability of being in the no-OSA class given the presence of the OSA diagnosis to 1, and zero otherwise. Results show that we are able to eliminate measurement error and perfectly recover the observed response probabilities on the other 14 symptoms for those with no-OSA, while allowing the estimation of 4 other classes for individuals known to have OSA. Two extensions are suggested and illustrated: (1) enumerating classes within each known group; and (2) incorporating criterion classification errors by lowering threshold constraints to reflect true and false positive rates. We provide application and simulation to highlight some of the issues that arise with these extensions.

Parallel analysis with a new decision rule

Thursday, 27th July - 14:30: Categorical Data Analysis (Margaret Brent) - Oral

Mr. Ahmet Guven (Augusta University), Dr. Ashley Saucier (Augusta University), Dr. Nicole Winston (Augusta University), Dr. Andria Thomas (Augusta University)

Green et al. (2012) proposed a Revised Parallel Analysis (RPA) due to concerns about comparing the remaining eigenvalues in traditional PA. The RPA considers modeling prior factors in evaluating *k*th + 1 factors. Green et al. (2016) evaluated the relative accuracy of the RPA for binary data by imposing thresholds and computed tetrachoric correlation matrices analyzed using principal axis factoring. Tetrachoric correlation based on a two-stage approach (Olsson, 1979) is not feasible with small samples (e.g., Lorenzo-Seva & Ferrando, 2021), resulting in unstable parameter estimates in factor analysis. In addition, the traditional decision rule ignores the effect of the ratio of *k*th eigenvalue to *k*th +1 eigenvalue in the original data. Therefore, we included a new decision rule and Bonett and Price's (2005) computation to approximate tetrachoric correlation to improve the performance of the RPA. The proposed rule involves comparing the ratio of *k*th eigenvalue to *k*th + 1 eigenvalue in original data with the ratio of 95th percentiles of *k*th eigenvalues to 95th percentiles of *k*th + 1 eigenvalues. We called this PA variant Revised Tetrachoric PA with Principal Component Analysis (RTPA-PCA). To evaluate the relative accuracy of the RTPA-PCA, we have conducted simulations by mimicking the one-factor model, the bi-factor model with one general and one specific factor, and the perfect-cluster model with two correlated factors defined in Green et al., (2016), and also included real data. We will present the full results at the conference.

Model size effects on measurement invariance testing with ordinal indicators

Thursday, 27th July - 14:45: Categorical Data Analysis (Margaret Brent) - Oral

Ms. Nana Amma Asamoah (Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas), Dr. Xinya Liang (Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas)

Model size is a critical factor that influences the sensitivity of fit indices in confirmatory factor analysis (CFA). Research on model size suggested that various types of fit indices responded differently to the model complexity; however, little is known about the model size effects in measurement invariance testing, especially with categorical indicators. In this study, we use simulations to investigate the sensitivity of fit indices to model size in multigroup CFA analysis of metric and scalar invariance with ordinal indicators. This study extends current research (e.g., Meade et al., 2008; Shi et al., 2018; Cao and Liang, 2022) on the topic by considering model complexity and ordinal indicators. The fit indices studied include the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) because they are some of the most ubiquitous fit indices used as standards for evaluating model fit, including in measurement invariance research.

Model size in this study is identified by the number of factors and the number of indicators per factor. Other simulation conditions include number of response categories, sample size per group, and magnitude and location of invariance. It is our goal that the results of this study will first show the performance of fit indices under the conditions tested and, second, highlight the importance of considering model size and other design factors in the interpretation of fit indices from multigroup CFA analysis.

A tailored sensitivity analysis procedure for the social sciences

Thursday, 27th July - 13:45: Missing Data (Juan Ramon Jimenez) - Oral

Dr. Brenna Gomer (Utah State University)

The problem of missing-data mechanism uncertainty can be approached in a variety of ways, one of which is sensitivity analysis. Many existing approaches require the user to specify a range of values for a sensitivity parameter which is designed to reflect differences in the assumed missing-data mechanism or the severity of the departure from the MAR mechanism. However, in the behavioral sciences, it may be of greater interest to compare results under different assumptions of missing-data mechanisms and missingness relationships. This type of approach falls outside of the typical sensitivity analysis procedures. In this talk, I present a novel hypothesis testing framework for conducting sensitivity analysis that quantifies the stability of statistical results when the assumed missing-data mechanism and missingness relationships are varied. Results are shown from Monte Carlo simulation studies that highlight the performance of the proposed method under various proportions of missingness and true underlying missing-data mechanisms. Results suggest that the hypothesis testing framework works as intended and may be useful for conducting sensitivity analysis in the social sciences.

Pay attention to ignorable missingness! How variations in the missing at random mechanism affect efficiency loss in parameter estimates.

Thursday, 27th July - 14:00: Missing Data (Juan Ramon Jimenez) - Oral

Dr. Lihan Chen (McGill University), Dr. Victoria Savalei (University of British Columbia), Dr. Mijke Rhemtulla (University of California, Davis)

The *missing at random* (MAR) assumption requires missingness be independent of the data once conditioned on observed data, but it does not specify the details of the conditioning relationships involved. For instance, under MAR, *X* may be missing in cases where: 1) *Y* values are high, 2) *Y* values are *either* high *or* low, or 3) *Y* values are average. MAR can also vary in many other aspects, such as how strongly the conditioning variables are related to missingness, and in which variables the missingness occurs. The details of MAR can often be safely ignored under modern missing data techniques for the purpose of consistent parameter estimates and valid statistical inferences. However, the *efficiency* of parameter estimates can differ among MAR variations, *even when* the rate of missing data is the same; this can result in unreported discrepancies in the power of statistical tests. The *fraction of missing information* (FMI) is a direct measure of efficiency loss due to missingness under MAR, which captures the impact of MAR variations. Using estimates of FMI at the population level via a *full information maximum likelihood* approach (Savalei & Rhemtulla, 2012), we explored how efficiency loss differed in regression coefficients under a wide range of MAR variations. We demonstrate how efficiency loss due to MAR is highly complex and not always intuitive, emphasizing the need not only to report FMI in empirical research, but also to further investigate the impact of MAR in methodological research. Implications on planned missingness designs are also explored.

The performance of strategies for handling non-effortful responses in equating

Thursday, 27th July - 14:15: Missing Data (Juan Ramon Jimenez) - Oral

Mr. Juyoung Jung (University of Iowa), Prof. Won-Chan Lee (University of Iowa)

The process of equating test scores in item response theory (IRT) helps to adjust the difficulty level of different test forms so that scores obtained on these forms can be used interchangeably (Kolen & Brennan, 2014). The accuracy of equating can be affected by many factors, and one such factor is the quality of estimated item parameters when non-effortful responses (NERs) exist in the data. When detected, NERs can be treated as ignorable, incorrect, or missing. Different approaches for managing NERs are expected to have different effects on the estimation of item parameters. In this study, we evaluate the effect of various approaches for handling NERs on IRT equating. NERs are detected based on item response time using computer recording. Flagged NERs are handled in three different ways: they are ignored, replaced by incorrect responses, or treated as missing. It is essential to consider the assumptions and limitations of these methods and choose the most suitable one for specific contexts. We investigate the performance of these strategies on the accuracy of equating through a simulation study with varied conditions of sample size, non-effortful prevalence, and non-effortful severity.

Estimators of the AIC and BIC in multiply imputed data

Thursday, 27th July - 14:30: Missing Data (Juan Ramon Jimenez) - Oral

Dr. Joost Van Ginkel (Leiden University), Dr. Dylan Molenaar (University of Amsterdam)

Statistical techniques that use Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) as fit measures for model comparisons, oftentimes handle missing data using full information maximum likelihood (FIML). Normally, FIML does not account for bias resulting from missingness depending on auxiliary variables that are not part of the statistical model of interest. Additionally, FIML only resolves the missing-data problem for the specific statistical model but not for other statistical techniques that may be applied to the same dataset. Consequently, for other techniques missing data are treated in a different way (for example, by deleting cases) and results of statistical techniques applied to the same dataset are no longer comparable regarding sample size and possible sources of bias. Multiple imputation for handling missing data could resolve both of the above problems. However, currently no estimators of the AIC and BIC in multiply imputed data exist, other than averaging the measures across imputed datasets. In the current research two different estimators of the AIC and BIC in multiply imputed are proposed, next to averaging. The estimators are studied in a simulation study in the context of factor analysis, multilevel analysis, and latent class analysis. The results show that the new estimators select the correct model more frequently than averaging does. Thus, both options may be good candidates to include in future versions of software packages.

Missing data in discrete time state-space modeling of EMA data

Thursday, 27th July - 14:45: Missing Data (Juan Ramon Jimenez) - Oral

Ms. Lindley Slipetz (University of Virginia), Dr. Teague Henry (University of Virginia), Mr. Ami Falk (University of Virginia)

This paper's purpose is to understand the impacts of types of missingness on the analysis of EMA-like data and offer guidelines for the use of missing data imputation methods in those cases. With EMA, missing data is pervasive as participant attrition is common. Such missingness can follow the traditional missingness patterns (i.e., MCAR, MAR, or MNAR), but there are missingness mechanisms particular to time series. TMAR occurs when the missingness depends only on the time variable in the model. ATMAR occurs when the missingness on a variable at time t is dependent upon the previous values of the variable. Any EMA study must have a solution to the issues created by these missingness mechanisms.

We performed a Monte-Carlo simulation of the impacts of missingness mechanisms (MCAR, MAR, TMAR, ATMAR, and MNAR at 15% and 30% missingness) on the modeling of EMA-like synthetic data, comparing the ability of several missing data imputation techniques (i.e., Kalman filter, MICE, and the EM algorithm) in the timeseries case. We found, first, that there is greater bias with greater percent missingness. Second, the parameters associated with missingness show more bias than the complete parameters. Third, increasing the strength of the true parameters results in more bias. Fourth, the Kalman filter performed well, while MICE performed poorly. Finally, the missingness mechanisms can be grouped with less bias and variability being associated with MCAR and TMAR and more bias and variability associated with MAR, ATMAR, and MNAR. These trends are seen in each of the parameters' recovery.

Decomposition of DIC for cognitive diagnosis model

Thursday, 27th July - 13:45: Diagnostic Classification Models (Thurgood Marshall) - Oral

Ms. Zhiduo Chen (University of Connecticut), Dr. Xiaojing Wang (University of Connecticut)

With the advances of computer-based tests, not only the item response but also the response time of examinees can be recorded, which could be an important source to refine our inference on the latent ability. A key point in modeling this is understand how the response time can help to evaluate the response accuracy in cognitive diagnosis model. Then I propose a new model comparison criterion based on the decomposition of deviance information criterion (DIC). The proposed model selection criteria can quantify the improvement on the fit of item responses due to incorporating the response time in a multidimensional hierarchical model based on Bayesian Markov chain Monte Carlo method. Also, a new DIC is constructed which is more stable and avoid the calculation of complicated integral. Simulation studies are conducted to examine the empirical performance of the joint model and model selection criterion.

Investigating the effect of Q-matrix misspecification on estimating cognitive diagnosis models for small sample sizes

Thursday, 27th July - 14:00: Diagnostic Classification Models (Thurgood Marshall) - Oral

Ms. Bea Margarita Ladaga (University of the Philippines School of Statistics), Dr. Kevin Carl Santos (University of the Philippines College of Education)

One of the critical components in cognitive diagnosis models (CDMs) is the Q-matrix which enumerates the necessary attributes per item. Multiple studies have shown that the specification of Q-matrices influences the fit of CDMs. This study explored the impact of Q-matrix misspecification for small sample sizes under the generalized deterministic input, noisy "and" gate (G-DINA) framework. A simulation study is performed to examine the effect of the Q-matrix misspecification on the attribute- and vector-wise classification rates. The study manipulated three factors: sample size, test length, and attribute correlation. For the Q-matrix misspecifications, three items were under-specified, over-specified, or both for a total of eleven modified Q-matrices with misspecifications. Three different estimation methods are compared: EM + monotonicity constraint (EMM), Bayes modal (BM), and BM + monotonicity constraint (BMM). Preliminary results show that the type of Q-matrix misspecification had varying levels of impact on the proportion of correctly classified examinees for each latent class. Compared to under-specification, over-specification of the attributes tends to result in a more significant decrease in the classification rate. However, under the same type of misspecification, higher sample sizes still resulted in a higher classification rate compared to smaller sample sizes.

Improving cognitive diagnosis in small samples with catalytic priors

Thursday, 27th July - 14:15: Diagnostic Classification Models (Thurgood Marshall) - Oral

Mr. David Arthur (Purdue University, West Lafayette)

Cognitive Diagnostic Models (CDMs) are powerful tools for providing personalized, formative feedback. However, despite their utility, these models are often only used with large sample sizes due to known model fit and estimation issues with small samples sizes. Bayesian methods have the potential to outperform other maximum likelihood based methods in these situations, but require the selection of prior distributions for model parameters. We show how catalytic priors can be used to guide the specification of prior distributions for parameters of a popular CDM, the deterministic inputs noisy "and" gate (DINA) model, and illustrate its effectiveness via simulation studies and real data analysis. We show that the proposed method leads to more accurate parameter estimates and increases classification accuracy of individual attribute profiles for sample sizes as small as 30 and as large as 500 and discuss how the method can be adapted to match performance of other methods for larger sample sizes.

Identifiability of Cognitive Diagnosis Models with polytomous responses

Thursday, 27th July - 14:30: Diagnostic Classification Models (Thurgood Marshall) - Oral

Ms. Mengqi Lin (University of Michigan, Ann Arbor), Dr. Gongjun Xu (University of Michigan, Ann Arbor)

Cognitive Diagnosis Models (CDMs) are a powerful tool that allows researchers and practitioners to learn fine-grained diagnostic information about respondents' latent attributes. There has been a growing interest in the use of CDMs for polytomous response data as more and more items with multiple response options become widely used. Similar to many latent variable models, the identifiability of CDMs is critical for accurate parameter estimation and valid statistical inference. However, the existing identifiability results are primarily focused on binary response models and have not adequately addressed the identifiability of CDMs with polytomous responses. This paper addresses this gap by presenting sufficient and necessary conditions for the identifiability of the widely used DINA model with polytomous responses, with the aim to provide a comprehensive understanding of the identifiability of CDMs with polytomous responses and to inform future research in this field.

Partially confirmatory Q learning for distinguishable attribute importance in the compensatory CDMs

Thursday, 27th July - 14:45: Diagnostic Classification Models (Thurgood Marshall) - Oral

Ms. Yunting Liu (berkeley), Dr. Yi Chen (Teachers College Columbia University), Mr. Mingfeng Xue (University of California, Berkeley)

Exploratory and confirmatory techniques of Q matrix specification are at opposite ends of a continuum when it comes to specifying cognitive diagnostic models (CDMs). The majority of previous research focuses on the dichotomous Q matrix, ignoring the distinguishing role of attributes in problem-solving. In this paper, we propose a reparameterized additive CDM (RA-CDM) model for dichotomous responses based on Q^* matrix (i.e., continuous Q matrix). Different from conventional dichotomous or polytomous Q matrix, Q^* matrix indicates the relative importance of each attribute for solving a specific item. Specifying an expert-defined binary Q matrix as a prior and Dirichlet distribution as the likelihood, we can estimate continuous matrix in RA-CDM, which uses the Gibbs sampling technique in Bayesian frameworks. Using a simulation study, we examined performance of parameter recovery across different sample sizes, different test lengths, and various numbers of attributes, as well as how the prior is specified — in a parsimonious or comprehensive manner. Results showed that the proposed method works better than the Generalized DINA model(G-DINA) in terms of classification results and item parameters, however, the improvement shrinks as sample-size and item number become larger. As an empirical example, a math assessment, fraction-subtraction data, was analyzed using the proposed model. We report evidence showing that the first attribute in fraction-subtraction data is of the highest relative importance and provide suggestions on amendment for future test design. In short, the developed methodology contributes to the development of CDMs and broadens their applicability to re-address test/item design issues.

Digital transformation-virtual standard setting

Thursday, 27th July - 16:15: Symposium: Digital Transformation-Virtual Standard Setting (Colony Ballroom) - Symposium Overview

Dr. Xinhui Xiong (Educational Testing Service)

Standard setting is critical for educational assessment because it defines cut score(s) that differentiate students from one category to another. The categories could be simply pass or fail, or various performance levels such below basic, basic, qualified, highly qualified, etc. Various standard setting methods have been proposed for multiple choice questions (e.g. Angoff, Bookmark, etc.) and constructed-response questions (e.g. Question-by-questions, analytical judgement method, etc.) over the past decades. No matter what method used, collecting judgements made by subject matter experts is a critical part of the standard setting in addition to the statistical analysis part. Traditionally, subject matter experts are convened to one location and are trained with necessary materials regarding the test and the test questions. After the training, they provide their initial opinions from different angels based on the specific standard setting method used. This part is highly interactive and of course it is not the only part in a standard setting procedure that involves frequent interactivities between humans. Nevertheless, COVID-19 changes the whole world of educational assessment, seemingly overnight, as living rooms turned into test centers and technology turned into one of the primary solutions for keeping learning and testing alive. How to shift the highly interactive “judgement part” in standard setting procedure to be virtual is key to a successful virtual standard setting.

In this symposium, three different implementations on virtual standard setting will be presented. Pros and cons of each method, and possible enhancements in the near or far future will be discussed.

Web-based standard setting for a credit-by-examination program

Thursday, 27th July - 16:30: Symposium: Digital Transformation-Virtual Standard Setting (Colony Ballroom) - Symposium Presentation

Dr. Weiling Deng (Educational Testing Service)

A credit-by-examination program sets its passing scores virtually through a two-stage process. First, a standard setting study is conducted over the internet in a month's time, with content experts and psychometricians leading a panel of 15-20 judges through various steps. The provisional cut scores from Stage 1 as well as the process through which they are derived are documented in a report and reviewed by the Test Development Committee in consultation with content and measurement experts to determine the final cut scores.

The panelists complete a sequence of steps at their own pace to learn the modified Angoff method for rating exam questions. However, they are required to attend two live discussions in real time to ensure quality training and timely resolution of questions or concerns. In the first live discussion, they collectively define the knowledge, skills and abilities of a minimally qualified candidate. In the second, everyone is encouraged to share the rationale behind their initial ratings for selected items on which their opinions differed most. Facilitators use threaded topics to guide the discussion, and everyone types out what they want to say. This format is chosen so that a record of the conversations is kept, and latecomers can still benefit from carefully reading all the posts. Compared to phone or video conferencing, this is more egalitarian and reduces the chance for aggressive personalities to dominate the discussion. As such, the web-based standard setting allows all panelists to bring their knowledge and experience to bear in a reasonable way.

A hybrid virtual standard setting implementation for statewide assessments

Thursday, 27th July - 16:45: Symposium: Digital Transformation-Virtual Standard Setting (Colony Ballroom) - Symposium Presentation

Dr. Jiawei Xiong (Pearson), Dr. Jennifer Galindo (Pearson)

The COVID pandemic has unprecedentedly affected educators and psychometricians to rapidly engage in ubiquitous virtual educational activities. The consequential impact may require dramatic changes in how the standard setting is delivered. Pearson's virtual standard setting meetings use a hybrid approach to successfully run online standard setting meetings for statewide assessments with several methodologies, including the Bookmark and Modified Angoff. The website can organize both the meeting protocol and the reference documents as well as judgment surveys and feedback data. Zoom helps panelists to engage with each other over several rounds of discussions, and to use the breakout room feature to replicate the small group discussions. Facilitators can show or hide each section on the website to control the meeting pace. The website also serves as a central location and a repository for documents and feedback data to be shared with the panelists and facilitators. It also serves as the central location for collecting panelist data and allows for streamlined data processing because all data is entered in a similar format. Evidence of validity about the standard-setting process is also collected through evaluation surveys.
The hybrid virtual standard setting allows:

- The panelist to discuss effectively while accessing necessary materials simultaneously.
- Standardized videos to be shared across committees for training purposes.
- The facilitator controls the pace by releasing the current meeting section and hiding other sections.

The final presentation will discuss the pros and cons of hosting standard settings in the virtual environment such as item security, recruitment, and staffing requirements.

A video-based implementation of the Bookmark method for a college placement testing program

Thursday, 27th July - 17:00: Symposium: Digital Transformation-Virtual Standard Setting (Colony Ballroom) - Symposium Presentation

Dr. Luz Bay (College Board), Dr. Liam Duffy (Pearson)

In 2018, the College Board signed a contract to develop a college placement test for a large southern state that required a standard setting meeting using the *Bookmark Method* to be implemented in the summer of 2020. Alas, the pandemic! The College Board halted all large meetings and the state disallowed all indoor meetings with 50 or more individuals attending. Pros and cons of different options were weighed including postponing until after the pandemic. The wisdom of sticking to the original schedule with a virtual implementation prevailed.
The previously written in-person standard setting plan was modified for a video-based implementation using Zoom. Other applications such as WatchDox for sharing secure materials such as the Ordered Item Booklet (OIB) were used. Rating and evaluation data were collected using MSForms. The modification of the in-person implementation of the Bookmark method to a video-based virtual method was done with our eyes on procedural validity.
Because the standard setting was a previously planned in-person implementation that had to be adjusted to an online implementation, we were mindful that the adjustments

- Minimized the impact of change in mode from in-person to online
- Elicited the desired behavior of all those involves
- Maintained or enhanced test security

This paper will describe the video-based virtual implementation of setting cut scores. The focus will be on the adjustments made when the pandemic forced a virtual. Key to success will be emphasized.

Implementing dynamic IRT models to account for response strategy variability

Thursday, 27th July - 16:15: Item Response Theory (Atrium) - Oral

Dr. Clifford Hauenstein (Johns Hopkins School of Medicine)

Aptitude test scores are typically interpreted similarly for examinees with the same overall score. However, research has found evidence of strategy differences between examinees, as well as in examinees' application of appropriate procedures over the course of testing. Thus, interpretations of performance are often obfuscated by both inter-examinee and intra-examinee variance in strategies and cognitive processes. This is especially true early in the testing window, when examinees are still becoming familiar with the demands of the task. Obtaining valid and reliable measures of ability is therefore contingent upon the development and application of psychometric models that can simultaneously account for these sources of variance in response data.
In light of these concerns, we consider early mixture IRT models, as well as more contemporary approaches that jointly consider both accuracy and response time data, in identifying latent clusters of examinees who may invoke different response strategies. We additionally propose a set of hidden Markov and state space IRT models that simultaneously consider accuracy and response time data in order to extract individual differences in cognitive processes, and additionally capture how these processes may shift over time as test familiarity increases. The utility of these different approaches is evaluated via improvement in validity coefficients, as indexed by correlations between corrected performance estimates on tests of ability (non-verbal reasoning, spatial ability, mathematical problem solving) and a set of measures theoretically related to intelligence. Practical implications for improving procedures in testing are discussed, with an emphasis on implementing classes of dynamic IRT models.

Comparing ability parameters in performance factor analysis and item response theory using Kullback-Leibler divergence

Thursday, 27th July - 16:30: Item Response Theory (Atrium) - Oral

Mr. Amirreza Mehrabi (Purdue University, West Lafayette), Dr. Ozge Altintas (Purdue University, West Lafayette), Dr. Jason Wade Morphew (Purdue University, West Lafayette)

Testlet based tests are designed to enhance the efficiency of assessments by grouping related items. Testlets are widely used in practice including textual passages, graphical data representations, musical excerpts, or numerical tables. The presence of a common stimulus, their connection can impact test takers' performance due to contextual effects. This can lead to violations of local item dependence, or testlet effects. There are different Testlet Response Theory Models (TRTMs) using different estimation methods to deal with local item dependence issues due to testlets in psychometric literature. The comparative investigation of these estimation methods is significant in terms of identifying their relative strengths and weaknesses. This study aims to demonstrate the testlet-specific parameters estimation method called the Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bound (L-BFGS-B) optimization method for optimizing the MLE as the estimator via the Performance Factor Approach (PFA). Its performance will be compared with the estimation methods commonly used in TRTMs. PFA can provide detailed information regarding the ability and item parameters estimation for testlet based tests by using the response patterns and taking into account the local item dependence within a testlet. The results of this study provide evidence of the effectiveness of the method in assessing item difficulty, discrimination, and individual student abilities in testlet-based tests. Additionally, the method is expected to provide valuable insights into the informative nature of each attribute through the evaluation of Fisher's Information Function, which can improve our understanding of the utility of the method.

Revisiting the 1PL-AG item response model: Bayesian estimation and application

Thursday, 27th July - 16:45: Item Response Theory (Atrium) - Oral

Dr. Jorge Bazán (University of São Paulo), Dr. Paula Fariña (Universidad Diego Portales)

The 1PL-AG model by San Martin et al (2006) was proposed to account for the effect of ability on guessing behavior in multiple choice items. Identifiability of this model and of the 1PL-G, one special model from it, have been studied;and applications considering calibration of items and Computerized Classification on Testing were developed too.

This paper presents a Full Bayesian approach of the 1PL-AG model and proposes alternative models as Normal-Ogive (1PN0-AG) e 1PNO-G. Using MCMC estimation we showed the accuracy of parameter recovery and applications in educational data comparing this model against alternative models illustrate the benefits of this approach.

Extreme and midpoint response styles: Two sides of the same coin?

Thursday, 27th July - 17:00: Item Response Theory (Atrium) - Oral

Mr. Martijn Schoenmakers (Tilburg University), Dr. Jesper Tijmstra (Tilburg University)

Response styles, the tendency of participants to respond to items regardless of the item content, have frequently been found to decrease the validity of Likert-type questionnaire results. While many models have been proposed to model and compensate for these response styles, it is still not entirely clear how these response styles relate to each other. Specifically, it is not always clear whether extreme responding (the tendency to endorse the extreme questionnaire responses) is the opposite of midpoint responding (the tendency to endorse the middle questionnaire responses), or whether these response styles are two separate dimensions. How these response styles are modelled influences the estimation complexity, parameter estimates, and substantive trait estimates of IRT models. In this paper, we thus examine whether it is possible to empirically distinguish extreme and midpoint responding as being either two separate dimensions or being two opposite sides of a single dimension under the multidimensional nominal response model using the AIC and BIC. Furthermore, we assess under what circumstances (factors sample size, test length, number of substantive dimensions, response style strength, and response style correlation) this assessment is possible and what the degree of error of this assessment is. Results indicate good performance for the AIC and BIC in a null condition and with extreme and midpoint responding as a single dimension, but worse performance for extreme and midpoint responding as two dimensions, depending on the considered factors. Recommendations for future practice are made and a hybrid approach combining the BIC and AIC is proposed.

Incorporating level-specific covariates in structural equation models of social-network data

Thursday, 27th July - 16:15: Network Analysis (Benjamin Banneker) - Oral

Ms. Aditi Manoj Bhangale (University of Amsterdam), Dr. Terrence D. Jorgensen (University of Amsterdam)

The social relations model (SRM) is applied to examine dyadic data within social networks. Multivariate SRMs have been modelled using linear mixed models but are insufficient to estimate structural equation models (SEMs) of complex theories. The social relations SEM (SR-SEM) combines the SRM and SEM, allowing researchers to fit more complex models and test a number of measurement-related and structural hypotheses about associations among SRM components. However, incorporating level-specific covariates, such as age, gender, or relationship quality, as predictors and outcomes of round-robin variables remains a challenge in the SR-SEM. In this study, we propose a two-stage estimation approach to easily incorporate level-specific covariates into SEMs of round-robin variables. Stage-1 of the two-stage estimator is Markov chain Monte Carlo estimation of unrestricted summary statistics of SRM effects. Stage-2 is maximum likelihood estimation of constrained SEMs using the Stage-1 summary statistics of SRM effects as input data. Uncertainty about the Stage-1 estimates is incorporated to adjust Stage-2 SEs and test statistics, analogous to fitting SEMs to polychoric correlations estimated for ordinal data. We introduce a new R package, lavaan.srm, to apply this estimation technique using rstan for Stage-1 and lavaan for Stage-2. We designed a simulation study to evaluate the accuracy and efficiency of estimated level-specific covariate effects in a round-robin SR-SEM. We assess the relative bias, root mean-square error, and coverage rates of the resulting parameter estimates and evaluate Type I error rates of the model-fit test statistic.

Keywords. Social relations model, structural equation model, two-stage estimation, level-specific covariates

Social network construct measurement error: An IRT-based latent space model

Thursday, 27th July - 16:30: Network Analysis (Benjamin Banneker) - Oral

Ms. Yishan Ding (University of Maryland, College Park), Dr. Tracy Sweet (University of Maryland, College Park)

The current body of literature on measurement error in social network analysis has primarily focused on proxy measurement error, which refers to inadequate or inaccurate observations of proxy measurements of social relationships. However, construct measurement error, which is a key concern in modern psychometric study, has been given less attention in social network studies. Construct measurement error is particularly relevant for social network relationships that are difficult or impossible to observe explicitly, such as friendships, and should be conceptualized as latent constructs.

To address the construct measurement error, researchers have long suggested using multi-item scales to measure social relationships (Marsden, 1990). However, there is a lack of methods for multivariate social network analysis using multi-item measurements. Even when network tie data is collected from several items, common analysis strategies include selecting a representative item or treating each item as a separate network and performing independent analyses.

To accommodate the construct measurement error in social network analysis, this study proposes a new model, termed IRT-LSM, that integrates an item response theory (IRT) model into a latent space model (LSM). The proposed method incorporates an IRT model to capture the latent social network relationship strength. We present three simulation studies to examine: model feasibility and impact of construct measurement error; a variety of data-generating models; and the effects of item parameter distribution and selection.

Network approaches to clinical assessment data: LSIRM vs. network psychometrics

Thursday, 27th July - 16:45: Network Analysis (Benjamin Banneker) - Oral

Ms. Ludovica De Carolis (University of Milano-Bicocca), Prof. Minjeong Jeon (University of California Los Angeles)

Questionnaires for measuring symptoms of mental disorders are critical tools for psychologists and psychiatrists. The extraction of information from clinical assessment data by suitable statistical models and clear outputs to visualize are both essential for generating meaningful clinical insights.
For this purpose, in the present study, we apply the item response latent space model (LSIRM, Jeon et al., 2021), a recently developed network approach for assessment data, to clinical symptom data in order to gain new insights concerning the interactions between patients and individual symptoms.
Importantly, we compare the LSIRM for binary data to Network Psychometrics (Epskamp et al., 2018), a well-established network approach for psychometric data.
In addition to conceptual differences, simulation and real data applications will be used to discuss similarities and differences between the two network approaches for clinical symptom data.

Structured factor analysis: A data matrix-based alternative approach to structural equation modeling

Thursday, 27th July - 16:15: Factor Analysis (Prince George) - Oral

Mr. Gyeongcheol Cho (McGill University), Dr. Heungsun Hwang (McGill University)

Jöreskog's covariance-based approach (JCA; Jöreskog, 1978) has been considered a standard method for structural equation modeling. However, JCA is prone to the occurrence of improper solutions and cannot make probabilistic inferences about the true factor scores. To address the enduring issues of JCA, we propose a data matrix-based alternative, termed structured factor analysis (SFA). Given a data matrix of indicators, SFA begins by estimating both measurement model parameters and factor scores by minimizing a single cost function via an alternating least squares algorithm, which mathematically guarantees convergence to proper solutions. It then employs the factor score estimates to estimate structural model parameters. Once all parameters are estimated, SFA further estimates the probability distribution of the factor scores that can generate the data matrix of indicators, which can be used for probabilistic inferences about the true factor scores. We investigate SFA's performance and empirical utility through simulated and real data analyses.

Comparison between Bayesian and frequentist regularization in factor analysis

Thursday, 27th July - 16:30: Factor Analysis (Prince George) - Oral

Ms. Lijin Zhang (Graduate School of Education, Stanford University), Dr. Xinya Liang (Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas), Prof. Junhao Pan (Department of Psychology, Sun Yat-sen University)

The application of regularization in factor analysis has been more and more popular in recent years. Methods like ridge, lasso, and adaptive lasso have been adopted to detect unspecified cross-loadings or explore the main loadings. Previous research demonstrated that lasso performed better in maintaining a simple model structure compared to ridge. However, these comparisons have been limited within Bayesian or frequentist framework, the difference between Bayesian and frequentist regularization in factor analysis remains unclear. Theoretically, Bayesian ridge can be equivalent to the frequentist ridge. But Bayesian regularization might be different from frequentist regularization in terms of quantifying the uncertainty of estimates, shrinking nuisance parameters exactly at zero, etc. In the current study, we compared ridge, lasso, and adaptive lasso with Bayesian and maximum likelihood estimation in confirmatory factor analysis. The aim of this research is to 1) investigate the similarity and difference in parameter estimation and variable selection of the same regularization under frequentist and Bayesian frameworks, and 2) explore whether the differences between various regularization methods remain similar under frequentist and Bayesian frameworks. We manipulated model size, sample size, effect size, factor correlation, and the number of non-zero cross-loadings per factor in the simulation study. The performance of different methods in identifying cross-loadings under various modeling conditions were investigated. We further discussed practical implications based on the findings in the current study.

InterModel Vigorish for model comparison in CFA with binary outcomes

Thursday, 27th July - 16:45: Factor Analysis (Prince George) - Oral

Ms. Lijin Zhang (Graduate School of Education, Stanford University), Prof. Benjamin Domingue (Graduate School of Education, Stanford University)

Confirmatory factor analysis (CFA) has been widely used to assess the fit of a theoretical measurement model to observed data. Here, we introduce a novel index, InterModel Vigorish (IMV), which measures the value (which is expressed in units that can be thought of as dollar amounts) of using one model over another based on prediction accuracy. We extend the IMV into CFA models with binary outcomes and demonstrate that IMV provides a unique perspective for model comparison. Three simulation studies were conducted to evaluate its effectiveness in model selection and compare IMV with traditional fitting indices. Results showed that IMV provides both model-level and item-level information for detecting model misspecification. It is also not sensitive to changes in sample size. The IMV and traditional fitting indices differ in what they evaluate: traditional indices (e.g., CFI, TLI) focus on the fit to the current dataset, while IMV focuses on model predictions and can penalize overfitted models. An empirical analysis illustrates the utilization of the IMV in practice. The IMV has practical implications for researchers and practitioners using CFA models and can be extended to structural models in future research.

Estimating the Completely Oblique 2-parameter Bifactor model

Thursday, 27th July - 17:00: Factor Analysis (Prince George) - Oral

Mr. Denis Federiakin (Department of Economic Education, Johannes Gutenberg University of Mainz), Prof. Olga Zlatkin-Troitschanskaia (Department of Economic Education, Johannes Gutenberg University of Mainz)

Bifactor models are common for modeling complex composite constructs (e.g. competence). However, researchers typically assume that all person parameters are orthogonal in these models, which can result in not meaningful model interpretations. Thus, to allow for estimation of non-zero correlations between all construct facets/dimensions, the authors (under review) proposed a new model – the Completely Oblique Rasch Bifactor (CORB) model. They proved that this model is identified, (in particular) if at least one item is shared between every pair of the specific factors (so-called "S-structure" of test dimensionality). However, like all Rasch models, the CORB model requires the constraint that all item discrimination parameters be set to the same value.
This presentation demonstrates that the CORB model can be treated as the first step in the estimation of the Completely Oblique 2-parameter Birnbaum Bifactor (CO2B) model. After the CORB model has converged, estimates of the correlations between person parameters can be constrained in the traditional 2PL bifactor model, and all discrimination parameters can be estimated as usual. Using a simulation study of the S-structure of test dimensionality and a real data example of the Systematic and Heuristic Information Processing (Schemer et al., 2008) scale, we show that the CO2B model provides a better approximation of the data-generating parameters than the orthogonal bifactor model. We also show that the most problematic case occurs when shared items have opposite discrimination parameters (positive and negative) on different specific factors. In such cases, we recommend dropping those items from the scale.

Controlling false discovery rate for exploratory factor analysis model

Thursday, 27th July - 17:15: Factor Analysis (Prince George) - Oral

Ms. Xinyi Liu (London School of Economics and Political Science), Dr. Yunxiao Chen (London School of Economics and Political Science), Prof. Irini Moustaki (London School of Economics and Political Science)

Exploratory factor analysis models aim to accurately reflect the relationship between indicators and latent factors. The false discovery rate (FDR) control problem is a crucial issue in this context, as it aims to limit the proportion of items that we falsely assume to load on certain factors when they do not. The proposed method based on mirror statistics overcomes limitations of popular methods, such as low power or the requirement for accurate estimation of the marginal distribution of latent factors.

The method uses data splitting and applies different rotation methods to each set of data. This approach allows for theoretical guarantees for controlling FDR when the true loading matrix is sparse, and the number of items approaches infinity. To further stabilize variable selection, we use the multiple data-splitting (MDS) method.

Numerical experiments demonstrate the effectiveness of the proposed method in controlling FDR and achieving high statistical power. This method has broad applicability to a wide range of studies that use exploratory factor analysis models.

Social network mediation analysis using degree-corrected stochastic block model

Thursday, 27th July - 16:15: Causal Inference (Margaret Brent) - Oral

Mr. chunyang zhao (Northeast Normal University), Dr. Xue Zhang (Northeast Normal University)

Social network mediation analysis, which quantifies the social network's mediation effect between social actors' independent variables and their outcomes of interest, has received much attention recently (e.g., Sweet, 2019; Liu et al., 2020; Che et al., 2020). However, existing models either unconvincingly choose the number of mediators or can't explain mediator defectively. To this end, we proposed a new social network mediation model in which the degree-corrected stochastic block model (DCSBM) allowing different connecting tendencies of each node was employed as the network part. By introducing community detection, the chosen number of mediators became more persuasive. We improved the interpretability of each mediator via choosing each node's connecting tendency vector. Network model parameters and the corrected parameters were estimated using maximum likelihood estimation (MLE) and method-of-moments (MOM), respectively. Simulation studies were conducted to evaluate the performance of the new model. The manipulated conditions included different sample sizes, numbers of communites, direct effects and indirect effects. A real data example to analyze a middle school students' friendship network was given in the end for an illustration.

A new causal mediation approach based on observational mediation modeling and instrumental variable regression

Thursday, 27th July - 16:30: Causal Inference (Margaret Brent) - Oral

Ms. Zhiming Lu (Sun Yat-sen University), Dr. Zijun Ke (Sun Yat-sen University)

Causal mediation analysis of data from randomized studies has recently been recommended for strengthening causality in research aimed at exploring the mechanism underlying a treatment effect. A typical experimental design consists of two experiments with the mediator measured in the first and manipulated in the second experiment. In this way, causal inferences can be drawn regarding the relationship between the mediator and the dependent variable. However, there are two concerns with existing approaches. First, to make the results from the two experiments comparable, the mediator in experiment 1 and the mediator in experiment 2 are conceptually deemed as the same variable. This assumption is often problematic because the mediators of interest in social psychology are often latent, continuous psychological constructs, which cannot be equated with the manipulation (e.g., a writing task for priming) of such a construct. Second, existing approaches largely overlook the possibility of other omitted mediator(s) covarying with the mediator of interest. This type of confounding cannot be eliminated by experimental manipulation of X and M, and ignoring such confounding can lead to biased estimates of the indirect effect. To address these concerns, we propose a new causal mediation approach for randomized studies based on the traditional observational mediation modeling framework and instrumental variable regression. We conduct Monte Carlo simulation studies to assess the proposed method's performance in parameter estimation and hypothesis testing compared to the existing methods, including Imai's causal mediation analysis, as well as the widely used analytical approaches for experimental-causal-chain and moderation-of-process designs.

Longitudinal mediation models: Understanding the impact of confounders and colliders

Thursday, 27th July - 16:45: Causal Inference (Margaret Brent) - Oral

Ms. Ziwei Zhang (University of Minnesota - Twin Cities), Dr. Nidhi Kohli (University of Minnesota - Twin Cities)

Mediators (M), confounders, and colliders are three types of variables that researchers usually consider for selecting covariates when identifying the relationship between the independent variable(s) (X) and the dependent variable (Y). Confounders are variables that have influences on both X and Y. In contrast, colliders are variables that are affected by both X and Y. Ignoring confounders and or controlling for colliders in a model can lead to biased model estimation. In educational and psychological fields, researchers are often interested in studying longitudinal mediation mechanisms (e.g., reading achievement growth affects science achievement growth via math achievement growth). In prior literature, researchers have explored the impact of ignoring confounders for latent growth curve mediation models (LGCMM) where both M and Y are repeated measures that follow a linear trend. However, no research has yet studied confounder effects on longitudinal mediation models for nonlinear patterns of change, nor collider effects on such models. Thus, the primary objective of the current study is to evaluate the performance of longitudinal mediation models for X, M, and Y where all the three concerned variables follow linear, and nonlinear trends (e.g., quadratic function, piecewise function with unknown random knots), in light of overlooking issues pertaining to confounders and or colliders. This study uses random effects models (REM) instead of LGCMM framework. This is because unlike LGCMM, REM can directly model intrinsically nonlinear functions (e.g., exponential, logistic) without requiring any transformations or reparameterizations. Lastly, the study will only focus on time-invariant confounders and colliders of M-Y relations.

Examining instrument relevance when there are multiple endogenous predictors: A new Index

Thursday, 27th July - 17:00: Causal Inference (Margaret Brent) - Oral

Dr. Zijun Ke (Sun Yat-sen University), Ms. Xin Tan (Sun Yat-sen University), Ms. Zhiming Lu (Sun Yat-sen University)

The instrumental variable (IV) method has been increasingly believed to be one of the most powerful tools to make causal inference in observational studies. One major challenge is how to select a set of theoretically and statistically qualified instruments. The question remains largely open when there are multiple endogenous predictors. It has been widely believed that with multiple endogenous predictors, it is important to ensure that "Z (instrumental variables) have components important to X1 (the focal endogenous predictor) that are linearly independent of those important to X2 (the rest X variables)" (Shea, 1997, p.348). Shea's index (1997) and Cragg-Donald (Cragg & Donald, 1993) statistic have been proposed to assess the instrument relevance in this situation. We show from the perspective of the structural equation modeling approach to the IV causal inference that there exist scenarios that the IV method works satisfactorily whereas both existing statistics indicate the lack of instrument relevance and thus the failure of the IV method. We thus develop a new measure to better assess the quality of instruments in this situation: singular value ratio (SVR). We use two simulation studies to examine the performance of the proposed measure SVR and compare it with the two existing statistics. We conclude the study with a discussion on the limitations of the newly developed measure and practical guidance for applied researchers.

Investigating weight constraint methods in causal-formative indicator modeling

Thursday, 27th July - 17:15: Causal Inference (Margaret Brent) - Oral

Ms. Ruoxuan Li (University of Notre Dame), Prof. Lijuan Wang (University of Notre Dame)

Causal-formative indicators are often used in psychology. To achieve identification in causal-formative indicator modeling (CIM), constraints need to be applied. A conventional method (CM1) is to constrain the weight of a formative indicator to be 1. The selection of which indictor to have the fixed weight, however, may influence the inference of the structural path coefficients from the causal-formative construct to outcomes. Another conventional method (CM2) is to use equal weights (e.g., 1) and assume all indicators equally contribute to the latent construct, which can be a strong assumption. To address the limitations of the conventional methods, we proposed an alternative constraint method, in which the sum of the weights is constrained to be a constant. We analytically studied the interpretations and relations in the structural path coefficients from the constraint methods, and found that the proposed method can yield better interpretations of the path coefficients. Simulation studies were conducted to compare the performance of the weight constraint methods in CIM with one or two outcomes. Results show that higher relative biases in the path coefficients were observed from the conventional methods compared to the proposed method. Specifically, the performance of CM1 depends on which indicator has a weight of 1 and the performance of CM2 is poor when formative indicators have unequal weights and variances. The proposed method had ignorable bias and satisfactory coverage rates in most studied conditions. This study emphasizes the importance of using an appropriate weight constraint method in CIM.

Addressing publication bias and uncertainty for power analysis: A hybrid classical-Bayesian approach

Thursday, 27th July - 16:15: Bayesian Methods (Juan Ramon Jimenez) - Oral

Ms. Winnie Wing-Yee Tse (University of Southern California), Dr. Mark H. C. Lai (University of Southern California)

While power analysis often requires the unknown population effect size, the conventional practice is to replace this unknown value with the best educated guess, which entails uncertainty. However, ignoring uncertainty and using a single best guess have been found to underestimate the sample size requisite, hence yielding a study design with low statistical power. To address uncertainty, past research has developed hybrid classical-Bayesian (HCB) approaches, which perform classical power analysis within the Bayesian framework. In essence, these approaches incorporate the prior distribution, which represents researchers' prior belief about the effect size, and determine sample size based on the resulting power distribution. The recommended choice of the prior distribution is a normal distribution based on the observed effect size estimate from the literature. Nonetheless, the observed effect size is likely an overestimate when only statistically significant findings are published, an issue known as publication bias. Such normal priors do not represent our belief about the effect size if we suspect that publication bias exists, and could result in an underestimate of the sample size needed if publication bias does exist. Therefore, in this study, I propose an HCB approach for power analysis to adjust for both uncertainty and publication bias. The proposed approach accounts for the truncated distribution of the effect size when only significant results are published. In the talk, I will present the results of a simulation study that evaluates the ability of the proposed approach in achieving the desired level of average power.

Alignment with Bayesian Region of Measurement Equivalence (ABROME) approach for multiple groups comparisons

Thursday, 27th July - 16:30: Bayesian Methods (Juan Ramon Jimenez) - Oral

Ms. Yichi Zhang (University of Southern California), Dr. Mark H. C. Lai (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, few studies evaluate the practical implications of noninvariance. The recently proposed *Alignment with Bayesian Region of Measurement Equivalence* (ABROME) approach quantifies the impact of partial invariance on the observed composite scores across groups, and allows researchers to use this index to directly support MI. Under the ABROME framework, researchers first compute *highest posterior density intervals* (HPDIs), which contain the most plausible values for expected group differences in total scores due to noninvariance, then compare the HPDIs with a predetermined range of values on the metric of total scores that are practically equivalent to no bias across groups (i.e., the *region of measurement equivalence*; ROME). One limitation of the ABROME framework is it evaluates MI across only two groups. However, applied researchers are often interested in assessing MI across more than two groups (e.g., ethnicity). Thus, the current study extends the ABROME framework to multiple groups by computing the cumulative expected group difference due to noninvariance and comparing this with the preset ROME. If noninvariance is found, pairwise group comparisons could be conducted to determine sources of noninvariance. The proposed procedure is illustrated using an empirical example that investigates the MI of a college-related alcohol beliefs scale across ethnic groups. Results show among all groups, the impact of noninvariance is largest between Asian and Hispanic participants, highlighting that the ABROME framework offers information beyond what conventional approaches can provide.

Estimating data saturation in qualitative research using approximate bayesian computation

Thursday, 27th July - 16:45: Bayesian Methods (Juan Ramon Jimenez) - Oral

Mr. Jinghao Ma (Waseda University), Prof. Hideki Toyoda (Waseda University)

Data saturation is a critical concept in qualitative research, referring to the state in which all relevant data in a study have been fully collected. In most cases, determining when data saturation is reached is subjectively decided by the researcher. In recent years, some studies have attempted to develop quantitative measures of saturation.

Toyoda, MA & Ohashi (2022) proposed a new method for estimating saturation using the sampling without replacement zipf distribution. The zipf distribution is a frequency-rank distribution. So estimating the parameters of the zipf distribution needs the Frequency and Rank of the data. However, in most qualitative research, only the frequency of each category can be observed. Through simulations, we found that assuming the rank of the data based on the observed frequency can overestimate saturation. Therefore, we used the approximate bayesian computation (ABC) method proposed by Pilgrim & Hills (2021) to estimate the parameters of the sampling without replacement zipf distribution. We demonstrated through simulations that the ABC method provides more accurate estimates of saturation compared to previous methods.

Identifiability and estimability of Bayesian linear and nonlinear crossed random effects models

Thursday, 27th July - 17:00: Bayesian Methods (Juan Ramon Jimenez) - Oral

Mrs. Corissa Rohloff (University of Minnesota - Twin Cities), Dr. Nidhi Kohli (University of Minnesota - Twin Cities), Dr. Eric Lock (University of Minnesota - Twin Cities)

Crossed random effects models (CREMs) are particularly useful in longitudinal data applications because they allow researchers to account for the impact of dynamic group membership on individual outcomes. However, no research has determined what data conditions need to be met to sufficiently identify these models, especially the group effects, in a longitudinal context. This is a significant gap in the current literature as future applications to real data may need to consider these conditions to yield accurate and precise model parameter estimates, specifically for the group effects on individual outcomes. Furthermore, there are no existing CREMs that can model intrinsically nonlinear growth. The goals of this study are to develop a Bayesian piecewise CREM to model intrinsically nonlinear growth and evaluate what data conditions are necessary to empirically identify both intrinsically linear and nonlinear longitudinal CREMs. This study includes an applied example that utilizes the piecewise CREM with real data and three simulation studies to assess the data conditions necessary to estimate linear, quadratic, and piecewise CREMs. Results show that the number of repeated measurements collected on groups impacts the ability to recover the group effects. Additionally, functional form complexity impacted data collection requirements for estimating longitudinal CREMs.

Pairwise likelihood limited information goodness of fit tests for factor models

Thursday, 27th July - 16:15: Model Fit (Thurgood Marshall) - Oral

Prof. Irini Moustaki (London School of Economics and Political Science), Dr. Haziq Jamil (Universiti Brunei Darussalam)

Limited information goodness of fit (GOF) tests have gained recognition in the literature for high-dimensional multivariate categorical data analysis. Sparsity issues in the ensuing contingency tables impair the dependability of GOF tests but can be circumvented by considering summary statistics involving univariate and bivariate residuals. Prior work in this area for factor models have focused mainly on maximum likelihood estimation, which itself can be computationally intensive when fitting large and complex models. This present work examines limited information GOF tests when composite likelihood estimation, specifically pairwise likelihood estimation, is used instead. Pairwise likelihood estimation offers a beneficial trade-off between computational efficiency and modelling accuracy in factor models, and hence we wanted to examine the performance of limited information GOF tests under this framework. The tests under consideration are based on the Pearson chi-squared test statistic and the Wald test statistic. We propose modifications to each of these tests with the aim of further reducing computational complexity. We then extend our findings beyond independent sampling to situations where complex sampling procedures (with known weights) are employed.

Using item scores and response times in person-fit assessment

Thursday, 27th July - 16:30: Model Fit (Thurgood Marshall) - Oral

Ms. Kylie Gorney (University of Wisconsin-Madison), Dr. Sandip Sinharay (Educational Testing Service), Dr. Xiang Liu (Educational Testing Service)

Joint models for item scores and response times are becoming increasingly popular in educational and psychological testing. In this paper, we propose two new person-fit statistics for such models in order to detect aberrant behavior in item scores and/or response times. The first statistic is computed by combining two existing person-fit statistics: one for the item scores, and one for the item response times. The second statistic is computed directly using the likelihood function of the joint model. Using detailed simulations, we show that (a) the empirical null distributions of the new statistics are very close to their theoretical null distributions, and (b) under our simulation conditions, the new statistics tend to be more powerful than several existing statistics for item scores and/or response times. A real data example is also provided using data from a licensure examination.

Latent class analysis with measurement invariance testing: Simulation study to compare overall likelihood ratio vs residual fit statistics based model selection

Thursday, 27th July - 16:45: Model Fit (Thurgood Marshall) - Oral

Dr. zsuzsa BAKK (Leiden University)

In latent class (LC) analysis a standard assumption is conditional independence, that is the indicators of the LC are independent of the covariates given the LC variable. Several approaches have been proposed for identifying violations of this assumption, and modeling direct effects. The recently proposed likelihood ratio based MIMIC test (Masyn, 2017) is compared to residual statistics (BVR and EPC statistics (Oberski et al., 2013; Vermunt & Magidson, 2016)) for identifying nonuniform direct effect of covariates on the indicators of the LC model. The simulation study results show that the LR test correctly identifies direct effects more often than the residual statistics, this at the price of having also a higher false positive rate. A real data example illustrates the use of the three procedures. Overall the combined use of the two type of statistics is recommended for applied research.

Necessity of model selections in CD: The absolute fit indices versus the general classification methods

Thursday, 27th July - 17:00: Model Fit (Thurgood Marshall) - Oral

Ms. Hyunjee Oh (University of Minnesota - Twin Cities), Prof. Chia-Yi Chiu (University of Minnesota - Twin Cities)

The study aims to assess the necessity of using the absolute model fit indices in cognitive diagnosis models (CDMs) in contrast to the generalized DINA (G-DINA) model and the general nonparametric classification (GNPC) method. Various conditions were formed with a special focus on sample size. Three absolute model fit indices, including SRMSR, RMSEA2, M2 were considered. The simulation results showed that when samples were small or when samples were large and examinees' attribute profiles were uniformly distributed, the GNPC method appeared to be the best choice. However, when samples were large and examinees' attribute profiles followed the multivariate normal threshold model, fit indices were preferred. The findings of this study can serve as a practical guideline on the selection of the most appropriate approach to maximize the classification rates in CD based on the sample size, data generation, and examinees' attribute profiles.

A re-evaluation of cross-validation methods for psychological time series data

Thursday, 27th July - 17:15: Model Fit (Thurgood Marshall) - Oral

Prof. Siwei Liu (University of California, Davis), Ms. Di Jody Zhou (University of California, Davis)

Cross-validation (CV) methods are widely used to evaluate the predictive performance of statistical models. Originally developed for cross-sectional data, which consist of independent observations, these methods have been less studied in the context of time series analysis, especially in psychology. Bulteel et al. (2018) evaluated the relative performance of five CV methods for psychological time series data (e.g., ecological momentary assessments, physiological and neuroimaging data). However, they did not examine whether performance was affected by characteristics of the data or the model fitted to the data. We present here a simulation study to re-evaluate two CV methods (ten-fold CV and blocked CV), specifically, whether and how they are affected by characteristics of the data and the fitted models. We show that CV methods tend to underestimate the prediction error of the person-mean and lag-1 autoregressive [AR(1)] model, and overestimate the prediction error of the lag-1 vector autoregressive [VAR(1)] model, especially when the number of observations is small. In addition, the error of CV varies with characteristics of the data, such as the magnitude of the autoregressive and cross-lagged parameter values. Finally, we find that CV methods are generally preferred over fit indices for selecting the best predictive model; however, in some scenarios they tend to incorrectly select a more parsimonious model (e.g., selecting AR(1) over VAR(1) while VAR(1) has better predictive performance). These results provide a fuller picture of the performance of CV methods in time series analysis. We will discuss the implications of our results in psychological research.

Some pathologies of psychometrics: Philosophical perspectives

Friday, 28th July - 09:00: Spotlight Session: Perspectives on Psychometrics (Colony Ballroom) - Oral

Dr. Mark Wilson (University of California, Berkeley)

In this paper, I utilize a new philosophical framework developed as a foundation for measurement to seek to understand some well-known frustrations with psychometrics. The framework is called “Measurement across the sciences” (MATS; Mari et al., 2021), and it describes six aspects of the foundations of measurement that taken together, can supply both necessary and sufficient conditions for justifying measurement practices. The “pathologies” in question expand upon the earlier contention by Michel about the unscientific nature of contemporary psychometric foundations and practices. The pathological practices include (a) the poverty of the conceptual underpinnings of the property being measured, (b) the lack of systematicity in developing items with which to observe the property, (c) the lack of definitional coherence in categorizing of the item responses, (d) the obscurity of the scoring procedures for the responses to the items, (e) the under-justification for using a (weighted) sum-score to estimate respondents scale locations and (f) the absence of an evidentiary basis for interpreting respondent locations on the quantitative latent variable. In the paper, the initial step will be to explicate the MATS framework, in tandem with the BEAR Assessment System (BAS) and illustrate its application to an example measurement system for Scientific Argumentation. Second, the framework will be recruited to explicate these frustrating practices in typical psychometric applications and discuss their origins and relationships (or, lack of relationships). Third, the framework will be utilized to describe scientifically-sound resolutions to these practices, and especially the inherent linkages between the parts of the framework.

Was psychometrics a mistake? Criticism from metrology and axiomatic measurement.

Friday, 28th July - 09:35: Spotlight Session: Perspectives on Psychometrics (Colony Ballroom) - Oral

Dr. Keith Markus (John Jay College of Criminal Justice, CUNY)

Application of metrology to testing in the behavioral sciences has brought a new wave of criticism of psychometrics continuous with the criticism associated with axiomatic measurement theory. The basic criticism is that standard units are not defined for constructs and thus a metric construct is assumed rather than tested by psychometric models used to evaluate tests. This claim is sound as far is it goes but often a further conclusion is drawn that psychometrics is therefore inherently flawed and must be replaced by metric tests. The result is a lopsided stalemate in which the testing community is happy to incorporate any useful insights that axiomatic approaches can offer but enthusiasts of axiomatic approaches are focused on what they present as a paradigm shift away from psychometrics. Attention to three principles can help improve the quality of argumentation in this critical literature: internal criticism (criticizing a program by its own goals), charity (criticizing the most charitable interpretation) and parity (recognizing when parallel arguments cut both ways). Contributors to this critical literature can make more effective contributions by setting aside the artifice asserting a new paradigm. Specifically, three features of this critical literature limit its effectiveness to a broader audience: Accounts of current testing practice fail to explain its current success. Arguments for radical change in testing practice remain disconnected from test construction itself. Despite claiming the mantle of realism, arguments remain insufficiently self-reflective regarding implicit empiricist assumptions.

Bridging the gap between probabilistic perception and generalization behavior with a computational model

Friday, 28th July - 09:00: Longitudinal and Multilevel Models (Atrium) - Oral

Mr. Kenny Yu (KU Leuven), Prof. Francis Tuerlinckx (KU Leuven), Prof. Wolf Vanpaemel (KU Leuven), Dr. Jonas Zaman (KU Leuven)

Human generalization research seeks to understand the processes that underlie the transfer of prior experiences to new contexts. To capture variations in the generalization process, researchers often rely heavily on descriptive statistics and has assumed a single generalization mechanism, which overlooks individual differences on perceptual sensitivity and may lead to biased conclusions about the generalization process. In this study, we propose a computational model that accounts for the effects of noisy perceptual mechanisms on generalization behavior. The proposed computational model allows researchers to better understand the causes of variations in generalization behavior for a given stimulus, providing a less biased measure of generalization tendency. This is important for distinguishing individuals with overgeneralization tendencies from those with perceptual acuity. By integrating both perceptual and generalization data, our approach provides a more nuanced understanding of generalization behavior, bridging the gap between individual perceptual sensitivity and generalization behavior. This has the potential to inspire future studies to explore more about the complex interplay between probabilistic perception and generalization behavior, ultimately leading to more accurate and effective approaches for treating individuals with overgeneralization tendencies in clinical settings.

Comparison of multilevel vs. standard prediction algorithms on nested data

Friday, 28th July - 09:15: Longitudinal and Multilevel Models (Atrium) - Oral

Ms. Brennan Register (University of Maryland, College Park), Dr. Tracy Sweet (University of Maryland, College Park)

The organization of data from our education system inherently requires a nested structure. For instance, students are in classrooms, which themselves are within schools, schools are within school districts, and so on. Therefore, many studies in educational research involve the collection and analysis of multilevel data. To answer important questions about education policies, practices and student outcomes, a variety of machine learning algorithms can be used to analyze large-scale datasets. For example, researchers can use classification algorithms to predict which students will graduate from undergraduate studies in a timely fashion, which students will pass a course, or identify STEM students at risk of dropping out of their course of study. Recently, machine learning has become increasingly common in education research; however, many of the most commonly used algorithms were built for non-nested data and very little research has been done comparing the prediction performance of standard prediction algorithms with that of their multilevel counterparts.Additionally, few studies were motivated by and applied to education-related data. Therefore, additional research is needed to determine under what conditions, if any, multilevel prediction algorithms improve upon the predictive performance of standard prediction algorithms in the context of nested data.

The current study will address this gap in the literature using Monte Carlo simulation and empirical data analysis to compare the predictive performance of several multilevel prediction algorithms and several standard predictionalgorithms. Results from this study will aid education researchers in selecting appropriate prediction techniques when working with multilevel data.

Fitting differential equation models to ILD with numerical optimizer

Friday, 28th July - 09:30: Longitudinal and Multilevel Models (Atrium) - Oral

Dr. Yueqin Hu (Beijing Normal University), Mr. Qingshan Liu (Beijing Normal University), Ms. Minglan Li (Beijing Normal University)

Differential equation models can be used to study univariate or multivariate dynamic systems, such as the self-regulation process of emotions, the bidirectional relationship between stress and sleep quality, and parent-child interactions. Based on intensive longitudinal data, this study proposed a one-step method for estimating the parameters of differential equation models, i.e., the model parameters were obtained by fitting differential equations directly to intensive longitudinal data using a numerical optimization method. Simulation studies supported the performance of this method, which had less bias than the two-step estimation methods and comparable accuracy to the one-step estimation method using the analytic solution of differential equations. The numerical estimation method proposed in this study can also be applied to multivariate and nonlinear models for which analytic solutions are usually not available. We have also compared different optimization algorithms, including Nelder-Mead, CG, BFGS and L-BFGS-B. Simulation results suggested that the L-BFGS-B method outperformed the other methods in terms of estimation accuracy and robustness to initial guesses. We have developed an R package and a Python package to facilitate the implementation of numerical methods for differential equation models. These packages currently support up to second order differential equations. Illustrative examples on empirical behavioral data were also provided. Future extensions under consideration include latent variables and multi-level models.

Sample planning for detecting cross-lag effect in longitudinal studies with ordinal outcomes

Friday, 28th July - 09:45: Longitudinal and Multilevel Models (Atrium) - Oral

Ms. Sijing (SJ) Shao (University of Notre Dame), Ms. Ziqian Xu (University of Notre Dame), Dr. Ross Jacobucci (University of Notre Dame)

Repeated measured data is a powerful tool and plays an important role in fields like psychology. At the designing stage of studies, the number of subjects and frequency of assessments must be determined by researchers. Insufficient samples can lead to inconclusive results, while over-sampling can waste the limited resources from researchers and participants. Thus, power calculation along with sample planning is crucial in studies to avoid these consequences. Comprehensive approaches for power calculation for repeated measures designs have been proposed and reviewed in last decades. However, many of these methods considered normally distributed outcomes while very often, outcomes are measured on an ordinal scale. Even though there is a growing recognition of the importance of specifying ordinal distributions in multilevel models (Bauer and Sterba, 2011), few studies have investigated sample size planning when the ordinal outcomes are correctly specified. Furthermore, cross-lag effect, which is the effect of one variable to another in the next time point, is vital when longitudinal data are collected to capture the dynamics in process. Multilevel autoregressive models specifying cross-lag effects are considered to serve this purpose. In this talk, we will discuss sample planning under multilevel autoregressive models framework in which the cross-lag effect is of interest. An empirical study is presented for motivation of the study. In addition, simulations with varying conditions of assessment frequency, desired effect size, and levels of violations of normality assumptions in the random effects of autoregressive and cross-lag effects are performed to determine sample sizes when outcomes are ordinal.

The many reliabilities of affective dynamics

Friday, 28th July - 10:00: Longitudinal and Multilevel Models (Atrium) - Oral

Dr. Sebastian Castro-Alvarez (University of California, Davis), Prof. Laura F. Bringmann (University of Groningen), Prof. Siwei Liu (University of California, Davis)

Reliability is a key concept in psychology that has been broadly studied since the introduction of Cronbach's alpha, which is a measure of the internal consistency of a test. Despite its importance, this is a topic that is relatively understudied when dealing with intensive longitudinal data. In particular, when studying the psychological dynamics of affective states, there is no warranty that intensive longitudinal measurements are reliable. Given this, empirical researchers need tools to study and report the reliability of the scales used in intensive longitudinal research. In recent years, different approaches to estimate the reliability of the scales and the items used when studying psychological dynamics have been proposed. However, the advantages and disadvantages of each of these methods are unclear, making it difficult to determine when a certain approach would be preferred over the others. Specifically, these diverse approaches estimate reliability indices based on statistical models such as linear multilevel analysis, vector autoregressive models, and dynamic factor models. Furthermore, while some methods suggest estimating one reliability index for the scale that applies to the whole sample, others estimate specific reliability indices for each individual in the sample. This wide variety of approaches can provoke some confusion for empirical researchers. Therefore, we aim to highlight the advantages and disadvantages of each of the available methods used to estimate the reliability of intensive longitudinal data. We also showcase their use with empirical data.

The implications of IRT calibration and IRT scoring choices on group differences for large scale language exams

Friday, 28th July - 09:00: Item Response Theory (Benjamin Banneker) - Oral

Dr. YoungKoung Kim (College Board), Dr. Tim Moses (The College Board)

When the analysis sample for a large-scale test includes subgroups which have distinctive characteristics, the selection of sample can affect the results of IRT calibration and scoring procedures can affect the characteristics of the resulting scores. The calibration and scoring of the items from a language exam is a good example since the exam is often taken by 'heritage speakers' although the exams are designed to measure the language ability for those who do not speak the language. The goals of the current study are to examine the impact of sample selections for calibrations and the implications of IRT scoring method choices on examinee scores when there are qualitatively different groups in the data. Using large scale language exam data, the present study compares the results of calibration including heritage speakers with the results excluding heritage speakers to evaluate the stability of the calibrations. Also, TCC, MLE and EAP scoring are examined to see the impact of IRT scoring methods on the heritage speakers versus non-heritage speakers.

Specifically, the approaches taken in this study will be to obtain calibration results for a large-scale language test based on two analysis samples, total and non-heritage speakers, and to score the total group of examinees with MLE, EAP and TCC scoring approaches using both sets of calibration results. Then, the scores will be evaluated for characteristics of interest, namely mean group differences in heritage speakers vs. non-heritage speakers using each of the six possible sets of scoring and calibration results.

Accurately estimating the dimensional relationships with the three-tier IRT model

Friday, 28th July - 09:30: Item Response Theory (Benjamin Banneker) - Oral

Dr. Ken Fujimoto (Loyola University Chicago), Dr. Eunju Yoon (Loyola University Chicago), Dr. Matthew Miller (Loyola University Chicago)

Researchers in psychology often measure latent traits with complex dimensional structures, such as those with multiple layers of specific dimensions nested within general dimensions (e.g., a three-tier structure). One such example is acculturation and enculturation, where the general dimensions that form the first tier of the three-tier structure. These general dimensions have specific dimensions nested within them, like language and identity (i.e., the second tier of dimensions), which have dimensions nested within themselves, like speaking, entertainment, and literacy (i.e., the third tier of dimensions).

The correlations among the general dimensions are often of interest. Unfortunately, many researchers estimate these correlations using a simple-structured multidimensional item response theory (IRT) model or by simply calculating the Pearson correlation coefficient based on raw scores. The problem with these approaches is that they ignore the lower tiers of the dimensional structure, thereby artificially increasing the covariation among the items while the covariation among the general dimensions is not affected, leading to underestimating the correlations among the general dimensions.

Through simulations motivated by actual acculturation-enculturation data, we demonstrate that the correlation between the general dimensions of a three-tier structure can be substantially underestimated. When common approaches were used, the true correlation of $r = –.70$ was estimated to be a weak-to-moderate correlation (e.g., $r = –.4$). Such a drastic underestimation could alter substantive conclusions. We also present an expanded version of the Bayesian three-tier IRT model and demonstrate through simulations that this model can accurately estimate the true correlation between the general dimensions.

Information test of model fit assessment for multidimensional item response models

Friday, 28th July - 09:45: Item Response Theory (Benjamin Banneker) - Oral

Mr. Youngjin Han (University of Maryland, College Park), Dr. Yang Liu (University of Maryland, College Park), Dr. Ji Seung Yang (University of Maryland, College Park)

Goodness-of-fit (GOF) assessment is a crucial step in applications of any statistical models including item response theory (IRT) models. Several statistics have been widely used for evaluating the GOF of IRT models, either at the overall level (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2005; Reiser 2008) or the item level (Haberman et al., 2013; Orlando & Thissen, 2000). However, these statistics are difficult to compute for multidimensional item response theory (MIRT) models due to the involvement of high-dimensional numerical integration. This study suggests a GOF statistic for MIRT models that is computationally more efficient using the information matrix. GOF can be evaluated by the discrepancy between the cross-product form and the Hessian form information matrices (White, 1982). Ranger and Kuhn (2012) observed that this information test exhibits decent performance for unidimensional IRT models. The study extends the framework to multidimensional models. A distinguishing feature of the method is that it can be used in conjunction with the Metropolis Hastings-Robins Monro algorithm (Cai, 2010), which has been implemented in widely used IRT software packages and recommended for MIRT models with more than 4-5 latent variables. We conduct a Monte-Carlo simulation study evaluating the empirical Type I error rate and power of the information test.

Differential algorithmic functioning: A framework for evaluating fairness in algorithmic decision making

Friday, 28th July - 09:00: Differential Item Functioning (Prince George) - Oral

Dr. Youmi Suk (Teachers College Columbia University), Dr. Kyung T. Han (Graduate Management Admission Council)

The prevalence of algorithmic decision-making processes, often as part of Artificial Intelligence (AI) systems, in our society has brought significant attention to the potential fairness issues with these algorithms. Despite its importance, there has been a lack of research on how to effectively evaluate algorithmic fairness through the lens of psychometrics. This paper proposes a new framework for evaluating algorithmic fairness, called Differential Algorithmic Functioning (DAF), which is based on the well-established concept of Differential Item Functioning (DIF) in psychometrics. The DAF framework consists of three crucial components: a decision variable, a "fair" variable, and a protected variable (such as race or gender). Under the DAF framework, an algorithm can exhibit uniform DAF, nonuniform DAF, intersectional DAF, or neither (i.e., non-DAF). The concept of intersectional DAF is introduced to address intersectionality across protected variables. To detect DAF, we propose modifications of DIF methods, such as the Mantel-Haenszel test, logistic regression, and residual-based DAF. We demonstrate the efficacy of the DAF framework through a real dataset concerning grade retention algorithms in K-12 education in the United States.

An effect size and asymptotic test for differential test functioning

Friday, 28th July - 09:15: Differential Item Functioning (Prince George) - Oral

Dr. Peter Halpin (University of North Carolina at Chapel Hill)

In many applied research settings, it is of interest to know whether and to what extent Differential Item Functioning (DIF) may affect comparisons among groups' scale scores, notably group-mean differences or "impact". One way to address this question is by comparing estimates of impact under two different model specifications: (1) no items exhibit DIF versus (2) some items may exhibit DIF. The difference between these two estimates can be interpreted as an effect size for Differential Test functioning (DTF). In this presentation, the theory of M-estimation is used to derive the asymptotic null distribution of this effect size. The estimate and resulting test can be computed as a post-processing step following separate calibrations in each group (i.e., it is not required to pre-specify which items, if any, may exhibit DIF). The approach builds on recent work in robust scaling and robust approaches to DIF, which are reviewed in the presentation.

DIF statistical inference without knowing anchoring items

Friday, 28th July - 09:30: Differential Item Functioning (Prince George) - Oral

Dr. Yunxiao Chen (London School of Economics and Political Science), Mr. Chengcheng Li (University of Michigan, Ann Arbor), Ms. Jing Ouyang (University of Michigan, Ann Arbor), Dr. Gongjun Xu (University of Michigan, Ann Arbor)

Establishing the invariance property of an instrument is a crucial step for establishing its measurement validity. Measurement invariance is typically assessed by differential item functioning (DIF) analysis, i.e., detecting DIF items whose response distribution depends not only on the assessed latent trait but also on the group membership. Many DIF analyses require knowing several anchor items that are DIF-free in order to draw inferences on whether each of the rest is a DIF item. When no prior information on anchor items is available, or some anchor items are misspecified, item purification methods and regularized estimation methods can be used. The former iteratively purifies the anchor set by a stepwise model selection procedure, and the latter selects the DIF-free items by a LASSO-type regularization approach. Unfortunately, unlike the methods based on a correctly specified anchor set, these methods are not guaranteed to provide valid statistical inference. In this paper, we propose a new method for DIF analysis under multiple indicators and multiple causes (MIMIC) model for DIF. This method adopts a minimal L1 norm condition for identifying the latent trait distributions. Without requiring prior knowledge about an anchor set, it can accurately estimate the DIF effects of individual items and further draw valid statistical inferences for quantifying the uncertainty, where the inference results may not be obtained with item purification and regularized estimation methods. Numerical studies are conducted to evaluate the performance of the proposed method and we compare it with the anchor-set-based likelihood ratio test approach and the LASSO approach.

Leveraging language prompts to generate distractors for fill-in-the-blank items

Friday, 28th July - 09:00: IRT Applications (Margaret Brent) - Oral

Dr. Jiyun Zu (Educational Testing Service), Dr. Ikkyu Choi (Educational Testing Service), Dr. Jiangang Hao (Educational Testing Service)

Flexible test administrations (e.g., testing at home, continuous testing) post heavy demands for large and continuous supplies of new items. Automated item generation (AIG), in which computerized algorithms are used to create test items, can potentially alleviate this demand by increasing the efficiency of new item development. One challenge in developing multiple-choice item is to write effective distractors, which need to be incorrect yet attractive (Haladyna, 2004). A question is how to automatically generate distractors for language assessments when the number of successful examples is not large. In this paper, we proposed a prompt-based learning (Liu et al., 2021) approach for automatically generating distractors for fill-in-the-blank vocabulary items, which are one of the most common item types in language assessments. Prompt-based learning, which leverages language prompts in finetuning language models, has been found to be particularly effective in small-sample scenarios (Gao et al., 2021). Considering distractor generation as a language generation task, we develop prompts for distractors and finetuned a transformer-based pretrained language model (Radford et al., 2019). We illustrated this approach using data from a large-scale standardized English language proficiency assessment. Specifically, we studied the effects of different prompts and demonstrated the effectiveness of the proposed prompt-based learning approach by comparing features of generated distractors with those from a rule-based approach.

Estimating an individuals' contribution to small-group task performance

Friday, 28th July - 09:15: IRT Applications (Margaret Brent) - Oral

Dr. Patrick Kyllonen (Educational Testing Service), Dr. Jonathan Weeks (Educational Testing Service), Dr. Jiangang Hao (Educational Testing Service), Dr. Michael Fauss (Educational Testing Service), Ms. Emily Kerzabi (Educational Testing Service)

We conducted studies with over 800 4-person teams designed to evaluate approaches to measuring individual contributions to collective performance on collaborative tasks: Letters-to-numbers problem solving, hidden-profile decision making, and "new recruit" negotiation. We evaluated individual and team performance, with tasks administered face-to-face (approximately 200 four-person teams, 4 university sites) and online (approximately 600 remote four-person teams). Participants appeared in two-day sessions on two different teams, enabling the quantification of individual contributions to teams. Participants used Educational Testing Service's (ETS) Platform for Collaborative Assessment and Learning (EPCAL) for task administration. Analyses estimated individuals' contributions to team success, evaluated personality and ability contributors to success, and compared the nature of the conversations engaged in by successful versus less successful teams. We fit a cross-classified mixed effects model by task to identify the contributions to the team scores from each individual. Due to data sparseness and other causes these models sometimes failed to converge; thus we also fit fixed-effects models to predict (a) team score as a function of background variables, and (b) the difference in predicted team score for different teams, with the model residual serving as an indicator of the additional contribution of the individual beyond the prediction given by background variables, finding positively correlated model residuals. We thus propose two interpretations of *individual contribution to teams*—background and conditioned on background. Potential research applications include (a) a test of *teamwork skills*, for selection, diagnosis, and monitoring, and (b) training for specific collaborative skills such as sharing, acknowledgement, and negotiation.

Modeling supervised and unsupervised items for non-cognitive tests

Friday, 28th July - 09:30: IRT Applications (Margaret Brent) - Oral

Dr. Veronica Cole (Wake Forest University), Dr. Shyh-Huei Chen (Wake Forest University School of Medicine), Dr. Patrick Kyllonen (Educational Testing Service), Dr. Jiyun Zu (Educational Testing Service), Dr. Edward Ip (Wake Forest University School of Medicine)

In many non-cognitive tests using multiple choice items, there is not necessarily one “correct” response option. For example, in a situational judgment test (SJT) that measures emotional intelligence, a situation about your good friend moving to another state is presented, then the respondent is asked to select amongst several options about how to respond (e.g., spend time with other friends, or hope that your best friend will return). Often expert opinion is used for forming keys for scoring such items, resulting in “correct” and “incorrect” answers. However, even experts may not agree on all items in a test. One promising data-driven approach to addressing this ambiguity uses the nominal response model (NIRM) to score the items. Without a key, the NIRM is analogous to unsupervised learning in which an individual’s score reflects how much the individual is similar or dissimilar to the patterns of other respondents. The direction of the score derived from NIRM, however, may not align well with the construct of interest. The current study proposes the introduction of “supervised” items (i.e., in a mixed item response model), in which keys are prespecified, into a test. In the current analysis, the extent to which supervised items can be used to “anchor” the direction and improve bias was explored. We analyzed the results of an SJT in young adults to illustrate the concept. We further conducted simulation to assess the NIRM and the mixed model for various mixes of supervised and unsupervised items.

Comparing IRT-based models for recognition task data

Friday, 28th July - 09:45: IRT Applications (Margaret Brent) - Oral

Ms. Nana Kim (University of Minnesota - Twin Cities)

There have been recent methodological studies integrating the item response theory (IRT) modeling framework with other cognitively-based models for recognition memory task data. Thomas et al. (2019) demonstrated a link between signal detection theory (SDT) and IRT models and, more recently, Güsten et al. (2022) proposed a General Condorcet Model for Recognition (GCMR) that combines the Rasch model and the 2-High-Threshold (2HT) model. Using empirical datasets, I compare different IRT-based approaches to modeling item responses on exposure tasks that require distinguishing “new” from “previously seen” items. The results illustrate some psychometric properties of recognition task items that are often overlooked in cognitive models, such as varying item discriminations and the effects of item position. I suggest the potential presence of distinct statistical dimensions for correct identification of new versus previously-seen item types, possibly reflecting respondent-level response biases that have implications for the use of such measures in understanding the cognitive processes and individual differences measured by recognition memory tests.

Exploring the feasibility and effectiveness of using NLP for generating valid and reliable clinical skills assessment items: An expert review approach

Friday, 28th July - 09:00: Measurement Applications (Juan Ramon Jimenez) - Oral

Dr. Burhanettin Ozdemir (Prince Sultan University), Dr. Arwa AlSughayyer (Saudi Commission for Health Specialties)

The development of valid and reliable assessment tools is crucial for evaluating clinical skills in medical education. In recent years, natural language processing (NLP) technologies have gained attention as potential tools for generating assessment items, such objective structured clinical examinations (OSCEs), structured oral examinations (SOEs), and multiple-choice questions (MCQs). This study explores the feasibility and effectiveness of using GPT-3 and ChatGPT, a state-of-the-art language models developed by OpenAI, for item generation purposes.

Using a dataset of previously generated OSCE, SOE, and MCQs by the Commission for Health and Specialties, GPT-3 was fine-tuned on the task of generating new assessment items. The generated items were evaluated by a committee of medical experts responsible for item generation process for their relevance, validity, and appropriateness. The generated items were also checked for plagiarism and results showed a range between 6 to 25%.

The results show that GPT-3 can generate high-quality assessment items that are comparable to those generated by human experts. However, the models are only as good as the data they are trained on and may reproduce biases or errors present in the training data

Overall, the study demonstrates the potential of NLP technologies, for generating valid and reliable items in medical education. The use of subject matter experts in evaluating the generated-items and feedback messages can help ensure their quality and validity. Future research should continue to explore the use of these technologies in real-world assessment settings and consider ways to mitigate potential biases and errors in the generated-items.

Using residual analysis to evaluate invariant measurement in cross-cultural research: The case of mathematics behaviors

Friday, 28th July - 09:15: Measurement Applications (Juan Ramon Jimenez) - Oral

Ms. Cigdem Toptas (University of Georgia), Prof. George engelhard (university of Georgia)

The purpose of this study is to examine the invariance of a mathematics behaviors scale used in international educational research. The model-data fit can be examined by examining the differences between the observed data and measurement model. If the residuals have systematic patterns, then we plan to examine sources of misfit. Differential item and person functioning is usually evaluated for cognitive items, but these analyses are less frequently conducted for affective items that are included in cross-cultural studies. In addition to examining item fit, this study stresses the description of methods for evaluating person fit for affective scales.

Model-data fit of both items and persons are evaluated from the perspective of Rasch measurement theory using data from PISA 2012. Specifically, we are examining data from four countries (6,856 from Chile, 6,351 from Japan, 4,848 from Turkey, and 4,978 from the United States). The mathematics behaviors scale consists of 8 items, and the student responded in four categories. (1= “never or rarely”, 2= “sometimes”, 3= “often”, 4= “always or almost always”).

The preliminary analyses suggest that the items exhibit good model-data fit. Person misfit appears to vary across the four countries. The final study will provide additional information about model data fit within the context with affective scales.

Examining the influence of linguistic features of student writing on rater scores with latent profile analysis

Friday, 28th July - 09:30: Measurement Applications (Juan Ramon Jimenez) - Oral

Dr. Magdalen Beiting-Parrish (Department of Education), Dr. Sydne McCluskey (NWEA)

Currently, the gold standard for assessing student written work is to use human raters who assign scores to students, usually based on a rubric. These rubrics may give a holistic score or based on individual rubric-based traits. These human-scored essays are also frequently used to train automated scoring engines but humans may not know exactly why they assigned the scores they have or may be biased in their scoring, based on their past experiences (Weigle, 2013). Additionally, essays with a singular holistic score do not capture the intricacies of the examinee's knowledge/misconceptions nor all of the lexical/linguistic factors that went into the final score. The current research hopes to build a more nuanced understanding of features that underpin essay scoring by suggesting a more robust scoring process than simple rubric scoring using several essay sets from the Hewlett Automated Student Assessment Prize. Firstly, essay sets scored with one holistic score compared with individual trait scores will be analyzed using Latent Profile Analysis (LPA) to create more thorough insight into student writing ability. Next, Natural Language Processing (NLP) will be used to create a profile of some of the more common parts of speech, word counts, and level of linguistic complexity per student response. These variables will then be used to create a series of models predicting class membership with the aim of better understanding the lexical/linguistic features that may be implicitly contributing to student profiles of writing ability as well as may be influencing human scoring.

Intuitive discrete model for likert scale called GSD

Friday, 28th July - 09:45: Measurement Applications (Juan Ramon Jimenez) - Oral

Prof. Lucjan Janowski (AGH University of Science and Technology), Dr. Bogdan Ćmiel (AGH University of Science and Technology), Dr. Krzysiek Rusek (AGH University of Science and Technology), Mr. Jakub Nawala (University of Bristol)

The likert scale is popular in numerous research fields. A latent variable approach (e.g., ordered probit) is often used to model such data.In many cases, it is convenient or even necessary to avoid latent variable approaches (e.g., a sample size too small) in order to avoid the latent variable approach.To avoid this, an appropriate discrete distribution class is required. We proposed a family of discrete probability distributions with only two parameters. The parameters can be easily interpreted, similar to normal distribution parameters. We call the new class the Generalised Score Distribution (GSD). The proposed GSD class covers the entire set of possible means and variances for any fixed and finite support, regardless of the number of answers. Furthermore, the GSD class can be treated as an underdispersed continuation of a reparametrized beta-binomial distribution. The parameters of the GSD class are intuitive and easy to estimate using the moment or MLE method.

Impact of low quality data on psychometric properties of scale

Friday, 28th July - 10:00: Measurement Applications (Juan Ramon Jimenez) - Oral

Dr. Nivedita Bhaktha (GESIS), Dr. Clemens Lechner (GESIS)

Low quality data (LQD) can be defined as those responses in the sample where the respondents have not put in sufficient thought and effort into responding, leading to their data not reflecting their true position on the construct being measured. It is a major concern in online survey research involving multi-item scales. Consequently, there has been a surge in literature on methods for identifying LQD. While there are many methods to identify such responses, we are interested in assessing how LQD affects the psychometric properties of scale such as dimensionality, reliability, convergence of factor model, model fit indices (chi-squared, CFI, RMSEA, SRMR), and the correlation between factors. We conduct a simulation study to examine the extent and impact of LQD on the psychometric properties of scale. We simulate four different kinds of LQD – mid-point responses, extreme style responses, random responses, and a mixture of them. The different factors considered in the simulation study are – percentage of LQD in the sample, number of factors, number of items, strength of the factor loadings, and strength of the correlation between factors. We find that different kinds of LQD affect the psychometric properties differently. Higher percentages of LQD are more detrimental but in some cases, even a low percentage of LQD (10%) tends to affect the results significantly.

Development of an eleven-Item scale for measuring food insecurity

Friday, 28th July - 09:00: Measurement Applications (Thurgood Marshall) - Oral

Ms. Jing Li (Universality of Georgia), Prof. George engelhard (university of Georgia)

The purpose of this study is to develop an eleven-item scale for measuring food insecurity based on a subset of items used in the Household Food Security Survey Module (USDA). A polytomous Rasch model is used to calibrate the scale. The data are based on families with children who participated in the United States Current Population Survey Food Security Supplement (CPS-FSS) in 2019 (N=1,248). This study differs from previous research in several ways. First of all, a continuous scale based on Rasch measurement theory is developed. A continuous scale provides increased sensitivity to changes in the severity of food insecurity as compared to simply reporting food insecurity categories. Another feature of the new Scale is the content alignment between child and adult items on the scale. Also, our approach uses the ratings directly obtained from respondents based on a rating scale structure rather than dichotomizing items. Overall, the model fit the data very well with 64.2 percent of variance explained by the model. The scale also provides the opportunity to determine cut scores so households can be assigned to USDA food insecurity categories. A scoring table is provided to convert observed scores to a metric scale based on the Rasch model. This study has implications for research, theory, and practice related to the measurement of food insecurity as well as secondary analyses of food insecurity.

Improving national assessment system in Uganda by means of modern test theory

Friday, 28th July - 09:15: Measurement Applications (Thurgood Marshall) - Oral

Mr. Lutalo Bbosa Sserunkuuma (University of South Africa)

The National assessment system in Uganda has informed several high stake decisions in the country's education system. However, its reliance on classical test analysis impedes it from the opportunities item response analysis permits. The study aimed at leveraging modern test theory principles to improve the psychometric properties of the national assessment of progress in education (NAPE) instruments. The study used the July NAPE 2018 numeracy and literacy in English assessment data gleaned from 31362 and 30954 grade three and six learners respectively from 1558 primary schools selected from each of the 122 districts in Uganda. The Rasch partial credit model was implemented in Winsteps software version 4.5.1 to glean evidences for measures' construct-related validity, differential item functioning, testlets effects, standard setting, and equating. Results showed that the NAPE measures: were; Unidimensional, invariant, targeting, affected by testlets effects, malleable to Rasch standard setting and equating, had; high person and item reliability statistics (>= 0.90), admissible person separation statistics (>=3.00) and item separation indices (>= 85), positive polarity, locally independent items, well balanced person (-6.70 to 5.29 logits) and item measures (-4.60 to 4.49 logits), few items with DIF and inadmissible hierarchy, many non-monotonic items with sufficient average and category; measures and fit statistics, items and persons fitting the Rasch models usefully not perfectly, and measures had instances of construct under-representation, item redundancy, construct-irrelevant variance and response aberration patterns. NAPE measures can be used in future administrations with improvements in the grim areas noted to increase their test score use and interpretation.

Investigating a potentially culturally relevant NAEP reading text passage

Friday, 28th July - 09:30: Measurement Applications (Thurgood Marshall) - Oral

Dr. Saki Ikoma (American Institutes for Research), Dr. Xiaying Zheng (American Institutes for Research), Dr. Yifan Bai (American Institutes for Research), Dr. Yuan Zhang (American Institutes for Research), Dr. Markus Broer (American Institutes for Research)

Developing a Socioculturally Responsive Assessment (SRA) is an ongoing quest to advance equity and fairness in educational assessment. The concept of SRA has moved beyond the contexts of culturally sustaining classroom assessments and called for innovative designs of large-scale assessments.

According to Bennet (2022), one working definition of SRA is an assessment that includes problems that connect to the cultural identity, background, and lived experiences of all individuals, especially from traditionally underserved groups. Ongoing discussions that aim to align the key definitions and components of SRA to the contexts of large-scale assessments will certainly shape the future of the largest nationally representative and continuing assessment—the National Assessment of Educational Progress (NAEP).

The current study set out to explore how a NAEP 2013 grade 4 reading passage ("La Ñapa") that is potentially culturally relevant for Hispanic students, is related to the performance of Hispanic students and other student subgroups. Hispanic students' performance in this passage was compared to that of other paired passages via percentile ranks. The invariance of the items for this passage was examined using the Wald test in a two-group confirmatory factor analysis. The findings indicated that (1) Hispanic students performed better overall on the focal text passage compared to other text passages and (2) items associated with the focal text passage slightly favored Hispanic students over non-Hispanic students, while the overall impact on the assessment score was small. The current study contributes to ongoing discussions about the implementation of SRAs in large-scale assessments with empirical data.

Evaluation of the effect of test delivery modes to item difficulties in a high-stake medical test

Friday, 28th July - 09:45: Measurement Applications (Thurgood Marshall) - Oral

Dr. Luc Le (Australian Council for Educational Research), Dr. Van Nguyen (Australian Council for Educational Research)

The Graduate Australian Medical School Admissions Test (GAMSAT) is a cognitive test developed by the Australian Council for Educational Research (ACER) for the Consortium of Graduate-entry Medical Schools. GAMSAT consists of two writing tasks and two multiple-choice (MC) sections: Reasoning in Humanities and Social Sciences, and Reasoning in Biological and Physical Sciences. The test had been administered in paper-based (PB) until 2019. Then it was successfully moved into computer-based (CB) testing in 2020 due to the Covid 19 pandemic. It was important to make sure that the scaled scores from the two delivery modes were comparable. This study used GAMSAT response data from 8930 candidates in March 2019 (PB) and 9636 candidates in March 2021 (CB) to investigate the stability of item difficulties by the two test delivery modes. There were 23 and 28 common items in the two MC sections respectively between the two testing administrations. The Rasch model was implemented to calibrate items in each MC section from each test separately. Differential item functioning (DIF) for the common items was computed as the difference between the item difficulty estimates obtained from the two tests. Initial results showed that the response data from both fitted well to the Rasch model. There was very small variation in item difficulty estimates for the common items between the two tests. Only one item showed large DIF and few items showed modest DIF. Effects of gender and resittingcandidates who did both tests were also examined and discussed.

Examining intelligence assessment in autism, developmental delay, and language impairments

Friday, 28th July - 10:00: Measurement Applications (Thurgood Marshall) - Oral

Dr. Stephanie Northington (Mindful Mentality, LLC)

Background: The most utilized, and widely taught, intelligence measures are the Wechsler scales. While these examine a wide range of abilities, they do not accurately reflect the abilities of neurodiverse individuals. Research (Courchesne et al., 2018) demonstrates traditional ways of assessing intelligence yield lower overall scores in those diagnosed with language impairments, and these individuals may be labeled as "untestable" or misdiagnosed with intellectual impairments (Eagle, 2003; Fillipek et al., 1999). Furthermore, discrepancies noted when giving different intelligence measures, especially to children with autism spectrum disorders (ASD), indicate that there are strengths and weaknesses related to each individual measure that must be considered (Groundhuis et al., 2018; Groundhuis & Mulick, 2013).
Objective: To utilize a nonverbal measure to assess cognitive abilities in comparison to a traditional measure.
Methods: Data were collected in more than 70 individuals ages 3 to 18. All completed a diagnostic assessment; consent and assent were obtained. Measures included a nonverbal measure (Leiter-3) and a traditional measure (age-appropriate Wechsler scale).
Results: One-way between-groups ANOVA results indicated no differences between groups on the nonverbal measure ($p = .138$). However, all performed significantly better on the nonverbal measure, regardless of diagnosis ($p < .001$). Those diagnosed with language impairments (e.g., ASD, developmental delay, language disorders) performed significantly worse than others (e.g., attention deficit disorder, anxiety, etc.) on the traditional measures ($p < .001$).
Discussion: All individuals performed better on the nonverbal measure when compared to the traditional measure. Those with language impairments weren't accurately assessed with the traditional measure.

Abstract Book:
Posters

Comparing different correlation test methods

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Prof. Zhenqiu Lu (University of Georgia), Prof. Kehai Yuan (University of Notre Dame)

The Pearson product moment correlation is the most widely used statistic to explore the relationship between two variables. To test whether the correlation is from a population with a mean of zero, traditionally researchers utilize a parametric procedure that involves Fisher's z transformation of correlation and comparison it with a normal distribution. However, this method assumes that the data are normally distributed and also free of outliers and missing values, as well as having a sufficient sample size. These assumptions may not always be met in practice. Bootstrapping, a resampling technique, can provide more accurate solutions in cases where data are not normally distributed, or where there are outliers, missing values, or small sample sizes. Several bootstrapping methods exist for testing correlations, including bivariate bootstrapping, univariate bootstrapping, and bootstrap hypothesis testing.

In this project, we compare and examine the four methods mentioned above. First, we derive the distribution of the correlation for each method. And then, we use Monte Carlo simulations to verify our conclusions. We also explain the theoretical mechanisms involved in testing correlations and discuss the underlying assumptions of data distribution. Furthermore, we analyze the strengths and weaknesses of each method, identify their similarities and differences, and highlight the connections among them. Finally, we offer recommendations for selecting the most suitable method under different scenarios.

Reliability of assessment of residents' intraoperative performance: Using Generalizability Theory

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Ting Sun (The University of Utah), Dr. Stella Kim (University of North Carolina at Charlotte), Dr. Brigitte Smith (The University of Utah)

System for Improving and Measuring Procedure Learning (SIMPL) was developed to facilitate the assessment of residents' intraoperative performance after each procedure across their general surgery training. The study aims to estimate various sources of error variability in the assessment of residents' intraoperative performance using SIMPL and predict the extent to which this measure can be reliable with varying raters or procedures.
The proposed research is a retrospective observational cohort study. Participants are residents in the general surgery programs. Faculty ratings of trainee's autonomy, performance, and perceptions of case complexity were obtained from SIMPL. Mixed models are used to estimate variance components. Random effects in the models are raters and procedures, and fixed effects are the three items. The study employs the p × t × r design, where p refers to trainees, t refers to the procedures, and r refers to raters. Variance components of the random facet (i.e., procedures) and its corresponding 95% confidence interval will be estimated using ANOVA. Generalizability coefficients will be estimated using the estimated variance components to examine the generalizability of the scale across varying raters and procedures. Data will be analyzed using R package "gtheory" (Moore, 2016).
Providing reliable assessment of surgical trainees' performance is paramount because raters and assignment of procedures are varying. It is critical to guarantee that each rater has the consistent and proper use of the SIMPL tool for varying procedures. This study informs program directors of the use of SIMPL by disentangling various sources of measurement errors.

Estimating IRT parameters with machine learning approaches: A comparison with traditional maximum likelihood methods

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Hongyu Yang (University of Maryland, College Park)

Item Response Theory (IRT) models are commonly used to measure latent traits such as ability or personality. The traditional method for estimating IRT model parameters is the Expectation Maximization algorithm (EM). However, recent studies have shown that machine learning approaches such as Adaptive Boost, Gradient Boost, and Extreme Gradient Boost can provide similar or better results, especially when the assumption of normally distributed person ability is violated. In this study, we will apply these machine learning approaches to a simulated dataset of N = 1000 participants and compare their performance to EM. It is hypothesized that the machine learning approaches perform similarly to EM in most cases and outperform EM in scenarios where the assumption is violated. These findings may have implications for the field of psychometrics, as machine learning approaches may offer a more robust and flexible alternative to traditional methods for estimating IRT parameters.

Influence of topic-word matrix misspecification on semi-confirmatory Latent Dirichlet Allocation

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Jordan Wheeler (University of Georgia)

Topic models are statistical methods used to analyze textual data. The objective of topic models is to estimate clusters of words, referred to as topics, from a set of related text. The use of topic models in psychological and educational measurement is relatively new and is often used as an exploratory approach to analyze constructed-response items. Recently, a confirmatory approach known as a semi-confirmatory Latent Dirichlet Allocation (scLDA; Wheeler et al, 2021) was developed to allow researchers to specify topics, prior to estimation, through a word-topic loading matrix. A potential limitation of scLDA is that the word-topic loading matrix may be misspecified. That is, not all words for a topic will be correctly identified apriori. The degree to which misspecification of the topic-word matrix influences parameter recovery has yet been reported. Therefore, in this study, we investigated the effects of misspecification of the word-topic loading matrix on parameter recovery for the scLDA model through a simulation study. The simulation design manipulated five factors and crossed all levels for a total of 891 conditions. The five factors were: number of unique words (3 levels: 250, 500, and 750 words), average response length (3 levels: 25, 50, and 100 words per document), number of documents (3 levels: 100, 250, and 500), the number of topics (3 levels: 3, 4, and 5), and the misspecification rate of the word-topic loading matrix (11 levels: 0%, 10%, ...,100%). The results of the simulation revealed that various rates of misspecification impact the recovery of parameters differently.

Exploring the effect of parceling strategies on measurement invariance testing

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Chunhua Cao (The University of Alabama), Dr. Xinya Liang (Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas)

There has been a long-standing debate on the use of item parceling in structural equation modeling (SEM). When the number of items is large, item parceling has been utilized in both single-group and multigroup SEMs. Research has showed that parceling tended to improve model fit and reduce the bias in structural parameter estimates in SEM. However, for measurement invariance testing, Meade and Kroustalis (2006) showed that the use of item parceling instead of individual items decreased the power of detecting measurement noninvariance using the log likelihood ratio test (LRT). They fixed the number of items per parcel at four and used only LRT as the fit index. The current study purported to examine the effect of parceling strategies, including isolated parceling (put the invariant items or noninvariant items in the same parcels) and distribution parceling (mix the invariant and noninvariant items in the same parcel) with different numbers of items per parcel on measurement invariance testing using fit measures of LRT, comparative fit index (CFI), root mean square of error of approximation (RMSEA), AIC, BIC, and sample size adjusted BIC (SaBIC). The population model was a two-factor model with 24 items per factor. The design factors included the location of measurement noninvariance, the percentage and magnitude of measurement noninvariance, sample size, parceling strategies, and number of items per parcel. The performance of each fit measure under various parceling strategies will be discussed.

Detecting gender DIF in mixed-format assessment

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Xiuyuan Zhang (College Board)

Large-scale assessments often include both multiple choice (MC) items and constructed response (CR) items to improve test reliability. Differential Item Functioning (DIF) is routinely calculated for MC items, but not very often for CR items. To ensure the fairness of all items, our test program evaluated a few methods in detecting gender DIF for a mixture of MC and CR items. These DIF detection approaches included: 1) the extended version of SIBTEST for polytomous items, or the PolySIB procedure, 2) logistic regression, and 3) Rasch model or FACETS. MC items and CR times from a large-scale English assessment were analyzed using these three DIF methods. Given that DIF estimation could be influenced by the use of different matching criterion variables, three matching scores were examined in this study, including the section score for the particular item type (i.e. MC score for MC items and CR score for CR items), MC and CR total score, and the total score excluding the item itself. The results indicated that the three DIF detection methods yielded slightly different patterns on gender DIF for MC and CR items. Both empirical and logistic inferences will be discussed, such as what criterion variable is the most appropriate, minimum sample size requirements to produce reliable DIF estimation, the ease of application in operational time window, and the potential correlations between item content characteristics and gender DIF.

Relationship among measurement invariance, differential item functioning and mean comparison

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Bo Zhang (University of Wisconsin - Milwaukee)

When different groups are compared on a latent trait, it must be assumed the groups have been measured in the same way, or measurement invariance (MI) of group. This assumption can be violated by confounding background variables. Previous research has identified different types of MI violation (Vandenberg & Lance, 2000) and has proposed procedures to detect violations (Van de Schoot, Lugtig, & Hox, 2012).

For practitioners, an important question is what to do if MI is violated. Can groups still be compared? If not, is there any way their study can still be salvaged? This study aims to address these questions by examining the relationship among MI violation, mean comparison, and differential item functioning.

Research Design

MI violation will be simulated incrementally to identify two situations. First, MI is violated but mean comparison is robust. The goal will be to identify possible indicators for this benign violation. Second, MI is violated and mean comparison is not valid. Differential item functioning (DIF) will be conducted to identify items that may have caused the violation. Those items will then be removed to test whether MI can be restored, or mean comparison can be valid.

Simulation design will examine: type of violations (metric and scalar), level of violations (5 levels from .3 to 1.5 for item discrimination or difficulty change), test length (10 and 20), group difference (small, medium, and large). Tests will have two factors, simple structure, and likert items.

Results

Computer program is ready. Simulation can be completed in two weeks.

Testing CDM local independence assumptions using nested model selection criteria

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Athul Sudheesh (University of Texas at Dallas), Dr. Richard M. Golden (University of Texas at Dallas)

In the standard widely used Cognitive Daignostic Model (CDM), the probability an examinee correctly answers a question is determined by a subset of binary latent skills specified using a variable called the Q matrix. Thus, the binary response answers to pairs of questions are expected to be correlated if they are dependent upon the same latent skills and uncorrelated otherwise. Only a few studies have investigated the performance of testing such conditional independence assumptions in CDMs and those studies have used the chi-squared test of statistical independence (e.g., Lim and Drasgow, 2019). In this paper, we propose a method for checking conditional independence assumptions using model selection criteria such as AIC and BIC for comparing a two parameter logistic regression model that predicts the response to one question given the response to another question with a competing one parameter intercept only logistic regression model. More specifically, we generated simulated response data from CDM with a known Q matrix and tested independence assumptions using both the widely used chi-squared and less widely used nested logistic regression model decision rule. Using the simulated response data and the area under the ROC curve as a discrimination performance measure, we found that the nested logistic regression model showed better discrimination performance than the more widely used chi-squared independence test decision rule and this advantage was pronounced for smaller sample sizes. Finally, the relevance of these methods are discussed for testing CDM conditional independence modeling assumptions without making strong parametric modeling assumptions.

Impacts of item discrimination parameters on the uniform DIF

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Gamze Kartal (University of Illinois Urbana-Champaign), Prof. Jinming Zhang (University of Illinois Urbana-Champaign)

In the literature, DIF magnitude for the uniform DIF is generally introduced through the difficulty parameter differences between groups. Though, in this study, it is suggested to consider the impacts of the discrimination parameters. Thus, to understand the α-parameters' influence on the DIF magnitude, it is recommended to create the item characteristic curves (ICCs) of DIF items with different levels of item discrimination parameters and different levels of item difficulty parameter differences between groups. DIF magnitudes, small (b_F - b_R = 0.3), medium (b_F - b_R = 0.6), and large (b_F - b_R = 0.9), which are frequently used in previous studies, are preferred as item difficulty parameter differences for the uniform DIF. Four levels of item discrimination parameters will be chosen to examine whether α-parameters would change the area between the ICCs. Then, based on the differences, it will be recommended to introduce the uniform DIF through both α- and b-parameters.

Rapid online assessment of reading ability with computer adaptive testing

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Wanjing Ma (Graduate School of Education, Stanford University), Dr. Adam Richie-Halford (Stanford University), Dr. Amy Burkhardt (Stanford University), Mr. Klint Kanopka (Graduate School of Education, Stanford University), Ms. Clementine Chou (Stanford University), Prof. Benjamin Domingue (Graduate School of Education, Stanford University), Prof. Jason Yeatman (Graduate School of Education, Stanford University)

The Rapid Online Assessment of Reading Single Word Recognition (ROAR-SWR) is a web-based, non-adaptive lexical decision task that measures reading abilities without a proctor. Here we study whether item response theory (IRT) and computer adaptive testing (CAT) can be used to provide a more efficient measure. We first analyzed parameter invariance across four groups of students (N = 1,572) who differed in age, socioeconomic status, and language-based learning disabilities. The majority of item parameters were consistent across groups and the 6 biased items were removed. Next, we implemented a JavaScript CAT algorithm that is compatible with jsPsych, and conducted a validation experiment with 495 students in grades 1-8 who were randomly assigned to either a random-order or CAT version of the test. We found that CAT improved test efficiency by 40%: 100 CAT items were selected to produce the same standard error of measurement as 165 items in a random order. We then created a new version of ROAR-CAT that interleaves adaptive items and new items. By comparing scores from ROAR-CAT and an individually administered assessment in 32 first and second-grade public school classrooms, we found that ROAR-CAT had exceptional sensitivity (91%) and accuracy (89%) for identifying students with reading difficulties. Our findings suggest that the ROAR-CAT is a promising tool for efficiently and accurately measuring reading ability. Furthermore, our development process serves as a paradigm for creating other adaptive online assessments for schools.

Bayesian IRT estimation by using Stan and Jags

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Selim Havan (University of Illinois Urbana-Champaign), Ms. Gamze Kartal (University of Illinois Urbana-Champaign), Mr. Onur Demirkaya (Riverside Insights)

Item response theory (IRT) is a paradigm for examining the relationship between an examinee's answer to a particular test item and used instruments to score individuals on their abilities or latent traits. Although several R packages that implement IRT models have been developed, using Bayesian methods is limited. Stan and Jags are Bayesian statistical algorithms and supporting languages that implement the Hamiltonian Monte Carlo (HMC) with a no-U-turn sampler and Gibbs sampling implementation of *the Markov Chain Monte Carlo* (MCMC) method, respectively. Due to HMC's more effective exploration of the posterior parameter space, it is considered quicker than the Gibbs sampling; however, JAGS is easier to use. Since both algorithms have their advantages and disadvantages, this study will compare unidimensional and multidimensional IRT models with Stan and Jags using R software.

Bayesian model evaluation and local identifiability for growth mixture models

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Xingyao Xiao (University of California, Berkeley), Dr. Sophia Rabe-Hesketh (University of California, Berkeley)

Growth mixture models (GMMs) are useful in modeling heterogeneity in longitudinal data. However, GMMs can suffer from local likelihood identifiability (Dawid, A. P., 1979) issues in certain situations. Estimating GMMs using Markov chain Monte Carlo methods such as Hamiltonian Monte Carlo (HMC) can encounter difficulties in these cases. We provide illustrations of these issues and show why HMC struggles to explore the parameter space. We recommend using informative priors and tuning HMC parameters to prevent local likelihood non-identifiability.

Specifying Bayesian GMMs involves choosing from different versions of the likelihood, either conditional on the latent variables or marginal with respect to the latent variables. In software like Stan that do not allow sampling of discrete parameters, the marginal likelihood, with respect to discrete and continuous latent variables, is the only meaningful option if the posterior predictive distribution is of interest, as in Bayesian model comparison. We discuss marginal Watanabe–Akaike information criterion (WAIC) and Leave-One-Out Cross-Validation (LOO-CV) based in this marginal likelihood and demonstrate the use of this approach with simulated data and a real world data example.

Development of a real-time treatment recommendation system for mental disorders using generalized structured component analysis and Bayesian networks

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Gyeongcheol Cho (Department of psychology, McGill University), Dr. Younyoung Choi (Department of psychology, Ajou University)

Wearable medical devices offer a unique opportunity to collect personalized data from patients, enabling the generation of real-time, tailored treatment recommendations based on this data. Bayesian networks have been demonstrated as an effective statistical tool for this purpose. They can estimate the probability of patients having a specific disease or responding to a particular treatment based on their individual metrics and update these probabilities as new information becomes available, thus facilitating personalized treatment decisions.
However, building a network model that incorporates psychological constructs in the context of mental health presents a challenge for Bayesian networks. As these constructs are not directly measurable, multiple observed variables must be incorporated into the model as indicators of the constructs, resulting in a high-dimensional and less interpretable model.
To overcome this challenge, we propose a two-step procedure for constructing a personalized treatment recommendation system for mental disorders using generalized structured component analysis and Bayesian networks. In step one, generalized structured component analysis extracts low-dimensional and interpretable representations for psychological constructs. In step two, Bayesian networks employ those statistical representations to build its network model that recommends specific treatments for the mental disorder of interest. We demonstrate the effectiveness of our two-step procedure by developing a tool for prescribing cognitive behavioral therapy for insomnia (CBT-I).

Potential for action sequence network characteristics in predicting item performance

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Ni Bei (University of washington), Dr. Elizabeth Sanders (University of washington)

Increasingly, there has been interest in how "process data" might provide insights about how respondents' thinking informs item responses. Unfortunately, there are few available datasets that consistently collect more than response time on items. The present study thus draws on a unique dataset derived from a university-developed online mathematics game, *Riddle Books*. In these data, **students' actions** (e.g., dragging pieces of information to build bar models) are tracked in real time. Specifically, we propose to present a 2-stage approach for **evaluating how action sequence network information can be used to predict student responses** with the *Riddle Books* data. Using **first attempts on four items for N = 18 fourth graders**, in the first stage, we transformed students' action sequences into binary, directed action-by-action networks to calculate network-level statistics for each student, by item. In the second stage, we incorporated item-student action network characteristics as predictors in a multilevel logistic regression model (items within students) to predict item responses. Our results show that, although response time and ability were initially predictive of the likelihood of an incorrect item response, only network density (i.e., how many actions connected together) was uniquely predictive of item response ($Coeff$ = -1.70, $p < .05$). In other words, increased connections among actions were associated with a 15.45% decrease in the probability of answering an item correctly, controlling for student ability and reaction time. Future work includes using this method for predicting "engagement" and "persistence" using students' repeated and total attempts on items across the game.

Consequences of data leakage on reproducibility in machine-learning-based psychometrics

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Susu Zhang (University of Illinois Urbana-Champaign)

The growing popularity of machine learning methods (e.g., neural nets, high-dimensional regression models, and tree-based methods) in psychometrics is evidenced by their increasing application to measurement tasks, for instance, scale construction to optimize criterion-related validity, detection of aberrance such as insufficient effort responding and cheating, and automated scoring based on constructed responses or process data. Responsible conduct of measurement research and practice, which emphasizes score reliability, validity, and generalizability, preordains scrutiny of whether scores and conclusions derived from machine learning-based methods are reproducible and replicable. One common pitfall that jeopardizes reproducibility in machine learning is data leakage, when information from the test set, intended for objective evaluation of model performance on unseen data, is unintentionally included during the model's training process, for example, by performing preprocessing (e.g., oversampling on imbalanced data, missing data imputation) or feature extraction/selection before the training-testing data split. Leakage was found to consistently yield overly optimistic evaluation of model performance on the test set across many natural science, biomedical, and engineering disciplines (Kapoor & Narayanan, 2022), but the consequences of different types of data leakage in behavioral research are less studied. This study considers two common measurement tasks, namely scale construction for predicting a criterion variable and automated scoring based on process data, and evaluates the consequences of common types of leakage (i.e., missing response imputation, oversampling for imbalanced classification, measurement model parameter estimation, and feature extraction/selection before training-testing split) on the evaluation of test set performance. Analyses are based on both simulations and empirical examples.

Evaluating two-step estimation approaches for a multigroup APIM

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Emma Somer (McGill University), Dr. Carl Falk (McGill University), Dr. Milica Miočević (McGill University)

The Actor-Partner Interdependence Model (APIM) is a popular model used across the social sciences to model dyadic relationships. Structural equation modeling (SEM) can be used to investigate the relationships between latent variables in an APIM framework, thus allowing researchers to model individuals' influence on themselves and their partners. Recently, the structural-after-measurement (SAM; Rosseel & Loh, 2022) approach has been proposed as an alternative to SEM to alleviate some of the issues associated with model misspecification and small samples for maximum likelihood estimation. However, the performance of SAM and factor score regression (FSR), another two-step approach, for the APIM is unknown. We conduct simulation studies to evaluate the performance of SAM and the bias avoiding method for a multigroup APIM in small samples, and we compare the methods to SEM using maximum likelihood. In particular, we examine the influence of partial measurement invariance, reliability of the items, number of indicators, and measurement model misspecification on the bias, efficiency, Type I error, power, and coverage of the path coefficients for an APIM in the presence of correlated residuals. We implement a novel identification strategy for SAM in multigroup models, and we propose extracting standard errors and confidence intervals under the SAM framework using bootstrapping. Preliminary findings suggest SAM outperforms the bias avoiding method and SEM in terms of coverage and Type I error rates, particularly under low reliability conditions.

Measuring emotional intelligence unobtrusively and objectively: An eye-tracking and machine learning approach

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Wei Wang (The Graduate Center, CUNY), Ms. Liat Kofler (The Graduate Center, CUNY), Mr. Chapman Lindgren (The Graduate Center, CUNY), Mr. Qiwen Tong (The Graduate Center, CUNY), Ms. Amanda Murphy (The Graduate Center, CUNY)

Measuring non-cognitive individual differences—such as personality, trait-based emotional intelligence, vocational interests—has been long overwhelmingly relying on self-reports. The trait emotional intelligence plays a critical role from frontline services to executive leadership in the workplace. It predicts job performance beyond general mental ability and Five-Factor personality traits. Yet the measurement of emotional intelligence has been long critiqued for serious unresolved issues, including unreliability and bias inherited from the self-report nature. Leveraging psychophysiology and machine learning models, the current study examined a novel approach to unobtrusively and objectively measuring emotional intelligence. Specifically, we exposed 122 participants to images of 1) four emotional faces (neutral, happy, anger, and fear; randomly arranged), and 2) twelve face-crowds with varying ratios of happy-to-angry faces. We recorded participants' eye movements with a high-end eye-tracker and processed the eye-tracking data with the gazepath technology to extract hundreds of eye movement features, which were then fed into machine learning models to predict the emotional intelligence scores. Our results showed that this approach was able to achieve relatively high predictive accuracy. In addition, we found this approach was particularly powerful for measuring two facets of emotional intelligence: self-emotion appraisal and other-emotion appraisal.

Exploratory measurement modeling with Lasso: The role of measurement quality

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Youngwon Kim (University of washington), Dr. Elizabeth Sanders (University of washington)

Exploratory approaches for measurement modeling may be useful with large-scale survey and assessment data for which researchers have little theory to guide model selection. For example, the least absolute shrinkage and selection operator (Lasso) and adaptive Lasso (aLasso) algorithms have been successfully applied within the IRT framework for differential item functioning detection (e.g., Wang, Zhu, & Xu, 2022), as well as within the SEM framework for determining an optimal number of factors and data reduction (e.g., Huang, Chen & Wen, 2017; Jacobucci, Grimm & McArdle, 2016). Despite their potential usefulness, researchers may justifiably be concerned about obtaining non-generalizable sample-specific model results, particularly in circumstances where only cross-sectional data are available. Using Monte Carlo simulation, we investigate the accuracy of *regsem* Lasso and aLasso algorithms in fitting a 3-factor model with three levels of measurement quality (item-factor loadings = .4, .6, and .8). Our results show that aLasso outperformed Lasso, but that both algorithms can result in biased loadings for factors with weak measurement. Future work will expand the comparison to other software and algorithms.

Utilizing cluster covariates to estimate treatment heterogeneity in multilevel data

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Graham Buhrman (University of Wisconsin-Madison), Ms. Xiangyi Liao (University of Wisconsin-Madison), Prof. Jee-Seon Kim (University of Wisconsin-Madison)

One aim of educational research is to evaluate interventions developed to improve student learning and behavioral outcomes. Estimating an intervention's treatment effect is one way to evaluate its efficacy. This treatment effect is not always constant, and heterogeneity arises when not all students respond to interventions similarly, particularly between subgroups with varying characteristics. Sample differences in sociodemographic features or individual covariates can identify key subgroups, which can be used to estimate heterogeneity via the conditional average treatment effect (CATE) for individuals with those characteristics. Multilevel data is a common occurrence in educational research, and one feature of this data structure is the distinction between individual and cluster-level covariates. Cluster-level covariates also indicate important subgroups and may influence the treatment effect differently than individual covariates. However, many contemporary procedures that flexibly estimate CATE do not give special consideration to these cluster-level covariates and may be limited in their ability to account for noteworthy heterogeneity due to clustering. This study compares various methods for estimating CATE that 1) take different approaches to handling data clustering and 2) utilize either single- or meta-learner frameworks that allow for different degrees of flexibility when estimating individual treatment effects. We apply these methods to the synthetic data generated for the 2018 Atlantic Causal Inference Conference (ACIC), simulated, and large-scale assessment data to evaluate each method's accuracy, precision, and interpretability. We discuss how to more effectively utilize cluster-level covariates in the non-parametric estimation of CATE and directions for future research.

The effect of model size on fit indices in ordinal factor analysis models

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Yunhang Yin (University of South Carolina), Dr. Dexin Shi (University of South Carolina), Dr. Amanda Fairchild (University of South Carolina)

This study investigated the effect of the number of observed variables (*p*) on SEM fit indices in the context of ordinal data. The behaviors of three fit indices including the comparative fit index (CFI), the Tucker–Lewis index (TLI), and the root mean square error of approximation (RMSEA) were examined under ordinal factor analysis models with the unweighted least squares (ULS) estimation method. We manipulated various simulation conditions by varying the number of observed variables, levels of model misspecification, sample sizes, number of response categories, and magnitude of factor loadings. The results showed that given the same level of model misspecification, the population values of CFI, TLI, and RMSEA were generally stable as the model size (*p*) increased. At the sample level, as *p* increased, the sample CFI estimates became closer to their corresponding population values. The effect of *p* was more pronounced when the sample size was small, and/or the magnitude of factor loadings was low. No clear pattern was observed regarding the effect of *p* on the sample TLI and RMSEA. Drawing on our findings, we discussed the practical implications of our study and suggested directions for future research.

Bayesian variance component priors for small-sample multilevel models

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Liu Liu (University of washington), Dr. Elizabeth Sanders (University of washington)

Estimating variance parameters accurately in multilevel modeling (MLM) is a well-known problem when the number of units at higher levels is small (e.g., when there are many items but few students assessed). Compared with frequentist maximum likelihood, Bayesian analyses using Markov Chain Monte Carlo offer a promising approach for estimating MLMs with small higher-level sample sizes. However, the priors for variance parameters are more complicated and less intuitive than priors for coefficients, and there is yet to be consistent advice for choosing MLM variance component priors (e.g., Baldwin & Fellingham, 2013; Browne & Draper, 2006; Gelman, 2006; McNeish, 2016). Typical recommendations for prior choice include inverse-gamma, uniform, half-Cauchy, half-t (as a special case of half-*t*), inverse-Wishart, and Gaussian distributions (Baek et al., 2020; Gelman, 2006; McNeish, 2016; Moeyaert et al., 2017). Because variance component estimates play a crucial role in the standard errors for model regression coefficients, we propose to present results from a Monte Carlo simulation of a 2-level logistic MLM analyzed with different Bayesian priors, including uninformative and informative (correct vs. incorrect) to demonstrate how the choice of priors affects parameter inferences for varied level 2 sample sizes. The results can help researchers make more informed choices about their variance component priors in MLMs with small samples.

Investigating pre-knowledge and speed effects in an IRTree modeling framework

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Hahyeong Kim (University of Illinois Urbana-Champaign), Dr. Justin Kern (University of Illinois Urbana-Champaign)

There is a growing psychometric literature focusing on pre-knowledge. Much of this work has used data where prior exposure to items for persons in unknown. A better understanding of the effects of pre-knowledge can help psychometricians overcome the problems pre-knowledge can cause. This research aims to explore the effects of pre-knowledge with experimentally obtained data. To collect these data, we ran an online experiment manipulating pre-knowledge levels by exposing a varying number of compromised items to participants in practice sessions prior to test administration.

One growing modeling paradigm is using so-called IRTree models to embed cognitive theories for responding to items into the model. One such IRTree examined the role of speed on intelligence tests differentiating fast and slow test-taking processes (DiTrapani et al., 2016). To investigate this, they used a two-level IRTree model with the first level controlled by speed and the second level controlled by an ability trait. This allows for separate parameters at the second level depending upon whether the responses were fast or slow.

Building on this literature, we are interested in determining whether and how item pre-knowledge impacts item properties. In this paper, the effects to be studied include 1) whether pre-knowledge impacts the first-level IRTree parameters, affecting response time; 2) whether pre-knowledge impacts the second-level IRTree parameters, affecting response accuracy; and 3) whether the first-level response (i.e., fast or slow) impacts the second-level IRTree parameters. In all cases, an interesting sub-question to is whether any of these effects are constant across items.

A systematic review of Cognitive diagnosis modeling applications: Insights into future applications and areas for Improvement

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Xiyu Wang (Purdue University, West Lafayette), Prof. Yukiko Maeda (Purdue University, West Lafayette), Ms. Xiuxiu Tang (Purdue University, West Lafayette), Mr. Yuxiao Zhang (Purdue University, West Lafayette), Prof. Hua Hua Chang (Purdue University, West Lafayette)

Cognitive diagnosis modeling (CDM) has been a subject of substantive interest for the past decade. Compared to traditional measurement methods, CDM enables test organizers to show the mastery level of each attribute rather than providing an overall ability estimate. Due to this feature, the application of CDM will provide practical and actionable implications in student learning, including identifying unmastered skills, facilitating targeted instructions, and promoting individualized learning. Various CDMs, including DINA, DINO, NIDA, NIDO, and the rule space method, have been developed with different assumptions regarding how skills interact as well as where and how measurement errors occur.

Despite the rigor of methodological advances and its potential practical benefits, the applications of CDM at practical setting seems to be limited. Therefore, this systematic review will aim to provide an overview of the existing literature on current applications of CDM in education settings since its introduction in 1980s and to provide guidance for and insight into promoting future applications of CDM. We will also identify areas for improvement, and discuss the challenges associated with the wide implementation of CDMs and potential future directions. For this purpose, we will summarize methodological characteristics of the applied studies and exclude studies that mainly rely on simulation results and include real data only for the purpose of providing an example for applying a newly proposed model. The key characteristics we summarized will be the chosen CDM model, model fit, measured constructs, attribute name and number, Q matrix specification accuracy, sample size, reliability, and validity.

Test analysis method using piecewise linear ICCs

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Prof. Gen Hori (Asia University)

Classical test theory and item response theory each have their advantages, but because item response theory analysis requires a large number of examinees and specialized software, classical test theory analysis is often used in educational settings. On the other hand, classical test theory analysis is dependent on the test taker population and questions, and therefore, it is not possible to calculate scores based on a unified standard for tests with different test takers. In this study, we develop a test analysis method that combines the advantages of classical test theory and item response theory. Specifically, the authors will develop a test analysis method that approximates the item characteristic curve with the question correct response analysis chart proposed by the authors, and performs scoring and equating based on this curve. A method will be developed that combines the advantages of classical test theory (applicable even when the number of examinees is small, no special software is required) with the advantages of item response theory (scores can be calculated based on a unified standard independent of the examinee population and questions).

Correspondent mappings between psychological network models and latent factor models

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Chi-Yun Deng (National Chengchi University), Dr. Hsiu-Ting Yu (National Chengchi University)

There are two common approaches to modeling data in social science researches: latent variable model (LVM) and psychological network model (PNM). Typical LVM assumed observations reflect latent variables, observations are statistically independent of each other when condition on latent variables; whereas network models assumed direct relations exist between observations, so relationships among observations are reflected by a network structure. Past researches have investigated relationships between LVM and PNM by algebraic approach, simulation studies and empirical data. Studies have demonstrated the statistical equivalence between LVM and PNM; the algebraical relationships between the two models also have been shown. Studies in social science usually based on the viewpoint of LVM in data analysis and interpretation. This study investigates how data generated from PNM would be reflected by LVM. Data from PNM are generated with various scenarios. Factors under investigations are number of nodes, sample sizes, level of sparseness, and distinct network structures (i.e., specific adjacent block diagonal matrices). Generated data are then fitted by LVM with various dimensionality. The correspondences between PNM and LVM are examined and compared systematically. Their estimated parameters in both models are also evaluated and investigated. This study aims to explore the relationships between PNM and LVM under complex network structure to gain more insights in the relationships between the two modeling approaches.
Key words: latent variable model, network model, latent structure, simulation study

Empirical comparisons among models in detecting extreme response style (ERS)

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Hui-Fang Chen (City University of Hong Kong), Mr. Jianheng Huang (City University of Hong Kong)

Models within the framework of item response theory (IRT) models have been proposed to identify (ERS), and they can be categorized into three approaches. The first approach treats ERS as an explicitly additional dimension, other than the intended-to-be-measured latent ability, to influence item responses (e.g., the multidimensional nominal response model for response styles, MNRM). The second approach recognizes that ERS could lead to varied space between response categories individually, and a weighting parameter for the thresholds of each item category is set to estimate individual participant's ERS (e.g., the modified generalized partial credit model for ERS, the ERS-RPCM). The third approach incorporates a decision model into the IRT and uses a dichotomous indicator to indicate the presentence of ERS among participants (e.g., the tree model with dominance models or the unfolding tree model, the UD tree).

To facilitate the practical use of these approaches, the present study compared the performance of these approaches against generalized partial credit model using empirical datasets. Findings suggested that all approaches yielded similar latent ability estimates, with correlations ranging between .70 and .99, in which the UD tree yielded the weakest relationship with the latent ability estimates from other approaches (but still at moderate to high levels). Regarding ERS estimates, the absolute values of correlations between different models ranged between .61 and .88. The ERS estimate from the ERS-GPCM were negatively related to the ones from the other approaches. The UD tree yielded the highest reliability in ERS estimates.

Validity evidence for an ECE classroom observation tool

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Elaine Ding (World Bank), Dr. Adelle Pushparatnam (World Bank), Mr. Jonathan Seiden (Harvard University), Ms. Estefania Avenado Nino (World Bank), Dr. Ezequiel Molina (World Bank), Dr. Marie-Helene Cloutier (World Bank), Dr. Diego Luna Bazaldua (World Bank)

Governments have invested in increasing access to early childhood education (ECE), with global enrollment rates in ECE nearly doubling in the past 20 years. However, increased access is not always accompanied by parallel improvements in the quality of ECE services. We present psychometric results from a validity study on an ECE classroom observation tool used worldwide in diverse countries and contexts. The scores produced by the observation tool do not show ceiling or floor effects and can inform about the strengths and areas of opportunity to support ECE teachers in strengthening their teaching skills. All scores show positive correlations, meaning that improved teaching skills in one teaching area may result in further improvements in others. In addition, the observation tool empirically reproduces theoretical constructs using confirmatory factor analysis techniques and multiple plausible models were compared using model fit statistics. The factor analysis results are further corroborated using nonmetric multidimensional scaling techniques. The results show that this ECE classroom observation tool can provide actionable information to monitor ECE quality at the system level to inform policymakers about the status of service delivery. At the same time, teachers and other school stakeholders can use formatively use these scores for the continuous improvement of their classroom practice. Areas of future work include the use of advanced multilevel techniques to identify the impact of teaching quality on student learning outcomes.

Prior sensitivity of Bayesian SEM fit indices to model misspecification

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Ejike Edeh (University of Arkansas), Dr. Xinya Liang (University of Arkansas), Dr. Chunhua Cao (The University of Alabama)

The development of procedures for assessment of global model fit for Bayesian structural equation modeling (BSEM) is an emerging issue. Several Bayesian analogues to common frequentist fit indices have been developed, including Bayesian root mean square error of approximation (BRMSEA), Bayesian comparative fit index (BCFI), and Bayesian Tucker-Lewis index (BTLI). These BSEM fit indices are computed in terms of the posterior predictive model checking, in which a discrepancy function and degrees of freedom are calculated at each iteration of the Markov chain. The literature has shown that the interaction of sample size and prior specification had a large effect on BRMSEA, BCFI, and BTLI. With large sample size (n = 1000), BRMSEA, BCFI, and BTLI tended to be insensitive to prior specification, except when strongly informative but inaccurate prior was specified. However, research was limited in terms of examining the model complexity and type of misspecification.
This paper systematically evaluates the sensitivity of BSEM fit indices to model misspecifications with varying model sizes, prior choices, and sample sizes through a simulation study. Prior sensitivity of these BSEM fit indices was investigated using CFA models consisting of three or six factors measured by five or ten items, with different degrees of misspecification of factor covariance matrices and item cross loadings. Five prior specifications were investigated with varying levels of accuracy and informativeness by controlling the prior mean and variance. Practical implications for the choice of Bayesian fit indices will be discussed.

Bayesian generalized method of moments approach for estimating rank preserving models: A flexible approach for causal mediation analysis

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. ROBERTO FALEH (University of Tübingen), Prof. Holger Brandt (University of Tübingen)

Mediation analysis is a fundamental tool in empirical sciences, particularly in medical and social sciences, where intermediate variables play a crucial role in understanding treatment efficacy. The classical regression-based mediation analysis method proposed by Baron and Kenny has been criticized for its restrictive assumption of no-unmeasured-confounder, which requires that all confounders have been measured and incorporated into the model. This assumption can be difficult to satisfy in practice and violations lead to spurious results and make causal interpretation challenging.

One of the most prominent models relaxing the assumption of no-unmeasured-confounder is the Rank Preserving Model (RPM), introduced by Ten Have and colleagues. The RPM assumes that unobserved confounders do not interact with treatment or mediators. This assumption is often more plausible than the no-unmeasured-confounders assumption making the model relevant in less confining theoretical and empirical circumstances. To further generalize the model and weaken the assumptions required by the RPM, Zheng and colleagues proposed a more flexible model that can handle multiple mediators and multilevel interventions.

However, models using the RPM assumption have not been used extensively due to low power and inefficiency in many scenarios. The Bayesian Generalized Method of Moments (GMM) is proposed as a solution to improve the power and flexibility of the RPM.

The Bayesian Generalized Method of Moments has several advantages over classic frequentist approaches, including the possibility to directly derive standard errors for estimates as well as higher power, robust, and unbiased estimation of the model's parameters.

Evaluating math language intervention with non-parametric tests

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Menghe Xu (Beijing Normal University)

Formal mathematical language is essential for the dissemination of mathematical knowledge, yet many students struggle to write in this style during their learning process. Despite significant attention from education researchers towards promoting students' mathematical language expression, current high school students still exhibit weaknesses in this area, and practical interventions with strict assessment are rare.

This study aimed to evaluate the effectiveness of a teaching intervention on formal math language writing skills. We sampled 30 senior-grade students from two classes in a high school in Xian, Shanxi, China, and collected pre-intervention data in the form of scores on the mid-term mathematical test. We identified four types of problems in writing formal math language, including imbalances in specification and summary, incorrect logic, nonstandard language, and messy typesetting.

We proposed teaching intervention strategies and implemented them with the students for 3 months. Post-intervention data were collected in the form of scores on a grade-level mathematical exam. The results of descriptive and significance tests showed a significant improvement in students' formal math language writing skills following the intervention.

Applying tree-based models in social and behavior sciences

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Eunbi Sim (University of Georgia), Prof. Caleb Han (University of Georgia), Prof. Zhenqiu Lu (University of Georgia)

Techniques in data science, strategic decision-making, and quantitative analytics for synthesizing complex information among individuals, groups, and organizations have become increasingly popular in social and behavioral sciences. Parametric models are often employed to understand these dynamic interactions and predict outcomes. However, they face many challenges in a fast-changing, data-driven world, such as linearity, homoscedasticity, and normality, as well as multi-level structure and multi-contexts.

To address these challenges, this paper introduces non-parametric methods for classification and prediction that are fast, accurate, simple, and visual. First, the paper introduces concepts such as non-parametric statistics, machine learning, and ensemble learning. Terms such as nodes and partitioning used in modeling are explained. And then, the paper presents tree-based models (TBMs), including Decision Trees (DT) and Random Forests (RF). It provides detailed methodological techniques for conducting DT and RF in computer software, highlighting specific algorithms for models of DT and RF. The paper also compares different tree-based models. For example, DT is an open-box model but not robust, RF, built from a number of DT models, is more accurate but less interpretability. With non-parametric, supervised, machine learning approaches, TBMs are immune to multi-collinearity, missing values, and outliers. Additionally, they do not require assumptions for traditional parametric models, and significantly improve data pre-processing. Applications of the tree-based models are illustrated through a real data analysis of multi-level and multi-contexts data. Practical guidelines for researchers and professionals are also provided.

A hierarchical prior for Bayesian variable selection in regression model

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Anqi Li (University of Illinois Urbana-Champaign), Prof. Steven Culpepper (University of Illinois Urbana-Champaign)

Selecting subsets of variables has always been a vital and challenging topic in educational and psychological settings. In many cases, the probability that an interaction is active is influenced by whether the related variables are active. In this study, we propose a hierarchical prior with Bayesian approaches by using the deterministic inputs, noisy "and" gate model to account for a structural relationship between variables and their interactions. Specifically, an interaction is more likely to be active when all the associated variables are active and is more likely to be inactive when at least one associated variable is inactive. The proposed hierarchical prior is implemented in the widely used Bayesian variable selection approaches including stochastic search variable selection (George & McCulloch, 1993) and Dirac spike and slab priors (Mitchell & Beauchamp, 1988). Metropolis-Hasting algorithm is used to uncover the selected variables and estimate the coefficients. Simulation studies based on a real data example are conducted under different conditions and the performance of the proposed hierarchical prior is compared with the standard independent prior as well as the mixture of g-prior (Liang, Paulo, Molina, Clyde, & Berge, 2008).

Meta-analysis of fMRI studies related to mathematical creativity

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Zehua Cui (University of Maryland, College Park), Prof. SUNGYEUN KIM (Incheon National University)

Creativity is emphasized in the curriculum as a core competency that must be developed in school education to survive in the future society (OECD, 2018; UNESCO, 2015; WEF, 2016). Therefore, this study aims to suggest educational implications in the mathematics subject by identifying brain regions related to mathematical creativity through meta-analyses of fMRI data. As a first step to achieving this goal, preliminary meta-analyses were conducted through Neuroquery. Two meta-analytic brain maps based on search terms “mathematical” and “creativity” were downloaded, respectively, and subsequently viewed and analyzed using the bspmview and MRIcron software. Peaks in brain activities related to mathematical cognition were located in brain areas including bilateral (BL) inferior and superior frontal gyrus, inferior parietal gyrus, left (L) middle frontal gyrus, right (R) superior parietal gyrus, and L precentral gyrus. Brain activations associated with creativity peaked in BL cuneus, inferior, middle, and superior frontal gyrus, R inferior parietal gyrus, and L precentral gyrus. The two maps showed overlap in BL inferior parietal gyrus and inferior and middle frontal gyrus. BL precuneus and cuneus were uniquely related to creativity. Based on these results, teachers can be suggested to use diagrams, dynamic visual materials, figures, gestures, graphs, maps, pictures, and videos in the classroom. For future analysis, we will conduct a systematic review of neuroimaging studies on mathematical and creative thinking using online databases, and meta-analyses using GingerALE. Limitations and suggestions for future research will be discussed.
(This work was supported by the NRF grant funded by the MSTI (No.2022R1A2C1010310).)

Assessing the performance of latent growth and mixed-effect models for analyzing non-normal longitudinal data of depressive symptoms in older adults

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Evan Pham (University of Manitoba)

In psychology research, multilevel (ML) and latent growth curve models (LGCM) have traditionally been used to analyze longitudinal patterns of depressive symptoms. However, the violation of the normality assumption in longitudinal data is prevalent in this field. Previous studies have separately examined how these two methods perform on non-normal data, but there has yet to be an attempt to compare their results using the same dataset. This article aims to explore the differences in outcomes between ML and LGCM when the assumption of multivariate normal distribution across time points is violated. To illustrate the performance of ML and LGCM frameworks, a large-scale panel from the Health and Retirement study will be used to estimate depression trajectories of n = 11,433 seniors over seven-time points from 1996 to 2008. The results reveal that the parameter estimates are identical between ML and LGCM, but the robust standard errors are higher in ML than in LGCM. Thus, the study suggests that when the assumption of normality in longitudinal data is violated, different outcomes between ML and LGCM are possible due to changes in model specifications and estimations. This finding emphasizes the need for future studies to discuss the robustness of longitudinal methods and develop remedies for the violation.

Comparison of FIML and multiple imputation in proportional odds model

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Ji Li (Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas), Dr. Xinya Liang (Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas)

Missing data is common in social sciences data and has garnered an increasing attention in recent decades. Initially, research was focused on dealing with missing in continuous data, while more recently, the direction is geared toward missing in categorical data. The full information maximum likelihood (FIML) and multiple imputation (MI) are two popular methods for handling missing data. These methods have a wide range of applications, whereas little is known about them in logistic regression models, particularly in proportional odds (PO) models. The PO models, also known as ordinal logistic regression models, are used when the dependent variable has three or more ordered categories. The estimation of PO models assumes that the effect of the independent variable on the odds of being in a higher category of the dependent variable is equal across all response categories. The PO models are widely used in social sciences, medicine, and psychology.

The proposed study will focus on comparing FIML and MI in PO models. It will introduce how to specify PO models and apply them to estimate models with missing predictors and outcomes. An illustration will be provided for conducting FIML and MI in PO models using Mplus and R. A simulation study will be carried out to compare the performance of FIML and MI under a variety of conditions, including the sample size, percent of missingness, and model complexity. The study will contribute to understanding the strengths and limitations of missing data techniques in PO models.

Application of topic modeling techniques in meta-analysis studies

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Minju Hong (University of Arkansas), Dr. Sunyoung Park (California Lutheran University)

In many educational studies, to integrate the findings of the studies with similar research questions, meta-analysis has been used commonly. However, it mainly deals with the quantitative results, not considering the qualitative findings, such as the texts in the abstracts of the articles. Thus, the primary purpose of this study is to suggest a way to synthesize the meta-analysis studies based on their findings' textual formats.

We analyze the abstracts of 198 articles from two journals, the Review of Educational Research and Psychological Bulletin, published from 2008 to 2018. We select the articles which are quantitative meta-analysis studies. By using the latent Dirichlet allocation model, we extract four topics. Based on the most representative 10 abstracts and 20 words of each of the four topics, we interpret four topics as 'Genetic or Environmental Characteristics of Individuals', 'Students' Achievement with Intervention', 'Cognitive Psychology', and 'Behavior Psychology'. And we suggest how to label/categorize each of the meta-analysis articles into the index based on these four topics. In addition, we show the relationship between the extracted topic structure and the meta-analytic characteristics of the articles together.

The implications of this study would be as follows. First, the findings of this study could provide an overview of the quantitative meta-analysis studies during the current ten years, from 2008 to 2018. Second, this study could suggest a guide for a future researcher who considers applying the topic modeling techniques to meta-analysis studies.

Using an iterative MIMIC-interaction modeling for referent variable selection

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Cheng-Hsien Li (National Sun Yat-sen University), Dr. Guo-Wei Sun (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 sciences. 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.

Detection of multiple group differential item functioning for students with disabilities taking an English language proficiency assessment

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Kyoungwon Lee Bishop (WIDA at the University of Wisconsin-Madison), Dr. Hacer Karamese (WIDA at the University of Wisconsin-Madison), Dr. Laurene Christensen (WIDA at the University of Wisconsin-Madison), Dr. Grace Xin Li (WIDA at the University of Wisconsin-Madison), Dr. Edynn Sato (WIDA at the University of Wisconsin-Madison)

The study aims to explore differential item functioning (DIF) by disability groups in a large-scale alternate English assessment for students with cognitive and physical disabilities. The preliminary step in most DIF methods is to designate a reference group and focal group(s) and then detect item bias towards one group. However, it is challenging to conduct DIF studies in the context of this alternate assessment because there is no reference group. Furthermore, it is even more challenging to conduct DIF studies when the sample size is small and unequal, the ability distributions differ across groups, the test length is short, and items follow polytomous scoring. In this circumstance, simultaneous DIF across all groups (Kim et all., 1995) might be preferred over multiple pairwise DIFs (Ellis and Kimmel, 1992) to address no reference group and multiple groups challenge while controlling Type I error rate. This presentation illustrates how a comparison of item parameters (Thissen, Steinberg, & Wainer, 1993) and item response functions (IRF) (Wainer, 1993) of multiple groups work in real-world data. Furthermore, the initial insights into items exhibiting DIF and considerations for developing inclusive items for students with disabilities will be shared.

Using psychometric function in subjective ecologically valid video experiment

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mrs. Dominika Wanat (AGH University of Science and Technology), Prof. Lucjan Janowski (AGH University of Science and Technology), Mr. Kamil Koniuch (AGH)

Video quality is important aspect of telecommunication network since around 70% of traffic is generated by video services. Therefore, it is critical to know when video quality is acceptable or not. Such a question is very similar to a psychophysical question: “can people detect compression?” We constructed a subjective experiment where a video was steady degradate. A tester could fix the quality of a watched video by pressing the button. After the button press, the quality comes back to the best available. The fixing comes with a cost of a small break in the video continuity. Our goal is to find a quality threshold (visual sensitivity) for the population and investigates differences between testers. The data are similar to psychophysical tests, but without randomness and without chance to see the worst quality. People fix the quality earlier. We are going to compare two different data preparation methods, with and without data filling.

The balanced development of international education in China: An empirical study with USAD China 2022

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Jiaqi Zhang (University of Cincinnati)

The purpose of this study is to explore the situation of international education in China by comparing seven subjects from an international-wide competition (USAD) hosted by the SKT Education Group with participants from international schools in China in 2022. Using the analysis of variance (ANOVA) with the dataset, this study reveals a balanced development in different regions of China regarding gender and region. With rapid economic development, China can provide better infrastructure and other necessary school conditions. International education in China is believed to be promising since it offers students ways to interact with a different educational system and experienced teachers, and balanced development of international education will impel the efforts.

Evaluating the quality of USAD China 2022: An Empirical Comparison between the CTT and IRT

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Jiaqi Zhang (SKT Education Group), Mrs. Yanyi Ren (SKT Education Group)

Two popular statistical frameworks address measurement problems, namely Classical Test Theory (CTT) and Item Response Theory (IRT). This study empirically evaluates the quality of seven subjects from an international-wide competition (USAD) hosted by the SKT Education Group with underrepresented minority participants from China in 2022 by examining the similarities and differences in parameter estimation using these two frameworks. The preliminary results suggested that the CTT performs equivalent to the Rasch and 1PL models if only the location parameter is important. Still, the 2PL model outperforms all other models when item discriminations are of interest. Overall, the items on USAD China 2022 are well developed, while some items might need to be revised or removed.

Psychometric characteristic of Brief Child Abuse Potential Inventory (BCAPI) among teachers in Iran

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Zahra Gheidar (Behavioral Research Center of Shahid Behashti Medical University.), Prof. Alireza Zahiroddin (Shahid Behashti Medical University), Dr. hanieh Zahiroddin (Behavioral Research Center of Shahid Behashti Medical University.)

Introduction: Generally, a valuable assessment is possible by reliable and valid measurement tools. In the early studies The Brief Child Abuse Potential (BCAPI) Inventory is considered as an effective tool for identification of child abuse potential and interventions in therapy and prevention. The purpose of this research was to analyze psychometric information on the Brief Child Abuse Potential Inventory (BCAPI) on teachers in Iran.

Method :500 teachers in Private and public schools of Tehran selected randomly and they answered Brief Child Abuse Inventory by Blind way. Data analyzed by SPSS 25, and Confirmatory factor analysis (CFA) and Varimax rotation was employed to assess factors structure. The reliability coefficient of items was estimated by general formula of Cronbach's alpha.

Results: The results of study estimated a sufficient internal consistency α=0/736 for Brief Child Abuse Potential Inventory. The component Analysis with Varimax rotation was extracted 7 factors with %53/17of total variance.

Conclusion:

The results of this study indicated the Brief Child Abuse inventory (Ondersman, S, J. Chaffin, M.Simpson, S&Lebreton,J.2005) is able to measure physical child abuse risk in teachers. It can be a reliable and valid tool to use in schools, child protection services, therapeutic intervention, public policies and studies**.**

Keywords: BCAPI, Reliability, validity

An EQ questionare for children 3-7

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Zahra Gheidar (Behavioral Research Center of Shahid Behashti Medical University.), Prof. Alireza Zahiroddin (Shahid Behashti Medical University)

Introduction
According to evidence, developing an appropriate tool for measuring and evaluating human traits has always occupied human mind. Also, regarding the importance of childhood and its impact on emotional development, a more precise look at the early growth period is required. Evidences show, this is not possible unless the accessibility to reliable and valid tools, capable of measuring and evaluating these traits, is provided. Considering the time-consuming process of measuring children's emotional skills by available tools, and verbal-design of some of them that requires more active involvement from the examiner, this study intended to build a shorter fully-pictorial questionnaire to evaluate children in a more real and appropriate situation to make a more reliable assessment.
Method
Regarding the theoretical foundations of Goleman's EI, in the first and second experimental stages, a 26-item pictorial questionnaire with a sample size of 30 children and a 31-item pictorial questionnaire with a sample size of 370 children were prepared and implemented, respectively.
Result
Reliability of the questionnaire was obtained as 0.830 using Cronbach's alpha. Factor analysis results of 370 questioners showed none of the questions was deleted and four extracted factors explained 48.145% of variance, out of which 30.184% was accounted by the first factor. The second to the fourth factors covered 7.550, 3.388, and 5.023 percentages of variance.
Conclusion
This 31 item full pictorial questionnaire has sufficient reliability and validity for assessing children EQ in 3 till 7 year's age.

Information matrix test misspecification assessment in cognitive diagnostic models

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Reyhaneh Hosseinpourkhoshkbari (University of Texas at Dallas), Dr. Richard M. Golden (University of Texas at Dallas)

A probability model is correctly specified, when the covariance matrix of the maximum likelihood estimates can be estimated using either the first or second derivatives of the likelihood function. If the determinants of these two different covariance matrix estimation formulas differ this indicates model misspecification. Using this method to detect the model misspecification, Golden et al. (2013) introduced the Determinant Information Matrix Test (D-IMT). This paper investigated the effectiveness of the D-IMT in Deterministic Input Noisy And gate (DINA) cognitive diagnostic model across different sample sizes and levels of misspecification. In a cognitive diagnostic model, the relevant skills for answering a particular question are defined through a special binary matrix called the “Q-matrix”.

Towards this end, misspecified models were generated by randomly altering a percentage of the DINA model Q-matrix elements. Next, 100 D-IMT statistics were computed from 100 bootstrap data sets generated for each possible combination of five misspecification levels: none (0%), low (5%), medium (10%), high (15%), and very high (20%) and four different sample sizes (n=67, n=134, n=268, n=536). The empirical distribution of the 100 D-IMT statistics for each condition was then used to estimate the Type 1 error probability “p-value” that the D-IMT incorrectly detected misspecification for that condition.

For larger sample sizes (n=268, n=536), estimated D-IMT p-values were statistically significant for the four misspecified conditions but not for the correctly specified condition ($p < 0.1$) suggesting good discrimination performance. Future work is planned to further investigate discrimination performance using multiple decision thresholds and other models.

Testing structural equation models with Monte Carlo asymptotic covariance matrices

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Hsin-Yun Lee (National Taiwan University), Dr. Li-Jen Weng (National Taiwan University)

The asymptotic covariance matrix of polychoric correlation estimates plays an essential role in both estimation and inferences when fitting structural equation models to observed ordered categorical data. Conventional approaches to calculating the full asymptotic covariance matrix are vulnerable to sampling variabilities, yielding poor performances of the overall test statistics unless the sample size is extremely large. To improve its estimation, Monroe (2018) has proposed a Monte Carlo (MC) estimator for the asymptotic covariance matrix which can be more efficient and lead to better calibration for established test statistics. However, the application of the MC estimator requires that the sample polychoric correlation matrix to be positive definite. This requirement is seldom met when analyzing large numbers of asymmetrically distributed variables with few response categories and insufficient observations, making the MC estimator unattainable. The current research modified the MC estimator to overcome its primary limitation by incorporating the model-implied polychoric correlation matrix as a replacement of the sample-based counterpart. This approach is based on the concept of two-stage asymptotically distribution free estimating process (Yuan & Bentler, 1997) and is referred to as the two-stage MC estimator. A simulation study demonstrated that the two-stage MC estimator can be implemented in the conditions we manipulated, including those when the Monroe's MC estimator is not accessible. Furthermore, chi-square statistics obtained from the two-stage MC estimator were closer to expected values than those from the conventional and Monroe's MC estimators, particularly for large models with data of highly skewed distributions and small sample sizes.

A new mixture IRT model for rater-mediated assessments

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Hung-Yu Huang (University of Taipei), Dr. Su-Pin Hung (National Cheng Kung University)

When rater-mediated assessments are administered, the performance of ratees must be appraised by human raters, and challenges arise regarding whether ratees' competencies can be validly reflected by raters' judgments with respect to scoring rubrics. A rater's decision-making process in scoring tasks can be qualitatively described by a theoretical judgmental model from the perspective of rater cognition and quantitatively characterized by a congruent measurement model. Research on rater cognition has suggested that both impersonal judgments and personal preferences may have effects on the judgmental process of raters. In this study, a dual-process lens model that serves as a theoretical base for rater judgment is introduced, and a mixture dual-process rater (MixDPR) model that allows the cognitive process of a rater to be consecutively dominated by unfolding and cumulative item response theory (IRT) models is proposed. The simulation results indicate that most of the parameters can be satisfactorily recovered, better parameter recovery is associated with increasing the number of raters assigned to score ratees and the use of highly correlated latent traits, and ignoring mixed rating response functions by fitting a conventional multifaceted rater model to the simulated data results in biased estimation. A real writing assessment is presented as an empirical example to demonstrate the application and implications of the new model.

A comparison of IRT-based subscore reporting methods for an objective structured clinical examination

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Nai-En Tang (National Board of Chiropractic Examiners), Dr. Igor Himelfarb (National Board of Chiropractic Examiners)

Subscores reporting from a high-stake testing agency to candidates for licenses in the health professions is necessary for its diagnostic purpose. Failing candidates want to know their deficiency areas for future remedial work (Haladyna et al., 2016). Objective Structured Clinical Examinations (OSCEs) have been widely used in the health professions for their ability to measure different knowledge and skills necessary for competent clinical practice (Sim et al., 2015) and, thus, are thought of as multidimensional examinations (Pell et al., 2010). The Item response theory (IRT)-based models were suggested to be a valuable method to produce accurate subscores for multidimensional exams (Haberman et al., 2010). While there are different IRT models to produce subscores, more is needed about the differences among these methods. Thus, this study aims to apply different IRT models to reproduce subscores for OSCE and evaluate these methods based on the model fit, reliability, and precision of the estimated subscores.

A single administration of the chiropractic OSCE exam of 952 examinees was analyzed in this study. The chiropractic OSCE consisted of 30 stations and was built around two content domains: case management and chiropractic technique. Following Toland et al. (2017) research framework, four IRT-based models were used to produce subscores: unidimensional graded response model (GRM), unidimensional GRM fit to each content domain, multidimensional GRM, and bifactor GRM. The results showed that multidimensional GRM fits the exam well and has better reliability and precision of the estimated subscores. The interpretation of the subscores will be discussed in the paper.

Rasch analysis of the chiropractic Case Management Risk Scale

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Nai-En Tang (National Board of Chiropractic Examiners), Dr. Igor Himelfarb (National Board of Chiropractic Examiners)

Case management refers to tasks that determine and implement the individualized clinical treatment plan for each patient and monitor that the plan is effective (Summers,2015). Like other health professionals, it is essential for chiropractors to evaluate medical risks to a patient associated with nonperformance or a subpar performance of case management tasks. Understanding chiropractors' perceived risk to case management tasks is also critical for chiropractic research, training, and patient care.

The aim of this study was to examine and evaluate evidence of the validity of the Case Management Risk Scale (CMRS) in a sample of chiropractors in the U.S. A representative sample of chiropractors in the United States was recruited to participate in the 2020 Chiropractic Practice Analysis study. After deleting missing data, a total of 2,054 chiropractors voluntarily completed all case management risk survey items as a sample for this study. Rasch analysis was conducted to provide empirical support for the content, structural, substantive validity, and generalizability evidence of the CMRS items. Our results suggested that the 10-item CMRS is a good overall data-model fit that measures a single latent trait in a consistent way. The response categories for the CMRS function effectively in our examination of the thresholds and the category characteristics curves. Overall, the results suggested the interpretation of CMRS items was generalizable across gender, ethnicity, age, and years of professional experience, except for a few item categories. Suggestions for item revisions, limitations, and future research direction are discussed in the final paper.

“What if applicants fake their responses?”: Modeling socially desirable responding in an item response theory framework

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Timo Seitz (University of Mannheim)

Many psychometric models have been developed in recent years to account for the influence of response biases in rating scale data. Previous research on response bias modeling has mainly focused on response styles, which reflect preferences of respondents for certain rating scale categories irrespective of item content. The present master thesis, however, aims at modeling a different kind of response bias, namely socially desirability responding (SDR). The multidimensional nominal response model (NRM; e.g., Falk & Cai, 2016), which is a flexible item response theory (IRT) model that allows to model response biases whose effect patterns vary between items, served as the modeling framework. For an empirical demonstration, responses from N = 3046 job applicants taking a personality test under high-stakes conditions were modeled. Effect patterns of SDR were specified by fixing scoring weights of SDR in the multidimensional NRM to appropriate values that were collected in a pilot study. Results indicated that modeling SDR improved model fit over and above response styles and led to effectively adjusted estimates of substantive personality traits. Furthermore, the modeling of SDR was validated in a sample of job incumbents taking the personality test under low-stakes conditions. However, while relationships between the degree of SDR and several covariates were found, modeling SDR did not alter the predictive validity of personality measures. Implications for theory and practice as well as limitations and future research directions concerning the modeling of SDR are discussed.

Small sample methods in multilevel analysis

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Yasuhiro Yamamoto (The Joint Graduate School (Ph.D. Program) in Science of School Education Hyogo University of Teacher Education), Prof. Yasuo Miyazaki (Virginia Tech)

In multilevel analysis, Bayesian methods have been said to solve small sample problems such as non-convergence or inaccurate estimation results in frequentist methods by reflecting prior knowledge about the population in the prior distribution. However, there are criticisms against reflecting prior knowledge with small samples. When the prior knowledge is strongly reflected in the analysis, there are risks of losing objectivity (e.g., Hox et al., 2012) or making erroneous inferences (e.g., Depaori, 2014). In addition, there are situations where the researchers have little prior knowledge about the population to reflect it in the analysis. Thus, we proposed using Bayesian methods with weakly informative prior (BayesW) and examined their effectiveness. Specifically, we compared the performance of the BayesW, which specified the weakly informative prior for the level-2 standard deviation parameters and variance-covariance matrices with restricted maximum likelihood (REML) method implemented in *R*. The results showed that the performance of point estimates of fixed effect parameters and level-1 variance parameters was comparable between BayesW and REML, but the BayesW performed better than the REML for interval estimation for fixed effect parameters and variance components parameters with small sample data. Thus, it could be said that when the main interest of the research is on the fixed effect parameters, the BayesW is more effective for small sample data. On the other hand, the BayesW significantly overestimated the level-2 variance components parameters compared to the REML, suggesting the necessity of comparing both results when conducting multilevel analysis for small sample data.

On the performance of horseshoe priors for inducing sparsity in path analysis models.

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Ms. Kjorte Harra (University of Wisconsin-Madison), Dr. David Kaplan (University of Wisconsin-Madison)

The present work focuses on the performance of two types of shrinkage priors - the horseshoe prior and a recently developed regularized horseshoe prior - in the context of inducing sparsity in path analysis models. Prior research has shown that these horseshoe priors induce sparsity by at least as much as the “gold standard” spike-and-slab prior. The horseshoe priors are compared to the ridge prior and lasso prior, as well as default non-informative priors, in terms of the percent shrinkage in the model parameters and out-of-sample predictive performance. An empirical study using data from PISA 2009 reveals the clear advantages of the horseshoe priors in terms of both shrinkage and predictive performance. A simulation study reveals clear advantages in terms of shrinkage, but less obvious advantages in terms of predictive performance, except in the small sample size condition where both horseshoe priors provide noticeably improved predictive performance.

Evaluating SEs of parameter estimates in the 2PL model with exact parametric bootstrap

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Hau-Hung Yang (National Taiwan University), Mr. Che Cheng (National Taiwan University), Prof. Yung-Fong Hsu (National Taiwan University)

Assessing a participant's ability is one of the goals of IRT and IRT-based Computerized Adaptive Testing (CAT). Ever since Lord (1980), the commonly adopted indicator of the precision of the latent ability estimator is the Fisher information function (IF), which is the standard to select test items in CAT. However, the IF cannot reliably reflect the precision of the estimator when the number of items is small (Magis, 2014). Two alternatives to evaluate the standard error (SE) of the ability estimator were proposed: the Monte-Carlo parametric bootstrap (Liou & Yu, 1991) and the exact-SE method (Magis, 2014). In this study, we prove that Magis's method is an instance of the parametric bootstrap. In particular, different from the Monte-Carlo method, one can derive the exact bootstrap distribution (Efron, 1979) based on the algorithm. Furthermore, with the aid of sufficient statistics, Magis (2014) proposed an accelerated version of the algorithm to reduce the computational complexity under the Rasch model framework. We generalize the accelerated version to the 2PL model in this study and dub our generalization 'fast and exact parametric bootstrap' (FEPB). Moreover, we use FEPB to construct the appropriate confidence intervals under the 2PL model. Our simulation study shows that the SEs calculated through the FEPB have smaller biases and RMSEs than those calculated through IFs for both the maximum likelihood and weighted likelihood estimators.

The impact of generating model on preknowledge detection in CAT

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Dr. Jianshen Chen (College Board), Ms. Kylie Gorney (University of Wisconsin-Madison), Dr. Luz Bay (College Board)

Enhanced test security is one of the advantages of computerized adaptive tests (CATs) because different test-takers see different sets of items. However, item exposure can still threaten the validity of test score interpretations, particularly for high-stakes tests. Therefore, detecting test-takers with preknowledge and items that have been compromised is critical for maintaining test fairness and validity. In practice, the impact of preknowledge is unknown and may vary across persons, items, or both persons and items. In this paper, a real-data-based simulation study is conducted for a CAT using two recently developed generating models that allow the impact of preknowledge to vary across persons and/or items. After generating the data, recent preknowledge detection methods using item scores, item response times, and item scores and response times are adopted to identify test-takers with preknowledge. The performance of the methods is evaluated under the different generating models, both for the overall group and for different ability and speed subgroups.

Examining strategies to establish partial invariance models with modification indices for ordinal missing data

Tuesday, 25th July - 17:30: Welcome Poster Reception (The Hotel) - Poster

Mr. Po-Yi Chen (National Taiwan Normal University), Prof. Wei Wu (Indiana University Purdue University Indianapolis), Mr. Min-Heng Wang (Mount Sinai Health System)

Introduction. Measurement invariance is important for cross-group comparisons. If full invariance assumptions are rejected, researchers could identify non-invariant items and establish partial invariance models with modification indices (MFI). However, when data are ordinal and incomplete, the MFI of the ordinal estimators based on polychoric correlations (e.g., weighted least squares with mean and variance adjustment, WLSMV) did not perform well with the default deletion-based missing data techniques like pairwise deletion (WLSMV_PD). Thus, referring to the method proposed by Mansolf et al (2020) for continuous data, we investigated the potential of using their method to pool MFI from WLSMV with multiple imputation (WLSMV_MI) to build partial invariance models with ordinal missing data. **Method.** We generated data from two-group measurement models (six five-point indicators per group). Manipulated designed factors include various sample sizes (500, 1000, 2000), patterns of loading non-invariance (x4), and missing data rates (complete, 15%, 30%). **Results:** The results indicate the pooled MFI from WLSMV_MI could maintain the type one error rate at the nominal level in conditions that all items are invariant. Furthermore, it also has higher power to differentiate invariant and non-invariant items than the MFI from robust full information maximum likelihood (robust FIML) and WLSMV_PD in most missing data conditions. **Conclusions**: Our results support the validity of using the pooled MFI of WLSMV_MI when examining partial invariant assumptions.

Mansolf, M., Jorgensen, T. D., & Enders, C. K. (2020). A multiple imputation score test for model modification in structural equation models. *Psychological methods*, *25*(4), 393.

Authors Index