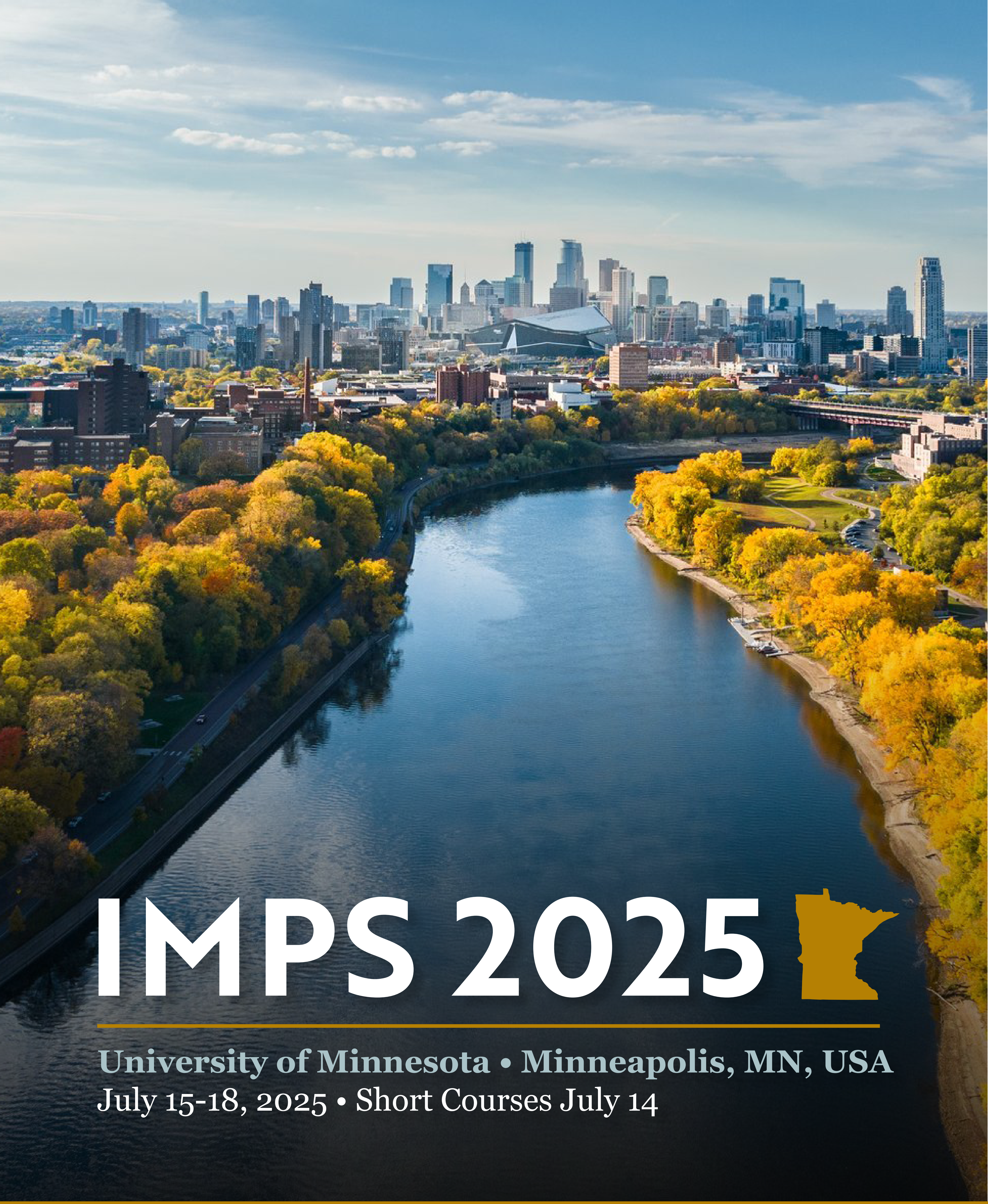
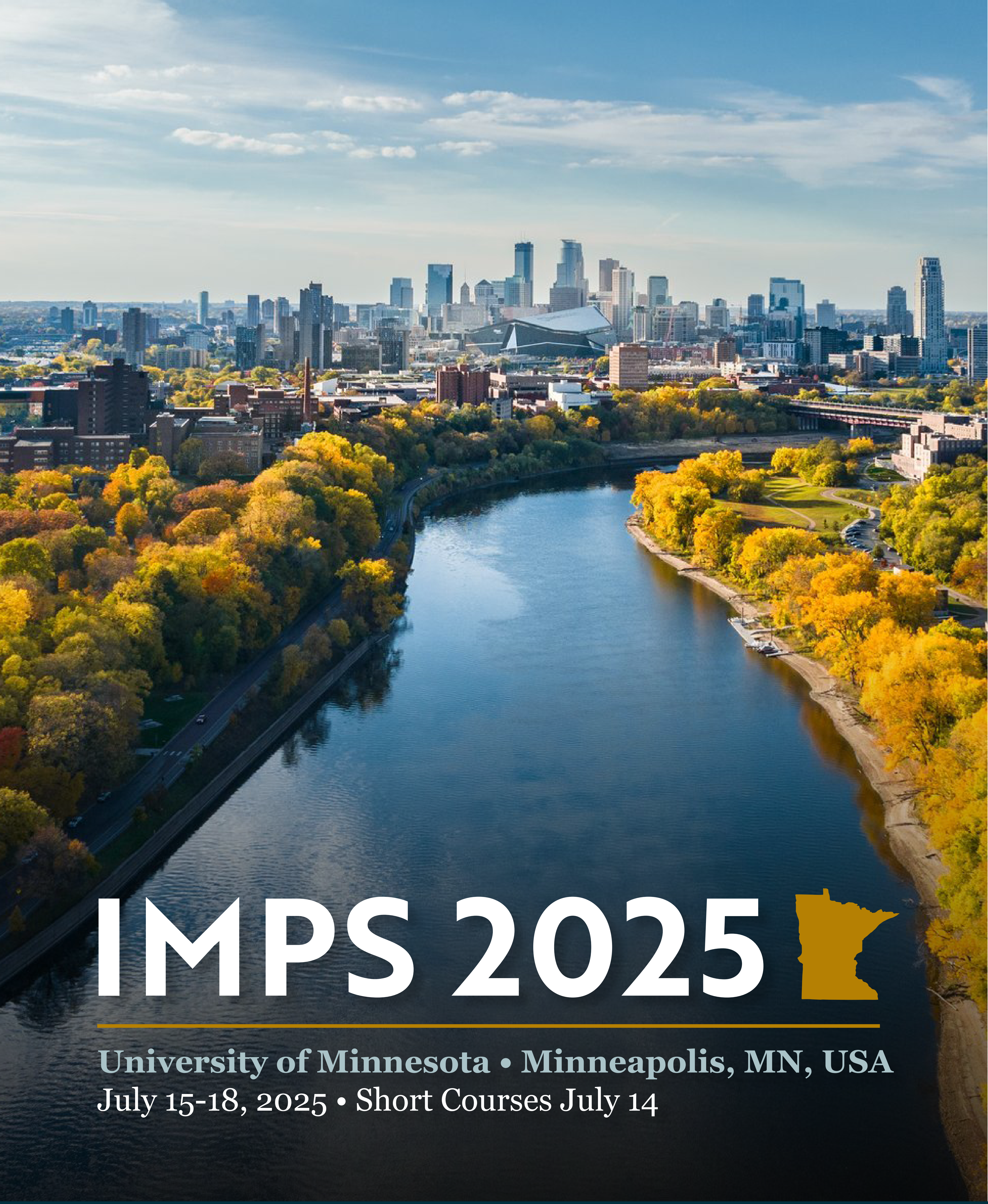


IMPS 2025

University of Minnesota • Minneapolis, MN, USA
July 15-18, 2025 • Short Courses July 14

ABSTRACTS

Table of Contents



IMPS 2025

University of Minnesota • Minneapolis, MN, USA
July 15-18, 2025 • Short Courses July 14

ABSTRACT BOOK: TALKS

A critical evaluation of similarity indices for psychometric research

Tuesday, 15th July - 10:45: Artificial Intelligence I (GH: Meridian 3-4) - Oral

Mr. Josiah Hunsberger (James Madison University), Dr. Paulius Satkus (Graduate Management Admission Council)

An increasing number of psychometric studies recently adopted machine learning and natural language processing (NLP) tools (Butterfuss & Doran, 2024; Micir et al., 2022). In enemy item detection, a common approach is to calculate cosine similarity on item embeddings (e.g., Meng & Li, 2024; Satkus et al., 2024). This is often justified by describing cosine similarity as "semantic similarity" or, by focusing on angular distance, which is said to ignore vector magnitude differences (Diallo et al., 2021). However, these fundamental assumptions are rarely questioned and relying on cosine similarity alone may obscure true relationships among items.

We conducted a simulation study manipulating item embedding characteristics derived from a sample of operational test questions. Specifically, we generated item embeddings from a multivariate normal distribution, varying embedding length, means vector, and variance-covariance structures. Our findings reveal that cosine similarity is sensitive to the magnitude of means vector, not merely their relative differences. For instance, embeddings with a correlation of .5 with a means vector of [-0.5, 0.5] (cos = .20) appeared less similar than those with [0,1] (cos = .35), despite having identical mean vector difference, variance, covariance, and length. This demonstrates that cosine similarity index is affected by absolute magnitude. Other indices, like Pearson's correlation, yielded results more consistent with theoretical expectations. Therefore, we caution against defaulting to cosine similarity for item embedding comparisons. Instead, based on our analysis of 13 similarity indices, we advocate for exploring metrics that better capture meaningful relationships among items, ultimately improving inferences about item similarity.

Enhancing item parameters prediction with transfer learning

Tuesday, 15th July - 11:00: Artificial Intelligence I (GH: Meridian 3-4) - Oral

Dr. Mingfeng Xue (University of California, Berkeley), Mr. He Ren (University of Washington)

Introduction

Traditional item calibration is labor-intensive and slows down item development due to the need for field testing. Machine learning (ML) offers a solution by predicting item parameters (e.g., difficulty, discrimination) from text features. Transfer learning, which refers to the technique of generalizing a model trained on one task to another, holds the promise of improving ML performance. This study explores transfer learning with large language models (LLMs) to improve prediction accuracy.

Method

We demonstrate the method by two personality scales (i.e., NEO-PI-R and 16-PF). First, IRT models are applied to calibrate item difficulty and discrimination for the two scales, respectively. Second, ML models for item parameter prediction are trained for each of the two scales using embeddings generated by the LLM. Third, we further apply the transfer learning technique to enhance the 16-PF item parameter prediction by training an additional layer on the ML model for NEO-PI-R. RMSE is used as the comparison metric.

Preliminary Results

Compared with the models purely trained on 16-PF items, the models from the transfer learning can increase prediction accuracy. We are evaluating the number of respondents that can be reduced due to prediction of transfer learning

Significance

It will enhance the item prediction accuracy with transfer learning and prove the machine learning model trained within one scale can be generalized to some extent to another scale under the same theme (e.g., personality assessment), which is critical for ML model training with small sample sizes.

Evaluating the capabilities of large language models in evidence synthesis

Tuesday, 15th July - 11:15: Artificial Intelligence I (GH: Meridian 3-4) - Oral

Ms. Yuchen Zhang (University of Toronto), Dr. Feng Ji (University of Toronto)

Introduction: Large language models (LLMs) like GPT-4o have received growing attention in scientific research due to their exceptional performance in tasks associated with natural language processing. Evidence Synthesis (ES), including systematic review and meta-analysis, is a systematic process that involves the integration of information from multiple empirical studies. Within it, data extraction is one of the most time-consuming and resource-intensive steps in extracting both qualitative and quantitative information, making LLMs promising to be applied.

Objective: We aim to empirically evaluate the performance (e.g., accuracy, reliability) under different strategies of GPT-4o in automating data extraction for meta-analysis.

Methods: We develop a Python-based automated workflow integrating GPT-4o API to extract data from 60 empirical studies for meta-analysis. Tailored prompt strategies, along with few-shot learning, and an iterative evaluation cycle are employed to improve the extraction performance.

Results: We find that, overall, our proposed workflow demonstrates a good performance in extracting qualitative data (e.g., participant demographics, survey methods) and simple quantitative data (e.g., sample size), with high accuracy. However, complex quantitative data (e.g., correlation coefficient) requires targeted refinement in prompts via few-shot learning to substantially enhance the accuracy of extracting correlation coefficients.

Impacts: This study highlight the viability of LLMs as tools to streamline ES workflows, offering a novel research pathway to reduce labor costs while maintaining its methodological rigor.

Item evaluation using LLM-respondents: A psychometric analysis with open-source data

Tuesday, 15th July - 11:30: Artificial Intelligence I (GH: Meridian 3-4) - Oral

Ms. Yunting Liu (University of California, Berkeley), Ms. Shreya Bhandari (University of California, Berkeley), Prof. Zachary Pardos (University of California, Berkeley)

Developing summative test items involves testing items with a large sample of respondents to measure their psychometric properties. However, collecting sufficient student responses is time-consuming and resource-intensive. In this study, we explore leveraging large language models (LLMs) to generate synthetic responses that mimic human characteristics, investigating whether LLMs can serve as a scalable alternative to human respondents in this process. We evaluate six LLMs—GPT-3.5, GPT-4, Llama 2, Llama 3, Gemini-Pro, and Cohere Command R Plus—alongside various ensembling methodologies. We utilize a dataset comprising algebraic items drawn from the OpenStax College Algebra textbook, along with human responses collected from college students on Prolific. Using Item Response Theory (IRT), we compare item parameter estimates derived from an augmented dataset (human + synthetic responses) to those obtained solely from human responses. Our results show that item parameters calibrated by LLM-generated responses exhibit strong correlations with human-calibrated counterparts (e.g., >0.8 for GPT-3.5). However, no LLM fully replicates human respondents due to the narrower proficiency distribution among synthetic responses. We further investigate strategies for data augmentation when human responses are limited, identifying sampling methods that enhance the Spearman correlation from 0.89 (human-only) to 0.93 (augmented dataset). We release all item content and synthetic responses under a Creative Commons license.

Careless responding in intensive longitudinal data: Effects on multilevel models

Tuesday, 15th July - 10:45: Multilevel Modeling (MAC: Johnson) - Oral

Ms. Alicia Gernand (RPTU University Kaiserslautern-Landau), Prof. Tanja Lischetzke (RPTU University Kaiserslautern-Landau)

Intensive longitudinal data provide a unique opportunity to model within-person dynamics alongside between-person differences. A commonly used approach is a multilevel model, which leverages person-centered variables as Level-1 predictors and person means as Level-2 predictors to disentangle within- and between-person effects. Despite their widespread use, the robustness of multilevel models against methodological biases such as Careless Responding (CR) remains largely unexplored. CR (i.e., responding to survey items without sufficient regard to the item content) has been found to significantly distort parameter estimates in statistical analyses commonly used in cross-sectional research. The aim of this study was to systematically examine the impact of CR in intensive longitudinal data on multilevel model parameter estimates at both the within- and between-person level. To this end, we conducted a simulation study by generating multilevel datasets with varied effect sizes and directions of within- and between-person relationships. CR was systematically introduced into the data by varying the affected variable (predictor, criterion, or both), CR type (random responding, straightlining), proportion of careless responders, and CR frequency. Results revealed that CR biased both within- and between-person variance estimates and distorted fixed slopes of Level-1 and Level-2 predictors. The extent of bias varied with the proportion of careless responders and CR frequency, while its direction depended on the true effect, affected variable(s), and CR type. These findings underscore the importance of implementing robust data quality checks in intensive longitudinal data research to mitigate CR-related biases and ensure reliable interpretations of within- and between-person effects.

The use of a multilevel multiple-indicators random-intercept cross-lagged panel model in college student goal pursuit

Tuesday, 15th July - 11:00: Multilevel Modeling (MAC: Johnson) - Oral

Dr. Hiroki Matsuo (Baylor University)

In social science and educational research where the nested data structure is more prevalent, multilevel modeling is often employed to estimate within- and between- effects separately (Goldstein, 2003). Curran (2003) mentioned the importance of the integration of structural equation modeling in the multilevel framework. Yet, the application of such techniques together can be challenging and has not been fully investigated, especially within longitudinal settings where the constructs are measured with multiple items. Therefore, the purpose of the current study is to demonstrate the effectiveness of a multilevel random-intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015) with multiple indicators, using an example examining temporal associations between goal patience and courage in a college student sample. Each item was nested within up to three individual goals and assessed across four measurement occasions. We examined measurement invariance simultaneously at within- and between- levels. Then, the measurement model was established at Level 1 to account for the within-level variability among goal contexts for each individual while Levels 2 and 3 together specified the RI-CLPM for individual goal pursuit, accounting for their trait-like stability. Equality constraints across measurement occasions were also added to establish the stationary model. In applied research where the construct can be measured within multiple contexts (i.e., goal pursuit), the results can be misinterpreted and lead to misguiding conclusions when the nested data structure is ignored. While future simulation-based research to investigate the utility of such modeling techniques is still necessary, the application of such models can suggest promising values.

What if sample size is the confound in multilevel models?

Tuesday, 15th July - 11:15: Multilevel Modeling (MAC: Johnson) - Oral

Mr. Michael Truong (York University), Dr. Xijuan Zhang (York University), Dr. David Flora (York University)

One of the primary strengths of multilevel modeling (MLM) is its ability to combine evidence from different clusters of data into a single regression model, using the size of each cluster to balance the cluster's role in the model. Here, we critically examine cases where cluster size (i.e., Level 2 sample size) plays a causal role in the true data-generating process (DGP). For example, if class sizes confound the effect of student motivation on final grades, how well can MLM control for this confounding and under what situations does MLM fail? To answer this question, we use a Monte Carlo simulation to compare model specifications that do and do not control for the effects of cluster size between frequentist MLMs and Bayesian MLMs. We predict that: (1) frequentist and Bayesian MLMs will exhibit comparable levels of bias; and (2) that the relative efficiency between frequentist and Bayesian MLMs will vary with the causal effects of cluster size. The results of this investigation may motivate researchers to consider and balance: (1) the goal of maximizing sample size; and (2) the open question of what effect different scales of sample sizes may play in their particular substantive context.

Estimating reliability using unbiased variance components in a nested design with group-specific distributions

Tuesday, 15th July - 11:30: Multilevel Modeling (MAC: Johnson) - Oral

Dr. Siyuan Marco Chen (Duolingo), Dr. J.R. Lockwood (Duolingo)

Intraclass correlations (ICC) have been developed to evaluate the reliability of scale scores in multilevel designs, where responses are nested in groups (e.g., teachers nested in schools). Existing ICC measures for multilevel designs often use variance components estimated from a linear mixed model to represent variability at each level. In doing so, these measures assume that an observation will share the same correlation with another observation in the same group, regardless of which group these observations come from. That is, the deviations of individuals are distributed with the same variability across all groups, and the effects of groups are sampled from a homogenous distribution of random effects. When groups are distributed differently, such as having differential magnitudes of random effect or residual variability, variance components estimated under the homogeneity assumptions may be inaccurate. This leads to ICCs that misrepresent reliability. We propose a method-of-moment approach that estimates means and variance components from group-specific distributions. This estimator yields unbiased estimates of variance components that can differ across groups. Relative to using traditional linear mixed models, ICCs based on this proposed approach better represent heterogeneity in the random effect distributions and within-group residual variabilities across many nested groups. To demonstrate the potential application of this approach, we evaluate the reliability of grades produced by Artificial Intelligence (AI) in a high-stakes English test, where the test takers come from different first-languages. This work helps researchers better understand the fairness of an assessment by evaluating if the test maintains its reliability across groups.

Multilevel model ICCs: Issues with reduced-model computation and full-model-based solutions

Tuesday, 15th July - 11:45: Multilevel Modeling (MAC: Johnson) - Oral

Ms. Yingchi Guo (University of British Columbia), Dr. Jason Rights (University of British Columbia)

Multilevel models (MLMs) are extensively used to analyze nested data structures and account for cluster-based dependency. When presenting MLM results, it is routinely recommended to report the *intraclass correlation coefficient* (ICC) or *variance partitioning coefficient* (VPC), which quantifies the degree of clustering. Though researchers typically have a full model of interest that contains predictors and random slopes, in practice, *ICC*s are usually computed from a simplistic model, such as a random-intercept-only model and/or a fixed-slope model. Despite the ubiquity of this practice, a key objective of this talk is to demonstrate that ICCs from reduced models can be problematic. Specifically, we will mathematically demonstrate that commonly used null-model and fixed-slope *ICCs* may not accurately quantify the proportion of variance nor the within-cluster correlation that they are thought to represent, especially in cases in which cluster means of level-1 predictors are not measured with sampling error. A complementary objective of this talk is to clarify the conceptual ambiguity in computing *ICC* for random-slope models by reviewing and generalizing existing measures. In doing so, we will integrate existing definitions of *ICC* and provide a taxonomy thereof, illustrate how the desired quantities thought to be derivable from reduced-model *ICCs* can instead be accurately obtained from the full model of interest, and provide novel extensions of *ICC*-based metrics. We will present Monte Carlo simulation results assessing the accuracy of our full-model-based measures relative to existing reduced-model ICCs and demonstrate the utility of our framework with empirical examples using open-source data.

Evaluating rapid guessing as noninformative about respondent proficiency

Tuesday, 15th July - 10:45: Methods for Aberrant Behaviors (MAC: Thomas Swain) - Oral

Prof. Daniel Bolt (University of Wisconsin - Madison)

The use of response time information to identify rapid guessing behavior has been a focus of much attention in educational measurement. Among the messy issues is the question of whether rapid guessing is noninformative regarding proficiency. We argue that attending to the relationships between rapid guessing behavior and item difficulty can provide evidence regarding lack of non-informativeness, and that evaluating such associations might regularly be considered when the non-informativeness assumption is made. Using a moderately low stakes mathematics test used to inform college course placement, we observe positive relationships between rapid guessing and item-response-theory (IRT)-based item difficulty even in the presence of extremely low response time thresholds for rapid guessing. Results suggest that respondent processing of personal item difficulty (as would theoretically be reflected by overall item difficulty + a person x item interaction) can occur well within time limits for classifying rapid guess responses.

Integrating disengagement identification into response style modeling: Response times matter

Tuesday, 15th July - 11:00: Methods for Aberrant Behaviors (MAC: Thomas Swain) - Oral

Mr. Jieyuan Dong (Beijing Normal University), Prof. Hongyun Liu (Beijing Normal University)

Response styles (RS) and disengaged responding (DR) are critical sources of bias in Likert-scale response data, yet their simultaneous control remains underexplored. Motivated by Ulitzsch et al.'s (2023) mixture model that pioneered distinguishing DR and RS, this study advances both theoretical rigor and practical implementation. Grounded in Tourangeau et al.'s (2000) cognitive theory of survey responding, we argue that response times (RTs) are essential for conceptually differentiating DR (shallow, trait-irrelevant processing) from RS (biased but trait-reflective). We then propose a general two-stage approach: (1) RT-driven identification of DR via threshold screening or distribution decomposition and (2) RS modeling that integrates the results of DR detection. Empirical validation using two large-scale survey datasets—one with item-level RTs and the other with screen-level RTs—demonstrates that methodological choices depend on RT levels. For item-level RTs, threshold-based dichotomous DR indicators are embedded as the first node in item response tree models. For screen-level RTs, DR probabilities derived from Gaussian mixture models serve as the sampling weights in the subsequent parameter estimation. Results show that incorporating RT-guided DR identification in RS modeling improves model fit over controlling either of them alone, and gains new insights into the analysis of survey responses. This work highlights RTs as a key tool for distinguishing DR and RS and offers adaptable workflows tailored to different RT availability, enhancing response validity in survey research.

Detecting aberrant responses using a general mixture model for cognitive diagnosis

Tuesday, 15th July - 11:15: Methods for Aberrant Behaviors (MAC: Thomas Swain) - Oral

Mr. Joemari Olea (University of the Philippines Diliman), Dr. Kevin Carl Santos (University of the Philippines Diliman)

Several CDMs introduced previously are proven to be special forms of the Generalized deterministic inputs, noisy "and" gate (G-DINA) model (de la Torre, 2011). However, the G-DINA model does not account for the heterogeneity that is rooted from possible existing subgroups in the population. Olea and Santos (2024) introduced a solution to this problem by creating a new model which incorporates G-DINA modeling with finite mixture modeling, aptly named the Mixture G-DINA model. In this paper, extensive simulation studies were conducted to examine the performance of the said model across different test-taking parameters such as number of items, number of examinees, item quality, and generating model as well as different test-taking behaviors such as cheating, creative responding, lack of motivation, plodding, speeding, cheating with randomness, and wrong response leakage. According to initial simulation results, the performance of the proposed model in correctly detecting aberrant responses are higher when the items are of good quality and there are higher number of items and examinees. The performance of the model is also shown to vary depending on the generating model. Lastly, the Mixture G-DINA model has a high detection rate for aberrant responses across different test-taking behaviors in certain cases.

A beta mixture model for careless respondent detection in Visual Analogue Scale data

Tuesday, 15th July - 11:30: Methods for Aberrant Behaviors (MAC: Thomas Swain) - Oral

Ms. Lijin Zhang (Stanford University), Prof. Benjamin Domingue (Stanford University), Prof. Leonie Vogelsmeier (Tilburg University), Prof. Esther Ulitzsch (University of Oslo)

Visual Analogue Scales (VASs) are increasingly popular in psychological, social, andmedical research. However, VASs can also be more demanding for respondents, potentiallyleading to quicker disengagement and a higher risk of careless responding. Existing mixturemodeling approaches for careless response detection have so far only been available forLikert-type and unbounded continuous data but have not been tailored to VAS data. Thisstudy introduces and evaluates a model-based approach specifically designed to detect andaccount for careless respondents in VAS data. To this end, we integrate existingmeasurement models for VASs (Noel and Dauvier, 2007) with mixture item response theorymodels for identifying and modeling careless responding. Simulation results show that the proposed model effectively detects careless responding and recovers key parameters, and highlights the unsuitability of the existing mixture factor analysis model for VAS data. Weillustrate the model's potential for identifying and accounting for careless responding usingreal data from both VASs and Likert scales. First, we show how the model can be used tocompare careless responding across different scale types, revealing a slightly higherproportion of careless respondents in VAS compared to Likert scale data. Second, wedemonstrate that item parameters from the proposed model, accounting for carelessresponding, exhibit improved psychometric properties compared to those from a modelthat ignores it. These findings underscore the model's potential to enhance data quality byidentifying and addressing careless responding.

Thresholding method for identifying anomalies in wellbeing time series data

Tuesday, 15th July - 10:45: Clustering Analysis (GH: Think 4) - Oral

Dr. Marie Turcicova (Institute of Computer Science, Czech Academy of Sciences), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences)

Identifying atypical patterns in digital behavior, physiological responses, and wellbeing metrics is essential for advancing psychometric research and intervention strategies. Anomaly detection techniques play a crucial role in (1) detecting deviations in digital behavior that may indicate excessive usage, social withdrawal, or digital addiction, (2) analyzing physiological and psychological responses, such as heart rate variability and stress indicators, to identify abnormal patterns related to technology use, and (3) monitoring responses to wellbeing interventions using digital technology by detecting atypical reactions that may require personalized approaches. To address these challenges, our research introduces an effective thresholding method for anomaly detection in multivariate data drawn from a normal distribution - a common model for many psychometric and mental health variables, as well as the distribution of residuals in various statistical models (e.g. regression models or models for time series). The proposed method provides a theoretically grounded threshold that is asymptotically optimal, ensuring that the number of misidentified anomalies approaches zero as the sample size increases. Our approach is compared to other common thresholding methods through simulation, and its performance is further demonstrated using real data.

Clustering experience sampled time series with deterministic and stochastic trends

Tuesday, 15th July - 11:00: Clustering Analysis (GH: Think 4) - Oral

Dr. Paul Wesley Scott (University of Pittsburgh)

Clustering time series(TS) offers a useful tool for capturing inter-individual variation with intensive longitudinal data. Clustering TS may be complicated in presence of both stochastic and deterministic trends. Further distortion may come from experience sampling. This project generates data with four deterministic trends (positive and negative linear; short and longer cycles) plus a near random walk exhibiting drift in a complete, equal interval time series(CTS) with n=1000,t=672 then subsamples to reflect experience sampling(ESTS) with 3(4hr) daily measures, skipping 4th 8hr interval (t=126). Dynamic Time Warping(DTW) is applied and evaluated in terms of recovering deterministic trends. We also use a feature-based clustering method where we automatically learn TS dynamics from SARIMA models allowing drift, LASSO to select TS dynamics to cluster on, rescale features, then apply LCA for clustering. LCA results are evaluated for class solution and association with original trend groups. Results are compared between CTS and ESTS. Visual inspection suggests DTW recovered linear trends well for CTS(56%) and ESTS(61%). In CTS, short cycles had fair recovery(44%), long cycles less so(29%); In ESTS, cycles were indiscernible. Cluster fit for CTS is better(DB*=0.92) than ESTS(DB*=1.08). LCA with BIC selected 2 classes for CTS and ESTS distinguishing shorter cycle from other trends with near-perfect accuracy in CTS(r_{pbs}=0.9151) and fair in ESTS(r_{pbs}=0.4742). Other trends were undistinguished in another class with low accuracy for ESTS($0.1<r_{pbs}<0.2$) and chance accuracy for CTS($r_{pbs}\approx 0.3$). Conflation of deterministic trends by stochastic trends provides explanation for DTW results, while 24hr period set for SARIMA provided best explanation for LCA results.

Comparing three strategies for clustering intensive longitudinal data

Tuesday, 15th July - 11:15: Clustering Analysis (GH: Think 4) - Oral

Dr. Yaqi Li (University of Oklahoma Health Sciences Center), Dr. Hairong Song (University of Oklahoma)

Clustering intensive longitudinal data (ILDs) aims to uncover distinct dynamic processes across multiple individuals or units. Three statistical strategies—shape-based, feature-based, and model-based clustering approaches—have been applied for this purpose. While each method has its own technical strengths and limitations, their relative performance remains largely unexplored, particularly in the social and behavioral sciences, where ILDs are often short, and sample sizes are small.

This simulation study compared these three strategies under varied conditions. Individuals' univariate dynamic processes were simulated using AR(1) and/or AR(2) models. The number of true clusters was set to be 2 with equal cluster sizes. Six factors were manipulated: (1) sample sizes, (2) lengths of ILDs, (3) between-cluster differences in dynamics, (4) deterministic trends, (5) (non-)stationarity, and (6) mixture of AR(1) and AR(2) processes. Performance was evaluated based on (1) recovery on the number of clusters, measured by the proportion of replicates correctly identifying the number of clusters, and (2) recovery on cluster membership, measured by the adjusted Rand index (ARI; 1 = perfect recovery).

Overall, the model-based approaches outperformed the others, achieving 92% accuracy in cluster number recovery and an average ARI of 0.76 for membership recovery. The shape-based approaches were highly accurate in identifying the number of clusters (97%) but performed poorly in recovering cluster membership (ARI =0.52). Lastly, the feature-based approach showed unacceptable performance under both criteria (69% accuracy in recovery of the number of clusters and ARI =0.36 for cluster membership). Further details and discussion of these findings will follow.

Understanding psychological temporal patterns: clustering time-series data and its challenges

Tuesday, 15th July - 11:30: Clustering Analysis (GH: Think 4) - Oral

Ms. Yuanyuan Ji (KU Leuven, University of Leuven), Dr. Marieke Schreuder (KU Leuven, University of Leuven), Prof. Ginette Lafit (KU Leuven, University of Leuven), Prof. Eva Ceulemans (KU Leuven, University of Leuven)

How people react to external stimuli and how their responses evolve over time have been widely studied in psychology. Examples include emotion dynamics after a daily event, the development of reading ability following an intervention, and stress responses elicited by noxious experimental stimuli. These temporal dynamics are considered important indicators of an individual's psychological state and are often measured through repeated assessments over time, resulting in time-series data. Clustering offers a promising approach for analyzing such data, as it captures not only the amplitude of the response (e.g., how strong an emotion is after an event), but also the shape of the temporal unfolding (e.g., a rapid peak followed by a slow decay or a gradual increase). However, several challenges arise. For example, these studies often capture relatively short time series (<10 data points), where time intervals may be uneven within a time series or vary between individuals. In self-report studies, moreover, missing data is common due to noncompliance. We will therefore conduct a simulation study to explore statistical solutions (e.g., imputation, smoothing) to these challenges. Furthermore, we will compare the performance of various clustering techniques, such as K-means, K- spectral centroid, and hierarchical clustering with different distance measures. The study aims to pave the way for the effective application of clustering approaches to psychological responses to external stimuli, ultimately enhancing our understanding of their temporal patterns over time.

Higher-order extended variational approximation to estimate latent variable models

Tuesday, 15th July - 10:45: Advances in Parameter Estimation (GH: Think 5) - Oral

Prof. Björn Andersson (University of Oslo)

Generalized linear latent variable models (GLLVMs) is a flexible class of models which supports joint modeling of continuous, count, ordinal, and binary variables with arbitrary structures among latent variables and observed variables. Estimation of GLLVMs is challenging when specifying complex models that include many types of observed variables and many latent variables. Various methods based on the variational approximation have been proposed in the literature to estimate GLLVMs, including an extended variational approximation based on the Laplace approximation. Such an approach can be applied to a large class of response functions but suffers from relatively high bias in some cases. To improve the estimation accuracy for this type of estimator, we derive an extended variational approximation estimator for GLLVMs based on a higher-order expansion and evaluate its properties. In the implementation, we analytically compute the higher-order derivatives required and exploit the sparsity of the model structure to realize computational efficiencies. The approach is compared against alternative methods, such as the computationally efficient but less accurate Laplace approximation and the gold standard method of adaptive Gauss-Hermite quadrature. We discuss the tradeoff between accuracy and computational efficiency with respect to different estimation methods and different model structures.

Resolving computational challenges of SEMs with big data using a divide-and-conquer approach to Bayesian synthesis

Tuesday, 15th July - 11:00: Advances in Parameter Estimation (GH: Think 5) - Oral

Prof. Katerina Marcoulides (University of Minnesota), Mr. Xinyu Liu (University of Minnesota), Ms. Hannah Hamling (University of Minnesota)

Analyzing structural equation models with massive datasets can pose major computational and parameter estimation challenges. For example, with large samples, standard errors of parameter estimates tend to be very small and generally suggest significance of parameter estimates that differ only trivially from zero. Similarly, even trivial model misfit can lead a researcher to a decision to reject a proposed model even when the model implies a good representation of the observed data. The purpose of this presentation is to introduce a novel procedure that can be used to resolve the computational challenges that arise when analyzing structural equation models with massive datasets. This study proposes the integration of a modified divide-and-conquer approach with Bayesian Synthesis data fusion methodology. The divide-and-conquer methodology refers to the popular computer science multi-step process that divides large datasets randomly into sub-data sets, while Bayesian Synthesis is a relatively new approach to data fusion where results from the analysis of one dataset are used as prior information for the analysis of the next dataset. In this manner, datasets of interest are sequentially analyzed until a final posterior distribution is created. This final posterior distribution incorporates information from all candidate datasets, rather than simply pooling resulting model estimates as is done in traditional divide-and-conquer type analyses. The proposed methodology is illustrated using a real data example. We will demonstrate that this approach can help address the computational challenges that arise with analyzing structural equation models with massive datasets.

Likelihood ratio tests with marginal maximum likelihood using Laplace approximations and adaptive quadrature

Tuesday, 15th July - 11:15: Advances in Parameter Estimation (GH: Think 5) - Oral

Ms. Lu Zhang (University of Oslo), Prof. Björn Andersson (University of Oslo)

Abstract: Generalized linear latent variable models are flexible models which enable individual or joint modeling of binary, ordinal, count, and continuous data. Estimation of these models is typically done with marginal maximum likelihood (MML) estimation which requires computing integrals without a closed form solution. To compute the integrals, different approximation methods can instead be used and in the literature Laplace approximations, adaptive and non-adaptive Gauss-Hermite quadrature, and stochastic methods have been proposed. In applied settings, likelihood ratio tests are often used to test hypotheses with respect to parameter restrictions. In this work, we examine the properties of likelihood ratio tests when estimation is done via approximate MML estimation. We consider first-order and second-order Laplace approximations, along with adaptive Gauss-Hermite quadrature with varying numbers of quadrature points. We conduct two independent simulation studies utilizing each approximation method to perform likelihood ratio tests, varying data types (binary, ordinal, and count data), sample sizes, and the number of observed variables. The first simulation uses hypotheses concerning the equivalence of slope parameters across the observed outcome variables, while the second examines restrictions across groups in a multiple group model. Our findings demonstrate that adaptive quadrature and second-order Laplace approximations provide reliable statistical inference across all settings with moderate to high sample sizes and a moderate number of observed variables. Meanwhile, the first-order Laplace approximation generally works well with the exception of binary data where the empirical size not equal to the nominal level with small or moderate numbers of observed variables.

An improved Satterthwaite (1941, 1946) effective df approximation

Tuesday, 15th July - 11:30: Advances in Parameter Estimation (GH: Think 5) - Oral

Prof. Matthias von Davier (Boston College)

This study introduces a correction to the approximation of effective *df* as proposed by Satterthwaite, specifically addressing scenarios where component *df* are small. The correction is grounded in analytical results concerning the moments of standard normal random variables. This modification is applicable to complex variance estimates that involve both small and large *df*, offering an enhanced approximation of the higher moments required by Satterthwaite's framework. Additionally, this correction extends and partially validates the empirically derived adjustment by Johnson and Rust, as it is based on theoretical foundations rather than simulations used to derive empirical transformation constants. Finally, the proposed adjustment also provides a correction to the estimate of the total variance in cases missing data have been replaced by multiple imputations such as in the case of plausible values in national and international large scale assessments.

Calibrated frequentist inference by stochastic approximation

Tuesday, 15th July - 11:45: Advances in Parameter Estimation (GH: Think 5) - Oral

Prof. Yang Liu (University of Maryland), Dr. Jonathan Williams (North Carolina State University), Dr. Jan Hannig (University of North Carolina at Chapel Hill)

Statistical inference about model parameters often relies on asymptotic confidence intervals (CIs), which are only trustworthy in large samples. When the sample size is limited, however, asymptotic theory may yield misleading conclusions. In this presentation, we explore a generic construction of conservative CIs based on the *p*-value function of the sample score statistic. Additionally, a smoothing approximation is introduced to make the *p*-value function more tractable numerically. Given sample data, we estimate the lower and upper confidence limits through the Robbins-Monro algorithm. We illustrate the proposed inferential procedure with two commonly used models in psychometrics: location-scale regression and linear normal factor analysis.

Classifying respondents' item-specific strengths with an interaction map approach

Tuesday, 15th July - 10:45: Item Response Theory I (GH: Think 3) - Oral

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

Latent space item response modeling (LSIRM; Jeon et al., 2021) has gained increasing attention as an innovative framework for visualizing local dependencies between respondents and test items. LSIRM places both items and respondents into a shared latent metric space, referred to as an interaction map, so that spatial proximity in the map reflects local dependencies—a feature that offers an intuitive diagnostic tool for uncovering respondents' item-specific strengths and weaknesses. Despite these advantages, methods for classifying and assessing such local dependencies remain limited. To address this gap, we propose three post-hoc algorithms that supplement the LSIRM framework for clustering: (1) a distance-based classification procedure that controls for respondents' overall trait levels; (2) a Bayesian approach to incorporate positional uncertainty into classifications; (3) a Bayesian resampling approach to incorporate distance uncertainty into classifications. We evaluate the diagnosis performance of these algorithms through simulation studies, in comparison to traditional methods, such as multidimensional IRT and diagnostic classification models. Results demonstrate that the proposed approach can reliably and effectively identify respondents' item-level proficiency nuances. Empirical applications to real-world assessments uncover subgroups of respondents with distinct respondent profiles, providing educators with finer insights into individual learning needs.

Unexplained item-person interactions due to heterogeneous item discriminations across individuals

Tuesday, 15th July - 11:00: Item Response Theory I (GH: Think 3) - Oral

Dr. Nana Kim (University of Minnesota), Dr. Minjeong Jeon (University of California, Los Angeles)

Traditional item response theory (IRT) models implicitly assume that individuals with the same ability level will have the same probability of a correct response to all items with the same item characteristics. However, such an assumption may not always hold true as there may be item-person interactions unexplained by the model parameters in IRT. While existing studies mainly focused on between-person variations in item difficulty for examining the unexplained item-person interactions, this study aims to illustrate the role of item discriminations in understanding such unexplained interactions. To this end, we use a recently proposed modeling approach called a latent space item response model (LSIRM) which produces an interaction map visualizing the item-person interactions remaining after accounting for main person and item effects in IRT models. Through a simulation study, we demonstrate how unexplained item-person interactions can arise from the heterogeneity in item discriminations across items and individuals, and further examine how such unexplained interactions may create bias in parameter estimation under traditional IRT models. Our results show that a specific pattern of unexplained item-person interactions is produced when item discriminations are heterogeneous across items and individuals, and further illustrate that fitting traditional IRT models may lead to biased estimates when there are such interactions present. We discuss implications of the results including the use of LSIRM for inspecting the presence of unexplained item-person interactions prior to fitting IRT models.

Impact of item locations on parameter recovery with the GGUM

Tuesday, 15th July - 11:15: Item Response Theory I (GH: Think 3) - Oral

Ms. Nicole Bonge (University of Arkansas), Dr. Ronna Turner (University of Arkansas)

Research has shown that unfolding item response theory (IRT) models perform well with non-cognitive data following both ideal point and dominance response processes under certain data conditions (Reimers et al., 2023). Previous research relating to parameter recovery for the generalized graded unfolding model (GGUM) has primarily involved items that are evenly spaced across the latent trait continuum. However, in practice, item locations may commonly be represented by sampling from a normal, uniform, or skewed distribution. To our knowledge, no studies have explored item and person parameter recovery for the GGUM with varied item location distributions.

This study builds on previous work (e.g., Roberts et al., 2002; Roberts & Thompson, 2011) to evaluate item and person parameter recovery with the GGUM under different item location distributions and sample sizes, using two, four, and six response options. Preliminary results suggest that items sampled from normal or uniform distributions yield better item parameter recovery than equally spaced items, though differences in estimation error diminish as the sample size and number of response options increase. In addition, estimation bias for both item and person parameters was minimal across most conditions. Person parameter estimates displayed acceptable estimation errors with four and six response options across all item location and sample size conditions, while two response options produced relatively large estimation errors. These results can help inform practitioners regarding the applications of the unfolding model to various test and item response formats.

Exposome burden scores to summarize environmental chemical mixtures: Creating a common scale for cross-study harmonization, report-back and precision environmental health

Tuesday, 15th July - 11:30: Item Response Theory I (GH: Think 3) - Oral

Dr. Shelley Liu (Icahn School of Medicine at Mount Sinai), Dr. Katherine Manz (University of Michigan), Dr. Jessie Buckley (University of North Carolina), Dr. Leah Feuerstahler (Fordham University)

Environmental health researchers are increasingly concerned about characterizing exposure to environmental chemical mixtures (co-exposure to multiple chemicals simultaneously). We discuss unsupervised approaches for quantifying an overall summary score or index that reflects an individual's total exposure burden. Sumscores and principal components analysis (PCA) are common approaches for quantifying a total exposure burden metric but have several limitations including the need for imputation, lack of accounting for exposure source heterogeneity, simplistic treatment of measurement error, and discarding information in pooled analyses. Meanwhile, item response theory (IRT) is a novel and promising alternative to calculate an exposure burden score that addresses the above limitations. It allows for the inclusion of exposure analytes with high frequency of non-detects without the need for imputation. It can account for exposure heterogeneity to calculate fair metrics for all people, through assessment of differential item functioning and mixture IRT. IRT also quantifies measurement errors of the exposure burden score that are individual-specific, such that it appropriately assigns a larger standard error to an individual who has missing data on one or more exposure variables. Lastly, IRT enhances cross-study harmonization by enabling the creation of exposure burden calculators to set a common scale across studies, and allows for the inclusion of all exposure variables within a chemical class, even if they were only measured in a subset of participants. Summarizing total exposure burden, through the creation of fair and informative index scores, is a promising tool as environmental exposures are increasingly used for biomonitoring and clinical recommendations.

Integrating multidimensional scaling into the multidimensional generalized graded unfolding model

Tuesday, 15th July - 11:45: Item Response Theory I (GH: Think 3) - Oral

Ms. Zhaoyu Wang (Georgia Institute of Technology)

This study enhances the Multidimensional Generalized Graded Unfolding Model (MGGUM) by integrating Multidimensional Scaling (MDS) to improve parameter estimates and reduce sample size requirements. The MGGUM often demands large samples, limiting its practical use in social science research with smaller datasets. By incorporating MDS scale values, derived with pairwise similarity judgments, our approach refines MGGUM parameter estimation within a hierarchical Bayesian framework.

Our approach employs item coordinates from an MDS configuration as the prior means for the item locations in the estimation of MGGUM. These informative items prior distributions must have a reasonable amount of impact on the ultimate solution of MGGUM item locations to be useful. To better determine the degree of impact that is optimal, we examine different levels of precision for the prior distributions to identify a level at which the informative prior distributions can improve the estimation of MGGUM item location parameters.

To evaluate this approach, we simulate item response and pairwise similarity data using previously estimated item parameters along with varying sample sizes. Additionally, we illustrate the hierarchical framework with real data from a physical attraction study that has previously been analyzed using the MGGUM with relatively noninformative prior distributions. Using a data set with responses from 1,224 participants, we test the approach on randomly selected subsamples with N=300 (small) and N=1000 (large) cases.

This new approach will expand MGGUM's applicability in psychological and social science research, enabling robust modeling of preference and attitude data in situations with more limited sample sizes.

Estimating marginal effects with zero-inflated Poisson models

Tuesday, 15th July - 14:30: Bayesian Methods and Their Applications (GH: Meridian 1-2) - Oral

Mr. Chendong Li (Texas A&M University), Dr. Oi-Man Kwok (Texas A&M University), Dr. Timothy Lawrence (Texas A&M University)

In psychological and medical research, count data often include a high frequency of zero outcomes that standard Poisson and negative binomial models cannot adequately address. The zero-inflated Poisson model improves on these approaches by separately modeling excess zeros and the count process. However, interpreting predictor effects for the entire population remains challenging because the model distinguishes between structural zeros and counts from individuals who may also generate zeros. This study introduces the marginalized zero-inflated Poisson (mZIP) model, which directly models the population mean count, allowing for straightforward inference on overall exposure effects (Long et al., 2014). By focusing on the marginal mean, the mZIP model provides parameter interpretations that reflect the entire population, similar to standard Poisson regression, while still accounting for zero-inflation. Using real-world data from Bullying, Sexual, and Dating Violence Trajectories from Early to Late Adolescence in the Midwestern United States, 2007 to 2013, we illustrate the estimation of the mZIP model with both frequentist and Bayesian methods in R. We also discuss the differences and advantages of these estimation strategies, concluding with limitations and future directions.

Probabilistic projections of country-level progress to the UN SDG indicator of minimum proficiency in reading and mathematics

Tuesday, 15th July - 14:45: Bayesian Methods and Their Applications (GH: Meridian 1-2) - Oral

Prof. David Kaplan (University of Wisconsin - Madison), Prof. Nina Jude (University of Heidelberg), Ms. Kjorte Harra (University of Wisconsin - Madison), Mr. Jonas Stampka (University of Heidelberg)

As of this writing, there are five years remaining for countries to reach their Sustainable Development Goals deadline of 2030 as agreed to by the member countries of the United Nations. Countries are, therefore, naturally interested in projections of progress toward these goals. A variety of statistical measures have been used to report on country-level progress toward the goals, but they have not utilized methodologies explicitly designed to obtain optimally predictive measures of rate of change as the foundation for projecting trends. The focus of this paper is to provide Bayesian probabilistic projections of progress to SDG indicator 4.1.1, attaining minimum proficiency in reading and mathematics, with particular emphasis on competencies among lower secondary school children. Using data from the OECD PISA, as well as indicators drawn from the World Bank, the OECD, UNDP, and UNESCO, we compare results using both Bayesian stacking and Bayesian model averaging, ultimately choosing BMA to obtain optimal estimates of the rate of change in minimum proficiency percentages then use that estimate to create probabilistic projections into the future. Four case study countries are also presented.

Fast and efficient robust Bayesian meta-analysis with spike-and-slab priors

Tuesday, 15th July - 15:00: Bayesian Methods and Their Applications (GH: Meridian 1-2) - Oral

Mr. František Bartoš (University of Amsterdam; Czech Academy of Sciences)

Meta-analyses are vital for synthesizing evidence across multiple studies, yet their validity is frequently undermined by publication bias. To address this issue, we developed robust Bayesian meta-analysis (RoBMA), which integrates multiple publication bias adjustment methods—such as selection models and PET-PEESE—using Bayesian model-averaging (BMA). Traditionally, RoBMA employed JAGS and bridge sampling to estimate individual models and combine them via BMA, providing a robust framework for evidence synthesis. In recent advancements, we expanded RoBMA to incorporate meta-regression and three-level models, enabling its application to a wide range of meta-analytic settings. However, these extensions substantially increase the model space, rendering traditional estimation methods computationally demanding. This talk reviews these methodological developments and introduces a significant enhancement: spike-and-slab model-averaging for both parameters and model components within RoBMA. Leveraging the spike-and-slab model-averaging markedly improves computational efficiency, facilitating applications to larger samples and more complex problems. The method retains RoBMA's robustness while overcoming the computational bottlenecks of expanded model spaces, delivering fast and reliable estimates. We illustrate the practical benefits of this innovation through examples and discuss its implications for advancing meta-analytic research.

An innovative way of estimating subgroups mean estimates: SABDB

Tuesday, 15th July - 15:15: Bayesian Methods and Their Applications (GH: Meridian 1-2) - Oral

Dr. Sinan Yavuz (University of Wisconsin - Madison)

Small Area Estimation (SAE) is a vital methodology to accurately and properly estimate subpopulations and disaggregated datasets when the direct estimation methods are not applicable. One example of a small area in an educational context is the subgroups with less than 62 students in the National Assessment of Educational Progress. Due to the imprecise estimates of mean achievement for these groups the National Center of Educational Statistics does not report the mean achievement estimates for this subpopulation, which are usually the underserved groups such as Black, Hispanic, and American Indian students.

To address this issue this study introduces a novel SABDB (Small Area Bayesian Dynamic Borrowing) approach for estimating the mean achievement of these subgroups. Viele et al. (2014) proposed the Bayesian dynamic borrowing (BDB) method to handle extensive historical or external, which was extended to large-scale assessments by Kaplan et al. (2022) to borrow information dynamically from the historical cycles of the same assessment. The BDB method borrows strength from more homogeneous datasets by automatically adjusting the hyperpriors of the joint prior distributions. In contrast, static borrowing occurs when the priors are fixed and don't account for the heterogeneity or homogeneity between the historical cycles of the assessment and the current dataset.

The current study extends BDB to spatial SAE and borrows information to estimate the subgroup of interest from more homogeneous other subgroups. The preliminary results show that the proposed approach produces less bias compared to non-model mean estimation. The final results will be shared during the presentation.

Multilevel heterogeneous factor model with fixed covariates

Tuesday, 15th July - 15:30: Bayesian Methods and Their Applications (GH: Meridian 1-2) - Oral

Ms. Meijin Lin (Sun Yat-sen University), Prof. Junhao Pan (Sun Yat-sen University), Prof. Edward Ip (Wake Forest University)

When repeated measurements are available, individual-specific factor models can be estimated. Ansari et al. (2002) introduced a multilevel heterogeneous model for confirmatory factor analysis (CFA) applied to repeated measurements, allowing varying means and factor loadings across individuals while maintaining an invariant factor structure. Pan et al. (2020) extended this approach with the Multilevel Heterogeneous Factor Model (MHFM), improving model fit by relaxing the residual covariance matrix's diagonality assumption using Bayesian Lasso. However, the MHFM does not incorporate covariates, which limits its practical applicability. To address this limitation, we propose the Multilevel Heterogeneous Factor Model with Fixed Covariates (MHFMC). By including fixed covariates, this extension enhances the power of the factor measurement model, allowing researchers to use the MHFMC as a tool for investigating hypotheses such as how specific individual-level characteristics influence factor structure. In this work, we introduce the MHFMC model, present posterior distributions of its parameters, and describe the estimation algorithm. Simulation studies are conducted to evaluate the performance of the proposed algorithm, using bias, posterior standard deviation, and root mean square to evaluate accuracy of parameter recovery. Finally, we apply the MHFMC model to a health science dataset, demonstrating its potential for more precise and flexible analysis in behavioral and clinical research.

Analyzing teacher mathematics skills using cognitive diagnosis modeling

Tuesday, 15th July - 14:30: Applications of Cognitive Diagnostic Models (GH: Meridian 3-4) - Oral

Mr. Nurym Shora (National Center for Professional Development "Orleu"), Mrs. Aidana Shilibekova (National Center for Professional Development "Orleu"), Mrs. Akmaral Zhumykbayeva (National Center for Professional Development "Orleu"), Mr. Baurzhan Yessingeldinov (National Center for Professional Development "Orleu")

This study applies Cognitive Diagnosis Modeling (CDM) to analyze mathematics teacher assessments and evaluate skill mastery across key domains: numbers, algebra, geometry, statistics and probability, and mathematical modeling. Using response data from over 3,000 teachers from diverse qualification categories, experience levels, schools, and regions, the Deterministic Inputs, Noisy "And" Gate (DINA) model was employed to classify teachers based on their mastery of specific mathematical competencies. Results indicate that the highest mastered skill is Number (76%), while Geometry (61%) is the least mastered. The probability of teachers mastering all five skills is 57%, suggesting that a significant proportion still struggles with certain mathematical domains. To validate the results, Item Response Theory (IRT) was used to estimate ability scores, allowing a comparison between cognitive skill profiles and traditional proficiency measures. Correlation analysis between IRT ability estimates and CDM mastery probabilities shows a moderately strong relationship, confirming that CDM effectively distinguishes teacher competencies and aligns well with IRT-based ability scores (de la Torre & Minchen, 2014). These findings highlight notable variations in skill mastery across teacher subgroups, providing insights for targeted professional development. The study suggests that CDM can enhance teacher assessment practices beyond conventional scoring methods by offering detailed diagnostic feedback on specific skill areas (Javidanmehr & Anani Sarab, 2017).

An empirical comparison of cognitive diagnosis models and multidimensional IRT

Tuesday, 15th July - 14:45: Applications of Cognitive Diagnostic Models (GH: Meridian 3-4) - Oral

Prof. Wenchao Ma (University of Minnesota), Dr. Hueying Tzou (National University of Tainan), Dr. Wenjing Guo (Pearson), Dr. Yi-Fang Wu (Cambium Assessment)

Cognitive diagnosis models (CDMs) aim to diagnose specific cognitive skills or attributes that a student has mastered or failed to master. It aims to provide detailed profiles of students' strengths and weaknesses. While some studies show CDMs can provide more diagnostic information than unidimensional item response theory (IRT) models, the comparison between CDMs and multidimensional IRT (MIRT) remains rare. CDMs and MIRT can differ in various aspects, but they both have the potential to capture students' proficiencies in a multidimensional space.

The goal of this study is to compare the CDM and MIRT in modeling students' responses using a set of real data. The data contains 1,037 middle school students' item responses to 31 multiple-choice items in a proportional reasoning (PR) test. The PR test was developed under the CDM framework, with an intent to measure six different skills.

We have applied both CDMs and MIRT models to evaluate their ability to estimate students' cognitive skills. The overall model-data fit was assessed using M_2 and $RMSEA_2$ indices, and the models were compared using various information criteria, such as AIC, BIC and SABIC. Additionally, we will compare the out-of-sample prediction accuracy of both models using cross-validation techniques. We will also examine the estimates of students' abilities/attributes and evaluate the reliability of these estimates. This comparison, based on real-world data, will complement existing methodological research and offer useful considerations for applied researchers in selecting the most suitable methods for diagnostic assessments.

The choice between cognitive diagnosis and item response theory: A case study from medical education

Tuesday, 15th July - 15:00: Applications of Cognitive Diagnostic Models (GH: Meridian 3-4) - Oral

Prof. Youn Seon Lim (University of Cincinnati)

Feedback is a powerful instructional tool for motivating learning. But effective feedback, requires that instructors have accurate information about their students' current knowledge status and their learning progress. In modern educational measurement, two major theoretical perspectives on student ability and proficiency can be distinguished. Latent trait models identify ability as a continuous uni- or multi-dimensional construct, with unidimensional item response theoretic (IRT) models presumably the most popular type of latent trait models. They report a single ability score that allows for locating examinees relative to their peers on the latent ability dimension targeted by the test. Latent trait models have been criticized for lacking diagnostic information on students' specific skills, their strengths and weaknesses in a knowledge domain. Cognitive diagnosis (CD) models, in contrast, describe ability as a combination of discrete skills (called "attributes") that constitute (partially) ordered latent classes of proficiency. The focus of CD is on collecting information about the learning progress for immediate feedback to students in terms of skills they have mastered and those needing study. CD has been underused in education; performance assessment still mostly relies on latent-trait-based methods. The motivation for the study reported here arose from the desire to conduct a side-by-side evaluation of the two seemingly disparate psychometric frameworks, CD and IRT. Data from a biochemistry end-of-term exam were used for illustration. They were fitted with multiple CD and IRT models, among them also HO-GDINA models that permit for a close approximation to several unidimensional IRT models.

Cognitive diagnostic models for measuring 21st century skills

Tuesday, 15th July - 15:15: Applications of Cognitive Diagnostic Models (GH: Meridian 3-4) - Oral

Ms. Kristin Lansing (University of Amsterdam), Dr. Andries Van der Ark (University of Amsterdam), Dr. Tessa van Schijndel (University of Amsterdam), Dr. Alina von Davier (EdAstra Tech)

Cognitive diagnostic models (CDM) have been successfully applied to estimating learners' strengths and weaknesses on various attributes based on their test responses (Ma & de la Torre, 2020). In this project we apply the CDM R-DINA (Najera et. al, 2023), a restricted deterministic input, noisy "and" gate parametric cognitive diagnosis model that functions as a comprehensive cognitive diagnostic model for classroom-level assessments, to measure 21st century skills (21CS), such as critical thinking, creative thinking, collaborative problem solving, and information and technology skills. Despite their importance and pervasiveness in the classroom, these skills are often embedded only implicitly, in curriculum and assessments. In this study, we explored a procedure for making these embedded skills explicit in existing assessments. Our procedure involves starting with the definition of certain 21CS. Next, for each item independent coders identify the 21CS required for a successful response, resulting in a Q-matrix. Finally, mastery of 21CS is estimated using the students' item scores, the Q-matrix, and the CDM RDINA. This procedure provides an estimate of skill mastery at the aggregate (classroom) level as well as an estimated mastery profile for each student. As a proof of concept, we applied the procedure to existing course assessments. Findings indicate that some 21CS are embedded in the assessments with high frequency and that the procedure provides valuable insights to support the development of 21CS, allowing for identification of students requiring support, remediation, or monitoring. We conclude with a discussion of challenges and opportunities for this procedure and future research.

Advancing behavioral pattern clustering with time-warped longest common subsequence method on process data

Tuesday, 15th July - 14:30: Process Data Analysis (MAC: Johnson) - Oral

Dr. Qiwei He (Georgetown University), Ms. Binhui Chen (Georgetown University)

The computer-based assessment not only enables the development of innovative item types but also the collection of a broader range of records in log files throughout human-machine interactions. These granular records, often referred to as *process data*, are typically stored in the form of an ordered sequence of multi-type, time-stamped events. This multidimensional sequential data necessitates new methods to integrate information from all dimensions simultaneously. In this paper, we introduce a new proximity computation method on process data, the time-warped longest common subsequence (T-WLCS), inspired by music retrieval techniques. The T-WLCS method is a revised version of the Longest Common Subsequence (LCS) algorithm, designed to account for temporal variations in sequences. It adapts LCS principles to allow for dynamic time warping (DTW), thus enhancing its robustness in identifying common patterns within time-dependent data. We illustrate this new method using process data from a Japanese sample in one Problem Solving in Technology-Rich Environments (PSTRE) item in the Program for the International Assessment of Adult Competencies (PIAAC). The T-WLCS was found better performed than DTW and LCS in identifying homogenous behavioral patterns in clustering analysis.

Understanding the evolution of individual response process: An exploratory approach

Tuesday, 15th July - 14:45: Process Data Analysis (MAC: Johnson) - Oral

Ms. Ruiting Shen (New York University), Dr. Klint Kanopka (New York University)

The massive use of computer-adaptive testing has contributed to revolutionizing psychometric analysis in terms of both data collection and analysis methodologies. The National Assessment of Educational Progress (NAEP) is one of the most well-known assessments that collect respondents' process data, which has been extensively analyzed (Bosch, 2021; Levin, 2021; Patel, Sharma, Shah, & Lomas, 2021; Zehner et al., 2020). Much of this work has primarily focused on efficiency estimation and training prediction models using mainly item-level feature engineering and person-level grouping techniques.

Some work focusing on exploring the data itself rather than training models has also focused on grouping respondents' behavior patterns on subsets of items (Wei, Zhang, & Zhang, 2024) through the use of clustering and relating these patterns to demographics and respondent ability. With a similar goal, we expand on this approach by first developing item-level behavior profiles using clustering-based methods (i.e., latent class analysis) for all items. Using these profiles, we explore individual trajectories from profile to profile over the course of the test. Combining these trajectories with latent abilities and speed estimated from item responses and demographic information, we characterize the respondents who comprise the individual trajectories. Preliminary results find item-to-item variation in the structure and interpretation of the behavior profiles but also distinct profiles within each item that represent differences in the underlying IRT-derived latent ability distributions. Understanding usage patterns for respondents of various ability levels, demographic backgrounds, and disability statuses may inform future item development, assistive tool design, and user interface for computerized assessments.

A copula-based joint model for item responses and response times

Tuesday, 15th July - 15:00: Process Data Analysis (MAC: Johnson) - Oral

Mr. Sunbeom Kwon (University of Illinois Urbana-Champaign), Prof. Susu Zhang (University of Illinois Urbana-Champaign)

Using response time as collateral information improves the accuracy and efficiency of ability estimates (van der Linden, 2010). However, the commonly used hierarchical approach (van der Linden, 2007) for modeling the dependency between item response and response time, which assumes multivariate normality of traits for speed and accuracy, often shows poor fit with empirical data. A consistent pattern observed across test datasets is greater variability in response time among lower-ability test-takers, indicating upper-tail dependency (Domingue et al., 2022).

This study proposes a copula-based joint model for item responses and response times. The copula framework allows for modeling trait dependencies beyond multivariate normal assumptions. The proposed model consists of two levels: first, the marginal distributions of item responses and response times are modeled using a two-parameter logistic model and a log-normal model, respectively. Second, the joint distribution of latent ability and latent speed is defined via a Gumbel copula, which accommodates upper-tail dependency. An EM algorithm with stochastic approximation is implemented for parameter estimation. A simulation study evaluates the proposed method's accuracy in ability estimation under various conditions, comparing it to traditional approaches. By better capturing the dependency structure between latent ability and speed, the proposed method is expected to outperform existing approaches.

This study is among the first to apply copula theory to model the dependency between latent variables, demonstrating its potential for modeling flexible dependencies such as tail dependency and non-normal associations. These findings can extend to broader applications in psychological and educational testing involving multivariate latent variable models.

Modeling process data using latent space models

Tuesday, 15th July - 15:15: Process Data Analysis (MAC: Johnson) - Oral

Dr. Tracy Sweet (University of Maryland), Dr. Xin Qiao (University of South Florida)

Previous studies on process data have treated such data as cross-sectional data but ignored the fact that students' behavioral patterns may change over time (i.e., the problem-solving process). Further, these studies have also aggregated students' actions across all students. Studying the change in behavior patterns as well as the variability in these patterns among students can be useful for diagnosis and remedial teaching. It can also shed light on possible improvements for assessment designs. Therefore, we developed a multilevel longitudinal latent space model (LLSM) for the analysis of process data from PISA 2012 problem-solving items where students "reset" their answers during the problem-solving process. To fit the multilevel LLSM, we created longitudinal network with three time points (i.e., first attempts, intermediate attempts, and final attempts) for each student. Network size at each time point was equal to the number of all possible actions in the item. We fit this model using a fully Bayesian Markov chain Monte Carlo algorithm.

We aim to investigate: 1) the change of the associations between some assessment characteristics and the probability of conducting certain actions over time; 2) variability across students in these associations; and 3) how a multilevel model compares to aggregating across all students. Based on the best fitting model, our preliminary findings suggested time-varying associations between some covariates and probability of carrying out certain actions and that a multilevel version of an LLSM is a better fitting model than the aggregate LLSM.

A utility-maximization framework for joint modeling of time and accuracy

Tuesday, 15th July - 15:30: Process Data Analysis (MAC: Johnson) - Oral

Dr. Weicong Lyu (University of Macau), Ms. Mingya Huang (University of Wisconsin - Madison)

The analysis of process data is gaining increasing attention in psychometric research due to its potential to uncover the cognitive processes underlying different item response patterns. Among various aspects, the trade-off between response time and accuracy in cognitive tests remains a key focus. Previous studies have shown that response time affects accuracy and vice versa. However, these studies often treat one variable as exogenous and the other as endogenous, imposing strong assumptions on the causal direction. In this study, we propose a utility-maximization model based on the 3PL model, which treats both response time and accuracy as endogenously determined, reflecting respondents' optimization of test performance under the constraint of limited total response time. The model offers intuitive predictions on how respondents allocate time across items, aligning with prior findings. Specifically, respondents are expected to spend more time on items that (1) are highly discriminative, (2) have low guessing probabilities, (3) exhibit high sensitivity of success probability to time spent, and (4) have success probabilities near 0.5 without guessing. To evaluate the model's practical utility, we apply it to data from the Trends in International Mathematics and Science Study (TIMSS). Our results show that the observed response times closely align with the model's predictions, with more proficient respondents adhering more closely to their theoretical predictions. This study introduces a novel framework to understand how respondents manage time and accuracy trade-offs under time pressure, offering deeper insights into cognitive processes during testing and advancing the analysis of process data in psychometrics.

Evaluation of partial measurement invariance under sparse ordinal indicators using induced Dirichlet threshold priors

Tuesday, 15th July - 14:30: Differential Item Functioning and Measurement Invariance I (MAC: Thomas Swain) - Oral

Mr. Fatih Ozkan (Baylor University), Mr. Jianwen Song (Baylor University)

Researchers commonly use multi-group ordinal confirmatory factor analysis to assess if a construct is measured on an equal basis across groups. In practice, however, some response categories are rarely—if ever—endorsed, and this creates model convergence issues or forces analysts to drop or merge categories. This study investigates if an induced-Dirichlet prior on threshold parameters can resolve these issues by stabilizing threshold estimation under sparse ordinal data in multi-group scenarios. With controlled simulation, we experiment with (1) the number of items with extremely sparse categories, (2) the degree of sparseness (e.g., 2% vs. 5% endorsement), and (3) the extent of partial measurement invariance—where thresholds differ for some but not other items between groups. We compare three threshold treatments: the induced-Dirichlet prior, sequential normal priors, and a baseline category-collapsing approach. We contrast parameter recovery (loadings, thresholds, factor means), model fit, and convergence statistics (effective sample sizes, R-hat) for both methods. The induced-Dirichlet prior should yield tighter credible intervals, fewer convergence issues, and improved detection of partial invariance, without losing infrequently endorsed categories. Such findings would push Bayesian multi-group methods for ordinal data forward by offering researchers a robust workbench for uncovering true cross-group differences—even when some categories are effectively empty. Beyond simulations, we describe how such procedures can be applied to real datasets where response style differences or cultural norms produce skewed distributions. In totality, this research strives to make evidence-based suggestions to analysts seeking to preserve the full ordinal scale and fully account for partial measurement invariance with sparse categorical responses

Statistical analysis of large-scale item response data under measurement non-invariance: A new alignment method and its application to PISA 2022

Tuesday, 15th July - 14:45: Differential Item Functioning and Measurement Invariance I (MAC: Thomas Swain) - Oral

Dr. Yunxiao Chen (London School of Economics and Political Science), Dr. Chengcheng Li (Microsoft), Dr. Jing Ouyang (The University of Hong Kong), Dr. Gongjun Xu (University of Michigan)

International Large-scale Assessments (ILSAs) collect valuable data on educational quality and performance across countries, enabling country groups to share effective techniques and policies. A key analytical tool for ILSAs is the Item Response Theory (IRT) model, which estimates performance distributions and group rankings. However, a major challenge in IRT calibration is that some items suffer from Differential Item Functioning (DIF), where different groups have different probabilities of endorsing items, controlling for individual proficiency. To address the challenge, we novelly propose an efficient approach to multi-group DIF analysis. This approach is statistically consistent, requires no knowledge about anchor items or reference groups, and provides uncertainty quantification for the group-specific parameters. The application to PISA 2022 demonstrates the effectiveness of proposed method.

Machine learning methods for differential item functioning: A systematic review

Tuesday, 15th July - 15:00: Differential Item Functioning and Measurement Invariance I (MAC: Thomas Swain) - Oral

Mr. Mubarak Mojoyinola (University of Iowa), Mr. Ahmed Bediwy (Univeristy of Iowa), Mr. Juyoung Jung (University of Iowa)

Machine learning (ML) methods are increasingly being used in Differential Item Functioning (DIF) detection due to their ability to handle complex data structures and non-linear relationships with high accuracy (Ebrahimi et al., 2021). However, these methods require large datasets, are prone to overfitting, and tend to be less interpretable compared to traditional techniques. To address the need for evaluating advancements in ML methods for DIF detection, we systematically analyzed studies which used ML methods across various assessment settings.

This study presents a comprehensive systematic review examining the application ML methods for detecting DIF in psychological and educational assessment. Following PRISMA guidelines, we analyzed publications from major databases including ERIC, Web of Science, EBSCOhost, PsycINFO, and PubMed/Medline. Our findings reveal a significant increase in ML-DIF research, with notable publication spikes in 2020 and 2023, demonstrating growing interest in this methodology.

The review categorizes studies by research type, with mixed-method approaches (combining simulation and applied methods) being most prevalent (13 articles), followed by applied research (11 articles) and methodological research (8 articles). We extracted data on measurement models, item types, sample sizes, and evaluation criteria used to assess the relative performance of ML methods against traditional DIF detection techniques.

This systematic review addresses a critical gap in psychometric literature by comprehensively evaluating ML approaches to DIF detection across diverse fields. The findings will inform best practices for detecting measurement bias in educational and psychological assessments while highlighting practical considerations for future methodological innovations in educational and psychological measurement research.

Investigating CogAT 7: Intersectional measurement invariance using ASEM

Tuesday, 15th July - 15:15: Differential Item Functioning and Measurement Invariance I (MAC: Thomas Swain) - Oral

Dr. Qingzhou Shi (Northwestern University), Prof. Joni M. Lakin (The University of Alabama)

The Cognitive Abilities Test Form 7 (CogAT 7; Lohman, 2011) is a comprehensive assessment designed to measure verbal, quantitative, and nonverbal reasoning abilities in students from kindergarten to grade 12. Previous research by Lakin and Gambrell (2012) investigated the CogAT 7's picture-based item formats with a bi-factor model analysis, particularly in younger grades (K-2), to understand their ability to capture specific reasoning domains alongside a general reasoning ability. Building upon this foundation, our study extended the exploration of the CogAT 7 by examining measurement invariance across gender, race, and socioeconomic status (SES) within a bi-factor model. Employing Alignment Structural Equation Modeling (ASEM; Asparouhov & Muthén, 2022), an extension of Alignment Optimization (AO; Asparouhov & Muthén, 2014), we analyzed Level 9 data from the 2010 national standardization of the CogAT 7, comprising 5537 third-grade students. Our findings revealed that 8% of the model parameters were identified as non-invariant, supporting the partial measurement invariance of the Level 9 items of the CogAT 7 in measuring general fluid reasoning, verbal, quantitative, and nonverbal abilities across diverse demographic backgrounds. The application of ASEM in bi-factor models showcased its ability to handle multiple groups simultaneously with complex models. These results contribute to our understanding of cognitive assessment practices, emphasizing the importance of developing and interpreting cognitive tests in ways that account for the multifaceted nature of student identities and experiences. Such insights are crucial for creating equitable assessment frameworks and supporting diverse student populations.

Can measurement invariance be established in a non-ergodic situation?

Tuesday, 15th July - 15:30: Differential Item Functioning and Measurement Invariance I (MAC: Thomas Swain) - Oral

Mr. YeongJin Jo (Yonsei university), Prof. JiHoon Ryoo (Yonsei university)

Statistical analyses assume that group-level results would apply to individuals, a concept related to ergodicity, where between-person variability is assumed to be identical to within-person variability. For ergodicity to be held, (1) stationarity—constant means and variances over time—and (2) identical dynamic systems—all individuals sharing the same statistical properties—must be satisfied. In Molenaar (2008), it was shown that, in the non-ergodic situations (i.e., heterogeneous populations), the impact is not well reflected by standard analysis models (e.g., ANOVA, regression, factor analysis models, etc.). These results were also reported in Molenaar (1997; 1999). In spite of this warning, ergodicity has been ignored. Similar and related issues can be found in measurement invariance. For example, factor analytic approach has been predominantly used to verify measurement invariance (MI) in longitudinal or cross-sectional studies. However, in longitudinal settings, verifying MI over time is known as a complex process, and it is even difficult to meet the conditions required for MI, which makes it challenging to apply in longitudinal models. In Molenaar (1999), it was found that the greater effect of heterogeneous populations, the more it influences model fit, which suggested that this effect is likely to impact the MI verification process.

This study combines two related issues and examines the effects of non-ergodicity and MI by applying Monte Carlo simulations. Furthermore, by analyzing cross-sectional and longitudinal MI validation processes, we aim to empirically demonstrate the challenges posed by heterogeneity and discuss the limitations of current MI validation methods.

From interval scales to scales with intervals

Tuesday, 15th July - 14:30: Conceptual Issues in Measurement (GH: Think 4) - Oral

Prof. Derek Briggs (University of Colorado Boulder)

The assumption that a scale has interval properties is implicit whenever a scale is being used to support inferences about magnitude. Two approaches commonly taken to evaluate the equal-interval assumption differ in their conceptual rationales. In an axiomatic approach one attempts to satisfy the conditions of additive conjoint measurement. In a pragmatic approach one focuses on the extent to which uses of the scale are sensitive to a variety of plausible non-linear transformations. I introduce and illustrate four principles that define a hybrid approach: (1) The meaning of scale intervals should build from a theoretical understanding of the construct of measurement. (2) The intervals between items should be invariant to the criterion used to locate them on the scale. (3) The generalizability of a reference distance should depend upon the strength of item location prediction by design variables. (4) The unit of measurement for a scale should always be interpretable relative to at least one item-based reference interval. The hybrid approach borrows from the axiomatic approach the idea that there need to be falsifiable criteria that can be used to evaluate the extent to which differences at various locations along a scale can be plausibly given an invariant interpretation; it borrows from the pragmatic approach in acknowledging that this invariance will only be approximate, and uses item mapping to establish a variety of candidate distances that differ in their substantive qualitative interpretations.

Tailoring educational assessments for varied student demographics: An initial exploration

Tuesday, 15th July - 14:45: Conceptual Issues in Measurement (GH: Think 4) - Oral

Dr. Sandip Sinharay (Educational Testing Service), Dr. Matthew Johnson (Educational Testing Service)

Standardized assessments are an integral part of modern civilization. However, there are rising concerns about inequities fostered by these assessments (e.g., Bennett, 2023; Buzick et al., 2023; Hughes, 2023). One alternative of standardized assessments is culturally responsive assessments, which are assessments that evaluate students' knowledge, skills, and understanding in a way that takes into account their unique cultural identities. This exploratory study details the adaptation of National Assessment of Educational Progress (NAEP) Grade 8 Mathematics items to make them more culturally responsive to Hispanic cultural contexts. Through a study conducted with several hundred eighth-grade students across several U.S. states, we evaluated the impact of cultural context on student performance and student behavior during testing. The paper concludes with considerations for future research and methodology refinements.

An update on Jaeger & Hendricks: Publishing in psychological measurement

Tuesday, 15th July - 15:00: Conceptual Issues in Measurement (GH: Think 4) - Oral

Mrs. Victoria Quirk (University of Illinois Urbana-Champaign), Ms. Hahyeong Kim (University of Illinois Urbana-Champaign), Ms. Xinchang Zhou (University of Illinois Urbana-Champaign)

Thirty years ago, Jaeger & Hendricks (1994) published an essential resource for graduate students and early career professionals in educational measurement, *The Publication Process in Educational Measurement*. Their paper examined the importance of publishing, the journal publication process, and strategies for selecting a journal in educational measurement.

While Jaeger & Hendricks (1994) remains a crucial guide, its narrow focus on educational measurement, reliance on expert opinion and informal editor communications, and lack of updates since its original publication underscore the need for a broader and more systematic literature review of contemporary psychometric publications.

Our literature review expands on the original article in three ways. First, we broaden the scope beyond educational measurement contexts. In addition to the original authors' chosen educational measurement journals, we also survey research from more broadly focused journals in psychological measurement, e.g. *Multivariate Behavioral Research* and *Structural Equation Modeling*. Second, we provide a systematic review of psychometric literature published in 2024. From these articles and information available on current journal websites, we outline a general model of the modern publication process in psychometrics, identify current trends in study types and keywords, and provide detailed metrics on the contributions and statistical complexity of articles across journals. Third, we update recommendations for current researchers, including guidelines for the digital publication process. By reviewing a broader scope of contemporary publications in psychometrics, we provide a comprehensive overview of the journal selection and publication process for a new generation of graduate students and early career professionals in measurement.

Dynamic value-added in school effectiveness: A model-free approach

Tuesday, 15th July - 14:30: Dynamic Value Added Models (GH: Think 5) - Oral

Dr. Sebastien Van Bellegem (LIDAM/CORE, UC Louvain)

This talk addresses the concept of school effectiveness through the notions of school and teacher effects, as well as value-added. These concepts have traditionally been studied within the framework of hierarchical linear models (HLM).

In this work, we consider longitudinal studies where multiple measurements at the pupil level are available over time. In this context, we introduce the notion of dynamic value-added, which evolves over time, and explore various modeling strategies to define and estimate it. The key questions we investigate are:

- What are the sources of variation in value-added over time? Do these variations reflect genuine changes in school performance, or do they stem from other factors?
- Does an increase in value-added indicate an actual improvement in school effectiveness?

To address these questions, we adopt a model-free definition of value-added, based on conditional expectations. By “model-free,” we refer to a definition that relies solely on properties of the statistical conditional distribution, making it well-defined even in the absence of an explicit statistical model.

When applying this approach to specific models, we employ autoregressive-like HLM modeling. We discuss different modeling strategies to answer the above questions and examine the assumptions required to identify and estimate value-added over time while distinguishing genuine school effects from other sources of variation. Finally, we illustrate our results using data from the Chilean education system.

Temporally dynamic, cohort-varying value-added models

Tuesday, 15th July - 14:30: Dynamic Value Added Models (GH: Think 5) - Oral

Dr. Ernesto San Martin (Pontificia Universidad Católica de Chile)

We aim to estimate school value-added dynamically in time. Our principal motivation for doing so is to establish school effectiveness persistence while taking into account the temporal dependence that typically exists in school performance from one year to the next. We propose two methods of incorporating temporal dependence in value-added models. In the first we model the random school effects that are commonly present in value-added models with an auto-regressive process. In the second approach, we incorporate dependence in value-added estimators by modeling the performance of one cohort based on the previous cohort's performance. An identification analysis allows us to make explicit the meaning of the corresponding value-added indicators: based on these meanings, we show that each model is useful for monitoring specific aspects of school persistence. Furthermore, we carefully detail how value-added can be estimated over time. We show through simulations that ignoring temporal dependence when it exists results in diminished efficiency in value-added estimation while incorporating it results in improved estimation (even when temporal dependence is weak). Finally, we illustrate the methodology by considering two cohorts from Chile's national standardized test in mathematics.

Bayesian approach to modeling dynamic group membership in nonlinear random effects models

Tuesday, 15th July - 14:30: Dynamic Value Added Models (GH: Think 5) - Oral

Prof. Nidhi Kohli (University of Minnesota), Dr. Corissa Rohloff (Human Resources Research Organization (HumRRO)), Prof. Eric Lock (University of Minnesota)

Crossed random effects models (CREMs) are a useful statistical approach for analyzing complex longitudinal data structures, allowing researchers to account for the influence of dynamic group membership on individual outcomes. However, there is a critical gap in the existing literature regarding the specific data conditions required to reliably

identify these models, particularly the group effects, in a longitudinal context. This gap is significant, as future applications to real data must account for these conditions to ensure accurate and precise model parameter estimates, especially concerning the group effects on individual outcomes. Additionally, existing CREMs do not accommodate intrinsically nonlinear growth patterns. Therefore, this study develops a Bayesian piecewise CREM for modeling intrinsically nonlinear growth and identifying the data conditions necessary for the empirical identification of nonlinear longitudinal CREMs. An applied example using

real data, along with three simulation studies, is presented to assess the data conditions for estimating piecewise CREMs. Results indicate that the number of repeated measurements per group significantly influences the ability to recover group effects.

Improving the evaluation of construct change over time: Comparing longitudinal moderated nonlinear factor analysis to the conventional first-order growth model

Tuesday, 15th July - 14:30: What Can Parameter Moderation do for You? (GH: Think 3) - Oral

Dr. Siyuan Marco Chen (Duolingo), Dr. Daniel J. Bauer (University of North Carolina at Chapel Hill)

Conventional growth curve models, often fitted to sum or mean scores of scale responses, do not account for potential changes in item measurement unrelated to construct growth (i.e., differential item functioning; DIF). An untested assumption is that the construct is stably measured over time. When this assumption is incorrect, estimates of construct change obtained from conventional growth models may be biased. To address this issue, we recently proposed a new, flexible second-order growth model based upon a longitudinal extension of moderated nonlinear factor analysis (MNLFA; Chen & Bauer, 2024) that allows for DIF from categorical or continuous covariates that may or may not vary over time (e.g., sex, age, age × sex). Further, we applied Bayesian regularization to evaluate DIF effects across multiple sources simultaneously without imposing item equality assumptions (i.e., anchor items). In this paper we present a simulation study to validate the model's performance in detecting DIF over time and between groups. Results indicate that the proposed approach effectively detects DIF without predetermined anchor items and avoids the biased growth estimates consistently observed for conventional models fitted to mean scores. We demonstrate the utility of the method in an empirical example on child externalizing behaviors.

Advancing a general framework of parameter moderation

Tuesday, 15th July - 14:30: What Can Parameter Moderation do for You? (GH: Think 3) - Oral

Dr. Ethan McCormick (University of Delaware), Dr. Gregory R. Hancock (University of Maryland)

Parameter moderation (e.g., MIMIC models, MNLFA) techniques are often applied as diagnostic tools for assessing the measurement properties of instruments. In both the history of their development and common implementation, these parameter moderation techniques focus on identifying individual parameters, items, or overall measures that show some sort of measurement non-invariance. While this granular focus serves several important goals in measurement science, the promise of parameter moderation for advancing our understanding of phenomena is much broader than individual parameters or measurement science alone. Here we develop a larger framework for parameter moderation, and its uses across a broad class of models, from simple path analytic regression models (e.g., mediation), to measurement modeling, and indeed whole-model information (e.g., the likelihood). Specifically, we demonstrate how the machinery of parameter moderation allows us to ask questions of moderation at different levels of abstraction, ranging from individual model parameters, to complex model-derived information, such as reliability and fit. Our treatment focuses on both practical technical approaches to estimate these effects reliably, as well as the conceptual landscape that these techniques open to investigation.

Individual variability as a moderator of latent structural relations

Tuesday, 15th July - 14:30: What Can Parameter Moderation do for You? (GH: Think 3) - Oral

Mr. Joshua R. Shulkin (University of Maryland), Dr. Gregory R. Hancock (University of Maryland)

Researchers in the social and behavioral sciences, and beyond, are often more interested in intraindividual or intragroup variability as the focus of their research questions rather than in individual or group means. For example, health researchers may wish to examine consistency in the number of hours of sleep adolescents get across a week, or military researchers might focus on homogeneity of squad members' trust in their team leader. Latent random variability models (Feng, 2023; Feng & Hancock, 2024) use multilevel structural equation modeling to parameterize variability (in measured or latent variables) as a latent variable, which allows variability to be incorporated as an outcome, predictor, or even mediator within a broader structural equation model. The current paper adds to the versatility of latent random variability models, employing the principles underlying Moderated Nonlinear Factor Analysis (MNLFA) to allow variability to act as a moderator of latent structural relations, using Bayesian estimation within the Blimp software package (Keller & Enders, 2021). This approach could be used to address questions such as whether the impact of adolescents' motivation on academic achievement is moderated by variability in mood. An example using empirical data is presented to illustrate the estimation procedures and interpretation of relevant moderation parameters.

Generalizing the specification and estimation of item heterogeneity models with parameter moderation

Tuesday, 15th July - 14:30: What Can Parameter Moderation do for You? (GH: Think 3) - Oral

Dr. Sanford Student (University of Delaware)

Recent studies have proposed studying the impact of interventions at the item level instead of or in addition to at the test level. Consider a typical randomized controlled trial in which a control and treatment group have had a relevant outcome measured via a test or survey consisting of multiple indicators of an underlying construct. The premise of these studies is that there is additional information to be learned about the items most or least influenced by the intervention, and that this information can be informative for understanding how, where, and to what extent the intervention works.

This paper focuses on the framework of Ahmed et al. (2024), who approach the estimation of item-treatment interactions through the lens of differential item functioning (DIF) via a multiple-group one-parameter logistic IRT model. This paper demonstrates that the models used in Ahmed et al.'s study are isomorphic with moderated nonlinear factor analytic models that can be estimated as multiple-indicator, multiple-cause models with linear and nonlinear constraints to capture moderation of model parameters. Basing model specification and estimation on an MNLFA approach instead of estimating the models as multiple-group IRT models enables several useful extensions including cluster-robust inference, differing item discriminations, nonuniform DIF representing item-specific variance impact, the extensive capabilities of model constraints in a software such as Mplus, and regularization of parameters to enable simultaneous estimation of a main intervention effect and deviations from this effect that represent item-level heterogeneity. A real data analysis from a randomized controlled trial illustrates these affordances.

On structural misspecifications in latent variable mediation analysis

Wednesday, 16th July - 09:00: Mediation Analysis (GH: Meridian 1-2) - Oral

Mr. Bing Cai Kok (University of North Carolina), Prof. Kenneth Bollen (University of North Carolina), Dr. Oscar Gonzalez (University of North Carolina), Mr. Alejandro Martinez (University of North Carolina)

Recent advances in causal inference have generalized mediation analysis to include non-linear interaction terms. Estimation, however, is complicated in the presence of latent interactions. This issue is particularly salient in psychology where a latent mediating construct is only indirectly accessible via a set of observed indicators. To remedy this, a multiple-group approach has been advocated for when the interventional variable is binary. Yet, this approach has been validated only in scenarios where the model structure is correctly specified. It is unknown how robust this approach is when structural misspecifications - such as missing paths or omitted correlated errors - are present in the mediation model.

In this work, we conducted a simulation study to examine if and how common structural mispecifications can affect conclusions about mediation in this multiple-group setting. Two different estimators were considered: (1) the standard Maximum Likelihood (ML) estimator and (2) the Model-Implied Instrumental Variable (MIIV) estimator. Results indicate that the standard ML estimator works well in the absence of structural misspecifications, but its performance degrades substantially (i.e. relative bias increases) when certain types of structural misspecifications are present. In contrast, we demonstrate how the MIIV estimator can perform well even in the presence of structural misspecifications. These results also extend to cases where models for ordinal data are considered.

Advancing multiple-group mediation using Bayesian regularization

Wednesday, 16th July - 09:15: Mediation Analysis (GH: Meridian 1-2) - Oral

Ms. Emma Somer (McGill University, Montreal, Canada), Prof. Milica Miocevic (McGill University, Montreal, Canada), Prof. Carl Falk (McGill University, Montreal, Canada)

Partial measurement invariance is a crucial prerequisite for comparing structural relationships across multiple groups. Recently, frequentist regularization approaches, such as the lasso and elastic net, have been extended to the measurement invariance framework and involve applying a penalty function to differences across item intercepts and loadings to improve the detection of non-invariant items. Despite their promising performance, frequentist regularization approaches may produce biased estimates and pose challenges for inference due to the unavailability of standard errors in some conditions. To address these limitations, Bayesian regularization methods have recently been extended to the differential item functioning framework. This study builds on previous work by evaluating the performance of Bayesian regularization approaches in terms of the bias, efficiency, and coverage of the indirect effect in a multiple-group single mediator latent variable model. We compare Bayesian regularization approaches – small-variance priors, Laplace priors, Bayesian adaptive lasso, spike-and-slab (SSP), and horseshoe priors. We vary the sample size (N = 200 and 500), proportion (1/3 and 2/3) and magnitude (small or large) of noninvariance, and the value of the indirect effect (*ab* = 0 or 0.144). Preliminary findings suggest that adaptive lasso and horseshoe priors produce lower bias and adequate coverage of the indirect effect under large proportions and magnitudes of noninvariance, whereas small-variance priors experience more difficulty. Additionally, small variance priors had biased latent means and intercepts under some conditions. The study provides recommendations for researchers estimating indirect effects in the presence of measurement noninvariance.

Nonlinear mediation model with Bayesian P-splines

Wednesday, 16th July - 09:30: Mediation Analysis (GH: Meridian 1-2) - Oral

Dr. Qijin Chen (Sun Yat-sen University), Dr. Siyi Wang (Sun Yat-sen University), Prof. Junhao Pan (Sun Yat-sen University), Prof. Edward Ip (Wake Forest University)

Traditionally, mediation analysis has been based on linear hypotheses, which, while useful, often fail to capture the complexity of human behavior. In many applications of behavioral science, linear models are inconsistent with the nonlinear dynamics observed data. For instance, when investigating whether emotional regulation mediates the relationship between work stress and job satisfaction, high emotional regulation may boost job satisfaction under low stress. However, under high stress, even strong emotional regulation may limit or reduce satisfaction. To address the limitations of linear mediation processes, the current research introduces a nonlinear mediation model. Within the framework of structural equation modeling, Bayesian P-splines are used to model the direct effect between the independent variable and mediator, the mediator and outcome, and the independent variable and outcome. The proposed approach does not require prespecifying a particular functional form for these relationships. To solve the complexity in estimating the nonparametric model, data augmentation and Markov chain Monte Carlo techniques are employed. We conducted simulation studies to evaluate the model's performance under various conditions, including different sample sizes and forms of direct effects. The results are compared with those from traditional parametric nonlinear models. Additionally, empirical data about emotion is used to demonstrate how the proposed model can be applied in real-world research.

Investigating the impact of cross-loadings on model fit and mediation inferences in the latent mediator model

Wednesday, 16th July - 09:45: Mediation Analysis (GH: Meridian 1-2) - Oral

Ms. Qiulin Lu (University of South Carolina), Dr. Amanda Fairchild (University of South Carolina), Dr. Dexin Shi (University of South Carolina)

Cross-loadings are a common phenomenon in psychological research. Extant studies have focused primarily on measurement models, like confirmatory factor analysis (CFA), and have shown that omitted cross-loadings can bias the inter-factor correlation. These results suggest that omitted cross-loadings might likewise bias structural paths in other structural equation models (SEMs), but this area remains under-investigated. Such an effect could impact the evaluation of third variable effects, like mediation, when estimated in a SEM framework.

Our study addresses this gap by examining the effects of cross-loadings in single, latent mediator model in a simulation study. We compare the performance of conventional SEM and Bayesian SEM (BSEM) approaches in the presence of omitted cross-loadings on the mediator through a simulation study with manipulation on the location, number, and magnitude of cross-loadings, as well as the magnitude of structural paths in the model. Study outcomes are model fit, estimation bias, Type I error rates, and power in relation to making mediation effect inferences. Results show that the additional modeling flexibility afforded by BSEM yields better statistical performance. Results provide insights into recommended strategies for managing cross-loadings in latent mediator models, as well as suggest directions for future research in more complex mediator models within simulation studies.

Longitudinal heterogeneous mediation analysis with latent mediators and a time-to-event outcome

Wednesday, 16th July - 10:00: Mediation Analysis (GH: Meridian 1-2) - Oral

Dr. Rongqian Sun (Shenzhen University), Dr. Xinyuan Song (Chinese University of Hong Kong)

This study proposes a semiparametric joint modeling approach to conducting mediation analysis with longitudinal multivariate data, addressing the complexities of time-varying latent mediators, a time-to-event outcome, and heterogeneity in the path-specfic effects across individuals. Time-varying latent mediators are characterized from multivariate longitudinal surrogates using factor analysis. We then introduce semiparametric Bayesian ensemble of trees into growth curve models to flexibly capture the unknown trajectory of the time-varying latent mediator and their associations with the exposure and individual covariates. The dynamics of the latent mediators are subsequently linked to the time-to-event outcome through a proportional hazards model with Bayesian additive regression trees. To detect potential heterogeneity, individual-specific interventional effects are identified on the scale of the logarithm of hazards and survival probability across direct and indirect causal pathways. We propose an efficient Markov chain Monte Carlo algorithm to estimate the conditional average interventional effects through the mediation formula. The empirical performance of the proposed methodology is validated through extensive simulation studies. We apply the proposed method to a real-world dataset to further demonstrate its utility in quantifying longitudinal mediation mechanism involving latent variables and potential heterogeneity.

Comparing symptom network structure across multiple psychiatric disorders

Wednesday, 16th July - 09:00: Network Models I (GH: Meridian 3-4) - Oral

Dr. Hao Luo (University of Waterloo), Prof. Björn Andersson (University of Oslo), Dr. Chris Perlman (University of Waterloo), Prof. John Hirdes (University of Waterloo)

The network approach to psychopathology conceptualizes psychiatric disorders as complex dynamic systems of causally connected symptoms and has recently gained prominence as an alternative to the traditional disease model. However, existing empirical investigations are often limited to small samples, a specific psychiatric diagnosis (particularly depression), or a small number of symptoms taken directly from pre-existing scales. Robust evidence on how symptom networks differ between disorders remains scarce. This study aims to generate large-scale evidence on symptom networks across multiple psychiatric disorders, using population-representative data from 325,166 individuals across three Canadian provinces (Ontario, Newfoundland and Labrador, and Manitoba). Data were collected between 2005 and 2024 through the interRAI Mental Health assessment, which is used in inpatient mental health settings to support care planning. The assessment, routinely administered at admission, discharge, every 90 days, and when a clinically significant change occurs, includes 33 mental state indicators that measure the frequency of symptoms observed in the past three days. Examples of symptoms are inflated self-worth, hyperarousal, irritability, obsessive thoughts, hallucinations, and delusions. We estimate network structures across psychiatric disorders (e.g., mood disorders, anxiety disorders, schizophrenia, and substance-related disorders) with L1 regularization by gender and age group. We examine network characteristics, including network connectivity and node centrality, and assess invariance in network structures, edge strength, and global strength with network comparison tests. The findings may have direct implications to the diagnosis and treatment of psychiatric disorders. The research also lays the foundation for future longitudinal investigation on within-subject networks across the continuum of care.

Degree distributions in psychological networks

Wednesday, 16th July - 09:15: Network Models I (GH: Meridian 3-4) - Oral

Dr. Jonathan Park (The University of California, Davis), Dr. Mijke Rhemtulla (The University of California, Davis)

Network analysis has emerged as a novel methodological tool in the social and behavioral sciences for its interpretability and relative simplicity. A key result of a network analysis is a set of estimated centrality indices that quantify the “importance” of individual nodes or symptoms to the network. The relevance of centrality indices has been debated in recent years due to results in support and against the utility of centrality-based measures for identifying meaningful intervention targets in psychological networks.

A missing component in this discussion is on a more fundamental aspect of networks: their underlying degree distributions. Formally, many real-world networks ranging from the electrical grid to biological systems self-organize into degree distributions such as power-laws, small worlds, or random graphs. The degree distribution of a network, then, plays a significant role in whether certain centrality metrics are effective or not as targets of intervention. However, despite this well-established precedent, the degree distribution of psychological networks is still relatively understudied. Further, simulation studies of network analysis do not often discuss their data-generating degree distributions and—often—make use of random sampling for populating the network structures. This may have implications for the relevance of simulation-based results to real-world data if psychological networks do not follow random degree distributions.

To address this gap, we conducted a simulation study where partial correlation networks were generated following the degree distributions. We then assessed the impact of network structure on stability and intervention metrics. Findings and implications are discussed.

Neighborhood selection in cross-sectional network analysis for ordinal data

Wednesday, 16th July - 09:30: Network Models I (GH: Meridian 3-4) - Oral

Mr. Kai Jannik Nehler (Goethe University Frankfurt), Mr. Martin Schultze (Goethe University Frankfurt)

Network estimation in cross-sectional data is most commonly performed using regularization techniques (e.g., Friedman et al., 2008; Williams, 2020). As an alternative, neighborhood selection based on node-wise regressions with model selection via the Bayesian Information Criterion (BIC) has been proposed (Williams et al., 2020). This approach was recently extended to account for missing data (Nehler & Schultze, 2024), thereby broadening its applicability in empirical research. However, existing evaluations have primarily focused on continuous, multivariate normally distributed data. Dedicated neighborhood selection approaches for ordinal data are still lacking (e.g., Isvoranu & Epskamp, 2023), despite the prevalence of such variables in psychological research. In this talk, we discuss the performance of neighborhood selection using BIC with ordinal data, considering both scenarios with and without missing values. Under full data conditions, a customized approach for ordinal data is compared to the existing estimation method, which assumes continuous variables (Williams et al., 2020). In conditions with missing values, we evaluate techniques based on maximum likelihood estimation and multiple imputation, both of which have already demonstrated strong performance with continuous data (Nehler & Schultze, 2024), testing their robustness in situations with ordinal data. The performance of the proposed methods is evaluated through a comprehensive simulation study that varies factors such as network size, number of observations, and proportion of missing data. Evaluation criteria include the accuracy of edge set identification as well as potential biases in estimated edge weights and strength values.

The impact of measurement error in dynamic network models

Wednesday, 16th July - 09:45: Network Models I (GH: Meridian 3-4) - Oral

Ms. Reeta Kankaanpää (University of Turku), Ms. Jill de Ron (University of Amsterdam), Ms. Ria H. A. Hoekstra (University of Amsterdam), Dr. Riet van Bork (University of Amsterdam)

Dynamic network models have become increasingly popular in understanding the development and treatment of mental disorders, as they allow for a more active view of symptom interactions over time. However, many of these models rely on single indicators per node, which can lead to biased estimates due to measurement error. While the impact of measurement error has been widely studied in cross-sectional networks, its influence on dynamic models remains unclear. This study aims to investigate the effects of measurement error on the estimation of network parameters in longitudinal models, focusing on temporal, contemporaneous, and between-subject relations. We hypothesize that measurement error will introduce bias across all levels of the network, with smaller bias observed in models that use multiple indicators per node. To test the hypothesis, we will conduct two simulation studies: one with time-series data (N=1) and another with panel data (N>1). These simulations will compare models that use single indicators to those that incorporate multiple indicators, such as average scores, factor scores, plausible values and latent variables. By examining how measurement error affects the estimation of temporal relations, which are crucial for identifying potential mechanisms for psychological interventions, our study will contribute valuable insights into the design of more accurate and robust longitudinal network models. The results of this study will offer recommendations for applied researchers, particularly in how to minimize bias by using multiple indicators when studying intervention mechanisms over time.

The invariance partial pruning approach to the network comparison in longitudinal data

Wednesday, 16th July - 10:00: Network Models I (GH: Meridian 3-4) - Oral

Mr. Xinkai Du (University of Oslo), Dr. Sverre Urnes Johnson (University of Oslo), Dr. Sacha Epskamp (National University of Singapore)

Network models in time-series and panel data have been powerful tools to investigate the dynamical relations among variables. A common goal of empirical research is to compare the networks of different groups, such as treatment and control, to understand how inter-variable relations are shaped by external factors. However, existing methods for comparing idiographic networks are restricted to global tests, which lack the capacity to identify the precise location of edge heterogeneity. Furthermore, there is a lack of easily applicable methods to compare networks from panel data where just a few time-points are available per person. We therefore present the invariance partial pruning (IVPP) approach, which first evaluates heterogeneity globally with the network invariance test, and then determine the exact locus of heterogeneity at the edge level with partial pruning. Through simulation studies, we discovered that network invariance test based on AIC performed well. BIC performed similarly well but also showed insufficient power to detect smaller true differences at small sample sizes, and LRT was prone to false discovery when detecting the edge differences in sparse networks. Comparison with the fully constrained model revealed superior performance than comparison with the fully unconstrained model. Partial pruning successfully uncovered specific edge difference with high sensitivity and specificity. We conclude that IVPP is an essential supplement to the existing network methodology by allowing the comparison of networks from both time-series and panel data, and also allowing the test of specific edge difference. We implement the algorithm in the R-package IVPP.

AI-GENIE: A simulation study on the fully automatic scale development methodology

Wednesday, 16th July - 09:00: Generative Psychometrics: Advancing Psychological Scale Development Through Large Language Models (MAC: Johnson) - Oral

Ms. Lara Russell-Lasalandra (University of Virginia)

The rapid advancement of artificial intelligence (AI), particularly large language models (LLMs), has introduced powerful tools for various research domains, including psychological scale development. This study presents a fully automated method to efficiently generate and select high-quality, non-redundant items for psychological assessments using LLMs and network psychometrics. Our approach called, Automatic Item Generation and Validation via Network-Integrated Evaluation (AI-GENIE), reduces reliance on expert intervention by integrating generative AI with the latest network psychometric techniques. The efficacy of AI-GENIE was evaluated through Monte Carlo simulations using the Mixtral, Gemma 2, Llama 3, DeepSeek, GPT 3.5, and GPT 4o models to generate item pools that mimic Big Five personality assessment. The results demonstrated improvement in item selection efficiency, with overall average increases of 9.78-17.80 and final values of 71.24 - 94.46 in normalized mutual information (NMI) in the final item pool across all models.

Empirical validation of AI-GENIE

Wednesday, 16th July - 09:00: Generative Psychometrics: Advancing Psychological Scale Development Through Large Language Models (MAC: Johnson) - Oral

Dr. Alexander Christensen (Vanderbilt University), Ms. Lara Russell-Lasalandra (University of Virginia), Dr. Hudson Golino (University of Virginia)

Simulation evidence for AI-GENIE demonstrates a consistent improvement to both the alignment of item content with the theoretical concepts as well as the psychometric robustness of the underlying structure. Despite this evidence, conclusions about the psychometrics are limited to the relationships in the embeddings. This talk discusses the empirical evidence for the psychometric robustness of the items generated by AI-GENIE. Five different large language models (LLMs) were used to generate five Big Five personality short form scales. Each scale was administered to an independent, nationally representative sample (in the U.S.; Ns = 1,000). The results indicate that the theoretical Big Five structures were robust for the majority of the LLMs. This finding provides empirical evidence that supports the connection between the meaning of item content in high-dimensional embeddings and how people interpret them. Broader conclusions are drawn about the application of AI-GENIE to other psychological phenotypes.

Optimizing LLM embeddings for automatic item development and validation

Wednesday, 16th July - 09:00: Generative Psychometrics: Advancing Psychological Scale Development Through Large Language Models (MAC: Johnson) - Oral

Dr. Hudson Golino (University of Virginia)

Large Language Models (LLMs) have shown promise in text clustering and dimensionality analysis through embeddings, yet their potential for optimization remains largely unexplored. We conducted a comprehensive simulation study to enhance the accuracy of LLM embeddings in trait mapping using Dynamic Exploratory Graph Analysis (Dynamic EGA). The simulation generated 200 items across 4 traits of Narcissistic Personality, randomly selecting 3-40 items per dimension. We analyzed 1,040,000 combinations across 260 embedding values (3-1300) in a 1536-dimensional space. Performance was evaluated using Total Entropy Fit Index (TEFI) and Normalized Mutual Information (NMI). Vector field analysis revealed complex dynamics between TEFI and NMI, with optimal performance occurring in regions of moderate TEFI values and NMI above 0.5. The number of items per dimension showed peak performance between 10-20 items, while embedding dimensions exhibited non-linear relationships with both metrics. A weighted scoring system prioritizing NMI (70%) over TEFI (30%) significantly outperformed traditional cross-sectional embedding approaches. The optimization demonstrated improved accuracy in concept mapping while maintaining structural stability, suggesting a promising direction for enhancing LLM-based text analysis methods.

Understanding the predictive capacity of admission test scores across the scale

Wednesday, 16th July - 09:00: Validity (MAC: Thomas Swain) - Oral

Mr. Pablo Espinoza (DEMRE, Universidad de Chile), Dr. Eduardo Alarcon-Bustamante (DEMRE, Universidad de Chile), Dr. María Inés Godoy (DEMRE, Universidad de Chile), Dr. David Torres Irribarra (Pontificia Universidad Católica de Chile)

Predictive capacity studies are typically conducted using a correlation coefficient or a linear regression coefficient, assuming a constant relationship between admission test scores and academic performance across the score scale. However, this assumption may obscure nonlinear patterns in the relationship, such as the stability of performance at higher scores or the difficulty of improving academic outcomes at lower scores.

This study focuses on predicting students' academic progress rate using a beta regression model, which effectively captures the continuous and bounded nature of the dependent variable. To assess the impact of test scores on the progress rate, we employ the marginal effect, which measures the rate of change in academic progress as scores vary. Unlike traditional approaches that summarize predictability with a single coefficient, this study identifies in which segments of the score scale the impact on academic progress is greater or smaller, providing a broader view of the relationship between both variables and enabling a better understanding of the connection between admission test scores and academic performance.

Decomposing the predictive capacity of a selection test

Wednesday, 16th July - 09:15: Validity (MAC: Thomas Swain) - Oral

Dr. Eduardo Alarcón-Bustamante (DEMRE, Universidad de Chile), Dr. María Inés Godoy (DEMRE, Universidad de Chile), Dr. Francis Tuerlinckx (Katholieke Universiteit Leuven)

When evaluating the predictive capacity of a test in the presence of groups, it is important to consider how group structure—such as differences in program selectivity, student composition, or the difficulty of each academic program—may influence the relationship between admission test scores and academic performance. These factors shape both individual outcomes and broader differences across programs, affecting how predictive relationships are interpreted.

Standard approaches typically compare correlation or regression coefficients across groups or estimate a single coefficient assuming a homogeneous population. However, comparing coefficients does not explicitly model how group composition influences predictions, while a single-coefficient approach overlooks structural differences that affect predictability.

We propose a method that explicitly incorporates group structure into the evaluation of predictive capacity. Using a decomposition of the marginal effect based on the law of total probability, we separate the impact of test scores into two components: a within-group effect, which measures how academic performance varies with test scores within each program, and a between-group effect, which captures how test scores explain differences in average predictions across groups and influence the probability of a student belonging to a particular program. Applying a beta regression model, we estimate these effects and assess in which parts of the score scale they are significantly different from zero. This approach provides a better understanding of how test scores impact academic performance within each group and how structural differences between groups influence predictability, revealing patterns that traditional methods fail to capture.

Psychometric measurement of forecasters using the wisdom of the crowd

Wednesday, 16th July - 09:30: Validity (MAC: Thomas Swain) - Oral

Ms. Jessica Helmer (Georgia Institute of Technology), Ms. Sophie Ma Zhu (University of British Columbia), Mr. Nikolay Petrov (University of Cambridge), Dr. Ezra Karger (Federal Reserve Bank of Chicago), Dr. Mark Himmelstein (Georgia Institute of Technology)

In most test situations, items have a fixed ground truth. However, in scenarios involving judgment under uncertainty, the ground truth may be noisy. The answer key in such situations is a random variable, and the optimal response is its expectation. One solution to reducing the noisiness of such answer keys is to use the wisdom of the crowd: the aggregate response from a group of people, rather than the ground truth. To demonstrate this, we simulated 1,000 forecasters, each defined by a skill parameter, theta, drawn from a normal distribution, which determined the expected distance between an item's expected outcome and their response. We generated 1,000 items, simulated each forecaster's distributional forecast on each item, and scored each forecast using a strictly proper scoring rule. We then repeatedly sampled groups of N = 1–64 forecasters and K = 1–64 items and scored their forecasts based on the absolute distance between their response and the aggregate forecast of the subgroup. We found that using the crowd aggregate as the scoring criterion correlated with forecasters' original skill parameter (theta) more strongly than the ground-truth scoring for groups of N = 4 or larger. This difference was especially prominent in combinations of larger samples of forecasters and fewer samples of items. We validate the usefulness of this approach using real data collected during a recent forecasting study. In tests involving uncertain outcomes, psychometric models should consider the wisdom of the crowd as an alternative criterion to the ground truth.

Latent structure discovery in 3D science CAT via matrix factorization

Wednesday, 16th July - 09:45: Validity (MAC: Thomas Swain) - Oral

Dr. Yi-Fang Wu (Cambium Assessment), Dr. Frank Rijmen (Cambium Assessment)

The purpose of this study is to provide construct validity evidence for a three-dimensional (3-D) science assessment administered through an item-cluster-based computerized adaptive test (CAT). Specifically, we examine the convergent and discriminant validity of the test structure by analyzing patterns in student-item interactions. Using response data from nearly 10,000 students in three grade levels, each of whom responded to 18 item clusters, we aim to determine whether students with similar abilities follow consistent test trajectories and whether the adaptivity mechanism preserves the intended construct structure.

To achieve this, we apply implicit matrix factorization techniques to uncover latent dimensions within the student x item-cluster response matrix, acknowledging missing data as inherent to the adaptive assessment. We use Alternating Least Squares (ALS) as a key factorization method while also employing Truncated Singular Value Decomposition (SVD) as a cross-validation step. The dual approach allows us to assess the stability of student representations, helping to verify that the identified latent structure is not an artifact of a single modeling technique. We further employ clustering techniques (e.g., K-means, hierarchical clustering) and dimensionality reduction methods (e.g., t-SNE, UMAP) to evaluate the coherence of student response patterns.

Analysis results will be visualized through latent student trajectories, where we compare factorized student representations across methods. Clustering and projection techniques will reveal whether item-cluster assignments align with ability estimates and intended constructs. Discrepancies between models may indicate construct drift, providing insights into how CAT-based item selection aligns with the intended test construct and supports validity in adaptive assessments.

Development of robust estimation methods for cognitive diagnosis

Wednesday, 16th July - 09:00: Cognitive Diagnostic Models I (GH: Think 4) - Oral

Dr. Daxun Wang (Jiangxi Normal University), Prof. Wenchao Ma (University of Minnesota), Prof. Yan Cai (Jiangxi Normal University), Prof. Dongbo Tu (Jiangxi Normal University)

Cognitive diagnosis enables the rapid, accurate, and detailed assessment of an individual's cognitive processes, processing skills, or knowledge structures, holding significant application potential in academic diagnostic evaluations. In practice, factors such as decreased examinee motivation, fatigue, and cheating can lead to anomalies in the data, which in turn can compromise the accuracy of cognitive diagnostic parameter estimation and the classification accuracy of examinees.

This study proposes a robust method to mitigate the impact of anomalous response data by weighting such data during the parameter estimation process, thereby enhancing the accuracy of cognitive diagnostic parameter estimation and the performance of the algorithm. The results demonstrate that robust cognitive diagnostic estimation methods can improve the precision of item parameter estimation and the accuracy of examinee attribute pattern classification.

Bayesian criterion-referenced diagnostic classification models

Wednesday, 16th July - 09:15: Cognitive Diagnostic Models I (GH: Think 4) - Oral

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

In this paper, we define and evaluate Bayesian methods for the estimating of scale score cutpoints following the two-stage approach of Templin et al. (2024). The basis for this paper is the finding that diagnostic classification models (DCMs) are coarse approximations to item response theory (IRT) models. Templin et al. (2024) showed that the location of the two-stage estimated cutpoints was proximal to the point of maximum separation of classes. Implications are that DCMs are but one choice of a method for determining a cutscore in an IRT model and that this choice utilizes the point of maximum separation. The location of the cutscore for maximum class separation in DCMs is seldom meaningful to applied researchers or practitioners. We demonstrate how Bayesian estimation of IRT models can be augmented to provide the simultaneous estimation of each two-stage cutscore with a prior distribution, providing DCM-like results. The process results in the estimation of the posterior distribution of each cutscore, which can be used to demonstrate the degree of uncertainty in each. We then extend this methodology so that classifications have substantive meaning with respect to defining what mastery is, such as having an 80% chance of answering all items correct. We then implement the mastery definition by imposing item parameter constraints into the Bayesian algorithm. We provide the details of the algorithm and demonstrate it with a simulation study and empirical data analysis comparing results when the cutscore is freely estimated versus when substantive constraints are made.

A Bayesian semi-parametric framework for cognitive diagnostic models

Wednesday, 16th July - 09:30: Cognitive Diagnostic Models I (GH: Think 4) - Oral

Mr. Michel Cordoba (Purdue University)

Cognitive diagnostic models (CDMs) are restricted latent class models designed to identify the discrete latent attributes of test respondents. Over time, many authors have proposed generalizations of CDMs as generalized linear models with a Bernoulli-distributed response variable. However, advancements in assessment technologies have highlighted the need to integrate diverse response data types into these models, particularly count-based and positive-valued distributions.

This study introduces a semi-parametric CDM framework that relaxes assumptions about the data-generating process. The proposed approach employs the biparametric quasi-likelihood family (Wedderburn, 1974), enabling model fitting without assuming a specific distribution for the response variable. The study explores the connection between the quasi-likelihood framework and the generalized Bayesian method to implement a Metropolis-Hastings within Gibbs algorithm and estimate the quasi-posterior distribution of the model parameters. To illustrate the method, the quasi-Poisson and quasi-binomial CDM-DINA models are introduced, followed by a simulation study evaluating their performance.

This methodology addresses overdispersion in count data models and can be naturally extended to various frameworks expressible as generalized linear models, including recently proposed approaches such as ExpCDM. Furthermore, by relaxing the distributional assumption of the data generator, this approach offers a base for exploring a novel Bayesian nonparametric framework for restricted latent class modeling.

Model-based differential item functioning detection in cognitive diagnostic assessments

Wednesday, 16th July - 09:45: Cognitive Diagnostic Models I (GH: Think 4) - Oral

Ms. Song Zhilin (Beijing Normal University), Prof. Ping Chen (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University), Ms. Qing Zeng (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University)

Differential Item Functioning (DIF) analysis is essential for ensuring measurement invariance across subgroups in major psychometric models, including Classical Test Theory (CTT), Item Response Theory (IRT), and Cognitive Diagnostic Models (CDM). In Cognitive Diagnostic Assessment (CDA), DIF analysis ensures fair comparisons between subgroups with identical attribute profiles. However, traditional DIF methods in CDA require two error-prone inputs: pre-defined comparison groups and anchor items. These requirements may introduce bias due to potential errors in these prior inputs, leading to misidentified DIF items and reduced diagnostic accuracy. To address these limitations, this study proposes an enhanced higher-order DINA model with three improvements: (1) automatic assignment of respondents to comparison groups based on latent classes, eliminating the dependency on pre-defined group specifications; (2) integration of L1-regularization for DIF detection without requiring anchor items, utilizing the sparsity assumption that few items exhibit DIF; (3) An Expectation-Maximization (EM) algorithm with optimization techniques to handle the computational challenges of L1 regularization. We conduct simulation studies to evaluate the performance of the proposed method. An application to empirical educational data further demonstrates its practical effectiveness in identifying DIF items.

Variable-length fully Bayesian adaptive testing and its stopping criteria

Wednesday, 16th July - 09:00: Computer Adaptive Testing I (GH: Think 5) - Oral

Dr. Luping Niu (National Council of State Boards of Nursing), Prof. Seung Choi (The University of Texas at Austin)

Currently, many computerized adaptive testing (CAT) systems rely on static item parameters derived from point estimates obtained during item pool calibration. However, ignoring the uncertainty in these estimates may cause an underestimation of error in trait estimation. A fully Bayesian CAT method addresses this problem by incorporating the uncertainty of item parameter estimates into the estimation of trait levels and has been compared to traditional CAT under fixed-length conditions.

This study explored the fully Bayesian CAT method in the context of variable-length CAT, where examinees receive a different number of items based on predefined stopping criteria. The study proposed several modified stopping criteria tailored to the fully Bayesian algorithm, such as the posterior standard deviation, change in trait estimates, and predicted standard error reduction (PSER). Additionally, the study introduced an approximated PSER technique to enhance computational efficiency and stability. By contrasting the fully Bayesian PSER rule with its approximated alternatives, the study demonstrated the potential for balancing accuracy and efficiency in practical settings.

Extensive simulations were conducted to evaluate the performance of these stopping criteria across different calibration sample sizes and ability levels. Results highlighted the stability and efficiency of the approximated PSER approach, making it a viable option for real-world applications. Lastly, the study pinpointed the most effective stopping criterion for the fully Bayesian CAT method regarding test length and estimation accuracy, as evidenced by comprehensive simulation results.

Two-level adaptive testing with polytomous items

Wednesday, 16th July - 09:15: Computer Adaptive Testing I (GH: Think 5) - Oral

Prof. Seung Choi (The University of Texas at Austin), Dr. Luping Niu (National Council of State Boards of Nursing)

Many computerized adaptive testing (CAT) methods used to assess patient-reported outcomes (PRO) have focused on one dimension at a time. However, in clinical research, evaluating a patient's health often requires multiple measures, each contributing to a multivariate perspective. The two-level CAT model for polytomously scored items offers a practical solution for measuring the related traits commonly found in psychological testing and PRO measurement. This study utilizes a real-world measure of mindfulness to demonstrate the effectiveness of this new approach. We illustrate how item selection within a domain and sequencing across domains function by comparing the two-level CAT algorithm with the traditional CAT method. The performance comparison between the fully Bayesian (FB) two-level model and the conventional EAP estimation method reveals intriguing results. The simulation study indicates that the two-level model outperforms its one-level counterpart in estimating accuracy across all five domains. Additionally, the FB algorithm provides a more accurate estimation of measurement error compared to the EAP algorithm, underscoring its advantages. This study suggests that the two-level CAT approach offers significant benefits in psychological testing and PRO measurement, enabling accurate estimations with fewer items in fixed-length tests. It invites further research to explore whether these advantages remain consistent in different testing situations, potentially expanding its applicability in other adaptive testing scenarios. Adopting this approach could transform how psychological assessments and health outcomes are measured, resulting in valuable outcomes in both research and practice.

Item parameter estimates with non-ignorable missing data patterns in CATs

Wednesday, 16th July - 09:30: Computer Adaptive Testing I (GH: Think 5) - Oral

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

Most estimation algorithms assume ignorability in the missing data pattern, that is the probability that a data value is missing depends only on the observed values. In the case of ignorability, marginal maximum likelihood (MML) estimates of item parameters are consistent. Data arising from adaptive tests are ignorable provided that the entire response pattern is included in the calibration. This attribute follows from the fact that adaptive testing algorithms determine item assignment (i.e., the missing data pattern) based only on responses to earlier items (i.e., the observed values). However, one could easily obtain situations where the ignorability assumption is violated, such as: 1) constructing an adaptive test from different sections, where items in a given section depend on responses to prior sections but sections are calibrated separately; or 2) omitting certain items from a CAT in calibration, such as those that evidence misfit or poor item statistics. Prior research has discussed the implications of non-ignorable missing data in the context of multidimensional MST models (e.g., Jewsbury & van Rijn, 2020). We generalize these results to the setting in which both items and persons are modeled as random samples from populations, and the calibration model may include regression models of latent item or person attributes on observable characteristics. We establish under what conditions missing data are or are not ignorable, and demonstrate how latent regression specifications can be used to restore ignorability in certain sequential CAT designs. We will include simulations to demonstrate the results and their implications.

Linear-on-the-fly testing (LOFT): Which design to choose?

Wednesday, 16th July - 09:45: Computer Adaptive Testing I (GH: Think 5) - Oral

Prof. Susan Embretson (University of Kansas), Dr. Alexander Thierbach (DCS Corporation), Dr. Julia Walsh (DCS Corporation), Dr. John Trent (Air Force Personnel Center), Dr. Thomas Carretta (Air Force Research Laboratory)

While advances in computerized adaptive testing (CAT) have addressed many of the drawbacks associated with traditional linear fixed-length high-stakes tests, the high costs of developing a large item bank often is a barrier for some organizations. An effective compromise between the limitations of traditional testing designs and the expenses of CAT is the linear-on-the-fly test (LOFT) design. LOFT is typically applied to the set of items developed for two or more parallel forms. LOFT permits dynamic, real-time selection of items or groups of items from the set, facilitating the creation of multiple test forms with specified test lengths. Trait estimates in LOFT are based on pre-calibrated item response theory parameters for the specific items selected for each examinee. There are various strategies for designing a LOFT. The most extreme approach involves random selection of items from the set of items for each examinee, which can lead to extreme differences in item parameters. A more controlled design utilizes item bins along with a technique called “dipping.” In this method, items are categorized into bins arranged by difficulty, and a computer algorithm randomly selects one item from each bin, resulting in more consistent and comparable test forms. Despite these innovations, there is a notable lack of literature on the impact of varying LOFT designs. The present study conducted a Monte Carlo simulation examining measurement precision and bias of several common LOFT designs, as well as parallel forms. The results from the study have important implications for the selection of the LOFT design.

Efficient online item parameter estimation in small-sample CAT using gradient descent methods

Wednesday, 16th July - 10:00: Computer Adaptive Testing I (GH: Think 5) - Oral

Dr. Zichu Liu (University of Georgia), Mr. Cong Cheng (University of Georgia), Dr. Yuan Ke (University of Georgia), Dr. Shiyu Wang (University of Georgia)

Online calibration, which involves embedding new items in operational tests and calibrating item parameters on-the-fly in computerized adaptive testing, offers an efficient and cost-effective approach to calibrating new items. However, many existing online calibration methods focus on using large samples to achieve precise item parameter estimation, a strategy that is often impractical in real-world scenarios where sample sizes in classrooms or schools are typically limited. This study addresses the challenge of small sample sizes by developing a practical and efficient approach to online item parameter estimation. Our proposed method is based on a family of gradient descent algorithms, incorporating three types of updating strategies: One-Time updating (OT), Multi-Time at Batch updating (MT-B), and Multi-Time at Individual updating (MT-I). Specifically, the OT method updates the parameters once per batch of students, the MT-B method updates the parameters multiple times per batch, and the MT-I method updates the parameters once per student. We conducted a series of simulation experiments to evaluate the performance of the proposed methods under various algorithm related factors (e.g., tunning parameters and batch size) and testing conditions (e.g., sample size, number of new items) reflecting small-scale CAT applications. The results show that, compared to existing methods, the proposed approach yields accurate item parameter estimates even with small sample sizes, demonstrating its significant practical applicability.

Joint consistency of a multidimensional nonparametric continuous response model

Wednesday, 16th July - 09:00: IRT estimation (GH: Think 3) - Oral

Mr. Mauricio Castillo (Universidad de la República (UdelaR)), Dr. Leonardo Moreno (Universidad de la República (UdelaR)), Dr. Laura Aspirot (Universidad de la República (UdelaR)), Dr. Pilar Rodriguez (Universidad de la República (UdelaR)), Dr. Mario Luzardo-Verde (Universidad de la República (UdelaR))

Ramsay (1991) aimed to estimate the Item Characteristic Curve (ICC) using kernel nonparametric regression, which can be considered a functional data analysis technique. This approach is justified since this model allows for greater flexibility in the shape of the ICCs, has fewer assumptions, and eliminates correlation issues by not estimating parameters. Douglas (1997) proved joint consistency in the one-dimensional case. Luzardo & Rodriguez (2015) focused on the dichotomous model where the trait is multidimensional and established conditions to obtain joint consistency. The extension to polytomous items is straightforward. Additionally, Luzardo (2019) extended the model to obtain an isotonic estimator. In this study, we present a multidimensional continuous-response nonparametric model and establish the conditions required including the choice of kernel functions, bandwidth selection, sample size, and test length, to ensure joint consistency in trait estimation and ICC estimation. To analyze joint consistency, we employ a method based on a triangular array of ICCs and ability estimates, considering the number of examinees as a function of the number of items. The asymptotic theory examines the adequacy of ICCs and ability estimates as test length approaches infinity. Additionally, we present simulations to assess estimation accuracy as a function of test length and sample size.

Generative adversarial networks for high-dimensional item factor analysis

Wednesday, 16th July - 09:15: IRT estimation (GH: Think 3) - Oral

Mr. Nanyu Luo (University of Toronto), Dr. Feng Ji (University of Toronto)

Advances in deep learning and representation learning have transformed item factor analysis (IFA) in the item response theory (IRT) literature by enabling more efficient and accurate parameter estimation. Variational Autoencoders (VAEs) have been one of the most impactful techniques in modeling high-dimensional latent variables in this context. However, the limited expressiveness of the inference model based on traditional VAEs can still hinder the estimation performance. We introduce Adversarial Variational Bayes (AVB) algorithms as an improvement to VAEs for IFA with improved flexibility and accuracy. By bridging the strengths of VAEs and Generative Adversarial Networks (GANs), AVB incorporates an auxiliary discriminator network to reframe the estimation process as a two-player adversarial game and removes the restrictive assumption of standard normal distributions in the inference model. Theoretically, AVB can achieve similar or higher likelihood compared to VAEs. A further enhanced algorithm, Importance-weighted Adversarial Variational Bayes (IWAVB) is proposed and compared with Importance-weighted Autoencoders (IWAE). In an exploratory analysis of empirical data, IWAVB demonstrated superior expressiveness by achieving a higher likelihood compared to IWAE. In confirmatory analysis with simulated data, IWAVB achieved similar mean-square error results to IWAE while consistently achieving higher likelihoods. When latent variables followed a multimodal distribution, IWAVB outperformed IWAE. With its innovative use of GANs, IWAVB is shown to have the potential to extend IFA to handle large-scale data, facilitating the potential integration of psychometrics and multimodal data analysis.

Some standard errors of polytomous item response theory models

Wednesday, 16th July - 09:30: IRT estimation (GH: Think 3) - Oral

Prof. Seock-Ho Kim (University of Georgia)

Computational procedures required to calculate the asymptotic standard errors of the parameters in the various polytomous item response theory models of the logistic form are described and used to generate values for some common situations. Sample sizes needed to obtain a set of stable estimates can be inferred for the respective polytomous item response theory models.

Weighted likelihood estimator and its standard errors for sequential IRT models

Wednesday, 16th July - 09:45: IRT estimation (GH: Think 3) - Oral

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

Sequential item response theory (SIRT; Tutz, 1990) models can be used in many contexts, for example modeling multiple-attempt procedures such as answer-until-correct. Recently, Lu et al. (2025) demonstrated that using a multiple-attempt procedure with SIRT models can increase item information and thereby shorten the test. Furthermore, applying a multiple-attempt procedure to computer adaptive testing is feasible, making it particularly advantageous to use SIRT models. In such contexts, especially when the test length may be short, the weighted likelihood estimator (WLE) often serves as a desirable alternative to maximum likelihood and Bayesian estimators due to its reduced bias. However, little research has addressed WLE for SIRT models. In this study, we first derive the WLE under sequential IRT models, along with its asymptotic standard errors (both the traditional version and that proposed by Magis, 2016). Second, we investigate the properties of WLE for SIRT models, with particular emphasis on the finiteness of the estimator. Third, we conduct a simulation study to illustrate the advantages and disadvantages of WLE, especially in the context of short test lengths.

Artificial neural networks excel in predicting missingness in psychometric data

Wednesday, 16th July - 13:30: Artificial Intelligence II (GH: Meridian 1-2) - Oral

Mr. Longfei Zhang (Beijing Normal University), Prof. Minjeong Jeon (School of Education & Information Studies, University of California Los Angeles), Prof. Ping Chen (Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University)

Missing data is pervasive in educational and psychological assessments. Naive remedies—such as discarding incomplete cases or leaving them as is—often degrade research validity and misinform subsequent decisions. Recent advances in Artificial Neural Networks (ANNs) have demonstrated their efficacy in prediction tasks by leveraging observed features to infer unknown values. Building on this potential, we developed an iterative ANN-based imputation method—Iterative Neural Network Imputation (IterNNImp)—to accurately predict missing entries in psychometric data column by column. Simulation studies showed that IterNNImp reliably estimated missing responses, yielding accurate item means, inter-item covariances, and person and item parameters under Item Response Theory (IRT). Furthermore, the method was validated using multiple empirical datasets across educational and psychological contexts. We conclude by providing guidelines for implementing ANN-driven imputation and discussing directions for refining this novel approach.

Using knowledge graphs to better understand test construction

Wednesday, 16th July - 13:45: Artificial Intelligence II (GH: Meridian 1-2) - Oral

Dr. Magdalen Beiting-Parrish (Federation of American Scientists)

Test blueprints are used to help developers ensure that all the necessary content domains and learning objectives are represented and that there are sufficient items. The traditional test blueprint is usually a document that represents the test items, but it is often hard to truly understand the concepts and skills within the test nor how they relate to each other in an easily digestible format. One tool that may help psychometricians to better understand how the content of the items and the domains they represent relate to each other is the Knowledge Graph (KG). These are similar to concept maps; however, unlike a traditional concept map in which the author determines the relationships between ideas, the text that represents the information is first converted into numerical representations that are used to help quantify the relationships between concepts to create the KG. This process helps to define existing relationships and can reveal relationships between concepts that were previously unknown. This paper proposes the use of KGs to visualize the underlying relationships between test items and their associated metadata with the aim of exploring three different questions: 1) Does the text of the items and their metadata represent the test blueprint? 2) What are the unexpected relationships between the items and are there items that represent multiple domains/skills? 3) Are KGs the same across tests that claim to measure the same content? To test these questions, a representative dataset that contains released items from grades 3-8 across 30 U.S. states is used.

Enhancing model generalizability for understanding CPS behaviors across different tasks: A large language model-based approach

Wednesday, 16th July - 14:00: Artificial Intelligence II (GH: Meridian 1-2) - Oral

Prof. Mengxiao Zhu (University of Science and Technology of China), Mr. Li Feng (University of Science and Technology of China), Dr. Mo Zhang (Educational Testing Service), Mr. Han Zhao (University of Science and Technology of China), Dr. Qizhi Xu (University of Science and Technology of China)

Collaborative problem-solving (CPS) competence is widely recognized as one of the most important competences in the 21st century, particularly in professions that require teamwork and collaboration. The collaborative process is inherently complex, involving both cognitive and social components. To better analyze and understand this process, it has become increasingly important to encode explicit behaviors into CPS skills. In the past, manual annotation was the primary method for coding CPS skills. However, manual annotation is not only time-consuming and labor-intensive, making it difficult to scale to meet the demand for large datasets and ensure timely annotations. As a result, researchers have developed various automatic annotation methods using traditional machine learning and deep learning approaches. While these methods may yield good results within the specific dataset they were trained on, they often fail to generalize well to new datasets. In this study, we propose a novel automatic annotation method for CPS that leverages large language models (LLMs). This approach shifts the CPS skills coding task from a discriminative framework to a generative one, offering a more flexible and scalable solution. Specifically, we fine-tune the LLM simultaneously on a set of diverse datasets. Additionally, we introduce a retrieval module into the prompt, allowing the model to link new CPS behaviors with previously seen CPS behaviors in training data. Our method achieves more accurate results on multiple CPS datasets, showing its effectiveness and generalizability.

Enhancing citation accuracy: Leveraging the NCBI API to verify and correct AI-generated references

Wednesday, 16th July - 14:15: Artificial Intelligence II (GH: Meridian 1-2) - Oral

Dr. Ting Wang (American Board of Family Medicine)

Academic citations generated by AI models often exhibit inaccuracies, including hallucinated references, incorrect author names, and improper formatting. These issues undermine the credibility of AI-assisted writing and pose challenges for researchers relying on automated citation tools. This presentation introduces a method that integrates the NCBI API to validate, retrieve, and correct citation formats, ensuring alignment with authoritative bibliographic databases.

Our approach follows a structured workflow: (1) Extract AI-generated citations, including author names, journal titles, publication years, volume/issue numbers, and DOIs. (2) Query the NCBI Entrez API using key metadata (e.g., DOI, PubMed ID, or title-based search) to retrieve the correct citation from an authoritative source. (3) Compare the AI-generated reference with the verified citation and identify discrepancies. (4) Automatically correct errors in authorship, journal names, page numbers, and formatting based on the retrieved data. (5) Output the final citation in AMA or APA style.

Preliminary testing on a dataset of 25 AI-generated citations revealed that fewer than half contained errors, with the most common issue being incorrect DOIs, followed by inaccurate author attributions and fabricated references. Our API-driven correction method resolved all formatting errors and accurately identified non-existent references, returning appropriate null results when citations were hallucinated. This approach enhances citation reliability by ensuring AI-generated references are verifiable and correctly formatted.

Beyond citations, this framework can be extended to fact-checking AI-generated outputs, such as legal citations and policy documents. Future work will refine search heuristics, integrate additional bibliographic APIs, and scale this method for broader adoption in research workflows.

Statistical curvature and algorithmic convergence in factor models

Wednesday, 16th July - 13:30: Structural Equation Modeling I (GH: Meridian 3-4) - Oral

Dr. Francis Tuerlinckx (KU Leuven, University of Leuven), Mr. Zihao Cao (KU Leuven, University of Leuven)

Common statistical models in psychometrics are examples of curved exponential family models (e.g., factor models, structural equation models, item response theory models). These models belong to a statistical manifold—a differentiable manifold endowed with a Riemannian metric given by the Fisher information. This manifold is a submanifold of the exponential family, and its curvature with respect to the ambient manifold can be quantified using the statistical curvature.

In this talk, we examine how the curvature of simple factor models influences the convergence properties of iterative algorithms. Specifically, we investigate whether greater statistical curvature systematically leads to slower or more unstable convergence, providing insights into the computational challenges posed by highly curved models.

A empirical Bayesian solution to the estimation of the covariance matrix of sample covariances in SEM

Wednesday, 16th July - 13:45: Structural Equation Modeling I (GH: Meridian 3-4) - Oral

Dr. Hao Wu (Vanderbilt University), Dr. Han Du (University of California, Los Angeles)

Real data are unlikely to be exactly normally distributed. Ignoring non-normality will cause misleading and unreliable parameter and standard error estimates as well as model fit statistics. For non-normal data, researchers have proposed a Distributionally-weighted Least Squares (DLS) estimator to combines the normal theory Iteratively Reweighted Least Squares (IRLS) estimation and Weighted Least Square (WLS) estimation. The key in DLS is to select an optimal weight to compute a weighted average of IRLS and WLS. To better estimate this weight, we propose an empirical Bayesian solution. When data were normal, DLS and IRLS provided similar Root Mean Square Errors (RMSEs) and biases of the standard error estimates, and they were smaller than those from WLS. When the data were elliptical or skewed, DLS generally provided the smallest RMSEs and biases of the standard error estimates. Additionally, the Type I error rates of the test statistic using DLS were generally around the nominal level.

Robust estimation of structural equation models

Wednesday, 16th July - 14:00: Structural Equation Modeling I (GH: Meridian 3-4) - Oral

Dr. Max Welz (University of Zurich), Prof. Patrick Mair (Harvard University), Dr. Andreas Alfons (Erasmus University Rotterdam)

Structural Equation Models (SEMs) are typically fitted to a given correlation matrix. Popular choices for this correlation matrix are the sample correlation matrix or the polychoric correlation matrix, both of which are usually estimated from a sample of Likert-type rating data. However, noisy or low-quality observations might be present in the data, such as (but not limited to) careless responses. We show that the presence of such responses can introduce a sizable bias in the estimated correlation matrix. We demonstrate that this bias is inherited by the SEM estimate, possibly leading to worse model fit and biased estimates of factor structure. As a remedy, we propose to use a polychoric correlation matrix estimated in a robust way, which has recently been developed by Welz, Mair & Alfons (2024). This robust estimator generalizes the commonly used maximum likelihood estimator of polychoric correlation while coming at no additional computational cost. We show through simulation studies and empirical applications that fitting a SEM to a robustly estimated polychoric correlation matrix can substantially improve SEM fit, enhance the accuracy of parameter estimates, and help identify potentially low-quality responses. In particular, we demonstrate how the fit of commonly used SEM estimators such as maximum likelihood or least-squares-based approaches like Diagonally Weighted Least Squares (DWLS) can be improved by using a robustly estimated polychoric correlation matrix. Our proposed procedure is implemented in the free open-source R package “robcat”, whose source is written in fast and efficient C++ code.

Bayesian item factor analysis when indicator variables have skewed marginal distributions

Wednesday, 16th July - 14:15: Structural Equation Modeling I (GH: Meridian 3-4) - Oral

Dr. Noah Padgett (Harvard University), Dr. Sonja Winter (University of Missouri)

Bayesian estimation approaches for item factor analyses (IFA) can provide a valuable tool to overcome difficulty in parameter estimation in frequentist approaches when the marginal response distribution is highly skewed. Depending on how the latent response distribution is parameterized, skewed marginal response distributions are problematic in Bayesian IFA models with binary factor indicators. The standard approach to parameterizing the latent response distribution uses a fixed residual variance making prior specification on factor loadings and factor (co)variances straightforward. However, research has shown that the total variance of the latent response distribution can be unstable when the marginal distribution of the observed indicators is highly skewed. We provide an alternative implementation of Bayesian IFA under a DELTA, unit total variance, parameterization of the latent response distribution. Our presentation will demonstrate how to parameterize the latent response distribution with a fixed total variance. The approach is expected to improve coverage rates of credible intervals for threshold parameters by constraining the parameter space

Regularized exploratory structural equation modeling for multiblock data

Wednesday, 16th July - 14:30: Structural Equation Modeling I (GH: Meridian 3-4) - Oral

Ms. Tra Le (Tilburg University), Prof. Katrijn Van Deun (Tilburg University)

The next-generation approach to behavioral research relies on intensive data collection from multiple disciplinary domains. Behavior and cognition are no longer studied from the psychological perspective only but also from other disciplinary perspectives such as environmental, social, clinical, and biomolecular. This often leads to so-called high-dimensional multiview data. In analyzing this type of data, it is of great importance to disentangle distinct mechanisms underlying each data block from common mechanisms shared by all (or multiple) data blocks. Current latent variable methods are not appropriate to address this challenge. To this end, we propose a Multiblock Regularized Exploratory Structural Equation Modeling method (RegularizedESEM). The method adopts the approximate factor model framework with hard cardinality constraint (instead of a penalized approach such as the group lasso) to impose simple structure across and within data blocks. That is, the model is estimated under the constraint that the loading matrix has a fixed specific-shared structure to identify the different types of mechanisms. In addition, exactly $\boldsymbol{K}$ loadings are imposed to be zero to encourage variable selection to ease interpretation. Both factor scores and loadings are estimated in an alternating optimization scheme. The performance of the proposed method is evaluated in an extensive simulation study. We also demonstrate the use of the method using a real-world dataset.

The hidden journey of affect dynamics: Bridging physiological and behavioral states through Markovian processes

Wednesday, 16th July - 13:30: Methods for Dynamic and Complex Data (MAC: Johnson) - Oral

Dr. Francesca Borghesi (Department of Psychology, University of Turin, Turin), Prof. Pietro Cipresso (Department of Psychology, University of Turin, Turin)

Emotions are among the most elusive aspects of human experience—fleeting and contingent upon stimuli, continuously shifting from one state to another. For centuries, literature, philosophy, and psychology have attempted to describe emotions separately, outlining their behavioral and physiological traits. Traditional frameworks like Ekman's discrete emotions and Russell's circumplex model provide static representations but fail to capture their transient nature. Recently, psychometrics has witnessed the emergence of Affect Dynamics, which aims to define emotions temporally. The predominant experimental design is the Experience Sampling Method (ESM), an intensive longitudinal data collecting technique where individuals rate their emotions multiple times per day, week, or month.

Our study introduces discrete-time Markov processes, a mathematical framework based on stochastic transition chains that estimate the probability of moving from one affective state to another, without retaining memory of prior transitions. We leverage Hidden Markov Chains (HMCs) to identify latent affective states and their transition probabilities. HMCs enable the discovery of underlying affective structures that govern the observed data, whether derived from physiological signals (Heart Rate Variability) or self-reported assessments. By applying HMCs separately to physiological and self-reported affective data, we investigate the degree of alignment between the two, assessing whether subjective experiences correspond to underlying physiological dynamics.

This study explores the use of Markovian modeling in Affect Dynamics to examine affective transitions, providing insights into short-term emotional fluctuations. The current approach may inform the development of computational models of affective states, with potential applications in emotion regulation and flexibility research and their psychophysiological underpinnings.

Optimizing ESM prompt timing with mobile sensing: Predicting non-compliance using behavioral and contextual data

Wednesday, 16th July - 13:45: Methods for Dynamic and Complex Data (MAC: Johnson) - Oral

Dr. Koen Niemeijer (KU Leuven, University of Leuven), Dr. Merijn Mestdagh (KU Leuven, University of Leuven), Prof. Peter Kuppens (KU Leuven, University of Leuven)

The Experience Sampling Method (ESM) is the gold standard for capturing emotions and experiences in daily life, but its frequent self-report place a high burden on participants. This burden can lead to decreased compliance, increased dropout rates, and biased data. Mobile sensing, which passively collects behavioural and contextual data via smartphones, offers a promising solution by reducing the likelihood of prompting participant at inopportune moments when they are less likely to respond. Thus, this study combines mobile sensing with ESM to investigate whether mobile sensing data can predict moments when participants are unlikely to respond to ESM prompts. We recruited 104 participants for a three-week ESM and mobile sensing study. Using the m-Path Sense app, we collected a wide variety of sensor data, including accelerometer readings, GPS location, and phone usage, alongside self-reported emotional states. We then developed models to predict whether a participant would respond to a prompt, using both person-specific and population-level approaches, with mobile sensing features—such as the presence of physical activity or location familiarity—as predictors. One methodological challenge in this study is the class imbalance caused by high compliance rates, with over 87% of prompts receiving responses. To address this, we employ resampling techniques and evaluate model performance using metrics robust to class imbalance, such as Matthew's correlation coefficient. Preliminary results suggest that both person-specific and population-level models exceed random chance. Remaining analyses will focus on identifying the behaviours most predictive of non-compliance and will lay the groundwork for implementing these algorithms in real-time ESM apps.

Interpreting parameters of dynamic regression models

Wednesday, 16th July - 14:00: Methods for Dynamic and Complex Data (MAC: Johnson) - Oral

Dr. Sigert Ariens (KU Leuven, University of Leuven), Prof. Ginette Lafit (KU Leuven, University of Leuven), Prof. Eva Ceulemans (KU Leuven, University of Leuven)

The analysis of time series of psychological data has become a popular topic in the recent methodological literature. Of much interest is the ability to investigate how specific contextual variables influence the psychological process over time. To investigate such research questions, structural dynamic regression models of various types (such as DSEM, RDSEM, SVAR,..) are often fit to the time series data.

Although these models have a lot of potential for answering research questions about psychological dynamics, it can be difficult to interpret the model parameters. In this talk, we provide a visual tour through a broad family of dynamic regression models, using model-implied impulse responses to show how the parameters determine the ‘shape’ of the dynamics. This information can be used to select a statistical model which appropriately translates the theoretical assumptions of the researcher. In addition, we introduce a user friendly Shiny application which can be used to understand the results of a dynamic regression analysis in a visual, intuitive way.

Accounting for interindividual differences in intensive longitudinal data from new samples using pre-learned embeddings in machine learning models

Wednesday, 16th July - 14:15: Methods for Dynamic and Complex Data (MAC: Johnson) - Oral

Dr. Sy-Miin Chow (Pennsylvania State University)

Embeddings are a powerful tool in machine learning for representing complex, high-dimensional data in a lower-dimensional space while preserving meaningful relationships. In the context of intensive longitudinal data with heterogeneity in change patterns, embeddings enable the transformation of such heterogeneity into structured, lower-dimensional representations of interindividual differences that facilitate personalized predictions, clustering, and downstream analyses. Using data simulated with a nonlinear predator-prey model with mixed effects in key dynamic parameters as predicted by covariates, I evaluate the utility of adding embedding layers in deep learning models to capture such individual differences, and the predictive efficacy of the trained models when applied to data from new samples. Results from predictions with training and independent test data sets, and the extent of variability of the optimized models across Monte Carlo replications are reviewed. Insights on interpretable ways to summarize the embedding results to enhance both theoretical understanding and practical applications of machine learning models are discussed.

Comparing fit indices for predicting difficulty of automatically generated items

Wednesday, 16th July - 13:30: Item Response Theory II (MAC: Thomas Swain) - Oral

Mr. Haoyang Yu (University of Kansas), Prof. Susan Embretson (University of Kansas)

Fit indices for explanatory item response theory (IRT) models have become increasingly important with the prominence of automatic item generation (AIG). Although AIG can reduce item production costs, it is unclear if empirical tryout can be avoided for new items. It is possible to use data from previous tryouts to estimate the relative impact of varying content features on item parameters and then the results to predict parameters for newly generated items. However, the observed level of predictability is an important factor to consider, as it can impact subsequent trait estimates (Pezeshki & Embretson, 2025). This study focuses on fit indices from different procedures for estimating item parameter predictability. Often item predictors and overall fit is obtained by applying multiple regression to item difficulty estimates obtained from previous test data that often includes relatively few items. Such results are likely to be unstable, due to the sample size being the number of items. In contrast, explanatory IRT estimates based on linear logistic test model (LLTM) and the delta fit statistic (Embretson, 1997; 2016) are full information estimates and likely to be more stable if applied to the data. In the current study, a simulation was conducted to compare the fit indices obtained from the multiple regression method versus the IRT based method. Sample size, test length and prediction levels were varied and the estimates of prediction levels and weights were compared. The IRT method, LLTM combined with the delta statistic, provided consistency better estimates of prediction levels.

Optimal designs for Thurstonian IRT models based on linear paired comparisons

Wednesday, 16th July - 13:45: Item Response Theory II (MAC: Thomas Swain) - Oral

Prof. Heinz Holling (University of Muenster)

The development of optimal designs for Thurstonian IRT models based on linear paired comparisons is an important topic for the application of these models in practice. Optimal designs of item pairs are characterized by combinations of those values of factor loadings which optimize predetermined criteria, such as the correlation between the estimated and true trait scores or the volume of the confidence ellipsoid of the trait scores. For many applications, e.g., the selection of personnel, paired comparison should consist of equally keyed items. This condition requires the development of novel types of optimal designs. Beyond properties of optimal designs developed in the literature so far, two more requirements must be given special consideration: (a) the restriction of the design region, and (b) the constraint that alternatives have to load on mutually distinct factors, respectively. In this talk, we present solutions for the optimal design problem which substantially outperform current methods in the literature in terms of precision and amount of paired comparisons required. Based on our results, trait scores can be easily estimated even if only factor loadings are known.

Estimating the primary dimensional correlations of nested-dimensionality data structures

Wednesday, 16th July - 14:00: Item Response Theory II (MAC: Thomas Swain) - Oral

Dr. Ken Fujimoto (Loyola University Chicago)

Item response data often represent complex dimensional structures such as nested-dimensionality structures, which is when nuisance dimensions are nested within primary dimensions. In these structures, researchers are oftentimes interested in the correlations among the primary dimensions. Unfortunately, these correlations are typically obtained by ignoring the nuisance dimensions (e.g., by calculating Pearson correlations on raw scores or using latent variable models based on simple dimensional structures), and these approaches could underestimate the magnitude of the correlations such that the substantive conclusions change (e.g., a correlation of –.88 estimated to be –.35), as demonstrated through simulations and analysis of acculturation and enculturation data (Fujimoto, Yoon, & Miller, 2022).

For this presentation, I will discuss the findings from a simulation study designed to investigate factors that contribute to the underestimation of the magnitude of the correlations among the primary dimensions when the nuisance dimensions are ignored. Real acculturation and enculturation data motivated the design of this study, and the Bayesian three-tier item response theory (IRT) model was used to analyze the data.

Two factors affect how much the magnitude of the correlations is underestimated: (a) the extent to which the nuisance dimensions are represented in the data relative to the primary dimensions and (b) the extent to which the orthogonal assumption holds for the nuisance dimensions (a common assumption of many nested-dimensionality IRT models). These findings highlight the importance of accounting for nuisance dimensions nested within primary dimensions when studying the correlations among the dimensions of substantive interest.

Asymptotic standard errors for reliability coefficients in item response theory

Wednesday, 16th July - 14:15: Item Response Theory II (MAC: Thomas Swain) - Oral

Ms. Youjin Sung (University of Maryland), Prof. Yang Liu (University of Maryland)

Recently, Liu, Pek, and Maydeu-Olivares (2024) introduced a regression-based framework for reliability, assuming an underlying latent variable (LV) measurement model. Within this framework, well-known classical test theory (CTT) reliability and proportional reduction in mean squared error (PRMSE) are defined as coefficients of determination in regression models, where LVs serve as either predictors or outcomes. Building on this framework, we classify reliability coefficients in item response theory (IRT) into CTT reliability and PRMSE and derive their asymptotic standard errors (SEs). In practice, reliability coefficients are estimated from sample data and are thus subject to sampling error. Despite the importance of quantifying this uncertainty, SE estimation for reliability coefficients has not been thoroughly explored in the IRT literature. Existing work primarily focuses on cases where reliability coefficients are expressed as transformations of item parameters, requiring variances and expectations in reliability formulas to be evaluated at their population values. Such calculations, however, are infeasible for long tests. In this study, we consider a more general scenario where sample moments replace population values in reliability estimation, introducing additional sampling variability beyond item parameter estimation. We derive asymptotic SEs for CTT reliability and PRMSE under IRT models that explicitly account for uncertainty from both sources: item parameter estimation and the use of sample moments. A simulation study evaluates the finite-sample properties of these uncertainty quantification measures under various test lengths and sample sizes.

Using recency to improve scoring in longitudinal assessments

Wednesday, 16th July - 13:30: Longitudinal Data Analysis I (GH: Think 4) - Oral

Mr. Aaron Myers (American Board of Internal Medicine)

Assessment designs intended to support instruction and learning in addition to mastery have garnered considerable attention from state assessment (i.e., through-course assessments) and professional certification (i.e., longitudinal assessments) agencies. Such assessment designs generally consist of multiple interim measurements of examinee ability throughout the course of an assessment cycle rather than a single end-of-cycle summative assessment. The interim assessment scores provide instructors and examinees with timely identification of weaknesses that facilitate targeted learning interventions. The scores across interim assessments may also be aggregated in some manner to produce a summative decision at the end of the assessment cycle.

Aggregation of scores across interim assessments can improve precision of measurement and mitigate error variance due to factors such as administration conditions and fatigue. Nonetheless, (unweighted) score aggregation may incorporate unintended bias in summative scores because—to the extent interim assessments promote additional learning—earlier assessment scores are less reflective of an examinee's end-of-cycle ability than more recent scores.

Weighting the longitudinally-collected interim assessment scores according to recency of administration may mitigate this unintended bias while also providing some advantages of score aggregation (e.g., more consistent score estimates). In the current study, I propose and evaluate three likelihood-based IRT proficiency estimators that incorporate longitudinally-defined item-level weights. Longitudinally-weighted extensions of maximum likelihood, Warm's weighted likelihood (1989), and maximum a posteriori estimators are compared to their unweighted counterparts and across different weighting schemes. Results and implications of score weighting will be discussed.

Cost-effective ESM studies: Integrating budget constraints into sample size decisions

Wednesday, 16th July - 13:45: Longitudinal Data Analysis I (GH: Think 4) - Oral

Mr. Jordan Revol (KU Leuven, University of Leuven), Prof. Ginette Lafit (KU Leuven, University of Leuven), Prof. Olivia Kirtley (KU Leuven, University of Leuven), Prof. Eva Ceulemans (KU Leuven, University of Leuven)

A crucial question when designing an Experience-Sampling-Method (ESM) study concerns the sample size needed, defined by the number of participants (N) and the number of measurement occasions per participant (T). Higher N and T benefit power, but also increase researcher and participant burden, and study cost. Although study design, operational expenses, participant incentives, and compliance rates sometimes strongly differ across research groups, current approaches for sample size planning rarely explicitly account for these constraints. In this study, we present a step-by-step framework that integrates ESM design and cost constraints into sample size planning through ESM-specific cost functions and power contour plots, enabling researchers to identify the most cost-optimal combination of N and T. We demonstrate this framework through an example where three researchers investigate the same research question with the same statistical approach but using different designs. The results reveal that varying cost constraints substantially return different optimal sample sizes, highlighting the importance of cost considerations in sample size determination. Hence, disregarding this factor can compromise study feasibility and/or lead to a waste of resources. Additionally, this framework highlights some limitations of current sample size methods that will be discussed.

Matrix decomposition SEM tree

Wednesday, 16th July - 14:00: Longitudinal Data Analysis I (GH: Think 4) - Oral

Dr. Naoya Todo (Tokyo Metropolitan University), Dr. Satoshi Usami (The University of Tokyo), Dr. Naoto Yamashita (Kansai University)

The Latent Growth Model (LGM) is a popular method for describing trajectories of changes of individuals in longitudinal studies. When unobserved heterogeneity in patterns of changes is expected, options such as the Latent Growth Mixture Model (LGMM) and the Structural Equation Model Tree (SEM Tree; or LGM-based tree) are available.

The SEM Tree is especially useful for classifying individuals based on external covariates and understanding the structure of heterogeneity from these covariates. Several packages (e.g., MplusTree) are already available for conducting this method. The potential limitations of the SEM Tree currently include computational time and improper solutions. Specifically, the SEM Tree may suffer from long computation times, especially when the number of covariates is large, and it may produce improper solutions in some nodes, partly because parameters in SEM/LGM are usually estimated by MLE.

Matrix Decomposition SEM (MDSEM), which has attracted attention in recent years, can be useful for estimating parameters in SEM in that it can effectively avoid improper solutions based on iterative algorithm.

In this study, we developed an MDSEM Tree that combines MDSEM with a decision tree to classify individuals based on covariates and latent growth curves estimated by MDSEM. Through a large simulation study, we will demonstrate that the MDSEM Tree is effective at avoiding improper solutions and offers faster computational times compared to SEM Tree/MplusTree.

Confidence interval-based determinations for optimal sample sizes and designs in a random intercept panel model

Wednesday, 16th July - 14:15: Longitudinal Data Analysis I (GH: Think 4) - Oral

Dr. Satoshi Usami (The University of Tokyo)

In longitudinal design, behavioral researchers are often interested in separating within-person variability from stable between-person (or stable trait) differences. In particular, in recent years there has been growing use of residual-level statistical models, which orthogonally decompose within- and between-level variances. Focusing on a random intercept panel model, in this paper confidence interval-based determination methods are investigated to find optimal sample sizes and design in the inference of within- and between-level variances. Using the derived asymptotic standard errors of parameter estimates, we propose a method to find the minimally required sample sizes to ensure the desired width of confidence interval, and a method to find the optimal pair of sample sizes and the number of time points that minimizes costs while maintaining the width of confidence interval below a certain threshold. An environment for determining optimal sample sizes and design based on the proposed method is provided via Shiny applications.

Examining item discrimination in growth measurement: A multilevel perspective

Wednesday, 16th July - 14:30: Longitudinal Data Analysis I (GH: Think 4) - Oral

Prof. Xiangyi Liao (University of British Columbia), Prof. Daniel Bolt (University of Wisconsin - Madison)

While multilevel modeling is widely used in longitudinal analysis to disaggregate between-person and within-person effects, the potential for measurement models to achieve the same has not yet been fully investigated. Building from an explanatory IRT approach, this study provides empirical evidence that, in the context of longitudinal item response datasets, an item can have distinct features in its cross-sectional discrimination and sensitivity to growth, separately reflecting between-person differences and within-person changes. Using real data from multiple forms of the Wisconsin Knowledge and Concepts Examination-Math (WKCE-M) administered in grades 3–8, this study suggests that the dimension(s) along which students grow may differ from the dimension(s) that distinguish them cross-sectionally, especially at early grade levels.

We then demonstrate that such effect can be generalized into a multilevel item response model that includes both cross-sectional and longitudinal item discrimination parameters. Through simulation studies, we show that traditional measurement models (i.e., Rasch, 2PL), commonly used for growth measurement, are misspecified because they fail to account for different forms of item discrimination. Although measurement model misspecification may not always be detected by fit statistics, it biases in the quantification of student growth at both the individual and aggregated levels. Our extended tool provides a straightforward way of investigating possible inconsistencies between the dimension(s) along which growth occurs and other dimensions (i.e., cross-sectional variability) that can lend practical insight into how students change over time.

Mapping out the Hexagon Measurement Framework in the human sciences

Wednesday, 16th July - 13:30: Perspectives from the Inaugural Meeting of the Society for the Study of Measurement (GH: Think 5) - Oral

Prof. Mark Wilson (University of California, Berkeley), Prof. Luca Mari (Università Cattaneo – LIUC, Castellanza (VA), Italy)

Recently the "Hexagon Framework" has been proposed as a unifying interpretation of measurement across the physical and human sciences (Mari et al., 2023). While the steps for using this Framework to perform a measurement are extensively described and discussed in that book, it only briefly discusses how the Framework can also be exploited as a guide to help organize and explain the process of *developing* a measuring instrument. The object of this paper is to lay out this process, particularly in the context of measurement in the human sciences.

First, we outline how the Hexagon Framework can be seen as a structure to interpret important advancements in the historical developments in physical measurement, taking Chang's (2007) account of the development of temperature measurement as an illustration. Second, we show how this structure is mirrored in a contemporary instrument development approach used in the human sciences: the BEAR Assessment System (BAS; Wilson, 2023). We describe how the BAS is overlaid on the Framework as the instrument development proceeds and may well iterate several times in that process.

Third, we use this underlay of the Framework to illustrate and exemplify how the BAS has been used to design and develop an example instrument in the human sciences, specifically to measure an educational achievement construct intended for use in schools: Scientific Argumentation. Finally, we make comparisons of the above account with the so called "norm-referenced" approach to measurement (Nunally & Bernstein, 1994) and also discuss some caveats and limitations of the account.

How uncertainty allows measurement to produce useful results

Wednesday, 16th July - 13:30: Perspectives from the Inaugural Meeting of the Society for the Study of Measurement (GH: Think 5) - Oral

Dr. Kent Staley (Saint Louis University), Dr. Hugo Beauchemin (Tufts University)

Investigators performing a measurement aim, among other things, to produce something useful for other inquiries. The tasks performed and resources used in measuring are underdetermined by the aims, even when conjoined with “background knowledge.” At various stages of the measurement process, for any given task and any given resource, investigators must make choices striking different balances among competing aims. In the philosophy of science, such underdetermination has been treated as an epistemological obstacle, to be overcome if possible and lamented if not. So it might seem here: the context of the user of a result might call for choices among underdetermined options making different tradeoffs among aims.

We show, however, that evaluating uncertainty converts underdetermination into an epistemologically significant asset. *Varying* the tasks and resources chosen in producing a measurement result and *analyzing* the consequences of such variation yields an interval that accounts for those consequences. Employing a pragmatist account of measurement as inquiry, and treating measurement, uncertainty, and sensitivity as three irreducibly connected concepts, we show how the evaluation of uncertainty in measurement in high energy physics is interwoven through an ensemble of data transformations aimed at producing a result that becomes useful, not just to some investigator who might have chosen differently, but to a wider range of investigators who might have made unanticipated choices. The basis for this view of measurement lies in general features of measurement practices that span disciplines, including psychology and the social sciences.

Re-interpreting the Weber-Fechner law as a probabilistic measurement model

Wednesday, 16th July - 13:30: Perspectives from the Inaugural Meeting of the Society for the Study of Measurement (GH: Think 5) - Oral

Prof. Robert Massof (Johns Hopkins University), Dr. William Fisher (University of California, Berkeley)

In 1932, the British Association for the Advancement of Science commissioned a committee chaired by physicist Alan Ferguson of 9 physicists and 10 psychologists to study the "possibility of quantitative estimates of sensory events" (Ferguson, 1932, 1938, 1940). One major issue was a critical flaw in Fechner's hypothesis that the logarithmic relationship between stimulus and sensation magnitudes is a consequence of the unverified assumption that changes in sensory magnitude are constant for all just noticeable differences (JNDs) irrespective of stimulus intensity. The failure of this flawed assumption led to the denial by the committee of scientific measurement as achievable in psychophysics and psychology. By dropping Fechner's assumption we were led to a general solution motivated by the Ferguson Committee's critique, firmly grounded in Weber's law for JNDs, and yielding experimentally testable candidates for the monotonic relationship between sensory experience and stimulus intensity. Metaphysical issues associated with reducing mental phenomena to physical are avoided by invoking the irreducible complexity of psychoneural congruency (Massof, 1985). This principle accepts discontinuities between conscious sensory experience and neurally processed stimulus energy in an overall context of coherent psychophysical structures. We thus adopt an epistemological framework analogous to those developed in multiple other disciplinary contexts, including hierarchically complex cognitive processes (Borsboom, et al., 2019; Fischer & Farrar, 1987), organizational issues (Woolley & Fuchs, 2011), and psychometric models (Forer & Zumbo, 2011; Leite & Commons, 2023; Snijders & Bosker, 1999; Sijtsma & Emons, 2013).

Evaluating the classification accuracy of an adaptive diagnostic test

Wednesday, 16th July - 13:30: Cognitive Diagnostic Models II (GH: Think 3) - Oral

Mr. Ahmed Bediwy (Univeristy of Iowa)

Achievement tests are widely used in educational settings to measure students' knowledge and skills. Computerized Adaptive Testing (CAT) has been shown to provide more accurate and efficient estimates of students' abilities compared to traditional linear tests. Adaptive diagnostic assessments further extend this approach by identifying students' strengths and weaknesses in specific content areas, often informing personalized instruction. These assessments are administered multiple times per year, classifying students into performance levels based on predetermined cut scores. However, classification accuracy—the extent to which these tests correctly assign students to the appropriate performance level—remains a critical consideration.

This study evaluates the classification accuracy of an adaptive diagnostic test through simulation. Specifically, we investigate classification consistency across subjects, content domains, grade levels, and testing windows. A simulation-based approach is used to examine the stability of classification decisions relative to students' true performance levels. Given that misclassifications are more likely for students near cut scores, we analyze their impact on instructional decision-making and adaptive learning systems.

Findings provide insights into the trade-offs between classification accuracy and the granularity of instructional decisions. By focusing on content domain scores, this study contributes to discussions on the stability of adaptive assessments, particularly in the context of through-year testing. The results have implications for educators, assessment developers, and policymakers seeking to balance precision with the need for personalized learning paths.

A Hamiltonian-Gibbs sampler with monotonicity constraints for diagnostic classification models

Wednesday, 16th July - 13:45: Cognitive Diagnostic Models II (GH: Think 3) - Oral

Dr. Alfonso Martinez (Fordham University), Dr. Jonathan Templin (University of Iowa)

We propose a novel Bayesian algorithm featuring monotonicity constraints for saturated diagnostic classification models (DCMs). The algorithm (Hamiltonian-Gibbs) synthesizes Gibbs sampling with Hamiltonian dynamics to efficiently explore the complex parameter spaces characteristic of diagnostic models. The Hamiltonian-Gibbs sampler partitions the parameter space into continuous and discrete parameter blocks. Hamiltonian dynamics, which models the movement of a particle in a physical system, is used to update continuous (item) parameters and Gibbs sampling is used to update discrete (attribute) parameters. A feature of the sampler is that it does not require marginalization of the attribute space, allowing attribute profiles to be directly sampled from their conditional posterior. Another feature is the embedding of a “particle-bouncing” mechanism within the Hamiltonian portion of the algorithm which enforces monotonicity constraints without requiring sequential updating steps. This mechanism also avoids the label-switching problem known to affect mixture models like DCMs. The properties of the algorithm are explored through two Monte Carlo simulation studies. Simulation study I (SimStudyI) explores the use of the algorithm across a wide variety of simulation conditions and simulation study II (SimStudyII) compares the sampler to the Holmes-Held auxiliary variable algorithm. Results from SimStudyI provide evidence of chain convergence within a few hundred iterations, acceptance probabilities of approximately 0.85-0.90, lag-1/lag-5 autocorrelations near 0.2/0.0, respectively, and sufficiently accurate item parameter recovery/profile classification rates. Results from SimStudyII provides evidence that the algorithm offers comparable performance to the Holmes-Held sampler with respect to item/attribute parameter recovery and an improvement with respect to chain convergence and autocorrelations.

Consistency theory of general nonparametric classification methods in cognitive diagnosis

Wednesday, 16th July - 14:00: Cognitive Diagnostic Models II (GH: Think 3) - Oral

Mr. Chengyu Cui (University of Michigan), Mr. Yanlong Liu (The University of Chicago), Dr. Gongjun Xu (University of Michigan)

Cognitive diagnosis models have been popularly used in fields such as education, psychology, and social sciences. While parametric likelihood estimation is a prevailing method for fitting cognitive diagnosis models, nonparametric methodologies are attracting increasing attention due to their ease of implementation and robustness, particularly when sample sizes are relatively small. However, existing consistency results of the nonparametric estimation methods often rely on certain restrictive conditions, which may not be easily satisfied in practice. In this article, the consistency theory for the general nonparametric classification method is reestablished under weaker and more practical conditions.

Increasing the flexibility of the MC-DINA model

Wednesday, 16th July - 14:15: Cognitive Diagnostic Models II (GH: Think 3) - Oral

Prof. Jimmy de la Torre (The University of Hong Kong), Prof. Wenchao Ma (University of Minnesota), Dr. Xiaopeng Wu (The University of Hong Kong), Mr. Zechu Feng (The University of Hong Kong)

Multiple-choice (MC) tests are widely used in educational measurement because they can be implemented easily and efficiently. However, MC data are typically analyzed as dichotomous data. This simplification, which ignores the diagnostic information in the distractors, is employed across different psychometric frameworks, including cognitive diagnosis modeling (CDM). The MC deterministic input, noisy “and” gate (MC-DINA) model, which allows for the distractors to be coded and analyzed, was proposed to overcome this limitation in the CDM context. However, the MC-DINA model has one important limitation – it requires the coded options (i.e., key and distractors) to follow a hierarchical linear structure, which limits the number of usable distractors. This work proposes an extension of the MC-DINA model, the flexible MC-DINA (fMC-DINA) model that relaxes the hierarchical linear structure requirement. In particular, the fMC-DINA model can be used with MC data provided the coded options result in latent classes that can be classified into unique latent groups. A such, the fMC-DINA model can accommodate more coded distractors. A simulation study was conducted to compare the two models when both hierarchical and nonhierarchical distractors are involved. Results show that, when high-quality items are involved, the improvement in the attribute vector classification accuracy from using the fMC-DINA model was 3.33%; however, when low-quality items are involved, the improvement jumps to 8.44%. These improvements can be traced to the 40% more usable distractors when fitting the fMC-DINA model. Results from other conditions and ways the fMC-DINA model can be further improved will be also presented.

Examining the occurrence of paradoxical scoring in cognitive diagnostic models

Wednesday, 16th July - 14:30: Cognitive Diagnostic Models II (GH: Think 3) - Oral

Mr. Tsuyoshi Kato (The University of Tokyo), Mr. Shun Saso (The University of Tokyo), Dr. Satoshi Usami (The University of Tokyo)

This study examined the frequency and conditions under which paradoxical scoring (PS) occurs in cognitive diagnostic models (CDMs) using both theoretical and simulation-based analyses. CDMs estimate individual's discrete latent traits (i.e., attributes), enabling multidimensional assessment. Previous studies have shown that PS—where answering a specific item incorrectly leads to an increased estimate of a latent trait in a certain dimension—can arise in multidimensional item response theory. Similarly, PS may also arise in CDMs. This study theoretically proves that PS involving an increase in attribute mastery patterns does not occur in nested relationships when using maximum likelihood or maximum a posteriori estimation under monotonicity constraints, though it may still arise in non-nested patterns. Simulation studies, using small sample sizes resembling classroom contexts, examines the effects of various factors (e.g., correlations among attributes, item quality, sample size, and number of items) on the occurrence of PS when applying DINA, DINO, and G-DINA models. As a result, while the increase in the number of estimated mastered attributes was relatively small, the increase in attribute mastery probabilities was salient. Consequently, a non-negligible proportion of individuals experienced PS in terms of increased mastery probabilities, particularly when attribute correlations and item quality were low. This phenomenon creates paradoxical feedback situations where students who correctly answer more items may receive diagnostic results indicating lower mastery of certain attributes, posing practical challenges for classroom implementation and educational feedback.

Latent propensity modeling of hint-seeking behavior in intelligent tutoring systems

Wednesday, 16th July - 15:15: Adaptive Assessment and Learning (GH: Meridian 1-2) - Oral

Dr. Hyeon-Ah Kang (The University of Texas at Austin), Ms. Ji Yun Lee (The University of Texas at Austin), Dr. Adam Sales (Worcester Polytechnic Institute), Dr. Tiffany Whittaker (The University of Texas at Austin)

Increasing use of tutoring software in education has called for analytic approaches that can identify aberrances in student behaviors. One of the pronounced behavioral patterns in intelligent tutoring systems (ITSs) is abusive use of a help system. In ITS, students can ask for a series of hints until they bottom out and obtain correct answers. The hinting system is devised to facilitate students' learning progression and completion of tutoring. It has been however used as an avenue for disengaged learners to get through ITS without meaningful learning. In this study, we present a psychometric model that can characterize students' propensity in bottoming out hints and demonstrate its application in predicting bottom-out behavior. We develop a latent propensity model applying students' interaction indicators (e.g., time on problems, first action) and propose model inference methods for highly sparse and stochastic item assignment design. A prediction model is developed based on recurrent neural networks applying three sets of feature variables: student features on the bottom-out propensity and interaction with the system, problem features (e.g., item type, content areas), and support features that characterize the help prompts (e.g., type, visual aids). A Monte Carlo simulation study is proposed to validate the performance of the inferential algorithm of the latent propensity model. The empirical performance of the prediction model is evaluated on log data from ASSISTments. Through the new latent propensity model and its application to predictive modeling, we demonstrate how conventional psychometric models can help enhance the prediction performance of machine-learned models.

Bayesian adaptive learning assessment for efficient skill acquisition

Wednesday, 16th July - 15:30: Adaptive Assessment and Learning (GH: Meridian 1-2) - Oral

Dr. Sangbeak Ye (Florida Atlantic University)

Adaptive learning assessments require both instructional sequencing and diagnostic precision to accurately track learners' evolving skill mastery. When multiple skills are assessed concurrently, many existing models—such as those used in intelligent tutoring systems—assume sequential and independent skill acquisition, leading to inefficiencies in assessment length and instructional adaptivity. This study introduces a Bayesian diagnostic framework for modeling multi-skill acquisition, accounting for hierarchical dependencies and simultaneous skill transitions. As new responses are observed, the framework updates posterior mastery probabilities across multiple skills in real time. Unlike conventional approaches that assess skills in isolation, this method integrates *a posteriori* item selection strategies, ensuring that assessments dynamically adapt to learner responses while minimizing redundant queries. The system formalizes adaptive item selection through a decision-theoretic model, balancing exploration and reinforcement in assessment sequencing. Thompson sampling is incorporated as one of multiple strategies to optimize information gain while maintaining assessment efficiency. By leveraging a Bayesian decision-theoretic approach, the framework optimally selects items that maximize information gain while minimizing redundancy, ensuring efficient assessment administration and accelerating learner skill acquisition.

Personalized adaptive, dynamic, and formative assessment in statistics education

Wednesday, 16th July - 15:45: Adaptive Assessment and Learning (GH: Meridian 1-2) - Oral

Dr. Wilco Emons (Tilburg University)

In this presentation, I will discuss the psychometric challenges involved in an educational innovation project aimed at improving statistics education in higher education in the Netherlands. The main goal of this project is to develop a personalized dynamic formative adaptive assessment tool. A key feature of the tool is that students receive statistical tasks within a substantive context of their own choice. For example, the student may choose to practice methods and statistics within the context of eating disorders. Furthermore, the tool will be designed not only to train students in performing the statistical analyses, but also on how to synthesize the results of their analysis into meaningful substantive conclusions. To achieve this goal, the tool integrates technology enhanced assessment methodologies – such as personalized contextual assessment, adaptive assessment, and dynamic automated item generation – with computational psychometrics and large language models to generate authentic tasks. In the presentation, I will outline the general design principles behind the tool, discuss psychometric challenges, present our solutions, and share initial experiences from the project.

Rationale and patterns of misclassified text in automated scoring systems

Wednesday, 16th July - 16:00: Adaptive Assessment and Learning (GH: Meridian 1-2) - Oral

Mr. Haowei Hua (The Culver Academies), Mr. Jiayu Yao (Anhui Polytech University)

Automated scoring systems have become a critical component of educational assessments and large-scale testing due to their efficiency, scalability, and consistency compared to human raters. These systems utilize advanced machine learning algorithms and natural language processing techniques to evaluate written responses. With the increasing popularity of large language models, generative AI also provides another approach for educators to evaluate writing assessments. While significant progress has been made to improve the system accuracy and consistency, challenges persist, particularly in cases of misclassified responses where predicted scores deviate from human raters' evaluations. Current research primarily focuses on improving model accuracy, interpretability, and fairness; however, the analysis of misclassified responses remains underexplored. This study investigates the text features of misclassified responses in AI-based and ML-based scoring systems, aiming to uncover the underlying rationale for scoring errors. Specifically, the state-of-the-art AI models, including ChatGPT-4o, Gemini, and other cutting-edge automated scoring frameworks have been employed, to analyze how different systems assign scores and where discrepancies arise. By implementing score error analysis, this research systematically identifies patterns and inconsistencies that contribute to misclassification, offering insights into linguistic, syntactic, and semantic features that impact scoring. This approach enhances the robustness, equity, and interpretability of automated scoring methodologies, contributing to the development of more reliable automated scoring systems.

Use factor-augmented regularized latent regression to analyze complex large-scale assessment

Wednesday, 16th July - 15:15: Machine Learning (GH: Meridian 3-4) - Oral

Mr. He Ren (University of Washington), Ms. Yijun Cheng (University of Washington), Dr. Chun Wang (University of Washington), Dr. Gongjun Xu (University of Michigan)

Introduction

Large-scale assessments (LSA) use latent regression (LR) to integrate students' background information for accurate and meaningful scores. Traditional LR uses principal components (PCs) that explain most predictor variation to reduce dimensionality and multicollinearity. However, there is no one-size-fits-all approach that can effectively guide the selection of PCs. We propose the factor-augmented regularized LR (FARLR) for a more interpretable and robust solution.

Method

The traditional LR with PCs is widely used in LSA, given by

$\theta=\mathbf{W}\Upsilon+\varepsilon$,

where $\theta\square\mathrm{R}^{n}$ is the ability vector for n students, **W** is an n×d matrix of the first d PCs from the students' covariate matrix **X** ($\mathbf{X}\square\mathrm{R}^{n\times d}$, $d<p$), ϒ is the regression coefficient vector, and ε is residual. However, d is typically chosen using empirical cut-offs. Too few PCs reduce congeniality, while too many threaten robust estimation (i.e., yield large standard errors) due to a small n/d ratio. To address this, we propose the FARLR method, incorporating both common factors and idiosyncratic components,

$\theta=\mathbf{f}\beta+\mathbf{u}\nu+\varepsilon$,

where $\mathbf{X}=\mathbf{f}\mathrm{B}^{\mathrm{T}}+\mathbf{u}$ is the approximate factor model, $\mathbf{f}\square\mathrm{R}^{n\times d'}$ represents common factors, and $\mathbf{u}\square\mathrm{R}^{n\times d}$ is the idiosyncratic residuals. Regularization is introduced on u to select significant components.

Preliminary Results

Simulations show that FARLR accurately selects significant idiosyncratic residuals with low Type I error and high power while ensuring accurate model prediction and statistical inference on key predictors.

Significance

The FARLR avoids subjective PC selection, ensures congeniality with diverse secondary analysis models, and enhances plausible value precision and model interpretability through implicit control of hidden confounding and variable selection.

Personalized predictive modeling with Bayesian nonparemetric ensemble learning

Wednesday, 16th July - 15:30: Machine Learning (GH: Meridian 3-4) - Oral

Ms. Mingya Huang (University of Wisconsin - Madison), Dr. Weicong Lyu (University of Macau), Prof. David Kaplan (University of Wisconsin - Madison)

Model uncertainty has been one of the major challenges in statistical inference and prediction, particularly for large-scale multilevel assessment data in social sciences. To address these uncertainty issues, current benchmark methods are ensemble learning which combine different information from each candidate model to achieve optimal final prediction, including Bayesian parametric approaches such as Bayesian Model Averaging (BMA) and the newly developed nonparametric machine learning method, Bayesian Additive Regression Trees (BART) (Yannotty et al., 2024). However, these methods rely heavily on the sum-to-one constraint, assuming that the optimal solution must be a convex combination of existing models, which may not hold in complex real-world settings. For instance, when the candidate model set is incomplete or misspecified, forcing the weights to sum to one (i.e. 100\%) can lead to nonoptimal predictions, as it requires the method to fully allocate the probability mass across potentially inadequate models. Thus, to address these limitations, we propose to use Varying Coefficients BART (VCBART) to relax the sum-to-one constraint by incorporating an unconstrained weighting scheme. Through both empirical and simulation studies, our finding indicates that VCBART can (1) obtain input-dependent weights that adapt to different data structures, (2) achieve optimal out-of-sample prediction at the individual level, and (3) be more computationally efficient. This study introduces a novel Bayesian nonparametric approach that can not only overcome the limitations of existing benchmark methods but also significantly improve predictive inference in large-scale assessment, providing a more personalized solution for handling complex multilevel data structures while accommodating individual differences.

A two-step imputation approach combining IRT and deep-learning methods in large-scale survey assessments

Wednesday, 16th July - 15:45: Machine Learning (GH: Meridian 3-4) - Oral

Dr. Usama Ali (ETS Research Institute), Dr. Peter van Rijn (ETS Global), Ms. Priyadarshini Dwivedi (Indian Institute of Technology, Kanpur)

We address the limitations of current plausible-value (PV) methodology in large-scale survey assessments by developing a two-step imputation method that combines item response theory (IRT) with generative adversarial imputation networks (GAIN). Our work builds upon promising results from earlier comparisons of IRT and deep-learning-based imputation methods. Previous studies indicated that deep-learning methods such as denoising autoencoders and generative adversarial networks showed promising results but also highlighted challenges with high rates of missing data. To address these issues, we plan to generate starting values using IRT-based imputation, which will then serve as input for the GAIN method.

Our objective is to develop a hybrid two-step imputation framework that integrates IRT and GAIN methods. The first step uses IRT models to provide initial imputations based on item features such as difficulty and discrimination, and on students' latent proficiencies. In the second step, the GAIN model is employed to refine these imputations through adversarial learning, allowing it to capture potentially more complex, non-linear relationships in the data. Also the performance of this new approach will be evaluated against oth traditional methods using data from PIRLS and TIMSS in which more than 80% of the item-response data is missing by design.

A successful hybrid method could lead to more accurate and efficient educational assessments. Moreover, the new method could offer a viable alternative to the current PV methodology, making results more accessible to secondary users and enabling more flexible reporting options. This could help improve assessment designs and reporting practices in educational research.

Generalized grade-of-membership estimation for high-dimensional locally dependent data

Wednesday, 16th July - 16:00: Machine Learning (GH: Meridian 3-4) - Oral

Dr. Yuqi Gu (Columbia University)

This work focuses on the mixed membership models for multivariate categorical data that are widely used for analyzing survey responses and population genetics data. These grade of membership (GoM) models offer rich modeling power but present significant estimation challenges for high-dimensional polytomous data. Such data take the form of a three-way (quasi-)tensor, with many subjects responding to many items with varying numbers of categories. Popular existing approaches such as Bayesian MCMC inference are not scalable to big data and lack theoretical guarantees in high-dimensional settings. We introduce a novel and simple approach that flattens the three-way (quasi-)tensor into a "fat" matrix. We then perform a singular value decomposition of this matrix to estimate parameters by exploiting the singular subspace geometry. Our fast spectral method can accommodate a broad range of data distributions with arbitrarily locally dependent noise, which we formalize as the generalized-GoM models. We establish finite-sample entrywise error bounds for the generalized-GoM model parameters. This is supported by our sharp two-to-infinity singular subspace perturbation theory for locally dependent and flexibly distributed noise, a contribution of independent interest. Simulations and applications to data from the American National Election Studies demonstrate our method's superior performance.

Multiple linked tensor factorization

Wednesday, 16th July - 15:15: Making Sense of High-Dimensional Data in the Psychological Sciences (MAC: Johnson) - Oral

Prof. Eric Lock (University of Minnesota), Mr. Zhiyu Kang (University of Minnesota)

In several fields, it is now common to generate high content data that are both multi-source and multi-way. Multi-source data are collectedfrom different high-throughput technologies while multi-way data are collectedover multiple dimensions, yielding multiple tensor arrays. Integrative analysis ofthese data sets is needed, e.g., to capture and synthesize different facets of complex biological or psychometric systems. However, despite growing interest in multi-source and multi-way factorization techniques, methods that can handle data that are both multi-source and multi-way are limited. In this work, we propose a Multiple Linked Tensors Factorization (MULTIFAC) method extending the CANDECOMP/PARAFAC (CP) decomposition to simultaneously reduce the dimension of multiple multi-way arrays and approximate underlying signal. We first introduce a version of the CP factorization with L2 penalties on the latent factors, leading to rank sparsity. When extended to multiple linked tensors, the method automatically reveals latent components that are shared across data sources or individual to each data source. We also extend the decomposition algorithm to its expectation–maximization (EM) version to handle incomplete data with imputation. Extensive simulation studies are conducted to demonstrate MULTIFAC's ability to (i) approximate underlying signal, (ii) identify shared and unshared structures, and (iii) impute missing data. The approach yields an interpretable decomposition on multi-way MRI and molecular data for a study on the effects of early-life iron deficiency on cognition.

Classifying alcoholism from electroencephalography data using high-dimensional logistic regression

Wednesday, 16th July - 15:15: Making Sense of High-Dimensional Data in the Psychological Sciences (MAC: Johnson) - Oral

Mr. Jong Won Lee (University of Minnesota), Prof. Nathaniel Helwig (University of Minnesota)

Electroencephalography (EEG) studies record electrical potentials from the scalp while a subject is at rest and/or participating in a task. The recorded electrical patterns can be used to understand various cognitive processes, as well as understand differences in cognitive processes between subject populations. A fundamental challenge of understanding individual and group differences in EEG data is the high-dimensional nature of the data. Typical EEG studies record high-resolution timeseries data from multiple electrodes on the scalp, which results in thousands (or tens of thousands) of data points recorded from each subject. Furthermore, subjects often participate in multiple trials of experimental conditions, and the subjects may be nested in different clinical populations. The high-dimensional nature of the data makes it a challenge to discern which combinations timepoints and electrodes are most relevant for understanding differences in the experimental conditions and/or subject populations. In past works, hypothesis testing approaches have often been used for this purpose. In this talk, we discuss how penalized logistic regression can be used to identify which electrodes and channels best distinguish visual evoked potential EEG data collected from different subject populations. Using open-source EEG data, we show that group elastic net (GRPNET) penalized logistic regression can identify a small number of time points and electrodes that distinguish alcoholics from controls. Furthermore, we show that the proposed approach provides more accurate and more interpretable classification results than have been reported in past works. The example demonstrates the potential of GRPNET for developing interpretable biomarkers that distinguish clinical subpopulations.

Capturing fluctuations in high-dimensional intensive longitudinal data

Wednesday, 16th July - 15:15: Making Sense of High-Dimensional Data in the Psychological Sciences (MAC: Johnson) - Oral

Prof. Katerina Marcoulides (University of Minnesota), Ms. Hannah Hamling (University of Minnesota)

While the technology for collecting intensive longitudinal data is readily available, the methodology to validly analyze such data is to some degree lagging behind; traditional methods are not ideally suited for handling the collected noisy and high-dimensional data. Although some recent methodological developments have been proposed in the literature and even incorporated into the new M*plus* Version 8, handling noisy and high-dimensional intensive longitudinal data largely remains a challenge, especially when it comes to methods capable of modeling individual fluctuations and individual differences.

The purpose of this presentation is to introduce a novel method for modeling changes in multivariate intensive longitudinal measurements of individuals over time. The method considers the analysis of intensive longitudinal data from the perspective of the study of individual differences and focuses on effectively modeling the process of change within each individual and across individuals. The proposed method blends various building blocks from spatial statistics to model individual changes over time and combines procedures for dimensionality reduction with time series. We will demonstrate that this combination not only provides an ideal way to visualize data but also to further model and predict its complex structure and growth pattern. The approach can be used irrespective of the frequency of data collection, the number of variables and dimensional complexity of the relationships being modeled or the theoretical growth curve underpinning the research question, and provides an alternative lens through which multivariate intensive longitudinal data can be examined.

Predicting attention problems from brain connectivity using high-dimensional Poisson regression

Wednesday, 16th July - 15:15: Making Sense of High-Dimensional Data in the Psychological Sciences (MAC: Johnson) - Oral

Dr. Kelly Duffy (University of Minnesota), Prof. Nathaniel Helwig (University of Minnesota)

Among the benefits of penalized regression is the ability to simplify a high-dimensional space by utilizing the technique to identify the most salient predictors within a large candidate set. The present work leverages the groupwise penalized spline regression implemented in the R package GRPNET (Helwig, 2025) to implement a penalized nonparametric Poisson regression model on a count variable, a symptom score of attention problems related to attention-deficit/hyperactivity disorder (ADHD) in a large pre-adolescent sample (*N* = 7979). From a large set of candidate predictors, including 78 within- and between-network brain correlations plus a number of demographic and family history risk factors, we calculate variable importance indices to identify the most relevant predictors. We highlight how the use of an optimal penalty, such as the minimax concave (MCP) or smoothly clipped absolute deviation (SCAD) penalties, can simplify the predictor space by identifying only a single brain network correlation of interest along with a handful of other risk factors. Importantly, this same network correlation has been identified in other large-sample studies using variable selection techniques, emphasizing the potential for penalized regression to increase the replicability of findings across unique samples and studies and thus to further the field's understanding of psychological processes.

High-dimensional regression and classification of psychological data

Wednesday, 16th July - 15:15: Making Sense of High-Dimensional Data in the Psychological Sciences (MAC: Johnson) - Oral

Prof. Nathaniel Helwig (University of Minnesota)

In this talk, I discuss how group elastic net (GRPNET) regression, in combination with tensor product penalized splines, provides a powerful framework for analyzing high-dimensional (HD) data in the psychological sciences. The proposed approach combines the strengths of two popular statistical modeling tools: (i) elastic net regression, which can be used for variable selection in HD applications of Generalized Linear Models (GLMs), and (ii) Generalized Additive Models (GAMs), which are flexible extensions of GLMs that use penalized and tensor product splines to model nonlinear effects. I begin with a brief overview of the group elastic net problem, as well as my algorithm for solving the problem (Helwig, 2025a). The remainder of the talk is focused on a discussion of practical implementation through the R package **grpnet** (Helwig, 2025b), which is available on CRAN. Details related to model specification, penalty selection, hyper-parameter tuning, and results interpretation will be discussed. Using real data examples, the talk demonstrates how GRPNET can be used for variable selection and smoothing in HD applications of GAMs and Generalized Additive Mixed Models (GAMMs). As I demonstrate, the GRPNET framework provides a bridge between classic statistical regression tools and modern machine learning methods. Specifically, GRPNET is prediction oriented (unlike classic methods) and produces transparent predictive rules (unlike many machine learning methods). Thus, the proposed GRPNET approach has the potential to offer genuine insights into the nature of functional relationships in HD psychological data. Extensions and future applications of the GRPNET framework will be discussed.

Vectorizing test constraints for faster automated test assembly

Wednesday, 16th July - 15:15: Practical Issues in Testing: Norming, Equating, ATA and fairness (MAC: Thomas Swain) - Oral

Dr. Anthony Shiver (Law School Admission Council)

Automated Test Assembly (ATA) can be used as a diagnostic tool to plan the development of an item bank or to evaluate the potential effects of imposing new test constraints for an existing testing program. These uses become less feasible the longer it takes the ATA software to produce solutions, as diagnostic problems often require the iterative evaluation of multiple solutions. In short, all else equal, faster is better. This talk will present a practical method for encoding statistical and content constraints as vectors. Vectorization of assembly constraints enables ATA implementations to take advantage of Single Instruction Multiple Data (SIMD) parallelization in modern CPUs and GPUs, leading to orders-of-magnitude increases in speed over naïve serial implementations of the same algorithm. Results for a vectorized implementation of a heuristic-based ATA algorithm, applied as a tool for exploring the effects of modifying IRT information constraints under an estimated reliability metaconstraint, will be discussed.

Integrating latent variables into test equating methods

Wednesday, 16th July - 15:30: Practical Issues in Testing: Norming, Equating, ATA and fairness (MAC: Thomas Swain) - Oral

Ms. Inés Varas (Pontificia Universidad Católica de Chile)

Equating is the most widely used linking method for adjusting scores from different test forms, ensuring they can be used interchangeably. These methods transform the scale of one test form to its equivalent on another based on score distributions, addressing differences in test difficulty. Since test-taker abilities may vary across different samples, equating methods consider various data collection designs to minimize bias. The transformation, known as the equating transformation (González & Wiberg, 2017), is typically estimated using continuous approximations of score distributions, despite scores being naturally discrete.

Varas et al. (2019, 2020) introduced the latent equating method, which preserves the discrete nature of equated scores by employing a latent representation of score distributions within a Bayesian nonparametric framework. However, this approach sacrifices the symmetry property of the equating transformation, a key characteristic emphasized in the recently proposed generalized kernel equating framework (Wiberg et al., 2024).

In this talk, we explore how the ordinal representation model used in the latent equating method can be integrated into the equating process to yield a continuous equating transformation. This exploration includes an evaluation of the proposed approach across different sampling designs. By combining the strengths of latent equating and kernel-based approaches, this study aims to enhance the robustness and applicability of equating methods in educational and psychological testing.

Analyzing identifiability in statistical models for test equating

Wednesday, 16th July - 15:45: Practical Issues in Testing: Norming, Equating, ATA and fairness (MAC: Thomas Swain) - Oral

Prof. Jorge Gonzalez (Pontificia Universidad Católica de Chile)

When using statistical models, statistical inference is concerned with the process of learning from data, while an identifiability analysis delineates the boundaries of what can be learned. Identifiability plays a critical role in statistical modeling, as it guarantees the essential conditions for making coherent inferences about the parameters of interest.

In this presentation, I will introduce a framework that conceptualizes test equating as formal statistical models and delve into the identifiability analyses of these models.

Regression-based norming of tests with small sample sizes

Wednesday, 16th July - 16:00: Practical Issues in Testing: Norming, Equating, ATA and fairness (MAC: Thomas Swain) - Oral

Dr. Nicolas Sander (German Federal Employment Agency), Dr. Erik Sengewald (German Federal Employment Agency)

Background: The norming of psychometric tests often faces the challenge of limited sample sizes, especially with specific populations such as blind individuals. This report examines the applicability and quality of regression-based norming compared to “classical” norming (area transformation) using the example of a test for assessing intelligence in visually impaired individuals utilizing categorical predictors (blindness, education level).

Method: Data from 561 individuals with and without visual impairments were analyzed. Regression-based norming was performed using Generalized Additive Models for Location, Scale, and Shape (GAMLSS), with particular comparison of the Beta Binomial (BB) distribution and the Box-Cox power exponential (BCPE) distribution. Model fit was assessed using quantitative (AIC, BIC) and graphical methods (QQ-plots, Worm plots). Additionally, a bootstrap method was used as one possible approach to calculate the norming error. The results were compared with classical norming.

Results: The BB model showed a better model fit than the BCPE model. The norming error tended to be smaller with regression-based norming than with classical norming. Validity, as measured by correlation with external criteria (other intelligence tests), did not differ significantly between the two norming methods.

Conclusion: Regression-based norming represents a valid alternative to classical norming, particularly with small sample sizes. The bootstrap method used here provides insight into the norming error but should be interpreted in the context of its exploratory nature. Future research should focus on optimizing model selection, evaluating different methods for determining norming errors, and considering extreme values.

Bayesian augmentation for real-time fairness monitoring in assessments

Wednesday, 16th July - 16:15: Practical Issues in Testing: Norming, Equating, ATA and fairness (MAC: Thomas Swain) - Oral

Mr. Shea Valentine (Harver), Dr. Chris Allred (Harver)

Fairness in assessments is a critical issue in education and employment contexts, yet traditional methods for bias detection often suffer from statistical instability and the multiple measurements problem. In this study we estimate the expected future adverse impact ratio through predictive Bayesian modeling, supporting our clients' risk mitigation strategies.

Previous work on this topic (e.g., Courey & Oswald, 2025) has used Bayesian methods to describe the credible range of latent adverse impact ratios. In the present study, we predict the likelihood of passing based on demographic variables, generating posterior pass rates for different groups. We derive key fairness metrics from these posterior distributions, including pass rate differences, Jensen-Shannon divergence, and impact ratios. Our analysis demonstrated that the estimated pass rate differences and JS Divergence can serve as reliable predictors of the range of credible underlying impact ratios. Like score augmentation, we borrow strength from other evaluation metrics for better impact ratio prediction.

Our approach can handle a continuous, around the clock, stream of assessment data with fairness evaluation conducted at regular intervals. We utilize the posterior samples to enable real time fairness monitoring. This approach mitigates issues associated with multiple hypothesis testing, statistical instability, and legal constraints. Our Bayesian inferential method provides a robust framework for continuous bias monitoring.

Our findings suggest that Bayesian methodologies offer a principled means of ensuring fairness in deployed assessments, supporting fair and equitable decision making, and providing the foundations for future research and policy making.

Machine learning approaches to item-level bias detection

Wednesday, 16th July - 15:15: Differential Item Functioning and Measurement Invariance II (GH: Think 4) - Oral

Dr. Brandon LeBeau (WestEd), Dr. Sarah Quesen (WestEd)

The Office of Management and Budget's recent revision to demographic data collection standards introduces new complexities in detecting measurement bias, as students can now select multiple categories from seven race/ethnicity options. Traditional differential item functioning (DIF) methods are limited in handling such intersectional data, often requiring pairwise comparisons and reference groups. This study evaluates a novel random forest (RF) approach for detecting measurement bias that eliminates these constraints.

Using Monte Carlo simulation, we compare the RF method against traditional approaches (Mantel-Haenszel [MH], logistic regression [LR], and item characteristic curve [ICC] methods) across various conditions. The simulation examines the impact of the percentage of items with bias (5%, 10%, 15%) and the magnitude of bias (0.1, 0.2, 0.3). Sample sizes, the number of items, and type of bias will be fixed at 1000 student responses, 30 items, and uniform bias, respectively. Data are generated using the graded response model, with intersectional group membership simulated via zero-truncated Poisson distribution to reflect realistic demographic patterns. The data will be replicated 1000 times. Each method proposed, MH, LR, ICC, and RF, will be fitted to each simulated data and will be a within-study factor.

Method performance is evaluated through accuracy rates, false positive/negative rates, and effect size recovery. The RF approach uniquely accommodates multiple group memberships without requiring reference groups or pairwise comparisons. This study contributes to the emerging field of computational psychometrics by providing empirical evidence for the effectiveness of machine learning methods in detecting measurement bias within complex, intersectional data contexts.

Detecting differential item functioning in forced-choice models with misspecification

Wednesday, 16th July - 15:30: Differential Item Functioning and Measurement Invariance II (GH: Think 4) - Oral

Dr. Jacob Plantz (Enrollment Management Association), Dr. Anna Brown (University of Kent), Dr. Jessica Flake (University of British Columbia), Dr. Keith Wright (Enrollment Management Association)

The Forced-Choice (FC) response style has the potential to reduce response bias from test-takers (Cao & Drasgow, 2019). This has made it a desirable choice for high-stakes educational and occupational settings where distorted responding is more likely to occur. In these settings the applicant pool may be diverse and issues with test fairness are of concern. One way of assessing the fairness of items on the test is by conducting Differential Item Functioning (DIF) analysis. Some DIF research has been conducted for FC tests (see Lee et al., 2021), however, these methods have not been examined in conditions representative of the real-world where model misspecification is likely. We assessed the efficacy of the free and constrained-baseline approaches to DIF testing when model misspecification (conceptualized as when the anchor contained DIF blocks) was present by conducting a simulation study with 336 conditions including: the size of DIF (small and large), sample size (1000 and 2000), sample size equality (focal and reference group equal or non-equal), the percentage of blocks with DIF on the test (40%, 50%, or 60%), and the number of traits (5 or 10). It also included conditions related to the size of the anchor (20% and 30% of all blocks) and the amount of misspecification (0%, 50%, or 100% of the anchor was misspecified). We discuss our findings and provide recommendations on when free and constrained baseline approaches perform well. We also provide practical recommendations based on our findings on how to examine DIF in FC tests.

A two-step method for detecting Differential Item Functioning

Wednesday, 16th July - 15:45: Differential Item Functioning and Measurement Invariance II (GH: Think 4) - Oral

Ms. Qing Zeng (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)

A well-known potential threat to test fairness in high-stakes testing environments is differential item functioning (DIF). Currently, several statistical methods have been proposed to detect DIF. However, these methods may inflate Type I error rate when calculating the test statistics for each item separately and performing multiple tests, thereby increasing the likelihood of false positives (i.e., items that are incorrectly flagged as exhibiting DIF). To address this issue, we put forward a novel two-step method that combines advanced machine learning techniques (e.g., Random Forest) with psychometric approaches. To evaluate the performance of the two-step method, we conducted multiple simulation studies under varying testing scenarios, comparing its results with those of traditional DIF detection methods (including logistic regression, Lord's Chi-Squared method, the Mantel-Haenszel method, and the standardization approach). Five factors—sample size, test length, proportion of DIF items, type of DIF, and the size of the DIF effect—were manipulated, and Type I error rate and statistical power were employed as the evaluation criteria. Simulation results showed that the two-step method controlled Type I error rate more effectively than the traditional ones, with higher power, especially when test lengths were long and the proportion of DIF items was large. On the other hand, the practical application of the two-step method was also discussed, aiming to evaluate its DIF detection performance in real scenarios. In conclusion, this innovative two-step approach offers more accurate and reliable estimates for DIF detection, thereby enhancing the fairness and validity of the testing process.

A neural network approach to small sample intersectional DIF detection

Wednesday, 16th July - 16:00: Differential Item Functioning and Measurement Invariance II (GH: Think 4) - Oral

Mr. Yale Quan (University of Washington), Dr. Chun Wang (University of Washington)

The 2014 standards for educational and psychological testing states that analysis of differential item functioning (DIF) is essential for establishing fairness in psychological and educational testing (AERA, APA, & NCME, 2014). Intersectional DIF extends traditional DIF methods to account for the complex ways in which students' identities (Crenshaw, 1990) interact to potentially influence how they respond to assessment items. Previous intersectional DIF research (e.g., Albano et al., 2024; Russel & Kaplan, 2021) has focused on adopting traditional DIF analysis tools, such as the Mantel Hanzel Test (Holland & Thayer, 1986) and the standardized D (Dorans & Kulich, 1986), for detecting intersectional DIF with large sample sizes.

Rather than relying on a single DIF test to identify biased items, we propose a novel neural network approach to small sample intersectional DIF detection that leverages information from multiple DIF detection tests. Through a rigorous Monte Carlo simulation study, we investigate both uniform and non-uniform DIF detection when either 3 or 10 groups are used with imbalanced group sizes. We also examine how neural network architecture and types of loss functions influence DIF detection results. From the simulation study, we conclude that our proposed intersectional DIF detection method is more effective when small group sizes are used (e.g., 50 students) as compared to traditional DIF detection methods. Our method also displays higher power to differentiate between uniform and non-uniform DIF, albeit with a slightly inflated Type 1 error rate, as compared to the latest multiple group Gaussian variational expectation-maximization model (GVEM; Wang, 2024).

DIF Detection in Ordinal Survey Data Without Pre-Specified Groups or Anchors

Wednesday, 16th July - 16:15: Differential Item Functioning and Measurement Invariance II (GH: Think 4) - Oral

Dr. Gabriel Wallin (Lancaster University)

Measurement non-invariance arises when the psychometric properties of a scale differ across subgroups, undermining the validity of comparisons. At the item level, this manifests as differential item functioning (DIF), where individuals with the same latent trait level respond differently to specific items. This paper introduces a statistical framework for detecting DIF in ordinal survey data, without requiring known group labels or anchor items. We propose a hybrid latent-class item response theory model to ordinal data using a proportional-odds formulation, assigning individuals probabilistically to latent classes. The model captures measurement non-invariance through class-specific intercept shifts and slope modifications. DIF effects are identified via an -penalised marginal likelihood function, assuming most items are invariant. Estimation is carried out using a tailored EM algorithm, and post-estimation confidence intervals provide uncertainty quantification for non-zero effects. Simulation studies demonstrate accurate recovery of both uniform and non-uniform DIF types and satisfactory interval coverage. Application to a widely used anxiety questionnaire reveals latent subgroups with distinct response patterns and identifies items that may bias group comparisons. The proposed framework enables principled assessment of measurement invariance in ordinal-scale survey instruments, even when comparison groups are unobserved or poorly defined and anchor items are difficult or impossible to pre-specify.

DSEM with missing not at random intensive longitudinal data

Wednesday, 16th July - 15:15: Missing Data (GH: Think 5) - Oral

Dr. Daniel McNeish (Arizona State University)

Intensive longitudinal designs are increasingly popular for assessing moment-to-moment changes in mood, affect, and health behavior. Compliance in these studies is never perfect given the high frequency of data collection, so missing data are unavoidable. Missing not at random (MNAR) data are particularly prevalent, especially with sensitive outcomes related to mental health, substance use, or sexual behavior. As a motivating example, a study on people with binge eating disorder that has large amounts of missingness in a self-report item related to overeating is considered. Missingness may be high because participants felt shame reporting this behavior, which is a clear case of MNAR and for which methods like multiple imputation and full information maximum likelihood are less effective. This proposal considers embedding a Diggle-Kenward-type selection model within a dynamic structural equation model to better accommodate MNAR intensive longitudinal data. Analysis of the motivating data show large differences between models assuming data are missing at random (MAR) or the MNAR proposed model, leading to different conclusions and inferential decisions from the two focal parameters in the model. Simulation studies are provided to demonstrate desirable statistical properties of the proposed model when the missing data mechanism in the population is known. Simulation results showed good parameter recovery and credible interval coverage when applying the proposed model to data that were simulated to have an MNAR mechanism, even with relatively (a) small samples (N = 50), (b) few timepoints (T=14), or (c) large amounts of missingness (40%). Limitations and future directions are also discussed.

Using factor scores estimated with missing data

Wednesday, 16th July - 15:30: Missing Data (GH: Think 5) - Oral

Dr. Ehri Ryu (Boston College)

This study investigates the statistical properties of factor scores estimated in a confirmatory factor analysis (CFA) model in the presence of missing data. How are the factor scores affected by missing data in the observed responses when the parameter estimates do not suffer severe bias, particularly with extremely sparse data? With missing-at-random missing data mechanism, full information maximum likelihood (FIML) estimation has been shown to perform well to produce unbiased estimates. Unless a separate exclusion criterion is applied, all observations with at least one non-missing value contribute to the FIML estimation. In multi-factor CFA models, it is possible that a factor score is estimated for cases with no observed indicator score for the factor, if they have non-missing responses to the indicators that load on the other factors. In other words, the factor scores are estimated with complete absence of observed data to measure the construct represented by the latent factor. A simulation study is conducted to empirically examine how the factor scores behave under various conditions: missing data mechanism (MCAR or MAR), missing patterns and proportion of missingness, communality of the indicators with missing values, correlation between the factors, and sample size. Two selection methods are explored to exclude poor factor score estimates.

Methods for pooling K-means clustering results in multiply imputed data

Wednesday, 16th July - 15:45: Missing Data (GH: Think 5) - Oral

Dr. Joost Van Ginkel (Leiden University), Dr. Anikó Lovik (Leiden University)

K-means clustering is a widely used technique to cluster cases in a dataset into a number of groups. When data are incomplete, missing data need to be treated prior to carrying out *K*-means clustering. Multiple imputation is a widely recommended procedure for dealing with missing data, which creates multiple plausible complete versions of the incomplete dataset. When applied to each of these imputed datasets, *K*-means clustering requires a method to combine the cluster solutions of the several imputed datasets into one overall cluster solution. Several combination methods have been proposed, such as majority vote, multiply imputed cluster analysis, and partitions pooling. These methods either come with practical problems, or try to resolve these problems using rather involved procedures. In the current presentation we propose two simple generalizations of the *K*-means clustering algorithm for complete data to multiply imputed datasets, which bypass all the problems that the other methods try to resolve. In a simulation study it is shown that the two newly proposed methods better recover the underlying cluster structure than the existing methods.

Optimal approaches for treating item-level missing data in composite-level models

Wednesday, 16th July - 16:00: Missing Data (GH: Think 5) - Oral

Dr. Victoria Savalei (University of British Columbia)

In many modeling contexts, the variables in the model are composites (e.g., sum scores) made up of components (e.g., items). For example, in regression and path analysis, scale scores are often used as proxies for the constructs of interest. In structural equation models (SEMs), parcels are sometimes used to reduce model size. However, when there is missing data at the item level, it can be difficult to deal with. Ad-hoc approaches (e.g., mean imputation, setting the composite as missing if any of the item scores are missing) are biased and inefficient, and item-level multiple imputation (MI) can be difficult to implement for the applied user. Recently, several superior analytic approaches have emerged: item-level two-stage maximum likelihood approach (TSML; Savalei & Rhemtulla, 2017a), which is the analytic equivalent of item-level MI; the fully efficient GLS approach (Savalei & Rhemtulla, 2017b); and the Pseudo-Indicator Model approach (PIM; Rose, Wagner, Mayer, & Nagengast, 2019), which uses full information maximum likelihood and is thus fully efficient. The first two approaches require custom programming in SEM software, and the last requires setting up a complex SEM that may be difficult for applied researchers. In this presentation I will review these analytic approaches on an example and provide an easy-to-use set of R functions to implement them. I will also report on the results of a simulation designed to compare these approaches empirically, for both path analysis and SEM with parcels. Extensions to nonnormal data will also be discussed.

Comparison of rating accuracy and rationales between AI ratings and human ratings of AP Chinese essays

Thursday, 17th July - 09:00: Automated Scoring (GH: Meridian 3-4) - Oral

Mr. Haowei Hua (The Culver Academies), Prof. Hong Jiao (University of Maryland), Ms. Dan Song (University of Iowa)

Automated scoring is one of the most successful applications of AI in large-scale testing due to its efficiency, scalability, and consistency. The advances in automated scoring are aligned with the advances in machine learning and natural language processing . Though the accuracy of automated scoring based on language models such as BERT has increased across the board, the inconsistency between AI raters and human ratings still persists for some scoring settings. With the recent advances in Large Language Models (LLMs), ChatGPT, and other generative AI chatbots have been explored for automated scoring. Given the strong reasoning capabilities, LLMs can produce rationales to support the score they produce. Thus, evaluating the rationales provided by human raters and AI raters may help to better understand the logic each type of raters apply in assigning a score. This study investigates rationales from human and AI raters to identify potential causes of scoring errors. Using one AP Chinese essay as an example, this study illustrates using LLMs including ChatGPT-4o, Gemini, and other LLMs to automatically score AP Chinese essays and producing the rationales for ratings. Quadratic weighted kappa (QWK) and Normalized Mutual Information (NMI) quantify scoring consistency, while Cosine similarity evaluates rationale similarity. Additionally, clustering patterns in rationales are explored through principal component analysis (PCA) based on the embedding derived from the rationales in addition to detailed scoring error analysis. The findings provide insights into the accuracy and 'thinking' of LLMs in automated scoring, contributing to the interpretability of scores from LLM-based automated scoring.

Evaluating the accuracy, reliability, and applicability of multiple large language models in automated scoring for writing assessments.

Thursday, 17th July - 09:15: Automated Scoring (GH: Meridian 3-4) - Oral

Mr. Henry S Makinde (University of North Carolina at Greensboro), Dr. Stephen Murphy (Focal Point), Mr. Mark Lynch (Learnosity), Mrs. Maria D'Brot (Focal Point)

Advancements in artificial intelligence have paved the way for scalable and cost-effective solutions in educational assessment, particularly in automated essay scoring. This study evaluates the performance of eight large language models including Claude Haiku, Claude Sonnet, GPT-4 variants (GPT4, GPT4omini, GPT4o), Multimodel, Standard Essay Model, and Advanced Essay Model in replicating human rater accuracy and consistency. Numerous prompt engineering techniques were used including Zero-shot, Few-shot, Chain of Thought and Prompt Chaining. Using a dataset of 850 human-scored responses from diverse linguistic and socio-economic backgrounds, we partitioned the data into training and testing sets, employing human scores as the benchmark for evaluation.

We conducted a quantitative analysis using key reliability metrics such as Quadratic Weighted Kappa (QWK), Intraclass Correlation Coefficient (ICC), Root Mean Squared Error (RMSE), and Pearson correlation. Additionally, we integrated natural language processing approaches for content authenticity by leveraging Levenshtein distance for plagiarism detection and dictionary-based methods for gibberish identification. These mechanisms effectively flagged problematic responses, they exhibited a propensity for false positives, indicating the need for further refinement.

Findings reveal that the Advanced Essay Model achieved the highest alignment with human scoring, demonstrating low RMSE and high adjacent agreement rates, with "almost perfect" scores on both QWK and ICC. However, all models, including the top-performing Advanced Essay Model, encountered challenges in replicating exact human scores, underscoring a persistent gap. This study emphasizes the potential of AI-driven scoring systems to complement traditional assessment, provided ongoing improvements in rubric design, model calibration are pursued to enhance fairness, accuracy, and reliability.

Improving automated scoring in reading assessments: Compress first, score next

Thursday, 17th July - 09:30: Automated Scoring (GH: Meridian 3-4) - Oral

Dr. Ji Yoon Jung (Boston College), Dr. Ummugul Bezirhan (Boston College), Prof. Matthias von Davier (Boston College)

This study explores the application of prompt compression in automated scoring (AS), using five constructed response items from the 2021 Progress in International Reading Literacy Study (PIRLS). Lengthy inputs, such as extensive reading passages and complex scoring guides, have hindered the scalability of AS in reading assessments. To address this, we propose a two-step approach for AS: first compressing passages and scoring guides, and then scoring student responses. Previous studies have shown that compression through large language models can effectively condense input length while retaining essential information (e.g., Jiang et al., 2023; Dinesjö & Floreteng, 2024). Specifically, we utilized OpenAI's GPT-4o to generate question-specific compressed summaries of reading passages and simplified scoring guides. Also, we developed a generalized scoring template for AS that can be applied broadly without the need to fine-tune passages or scoring guides for each item. Results indicate that reading passages and scoring guides can be compressed to approximately 18% and 15% of their original lengths, respectively, without compromising the performance of AS. Remarkably, despite this substantial compression, the AS achieved an accuracy of 92.87% and a kappa score of 0.8041 across all 27 participating countries in PIRLS, demonstrating robust multilingual performance. This study highlights the potential of prompt compression to improve the cost-efficiency and scalability of AS in international reading assessments.

Investigation of NLP and ML-based algorithms for automated essay scoring

Thursday, 17th July - 09:45: Automated Scoring (GH: Meridian 3-4) - Oral

Prof. JiHoon Ryoo (Yonsei university), Mr. Jeongheum Cho (Yonsei university), Mr. YeongJin Jo (Yonsei university)

Automated Essay Scoring (AES) has been developed along with the development of generative AI. Recent studies show that algorithms showing the efficiency for AES encompass Natural Language Process (NLP) and Machine Learning (ML) and those perform better than traditional methods such as regression-based ones or ML without NLP. There are many competitions and research outcomes published and thus, it is not hard to find comparative studies utilizing LSTM, Bi-LSTM, BERT, RoBERTa, Neural Pairwise Contrastive Regression (NPCR), etc. However, the State-Of-The-Arts (SOTAs) found among algorithms seem to be dependent on the evaluation indexes applied. In addition, most studies have been conducted essays written in English or using a machine translation.

First, this study examines the efficiency of the algorithms (also called models) across various evaluation indexes including Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Mean Squared Error (MSE) and Accuracy. Automated Student Assessment Prize (ASAP) dataset will be used. Second, this study examines the effectiveness of two Korean-based algorithms, such as KPF-BERT and KoBERT, for essays that are translated from the ASAP essays. It is hypothesized that AES of Korean (or other languages than English) essays would perform better with Korean-based algorithms than English-based algorithms with machine translations. In addition, we also consider the types of essays in terms of length and subjects so that the results can be generalized in various formats of essay scoring. Study results would not only inform the SOTA algorithm but also provide an option for AES within adaptive testing systems.

Evaluating strategies for handling label switching in Bayesian latent variable models

Thursday, 17th July - 09:00: Bayesian Methods and Their Applications (MAC: Johnson) - Oral

Dr. Lihan Chen (McGill University, Montreal, Canada), Prof. Carl Falk (McGill University, Montreal, Canada), Prof. Milica Miocevic (McGill University, Montreal, Canada)

Reflection invariance refers to when a latent variable model stays equivalent as a column of indicator loadings and corresponding path coefficients switch signs, i.e., they are multiplied by minus one. It commonly occurs when a latent variable model is identified using a *unit variance identification* (UVI) strategy, which does not place any constraints on the signs of indicator loadings. This can cause a *label switching* problem in the Bayesian estimation of latent variable models, as multiple chains may yield estimates that differ in signs, causing nonconvergence and invalid estimates. The *unit loading identification* (ULI) strategy eliminates label switching but can be dependent on the selection of appropriate reference indicators, so it may not always be the preferred identification strategy. Many different strategies under UVI have been used to handle label switching. In empirical literature, researchers sometimes use truncated priors or lognormal priors to constrain the signs of loadings. Erosheva et al. (2017) proposed a *relabeling* algorithm that relabel the posterior distributions from multiple chains *post hoc*. The *R* package *blavaan* adopts an “online” strategy that flips the signs of the loadings between iterations of the Bayesian estimation. We conducted a simulation study with a three factor model, varying sample sizes and loading strengths, and performed Bayesian estimation using UVI and ULI, as well as various label switching handling strategies, including different constrained priors strategies, post hoc relabeling, and online relabeling. We highlight the potential downsides of each approach, and provide recommendations for label switching handling for Bayesian latent variable models.

Prior sensitivity in Bayesian structure learning of Gaussian graphical models

Thursday, 17th July - 09:15: Bayesian Methods and Their Applications (MAC: Johnson) - Oral

Mr. Marwin Carmo (The University of California, Davis), Dr. Phillippe Rast (The University of California, Davis)

Understanding the conditional dependency structure among variables—such as symptoms of mental health disorders or personality traits—is a central goal in psychological research. This is often accomplished using Gaussian Graphical Models, where edges in the graph represent conditional dependencies between variables. In these regards, Bayesian methods for structure learning offer unique advantages, such as quantifying uncertainty and incorporating prior knowledge, but their results can be highly sensitive to the choice of prior distributions. Despite recent advancements in Bayesian algorithms—such as reversible jump MCMC, birth-death MCMC, spike-and-slab, and horseshoe methods—the impact of prior specifications on structure learning remains underexplored, particularly in psychological applications. In this study, we systematically investigate prior sensitivity in Bayesian structure learning by comparing the performance of state-of-the-art algorithms under a range of prior settings. Using simulated data, we evaluate how different priors (e.g., graph sparsity and precision matrix hyperparameters) influence the inferred graph structure and precision matrix. Our results address how sensitive the different Bayesian structure learning algorithms are to the choice of prior distributions and which algorithms are most robust or sensitive to prior misspecification. The study to provides practical recommendations for setting hyperparameters and advice on choosing priors based on sample size, dimensionality, and expected sparsity of the data.

Infinitesimal Jackknife standard errors for Bayesian quantile regression and other misspecified models

Thursday, 17th July - 09:30: Bayesian Methods and Their Applications (MAC: Johnson) - Oral

Dr. Sophia Rabe-Hesketh (University of California, Berkeley), Dr. Feng Ji (University of Toronto), Dr. JoonHo Lee (The University of Alabama)

Infinitesimal jackknife (IJ) standard errors (Giordano and Broderick, 2024) are frequentist standard errors for Bayesian estimators (posterior means) that do not require resampling but can be obtained from a single MCMC run. IJ standard errors are robust to model misspecification and can be adapted to take clustering into account. In standard Bayesian quantile regression, an asymmetric Laplace (AL) distribution is used for the likelihood. The AL likelihood is chosen merely because the corresponding maximum likelihood estimator is identical to the classical quantile regression estimator by Koenker and Bassett (1978). The model is misspecified because there is no reason to believe that the AL distribution is a plausible data-generating mechanism. While point estimation is consistent, credible intervals tend to have poor frequentist coverage. Yang et al. (2016) proposed an adjustment to the posterior covariance matrix that produces asymptotically valid intervals. However, we show that this adjustment is sensitive to the scale parameter of the AL distribution and can lead to poor coverage when the sample size is small to moderate. We therefore propose using IJ standard errors. Simulations and applications to real data show that the IJ standard errors have good frequentist properties, both for independent and clustered data. We provide an R package, **IJSE**, that computes IJ standard errors after estimation of any model with the **brms** wrapper for **Stan**. We believe that IJ standard errors will become as popular for Bayesian inference as “robust” standard errors (based on the sandwich estimator) for frequentist inference.

Are Bayesian regularization methods a must for multilevel dynamic latent variables models?

Thursday, 17th July - 09:45: Bayesian Methods and Their Applications (MAC: Johnson) - Oral

Mr. Vivato Vahatriniaina Andriamiarana (Eberhard Karls Universität Tübingen), Dr. Pascal Kilian (Eberhard Karls Universität Tübingen), Prof. Holger Brandt (Eberhard Karls Universität Tübingen), Prof. Augustin Kelava (Eberhard Karls Universität Tübingen)

Due to the increased availability of intensive longitudinal data, researchers have been able to specify increasingly complex dynamic latent variable models. However, these models present challenges related to overfitting, hierarchical features, non-linearity, and sample size requirements. There are further limitations to be addressed regarding the finite sample performance of priors, including bias, accuracy, and Type-I error inflation. Bayesian estimation provides the flexibility to treat these issues simultaneously through the use of regularizing priors. In this paper, we aim to compare several Bayesian regularizing priors (ridge, Bayesian Lasso, adaptive spike-and-slab Lasso, and regularized horseshoe). To achieve this, we introduce a multilevel dynamic latent variable model. We then conduct two simulation studies and a prior sensitivity analysis using empirical data. The results show that the ridge prior is able to provide sparse estimation while avoiding overshrinkage of relevant signals, in comparison to other Bayesian regularization priors. In addition, we find that the Lasso and heavy-tailed regularizing priors do not perform well compared to light-tailed priors for the logistic model. In the context of multilevel dynamic latent variable modeling, it is often attractive to diversify the choice of priors. However, we instead suggest prioritizing the choice of ridge priors without extreme shrinkage, which we show can handle the trade-off between informativeness and generality, compared to other priors with high concentration around zero and/or heavy tails.

Bayesian graphical models for factorial correlation estimation

Thursday, 17th July - 10:00: Bayesian Methods and Their Applications (MAC: Johnson) - Oral

Ms. Yifan Zhang (The University of Hong Kong), Prof. Jinsong Chen (The University of Hong Kong)

This research tries to use Bayesian graphical models in factorial correlation estimation to accommodate different factor structures.

In modern social science, many complex structures appear to accommodate rich data and psychological phenomena like bifactor models with multiple general factors and MTMM. Factorial correlation in such structures is partially correlated and sparse.

Psychometricians have introduced the graphical Lasso to the residual matrix in factor analysis, which leads to a sparse solution of the residual matrix and not necessarily diagonal (Chen, 2020; Pan et al., 2017). We consider three Bayesian graphical estimators (Lasso, horseshoe, SSP) for factorial correlation, in which horseshoe and SSP have been relatively unexplored in latent variable models.

Unlike the residual matrix and other graphical model scenarios, the sampling matrix in this study does not include observed variables (independent-centered observations) but latent factor scores. This Bayesian regularized factorial correlation estimator will be integrated into Partially Confirmatory Latent Variable Modeling (PCLVM, Chen & Zhang, 2024) using a block Gibbs sampler.

From preliminary simulation studies, we found that traditional methods and three regularized methods behaved similarly in the case of simple CFA with/without local dependence and minor factors. In multiple general multiple group factor models, graphical Lasso and graphical SSP have better convergency rates and standard errors. These two methods can significantly recognize all nonzero factorial correlations while the traditional method and graphical horseshoe cannot.

Finally, we plan to demonstrate the new factorial correlation methods by two Personality Inventory real-data examples.

Between-case incidence rate ratio: A design comparable effect size for count outcomes in single case experimental designs

Thursday, 17th July - 09:00: Causal Inference I (MAC: Thomas Swain) - Oral

Dr. Wen Luo (Texas A&M University), Dr. Haoran Li (University of Minnesota), Dr. Eunkyeng Baek (Texas A&M University), Mr. Chendong Li (Texas A&M University)

Single-case experimental designs (SCEDs) are essential for evaluating interventions in psychological, educational, and behavioral research. While the field has made significant progress in developing effect size measures for SCEDs, existing metrics, such as the between-case standardized mean difference (BC-SMD), have limitations when applied to count outcomes. Count data often violate assumptions of normality and homoscedasticity, presenting challenges for accurate effect size estimation. This study addresses the need for an alternative effect size measure tailored to count outcomes in SCEDs that is directly comparable to the incidence rate ratio (IRR) used in between-subjects designs. We define the between-case incidence rate ratio (BC-IRR) within a counterfactual framework and demonstrate its equivalence to the IRR in randomized controlled trials under specific conditions. Using real-world SCED data, we illustrate the estimation of BC-IRR via generalized linear mixed models (GLMMs). The study also discusses the assumptions underlying BC-IRR, its potential misuse, and its limitations while providing directions for future research.

Between-case incidence rate ratio for count outcomes in single case experimental designs: A Monte Carlo simulation

Thursday, 17th July - 09:15: Causal Inference I (MAC: Thomas Swain) - Oral

Dr. Haoran Li (University of Minnesota), Dr. Wen Luo (Texas A&M University), Dr. Eunkyeng Baek (Texas A&M University), Mr. Chendong Li (Texas A&M University)

Single-case experimental designs (SCEDs) have become increasingly reviewed and included in research synthesis to identify evidenced based interventions in psychological, educational, and behavioral research. Despite numerous effect size measures proposed to allow researchers to compare and synthesize findings in single case studies answering the same research question, there is a lack of design-comparable effect size measures tailored to count-based outcome metrics commonly encountered in SCEDs. This study systematically evaluates an alternative effect size measure, namely, between-case incidence rate ratio (BC-IRR), that is directly comparable to the incidence rate ratio (IRR) used in randomized control trials (RCTs). Using simulation studies, we evaluate the performance of generalized linear mixed models (GLMMs) and generalized estimating equation (GEE) regarding their accuracy and efficiency to estimate BC-IRR under various conditions. We also illustrate small sample corrections methods and delta methods for calculating standard errors of BC-IRR. To provide guidance for SCED researchers, we present a step-by-step analytical procedure using a real example, while discussing the limitations of this study and directions for future research.

Using regression discontinuity to evaluate language learner reclassification

Thursday, 17th July - 09:30: Causal Inference I (MAC: Thomas Swain) - Oral

Dr. Hirotaka Fukuhara (Pearson), Dr. Dipendra Subedi (Pearson), Dr. Anju Kuriakose (Arizona Department of Education)

This study uses Regression Discontinuity Design (RDD) to evaluate whether the proficiency cut score for exiting a language learner program is appropriately set based on empirical evidence. RDD helps determine whether the cut score is too high, unnecessarily delaying exit, or too low, leading to premature reclassification. Unlike traditional observational studies, which suffer from selection bias due to inherent differences between exiting and remaining students, RDD compares students only at the margin of exit eligibility, ensuring that post-exit differences are not driven by pre-existing characteristics (Robinson, 2011). Robinson (2011) highlights that traditional ELL exit criteria often result in over-retention or early exit. Using data from language assessment in a state, the current study examines the impact of reclassification on assessment outcomes, controlling for the binding proficiency score. We also performed a sensitivity analysis by adding demographic variables into the final linear regression model as covariates. The findings support the appropriateness of the current cut score for reclassification. While RDD is sometimes criticized for limited generalizability to students far below or above the cutoff, its strength lies in identifying impacts for those most affected by policy decisions (Imbens & Lemieux, 2008). By leveraging RDD in practical research such as this study, policymakers can optimize proficiency thresholds to balance adequate language support with timely integration into mainstream instruction.

Causal inference in high dimensional settings via sparse autoencoders for improved propensity scores estimation.

Thursday, 17th July - 09:45: Causal Inference I (MAC: Thomas Swain) - Oral

Mr. Roberto Faleh (Eberhard Karls Universität Tübingen), Ms. Sofia Morelli (Eberhard Karls Universität Tübingen), Prof. Holger Brandt (Eberhard Karls Universität Tübingen)

In many observational studies, traditional methods for causal inference struggle due to challenges such as high dimensionality, nonlinear treatment assignments, and residual confounding. To address these issues, we propose a novel method for estimating propensity scores in these complex settings. Our approach employs a Deep Neural Network integrated with a Sparse Autoencoder (e.g., Ghosh 2021), which is designed to extract low-dimensional, relevant features from complex, high-dimensional data. To create a more compact and structured latent representation that improves counterfactual prediction and enhances the estimation of the conditional probabilities distribution of the data, we implement Integral Probability Measures (IPMs), such as the Wasserstein distance, in the architecture.

With this extension, we reduce bias and enhance the accuracy of causal estimates, even in the presence of complex treatment assignment mechanisms. We explore the effectiveness of our approach through empirical evaluations on both synthetic and semi-synthetic real-world datasets.

Causal decomposition analysis with synergistic interventions: A triply-robust machine learning approach to addressing multiple dimensions of social disparities

Thursday, 17th July - 10:00: Causal Inference I (MAC: Thomas Swain) - Oral

Dr. Soojin Park (University of California, Riverside), Ms. Su Yeon Kim (University of California, Riverside), Dr. Chioun Lee (University of California, Riverside)

In this talk, we will introduce a new framework for causal decomposition analysis that addresses multiple intervening factors, allowing the evaluation of synergistic effects. In the United States, social disparities persist across race, gender, and socioeconomic status (SES). Most intervention strategies focus on a single domain and evaluate their effectiveness using causal decomposition analysis. However, a growing body of research suggests that single-domain interventions may be insufficient for individuals facing multiple forms of marginalization. While interventions across multiple domains are increasingly proposed, there is limited guidance on appropriate methods for evaluating their effectiveness. These scenarios often involve challenges related to model misspecification due to complex interactions among group categories, intervening factors, and their confounders with the outcome. To mitigate these challenges, we propose a triply-robust estimator that leverages machine learning techniques to address potential model misspecification. Specifically, within our new framework, we 1) define the effect of synergistic interventions in reducing disparities, 2) provide identification assumptions and results, and 3) develop a robust estimation method. We apply our method to a cohort of students from the High School Longitudinal Study (HSLS:2009), focusing on math achievement disparities between Black, Hispanic, and White high schoolers. Specifically, we examine the effects of equalizing the proportion of students taking Algebra I by 9th grade and benefiting from high-quality schools across racial groups, allowing their synergistic effects on reducing educational disparities. We anticipate that this proposed framework can be applied to a wide range of topics aimed at reducing social disparities through synergistic interventions.

Defining asymmetric item response theory

Thursday, 17th July - 09:00: Novel IRT Models (GH: Think 4) - Oral

Dr. Leah Feuerstahler (Fordham University), Prof. Jay Verkuilen (City University of New York), Mr. Fabio Setti (Fordham University), Mr. Peter Johnson (City University of New York)

Asymmetric item response theory (AsymIRT) is a subfield of item response theory (IRT) that challenges the dominance of the symmetric normal ogive and logistic shapes of item response functions (IRFs) for defining the relationship between latent traits and observed responses. In recent years, there has been a boom in AsymIRT research such that there is now a variety of models and motivations that encourage more widespread adoption of AsymIRT. For example, AsymIRT has been developed from the perspectives of (a) a philosophy of scoring, (b) tracking cognitive response processes, (c) considerations of the theoretical trait distribution, and (d) providing more flexibility in the modeling framework. Although Samejima (2000) provided a definition for symmetric item response models, not all models that fail to meet Samejima's definition address all motivations for AsymIRT. Therefore, the purpose of this talk is to identify the relevant features of AsymIRT models that correspond to different motivations and to provide a cohesive framework in which to select and apply AsymIRT to real data sets. We will frame the discussion around P', that is, the first derivative of an item response function, and whether P' is an even function around a point of symmetry. In addition to analyzing the features relevant to each motivation and how they are addressed by specific AsymIRT models, we will consider the extent to which P' may be used to quantify the amount of asymmetry in a given item response function.

Bayesian IRT for continuous measurement of student proficiency

Thursday, 17th July - 09:15: Novel IRT Models (GH: Think 4) - Oral

Dr. Erin Banjanovic (Curriculum Associates), Dr. Ted Daisher (Curriculum Associates), Dr. Logan Rome (Curriculum Associates)

In today's classrooms, students interact with technology on a nearly daily basis, including digital assessments, instruction, and learning games. Many of these activities present opportunities for measurement but do not elicit enough measurement information to produce reliable and valid scores. To continuously update student proficiency estimates over time requires a dynamic measurement model capable of using information across many activities where learning happens between activities.

In this paper, we present a Bayesian extension of item response theory (IRT) that borrows elements from Glicko, a paired-comparisons algorithm used to track chess player ratings over time. Glicko-IRT uses proficiency estimates across multiple time points, together with information from student growth trajectories to build a prior distribution for estimation of student proficiency at a given timepoint. The model then uses closed-form equations to update the posterior distribution, leveraging information from the prior and students' performance on the given activity.

Specifically, this paper explores several approaches to deriving the prior distribution for Glicko-IRT. Based on growth trajectories observed from a national interim assessment, we simulate item responses to fixed-form activities administered at several timepoints across an academic year. We compare three methods for deriving the mean of the prior distribution and four methods for deriving the variance of the prior distribution for (3 x 4) twelve total methods. For each method, we examine the relationship between true and estimated proficiency at each timepoint to determine the optimal method for deriving the prior distribution for Glicko-IRT.

Extending the quasi-Poisson IRT model: On choosing latent structure

Thursday, 17th July - 09:30: Novel IRT Models (GH: Think 4) - Oral

Dr. Nelis Potgieter (Texas Christian University), Dr. Xin Qiao (University of South Florida)

Building on recent advances in item response theory (IRT) for count data, this work extends the Quasi-Poisson model of Qiao and Potgieter (2025), who consider unbounded count responses. Specifically, this research broadens the scope of application by allowing for the selection of various latent ability distributions. Traditional count-based IRT models often impose rigid parametric assumptions on the observed counts. In contrast, the Quasi-Poisson model avoids these constraints by specifying only the conditional mean and variance given the latent ability. The key to this flexibility lies in the moment generating function (MGF) of the latent ability, which determines the unconditional means and variances, allowing for the deployment of a richer class of latent distributions.

In this presentation, we explore normal, Laplace, and skew-normal latent structures and develop a Generalized Method of Moments (GMM) approach for parameter estimation in the Quasi-Poisson framework. Simulation studies highlight how well GMM recovers parameters across different latent specifications, while empirical applications demonstrate its practical value in real-world assessments. We also address the challenge of assessing model fit, offering practical tools for evaluating latent structure choices in count-based IRT modeling. By extending the one- and two-parameter Quasi-Poisson frameworks, this work provides a principled and computationally efficient way to integrate flexible latent distributions into the analysis of count data.

Modeling dynamic test-taking behavior: A response time based HMM-IRT approach

Thursday, 17th July - 09:45: Novel IRT Models (GH: Think 4) - Oral

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

The increasing use of computerized testing provides access to item response times (RT) alongside item responses, offering insights into test engagement. Incorporating RT can enhance the accuracy of ability estimates, as faster responses may indicate less careful consideration, leading to rapid guessing and biased estimates of item and person parameters (e.g., Rios & Soland, 2020). Existing models that integrate RT and item response theory (IRT) include effort-moderated (EM-IRT) models (Wise & DeMars, 2006), mixture Rasch models (Meyer, 2010), and hierarchical mixture models (e.g., Wang & Xu, 2015; Molenaar et al., 2016, 2019). These approaches classify examinees as engaged or disengaged but assume static engagement throughout the test. However, test-taking behavior is dynamic, with engagement levels shifting across test sections.

To address this limitation, this study proposes an RT-HMM-IRT model, integrating RT, Hidden Markov Models (HMM), and the two-parameter logistic (2PL) IRT model to capture dynamic engagement shifts. This approach allows examinees to transition between careful engagement and rapid guessing states, providing a more nuanced understanding of test-taking behavior. Unlike prior models, it employs a fully Bayesian approach using the no-U-turn sampler (NUTS) for improved efficiency.

Monte Carlo simulations were conducted to assess the model's performance across test difficulty (easy, moderate, hard) and fatigue (high, mild, and low) levels, comparing its results with conventional RT-integrated IRT models. The accuracy in recovering model parameters was assessed using bias and root mean squared error (RMSE). Findings highlight the proposed model's advantages in capturing test-taking behavior dynamics and improving estimation accuracy over alternative models.

The two-parameter quasi-Poisson item response theory model

Thursday, 17th July - 10:00: Novel IRT Models (GH: Think 4) - Oral

Dr. Xin Qiao (University of South Florida), Dr. Nelis Potgieter (Texas Christian University)

It is common that cognitive, educational, and psychological assessments generate count data. Qiao and Potgieter (2025) proposed a one-parameter Quasi-Poisson item response theory (IRT) model that handles equidispersion, overdispersion, and underdispersion in count data. This model has several advantages. First, it provides more accurate statistical inference than existing methods, such as the Rasch Poisson Counts Model (RPCM) and the negative binomial IRT model (NBM), when their assumptions on count data dispersions are violated. Second, it includes dispersion parameters that are directly comparable to RPCM and NBM, thus facilitating interpretations when these models are used together. The existing Conway-Maxwell-Poisson IRT model that also handles heterogeneous dispersions, however, has dispersion parameters on a different scale than NBM. Third, the Quasi-Poisson model has superior computation efficiency than existing methods.

This research extends the Quasi-Poisson model (Qiao & Potgieter, 2025) to a two-parameter Quasi-Poisson IRT model that allows both varying discriminations and dispersions, aiming to better accommodate data from count data tests. It is a semiparametric model specifies the first two conditional moments for the count variables and derives marginal moments to estimate model parameters. Simulation studies demonstrate the two-parameter Quasi-Poisson model's efficacy in parameter recovery across different scenarios. Empirical data analysis further illustrates the application of the model in a real-world setting.

Development and application of the random effect diagnostic classification multilevel growth curve model

Thursday, 17th July - 09:00: Longitudinal Cognitive Diagnostic Models (GH: Think 5) - Oral

Dr. Kazuhiro Yamaguchi (University of Tsukuba), Dr. Haruhiko Mitsunaga (Nagoya University), Mr. Shun Saso (The University of Tokyo), Dr. Yuri Uesaka (The University of Tokyo)

Diagnostic classification models have been developed to assess students' learning status and provide remedial instructions. However, the impact of mastery or non-mastery of specific attributes on long-term learning development remains uncertain. If certain non-mastered attributes hinder the growth of mathematical ability, early intervention becomes essential. In this study, we developed a random effect diagnostic classification for multilevel growth curve (RDC-MGC) model to identify the specific effects of attribute mastery on the individual-level mathematics ability growth. The model was applied to arithmetic test data from second- to sixth-grade elementary students. Diagnosis was conducted in the second grade, and the effects of mastery on mathematics ability growth from the third to sixth grades were assessed. The results showed that attribute mastery in the second grade influenced both the intercept and slope of individual ability growth, highlighting the importance of early-stage diagnosis in supporting mathematical development.

Longitudinal designs for diagnostic models: Identification and estimation

Thursday, 17th July - 09:15: Longitudinal Cognitive Diagnostic Models (GH: Think 5) - Oral

Mr. TRUNG LE (University of Illinois Urbana-Champaign), Dr. Steven Culpepper (University of Illinois Urbana-Champaign), Dr. Jeffrey Douglas (University of Illinois Urbana-Champaign)

Recent studies on cognitive diagnostic models (CDMs) have extended the framework to longitudinal data. Various methods combined CDMs and hidden Markov models (HMMs) to assess changes in attributes over time, and to evaluate the effects of interventions and covariates. A requirement for model fitting, inference, and interpretation is that models are identifiable. Methods for investigating this assumption are made under a variety of conditions. In this paper, we outline general techniques for deriving identifiability conditions in common experimental design used in education research. Specifically, we consider three typical designs: counterbalancing, pretest/posttest single group design, and multiple-group longitudinal design. For each, we propose the identifiability constraints, and comment on the general techniques used to derive such constraints. We then focus on the multiple-group longitudinal design which is widely used for evaluating intervention effects. A general HMM model is introduced, and its parameters are estimated using a Gibbs Sampling algorithm under a Bayesian framework. We assess parameters recovery through a Monte Carlo simulation study and apply our model to a real dataset from a study which evaluates the effectiveness of two interventions relative to a control condition. The results demonstrate the flexibility of the model and its potential to offer new insights into learning processes compared to existing methods.

A statistical framework for dynamic cognitive diagnosis in digital environments

Thursday, 17th July - 09:30: Longitudinal Cognitive Diagnostic Models (GH: Think 5) - Oral

Ms. Yawen Ma (Lancaster University), Dr. Gabriel Wallin (Lancaster University), Dr. Anastasia Ushakova (Lancaster University), Prof. Kate Cain (Lancaster University)

This study introduces a Bayesian estimation framework to enhance cognitive diagnostic models (CDMs) by estimating the unknown Q-matrix directly from real-world data, addressing key limitations of existing models. The framework jointly estimates the Q-matrix, item parameters under Deterministic Input, Noisy “And” Gate (DINA) models, and student latent skill mastery states across multiple time points using Markov Chain Monte Carlo (MCMC) methods. By incorporating covariate effects, it provides insights into how students acquire skills and transition between mastery states. We applied this framework to log files from a digital learning environment. The empirical analysis shows that the model effectively captures Q-matrix that maps items to required skills, as well as students’ skill states and transitions over time, offering detailed insights into student learning trajectories. When applied to real-world data, the model demonstrates practical value in early educational settings. Rigorous validation through simulation studies confirms the effectiveness of the MCMC algorithm. Simulations highlight the robustness of this framework under various conditions, including different sample sizes, different number of items, and sparsity levels of Q-matrix. The results confirm that the Q-matrix can be recovered directly from the response data with promising accuracy, along with other key parameters. Overall, this study demonstrates the comprehensiveness of such data-driven approach to accurately capture transitions in learning trajectories using real-world data and simulations, offering valuable insights into student learning processes and advancing personalized interventions in digital learning environments.

Leveraging Computerized Adaptive Testing (CAT) to overcome teaching and learning challenges in gateway STEM courses at U.S. universities

Thursday, 17th July - 09:00: Computer Adaptive Testing II (GH: Think 3) - Oral

Dr. Hua-Hua Chang (Purdue University)

At most large universities in the U.S., introductory STEM courses are taught in massive lecture halls, often enrolling over a thousand students annually. This “one-size-fits-all” approach, typically lacking sufficient teaching staff and resources (e.g., TAs or graders), contributes to high DFW rates (grades of D, F, or Withdrawal), often reaching 30%-50%. The rates are even higher for underrepresented minority (URM) students; for instance, a recent Algebra and Trigonometry course at a Midwestern university saw a 66% DFW rate among URMs.

These “gateway” courses are critical prerequisites for STEM majors, and failing them can force students to switch majors or extend their time to degree completion. To address this challenge, we developed a CAT-based diagnostic testing tool that provides personalized assessments and targeted feedback. By leveraging CAT and cognitive diagnostic models, the tool helps instructors pinpoint students’ learning gaps while ensuring fairness by identifying biased items.

This innovative approach has the potential to significantly improve student outcomes and promote equity in STEM education. Additionally, this paper explores how CD-CAT can enhance Gen-AI applications in education, particularly in designing personalized learning pathways and delivering tailored insights. We discuss how CAT-based assessments can optimize individualized learning experiences while balancing detail and brevity in instructional feedback.

You can’t tuna fish, but can you tune a CAT?

Thursday, 17th July - 09:15: Computer Adaptive Testing II (GH: Think 3) - Oral

Mr. Joseph DeWeese (University of Minnesota), Prof. David Weiss (University of Minnesota)

The performance of computerize adaptive tests (CATs) can heavily depend on values of tuning parameters of its component algorithms, such as the standard error (SE) cutoff value used for a SE termination rule. However, these tuning parameter values are often chosen in an unsystematic manner, especially in the CAT variable-length termination criteria literature. This has resulted in comparisons of CAT termination methods that lack interpretability; it is unclear to what extent observed differences in CAT performance are due to differences in termination methods or in how (or if) the methods were tuned. This study provides a preliminary evaluation of a general framework for systematically selecting CAT component algorithms and tuning their parameter values. Because the CAT tuning parameter space is often of low dimensionality, a grid search across a specified subset of the parameter space is feasible, using simulated CAT data. The search is guided by a user-created objective function that incorporates the goals of CAT, such as jointly minimizing measurement error and test length. It is not expected that there is a single optimal objective function formulation, but rather the formulation of the objective function may vary depending on the use case and goals of the test. However, for a single comparison scenario, all comparisons are made using the same formulation of the objective function. This study explores different possible formulations of this objective function and their properties, and preliminary simulation results show that multiple formulations are viable, at least for the comparison of CAT termination methods.

Incorporating omission behaviors into computerized adaptive testing: A psychometric evaluation using IRTree models

Thursday, 17th July - 09:30: Computer Adaptive Testing II (GH: Think 3) - Oral

Ms. Lixin (Lizzy) Wu (University of Illinois Urbana-Champaign), Dr. Huan (Hailey) KUANG (Florida State University), Dr. Justin Kern (University of Illinois Urbana-Champaign)

Computerized adaptive testing (CAT) improves measurement efficiency by dynamically selecting items based on an examinee's ability level. However, traditional CAT methodologies, particularly those employing the unidimensional two-parameter logistic (2PL) item response theory (IRT) model, often overlook omission behaviors—either treating omitted responses as incorrect or applying simplistic imputations. These limitations can introduce systematic bias in ability estimation and compromise test score validity.

This study addresses these issues by integrating omission behaviors into CAT using an Item Response Tree (IRTree) framework, which decomposes response processes into distinct dimensions, such as omission tendencies and ability traits. Two IRTree models—a multi-unidimensional and a bifactor IRTree model—were implemented alongside a conventional unidimensional 2PL IRT model, which served as a baseline. A Monte Carlo simulation study was conducted to evaluate model performance across varying sample sizes (500, 1000, 2000), fixed-length CAT administrations (10, 20, 30 items), and item bank sizes (100, 200, 500). To ensure a rigorous assessment, each model also functioned as a data-generating model, with omission and correct response rates systematically controlled across conditions. After initial calibration, CAT simulations based on each model underwent 100 replications, with performance evaluated using bias, root mean squared error (RMSE), estimated-to-true trait correlations, test overlap rates, and item exposure distributions.

Results indicate that IRTree-based CAT models, particularly the bifactor IRTree approach, reduce bias and RMSE for omission dimensions while maintaining comparable ability estimation precision. These findings highlight the benefits of modeling omissions to enhance score validity and fairness, particularly in high-stakes assessments.

Post-hoc multiple comparison tests in adaptive measurement of change

Thursday, 17th July - 09:45: Computer Adaptive Testing II (GH: Think 3) - Oral

Mr. Raj Wahlquist (University of Minnesota), Prof. David Weiss (University of Minnesota)

To measure psychological change in an individual, IRT-based computerized adaptive tests can be administered at multiple occasions to determine if significant intra-individual change in the latent trait has occurred for that person. This procedure is known as adaptive measure of change (AMC), which is, specifically, the measurement and testing for significant individual psychometric change in estimated trait levels at two or more measurement occasions using null hypothesis significance testing (Finkelman et al., 2010; Tai et al., 2023). However, only omnibus tests have been implemented for AMC when data from more than 2 measurement occasions are present (Phadke, 2017; Tai et al., 2023). This necessitates the use of post-hoc comparison tests to show which pair of measurement occasions are psychometrically different. Because AMC uses null hypothesis significance testing there is a concern that doing multiple tests per person will inflate the family-wise Type 1 error rate. This study developed and evaluated different pairwise comparison tests to identify which pairs of θ were significant from each other, i.e., when significant change had occurred. Using Monte-Carlo simulation, results showed that the comparison tests were able to successfully control the family-wise Type 1 error rate with only a moderate decrease in power. Further, using an ANOVA framework, this study also was able to ascertain which variables had the most effect on the power and Type 1 error. These variables were type of item bank, amount of psychometric change, number of items, number of measurement occasions, and starting level.

Adapting transformers to wording-based item difficulty prediction

Thursday, 17th July - 13:30: Artificial Intelligence III (GH: Meridian 1-2) - Oral

Mr. Jan Netík (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University)

Accurate item difficulty estimation is a fundamental challenge in high-stakes educational test construction. Traditionally, difficulty is estimated through small-scale pretesting or expert judgments. However, pilot samples may not accurately represent the target population, estimates may exhibit large errors, and expert ratings are not sufficiently reliable. Alternatively, difficulty can be derived from item wording, but most approaches use various text features extracted from extensively preprocessed wording, leading to a loss of crucial morphosyntactic information. To address this, recent research (Netík et al., 2024) has explored fine-tuning transformer models that enable circumventing the preprocessing step, yet improvements over feature-based methods have been only marginal, presumably due to insufficient item wording representation.

This paper introduces architectural modifications to improve model performance by ensuring a more comprehensive representation of the item. We propose three key changes: (1) a multitask learning framework that jointly predicts item difficulty and classifies wording components to encourage balanced attention distribution, (2) an encoding strategy that processes wording parts separately before merging them into a unified representation, and (3) leveraging deeper layers of the encoder stack instead of relying solely on the last hidden states of the first token. Additionally, we compare these modifications with the few-shot learning capabilities of state-of-the-art large language models and explore the potential of using their predictions to generate training data for fine-tuning smaller models.

We aim to contribute to a more robust transformer-driven solution for item difficulty prediction, potentially reducing reliance on established yet limited approaches while improving the precision of difficulty estimates.

Automated cognitive feature generation: LLM applications in item difficulty modeling

Thursday, 17th July - 13:45: Artificial Intelligence III (GH: Meridian 1-2) - Oral

Mr. Mubarak Mojoyinola (University of Iowa)

This study explores the potential of Large Language Models (LLMs) to automate the generation of cognitive features for item difficulty modeling (IDM) in educational assessment. Traditional IDM approaches rely on cognitive features manually coded by trained raters, which limits scalability. While recent studies have utilized Natural Language Processing to extract textual features, cognitive features remain challenging to automate.

Using Claude 3.5 Sonnet with zero-shot prompting, we generated six cognitive features for 714 Grade 3-5 mathematics items: number of unknown parameters, computation steps, relative definition of unknowns, problem type (abstract/real-world), equation type, and depth of knowledge. These were combined with fourteen text-based features including word counts, readability indices, and digit counts to predict item difficulty using random forest models.

Results show that models incorporating both text-based and cognitive features achieved superior predictive performance (R^2=0.10) compared to models using only text-based (R^2=0.04) or cognitive features (R^2=0.04) alone. Feature importance analysis revealed that word character count, depth of knowledge, total character count, Flesch reading ease, and stem word count had the greatest impact on model accuracy.

This study demonstrates the potential of using LLMs to generate theoretically grounded cognitive features that would traditionally require extensive human effort. By automating cognitive feature extraction, we streamline the IDM process while enhancing its theoretical foundation, offering a promising approach for more efficient and scalable assessment development.

Differential embedding dimension functioning in natural language processing for psychological assessment

Thursday, 17th July - 14:00: Artificial Intelligence III (GH: Meridian 1-2) - Oral

Mr. Pengda Wang (Rice University), Ms. Ashley Sylvara (Kansas State University), Dr. Tianjun Sun (Rice University), Dr. Mikki Hebl (Rice University), Dr. Fred Oswald (Rice University)

Psychological assessment plays a crucial role in organizational research and broader social-behavioral sciences by providing data-driven insights. A key challenge in this domain has been ensuring measurement invariance across diverse groups, leading to the development of methodologies such as Multi-Group Confirmatory Factor Analysis (MGCFA) and Differential Item Functioning (DIF). These psychometric techniques, grounded in Factor Analysis and Item Response Theory, assess whether measurement tools function equivalently across populations. Their significance extends to fields such as personnel selection, education, and healthcare. With advancements in natural language processing (NLP) and machine learning (ML), embedding models have emerged as powerful tools in psychological assessment. These models convert textual data into numerical representations, capturing semantic relationships beyond traditional methods. However, NLP applications raise concerns regarding fairness and potential biases in machine-derived assessments. This study introduces Differential Embedding Dimension Functioning (DEDF), a novel approach to detecting bias in embedding models using psychometric techniques. We analyze data from 357 participants across undergraduate and working adult samples, who completed a Big Five personality assessment via AI chatbot interviews. By treating embedding dimensions as psychometric items, we employ measurement invariance analyses (MGCFA and DIF) to identify discrepancies across demographic groups. Additionally, qualitative analysis of text data offers insights into contextual influences on trait predictions. This work bridges traditional psychometrics and modern ML techniques, aiming to enhance the fairness and validity of NLP-based assessments in high-stakes settings.

Using a psychometric approach to design AI agents with personality under different scale formats

Thursday, 17th July - 14:15: Artificial Intelligence III (GH: Meridian 1-2) - Oral

Dr. Xijuan Zhang (York University), Ms. Muhua Huang (University of Chicago), Dr. Jessie Sun (Washington University in St. Louis), Dr. Victoria Savalei (University of British Columbia)

The Big Five Inventory-2 (BFI-2) by Soto and John (2017) is one of the most popular psychological scales for measuring personality. In this talk, I will present my two recent research papers on BFI-2.

In the first research paper, we converted the original BFI-2 in the Likert format into three alternative formats to address response bias and methods in the original BFI-2. Our findings revealed that while the Likert and alternative formats exhibit similar validity, the alternative formats—particularly the Expanded format—showed better psychometric properties, including enhanced factor structure, increased reliability, and reduced careless responding.

In the second research paper, we examined using the BFI-2 in the Likert and Expanded format to design Large Language Models-Based Agents (a.k.a., AI-agents) with different personalities. We found that compared to the Likert format, using the BFI-2 in the Expanded format makes it easier to assign different AI-agents with personalities. We validated the AI-agents by showing strong correspondence between human and AI-Agent answers to other personality tests and decision-making scenarios. This suggests that researchers could potentially use AI-agents as study participants. However, there are significant limitations to using AI-agents in research, which I will discuss at the end of the talk.

Enhancing systematic review efficiency: A generative AI-powered data extraction pipeline

Thursday, 17th July - 14:30: Artificial Intelligence III (GH: Meridian 1-2) - Oral

Ms. Xiyu Wang (Purdue University), Dr. Hua-Hua Chang (Purdue University), Dr. Yukiko Maeda (Purdue University)

Systematic reviews are critical for evidence synthesis, yet manual data extraction remains time-consuming and error-prone, limiting efficiency. Existing automated tools, such as Covidence, DistillerSR, and EPPI-Reviewer, assist with study screening but still rely heavily on manual input for data extraction (Ofori-Boateng et al., 2024). Moreover, keyword-based automation techniques often struggle to accurately capture context-dependent information (Ge et al., 2024). Advances in natural language processing (NLP) and generative AI present an opportunity to enhance data extraction with improved accuracy and efficiency.

This study proposes an R-based pipeline, designed as a lightweight app or webpage, integrating ChatGPT via API to automate structured data extraction in systematic reviews. The pipeline processes three key documents: (1) a coding sheet specifying the structured data points to ensure standardization across studies, (2) a full-text corpus of eligible articles serving as the primary data source, and (3) a background document providing domain-specific context, including definitions, terminology, and examples, to enhance AI interpretability. Structured, task-specific prompts guide the AI in retrieving targeted data while minimizing interpretative bias. To address token-length constraints, the pipeline employs a document chunking strategy, segmenting lengthy articles into semantically coherent sections to preserve context across processing iterations. Extracted data will be validated against human-coded results for accuracy and consistency.

By automating structured data extraction, this pipeline reduces manual effort, enhances standardization, and improves the reliability of evidence synthesis. Its integration into systematic review workflows supports scalability, allowing researchers to process large volumes of literature more efficiently while maintaining rigor in data extraction.

Exploring link prediction in social networks

Thursday, 17th July - 13:30: Network Models II (GH: Meridian 3-4) - Oral

Ms. Apoorva Verma (University of Maryland), Dr. Tracy Sweet (University of Maryland), Ms. Daria Smyslova (North Carolina State University), Dr. Daniela Castellanos-Reyes (North Carolina State University)

To model interactions or relationships among individuals, social network models can directly (e.g., Exponential Random Graph Models; Robins et al., 2007; Wasserman & Pattison, 1996) or indirectly (e.g., latent space models; Hoff, Raftery, & Handcock, 2002) accommodate the underlying dependencies among individuals in a group setting. In addition to modeling network ties, social network models can be used to predict tie values, also called link prediction. Link prediction can be utilized for imputing missing data (Hoff, 2008), evaluating model fit (Dabbs et. al, 2020), and for predicting ties among nodes in other networks or ties with a node new to the existing network. Thus, we explore the potential of using machine learning algorithms to predict binary network ties from node and edge covariates. We investigate the performance of standard machine learning algorithms in predicting math advice-seeking among school staff members in a U.S. district. We then compare these results with those obtained from latent space models to determine whether machine learning algorithms can produce more accurate link prediction in social networks than model-based methods.

Psychological networks as scale-free and small-world networks: Insights from large-scale survey data

Thursday, 17th July - 13:45: Network Models II (GH: Meridian 3-4) - Oral

Dr. Guangyu Zhu (The Australian National University)

Network approaches have gained increasing prominence in psychology; however, the structural attributes of psychological networks remain underexplored. As parts of biopsychosocial systems, psychological networks may share properties observed in neural, social, and semantic networks—such as scale-free topology, characterized by power-law degree distributions. This study uses data from seven large-scale psychology-related surveys to investigate whether psychological networks exhibit scale-free properties and other systemic attributes (e.g., small-world organization, modularity).

Results found that, while psychological networks do not strictly adhere to power-law distributions, they display a long-tail degree distribution, suggesting a hierarchical structure where most variables are sparsely connected, while a minority serve as highly connected hubs. Moreover, the networks exhibit robust small-world properties, including moderate to high global clustering coefficients (C = 0.25–0.51), short average path lengths (L = 2.23–8.09), and high small-world indices (O = 1.38–3.00). Additionally, the networks showed positive degree correlations (r = 0.06–0.34) and high modularity (Q = 0.33–0.92).

These findings reveal shared structural patterns in psychological networks, advancing our understanding of psychological systems as complex adaptive systems. They offer significant implications for both theoretical development and practical interventions, such as identifying key hubs within psychological networks and designing targeted interventions informed by network topology.

Estimating causal effects on psychological networks using item response theory

Thursday, 17th July - 14:00: Network Models II (GH: Meridian 3-4) - Oral

Mr. Joshua Gilbert (Harvard University), Dr. Benjamin Domingue (Stanford University), Dr. James Kim (Harvard University)

Network models in which each variable interacts with the others in a complex system have emerged as an important alternative to latent variable models in psychometric research. However, confirmatory methods for group network comparison are limited by practical constraints, such as the computational intractability of the Ising model in large networks. In this study, we demonstrate how to estimate causal effects on network state and strength when direct network estimation is not feasible by leveraging the mathematical equivalencies between the Ising model and item response theory (IRT) models. We demonstrate through simulation that a two-parameter logistic (2PL) explanatory IRT model can simultaneously recover causal effects on network state and strength. We first apply the method to a single empirical example of a vocabulary assessment from a content literacy intervention to demonstrate model building and interpretation strategies. We then replicate our approach with 72 empirical datasets from randomized controlled trials with item-level outcome data in education, economics, health, and related fields. Our results show that causal effects on network strength are both common and uncorrelated with effects on network state, suggesting that causal network models can provide new insight into the impact of interventions in the social and behavioral sciences.

Item pool maintenance in computer adaptive tests: A network approach

Thursday, 17th July - 14:15: Network Models II (GH: Meridian 3-4) - Oral

Dr. Klint Kanopka (New York University), Ms. Sophia Deng (New York University), Ms. Yining Lu (New York University)

Computer adaptive testing has delivered enormous advances in measurement efficiency and test security, though new issues persist. Maintaining a sufficiently large item pool requires constantly managing item exposure patterns, rotating out overexposed items, and calibrating new items. Additionally, parameter drift presents an additional maintenance task of pool recalibration, further complicated by potential bias in estimated parameters due to item delivery that prioritizes testing efficiency.

We propose representing the state of the item pool as a network. Here, individual items are nodes with edges connecting items with responses from the same individual, weighted by the inverse number of co-respondents. This builds a natural sense of distance, giving the shortest weighted path between two items the interpretation of an indicator of the magnitude of potential relative calibration error. Additionally, it provides a natural metric to optimize when selecting items to deliver for pool quality maintenance: network diameter, or the longest-shortest path between items. New item selection algorithms can balance pool maintenance and maximum information on the fly. Finally, network structure allows for the adaptation of message passing from graph neural networks to engage in local online item recalibration, combating parameter drift.

This presentation builds intuition for the structure, as well as the algorithms for item selection and information propagation. Through simulated data, we present results on the efficacy of our network-based method in reaching the stated goals of online calibration of new items, online recalibration of existing items, and more balanced exposure patterns alongside tradeoffs to necessary test length to maintain precision.

Using social network analysis to detect and interpret network collusion

Thursday, 17th July - 14:30: Network Models II (GH: Meridian 3-4) - Oral

Dr. Richard Feinberg (National Board of Medical Examiners)

Testing organizations routinely investigate if secure content has been compromised, giving future test-takers potential advanced access to exam content, known as item preknowledge. Traditional statistical methods for detecting preknowledge, specifically response similarity indices (Holland, 1996; Wollack, 1997; van der Linden & Sotaridona, 2006), focus on identifying unusual similarities between pairs of examinees, such as when one student copies off another. However, these indices are no longer aligned with modern collusion behavior, as prospective examinees can easily engage online and in unpredictable ways with emerging technologies. Recent research has begun to address this concern with methodologies that leverage response similarity probabilities among all pairs as a distance matrix in clustering algorithms to uncover group-level patterns of collusion (Wollack & Maynes, 2017; Below & Wollack, 2021; Eckerly, 2021).

In the present study, we extend this research by applying social network analysis to demonstrate a method to not only detect the presence of collusion communities but also relate explanatory factors to facilitate interpretation and, if warranted, action. Operational data from a high-stakes examination containing 2,130 examinees and 100 items will be manipulated to create several conditions of collusion, including the number of examinees and groups with preknowledge, the number of compromised items, and the intensity of the compromise. Discussion will include evaluation of results, approaches for network visualization, policy implications, and additional analysis steps to further explore detected collusion communities.

Assumptions in latent moderation: The role of measurement (non)-invariance

Thursday, 17th July - 13:30: Moderation and Mediation Analysis (MAC: Johnson) - Oral

Dr. Kaylee Litson (University of Houston)

Statistical interaction effects are common in psychology research because they allow researchers to examine the effect of a predictor to an outcome, conditional on the values of moderating variables. Although moderation is often evaluated among observed variables (Cortina et al., 2021), approaches for evaluating both the measurement and structural components of latent moderation are developing. When interaction effects are directly evaluated among latent variables using modern approaches, like product indicator or latent moderated structural equations, the underlying measurement properties of indicators are assumed equal across different levels of the moderator. This so-called measurement invariance assumption is untested and may be incorrect, since an estimated interaction effect may be due to differences in the measurement structure of the latent variables across levels of the moderator, and/or due to true differences in the structural paths among latent variables across levels of the moderator. Although measurement invariance is sometimes evaluated in the presence of categorical latent interaction effects (e.g., Litson et al., 2017), assumptions regarding measurement invariance have yet to be applied to continuous latent moderation, despite developments in approaches to evaluate measurement invariance across continuous variables (Hirschfeld et al., 2014; Molenaar, 2021). The goal of this project is to extend the framing around continuous latent moderation such that measurement invariance is a *testable* assumption in the evaluation of both continuous and categorical latent interaction effects. This paper will present a method with assumptions to be tested and apply the method to a dataset measuring personality and networking on career goals and outcomes.

MNLFA with three or more latent dimensions

Thursday, 17th July - 13:45: Moderation and Mediation Analysis (MAC: Johnson) - Oral

Dr. Noah Padgett (Harvard University)

Moderated nonlinear factor analysis (MNLFA) is readily applied to one or two latent factors. However, extending beyond two factors has been difficult due to the complexity of ensuring positive definiteness of the factor covariance matrix for every combination of person factors in the design matrix. I will present an approach to parameterizing the factor covariance matrix that decomposes the factor variances and factor correlations to allow for moderation in the variances, correlations, or both. Simulated data will be used to illustrate how the parameterization leads to differences across subgroups in the observed correlations among items. Implications for testing measurement non-invariance and impact will be discussed.

Bayesian nonparametric nonlinear moderation model

Thursday, 17th July - 14:00: Moderation and Mediation Analysis (MAC: Johnson) - Oral

Dr. Siyi Wang (Sun Yat-sen University), Dr. Qijin Chen (Sun Yat-sen University), Prof. Junhao Pan (Sun Yat-sen University)

Moderating effects explore how a third variable influences the relationship between two variables, offering deeper insights into psychological processes. Recent research has increasingly emphasized nonlinear moderating effects, such as quadratic terms, reflecting the complexity of psychological relationships. For example, parental expectations moderate the link between parental educational involvement and children's academic performance, but this effect is likely nonlinear—higher expectations do not consistently enhance the positive impact of parental educational involvement on children's academic performance. However, studying complex and dynamic moderators remains challenging with traditional parametric methods, which rely on rigid assumptions about the moderator's predefined form. In this research, the Bayesian nonparametric nonlinear moderation model, drawing on the idea of a varying-coefficient model, provides a more flexible approach. It allows for the modeling of nonlinear moderating effects that can change across levels of the moderating variable, thus accommodating more complex relationships without the need for a rigid parametric assumption. A simulation study was conducted to evaluate how factors such as sample size and the nature of the moderators influence the model's performance. Additionally, an empirical example about parenting behavior was presented to demonstrate how this approach can be applied in real-world research. By using a varying-coefficient model within the Bayesian framework, this method enhances the understanding of nonlinear interactions and provides a robust tool for analyzing complex moderation effects in psychological research, offering richer insights into the underlying mechanisms of behavior.

Cost-efficient sampling strategies for experiments detecting moderation and main effects

Thursday, 17th July - 14:15: Moderation and Mediation Analysis (MAC: Johnson) - Oral

Dr. Zuchao Shen (University of Georgia), Dr. Benjamin Kelcey (University of Cincinnati), Prof. Zhenqiu Lu (University of Georgia), Ms. Eunji Lee (University of Georgia)

Moderation analyses can answer questions about where, for whom, and under what conditions intervention effects are most salient. Together with the main effect analysis that can answer questions about whether interventions work or not, the results from moderation and main effect analyses provide the evidence base for effective policy development (Raudenbush & Liu, 2000; Spybrook et al., 2016). The literature has developed statistical power formulas for main and moderation effects (Dong et al., 2018; Dong, Kelcey, Spybrook, 2021; Dong, Spybrook, et al., 2021; Raudenbush, 1997; Raudenbush & Liu, 2000). However, the literature has not offered a framework that simultaneously considers both effects (moderation and main), cost, efficiency, and statistical power when designing experimental studies.

The purpose of the present study is to develop a jointly optimal design framework for cluster-randomized trials (CRTs) investigating both moderation and main effects. Such a framework enables researchers to identify the optimal sampling allocations by considering the cost of sampling and statistical power for both effects. This study is an important extension of the past work on optimal design (Shen & Kelcey, 2020, 2022a, 2022b). The present study has developed such a framework for CRTs investigating moderation and main effects, using the same ant colony optimization algorithm that has been shown to work as intended for the optimal design of CRTs investigating mediation effects (Shen et al., 2025). The proposed methods, including optimal sample allocation calculation and power analysis for both effects, will be implemented in the R package *odr* before the presentation.

Recent developments in dynamic fit index cutoffs for latent variable models

Thursday, 17th July - 13:30: Model-Data Fit (MAC: Thomas Swain) - Oral

Dr. Daniel McNeish (Arizona State University)

The prevailing approach to scale validation involves comparing factor analytic model fit indices to benchmarks suggested by Hu & Bentler (1999) like RMSEA < .06 or CFI > .95, as evidenced by the paper receiving over 135,000 citations. However, these benchmarks were derived from simulation and several studies have shown that benchmarks with desirable statistical properties will change depending on characteristics of the data or model being evaluated. Broadly using static cutoffs like RMSEA < .06 or CFI > .95 could potentially reward or punish certain models arbitrarily based on their characteristics rather than how well they actually fit. For nearly 20 years (e.g., Millsap, 2007), methodologists have suggested re-simulating cutoffs for each model to avoid possible issues with overgeneralizing static cutoffs. These suggestions mostly went unheeded until advances in computational power and open software facilitated computational approaches to re-simulate cutoffs. McNeish & Wolf (2023) developed dynamic fit index cutoffs as a way to replicate Hu & Bentler's simulation while adaptively replacing the simulation conditions with characteristics from the researcher's specific model. However, Hu & Bentler's simulation does not extend beyond CFA, maximum likelihood, and continuous responses whereas many scales use ordinal/Likert responses, consider a variety of factor structures, or alternative estimators. This talk discusses recent developments to extend dynamic fit index cutoffs beyond models and response scales considered in Hu & Bentler's original simulation so that benchmarks with desirable statistical properties can be derived for the broader types of models researchers consider and data structures that they often possess.

Cut-off for the deleted-one-covariance-residual case influence measure in covariance structure analysis

Thursday, 17th July - 13:45: Model-Data Fit (MAC: Thomas Swain) - Oral

Dr. Fathima Jaffari (Qiyas,ETEC,Saudi Arabia), Dr. Jennifer Koran (Quantitative Methods Program AT Southern Illinois University Carbondale)

The existence of unusual cases affects the modeling results yielded from fitting target models to the data. In Structural Equation Modeling (SEM), the influence of any case on the modeling results is calculated by case influence measures. The problem with these measures is that they are model based measures, and their performance is subject to specification error. A new model-free case influence measure, Deleted-One-Covariance-Residual (DOCR), was introduced to the SEM field by (Jaffari and Koran, 2023) as a model-free case influence measure that overcomes the problem of misspecification. this study aims to evaluate the adequacy of the cut-off set for the DOCR and compare its adequacy to those points resulted from narrowing the area of the dramatic increase in the influence of the cases, and the performance of DOCR in flagging the simulated target cases was compared by calculating the miss-rate of DOCR at those points when the sample size, proportion of target cases to non-target cases, and type of model used to generate the data are manipulated. The mean of the 100 replications for the miss rates and their 95% confidence intervals have been calculated using R package psych (Revelle, 2018). The results showed that with a cutoff value of (0.008), the *DOCR* recorded the lowest miss rate since about % 1.1 to % 28.3 of the simulated target cases were not flagged by *DOCR*, and these percentages were the same for other cut-offs set below (0.008), which supports setting the point (0.008) as a recommended cut-off for DOCR.

Robust methods for computing structural fit indices: A Monte Carlo investigation

Thursday, 17th July - 14:00: Model-Data Fit (MAC: Thomas Swain) - Oral

Dr. Graham Rifenbark (University of Wisconsin - Madison), Dr. Terrence Jorgensen (University of Amsterdam)

Researchers have proposed procedures to construct structural fit indices (SFIs) due to the shortcomings of global fit indices (GFIs) when evaluating the structural component of structural equation models (SEMs). We focus on test statistics and SFIs that stem from a two-step procedure given by Hancock and Mueller (2011): first, a confirmatory factor analysis is estimated and subsequently, a path analysis is fitted using the model-implied latent covariance matrix (Φ). This procedure results in a *pseudo-χ^2 test statistic* which can be used to construct *structural versions* of GFIs. This two-step SFI procedure ignores uncertainty about Φ, which elevates Type-I error rates, performing worse with less reliability indicators, more indicators per factor, and smaller samples (Rifenbark, 2019; Cao et al., 2023; Heene et al., 2024). As such, modeling and empirical solutions have been investigated to improve two-stage SFIs. Rifenbark and Jorgensen (2023) investigated the use of the *structural-after-measurement* (Rosseel & Loh, 2024) framework, while Zhang and Wu (2024) proposed post-hoc corrections that consider various computations of information and asymptotic covariance matrices of Φ. We use Monte Carlo simulation to compare the performance of these advancements and report results for varying levels of reliability, sample size, and levels of structural misspecification across various models; hypothesizing the new methods will result in nominal Type-I error rates (for χ^2) and 90% CI coverage (for indices based on it), compared to the uncorrected pseudo-χ^2 and its SFIs.

Can we rely on reliable parameter estimates?

Thursday, 17th July - 14:15: Model-Data Fit (MAC: Thomas Swain) - Oral

Dr. Niels Vanhasbroeck (University of Amsterdam), Dr. Kenny Yu (KU Leuven, University of Leuven)

Psychological researchers apply quantitative models to discover the structure of their construct of interest, typically relying on the fit of the model to the data to reach their conclusions. It has been argued, however, that fit of the model is not sufficient to shield one from model misspecification, that is from inaccurately representing the underlying psychological process of interest. Yet it is unclear to which extent misspecified models can be estimated reliably, and whether the reliability with which parameters can be estimated may signal such misspecification. In this work, we simulate and estimate a whole range of correctly specified and misspecified models, going from the inclusion/exclusion of interaction effects to the aggregation across heterogeneous populations to a linear/nonlinear structure of the model. For each of these models, we then compute the reliability of the parameter estimation for the two types of models and compared the results. We predict that misspecified models may still yield parameter estimates with acceptable reliability metrics when evaluated in isolation. Additionally, we predict that when compared directly to correctly specified models, these same misspecified models will exhibit distinctive reliability degradation patterns reflecting compensatory mechanisms. This dual perspective highlights that while misspecified models might appear adequate when judged solely on standard reliability metrics, comparative analysis against well-specified alternatives can reveal systematic differences that signal model-process misalignment.

A comparison of IRT model fit indices under different misfitting conditions

Thursday, 17th July - 14:30: Model-Data Fit (MAC: Thomas Swain) - Oral

Mr. Xinyu Liu (University of Minnesota), Prof. David Weiss (University of Minnesota)

In item response theory (IRT), several assumptions must be met to obtain appropriate and unbiased estimates of a person's ability level. These assumptions include: (a) a specific type of item response function accurately describes the data; (b) local independence holds; and (c) the dimensionality of the model is correctly specified, whether unidimensional or multidimensional. Violations of these assumptions, along with factors such as small sample sizes or multiple misfitting items, can result in model misfit. Misfitting models lead to biased parameter estimates, invalid person scores, or, in extreme cases, non-identifiable models.

Several fit indices have been developed to assess IRT model fit. The Pearson chi-squared statistic and the likelihood ratio G-squared statistic, are adapted from factor analysis and compare observed response patterns with expected response patterns. Alternative model fit indices adapted from factor analysis, such as RMSEA, SRMR, comparative fit index (CFI), and Tucker-Lewis index (TLI), have been proposed for evaluating IRT model fit. In addition to overall model fit indices, residual analysis serves as another approach for evaluating IRT model fit. Bayesian approaches have also been explored for assessing IRT model fit.

Although extensive research has examined various model fit indices, a comprehensive evaluation of their performance under different types of model misfit is lacking. The present simulation study aims to assess the effectiveness of various model fit indices across different types of model misfit and to determine whether specific fit indices can distinguish among different sources of misfit.

Adaptive tests and questionnaires using online survey tools

Thursday, 17th July - 13:30: Computer Adaptive Testing III (GH: Think 4) - Oral

Dr. Andries Van der Ark (University of Amsterdam), Ms. Lydia Dekker-Klein Nibbelink (University of Amsterdam), Mr. Goan Booij (University of Amsterdam)

Time constraints are a well-known problem in online surveys: There is often insufficient time to administer all the items necessary for optimal measurement. Measuring constructs using only a few items is undesirable, as these measurements generally have low reliability, which attenuates the estimates of the effects hypothesized by the researcher. Fixed-length computerized adaptive testing (CAT) has been proposed as a solution to mitigate this problem, as it allows for relatively precise measurement with relatively few items. However, many behavioral scientists use online survey tools such as Google Forms, LimeSurvey, Qualtrics, or SurveyMonkey for data collection, none of which support CAT. We propose a simplified version of fixed-length CAT that can be implemented in online survey tools, called Approximate Multi-Stage Testing (AMST). Depending on the amount of information obtained prior to data collection, a more or less sophisticated form of AMST can be employed. Using the linear administration of all items and fixed-length CAT as upper benchmarks, and the linear administration of a reduced test as a lower benchmark, we will investigate the bias and variance of measurements obtained using AMST with simulated data. We expect that AMST will yield the most precise measurements if all estimated item information functions are available prior to data collection. We also expect that if limited information is available (e.g., only item means reported in the literature), AMST will still be more precise than the linear administration of a reduced test. In the presentation, we will present the results and demonstrate AMST using Qualtrics.

Issues in calibrating post operational CAT data

Thursday, 17th July - 13:45: Computer Adaptive Testing III (GH: Think 4) - Oral

Dr. Eric Loken (University of Connecticut), Dr. Xiaowen Liu (Tianjin Normal University)

Adaptive testing, whether by multistage or item level designs, generates data with substantial missingness. The marginal properties of post-administration CAT data differ from those of standard linear testing data. Beyond the challenge of a large percentage of missing values, the observed item correlation structure will be different, and the items will have been answered by examinees with a much narrower range of abilities. The correlation between total score and item difficulty is also greatly reduced in CAT data. Nevertheless, the missingness in post-operational CAT data qualifies as "missing at random" (MAR) in the Rubin (1976) framework because it depends only on the observed responses driving the item selection. It is possible to recalibrate post-CAT response data and successfully estimate item and person generating parameters. However, as many researchers have pointed out (Jewsbury &van Rijn, 2020; Mislevy & Wu, 1996), some post-operational uses of such data violate the MAR assumptions. In particular, if the post-operational calibration omits items relevant to the adaptive item selection, as might happen if subscales calculated, then significant parameter bias may occur. We simulate analyses of post-CAT data that lead to significant parameter bias, and even a complete breakdown of the underlying factor. We also show that these problems never occur non-adaptive designs, even with high missingness. Our discussion speculates on other contexts in which analyses of data from adaptive testing could be biased.

Directionally-weighted loss functions for shaping multistage adaptive testing item modules

Thursday, 17th July - 14:00: Computer Adaptive Testing III (GH: Think 4) - Oral

Dr. Matthew Naveiras (Riverside Insights), Dr. Onur Demirkaya (Riverside Insights), Dr. Jongpil Kim (Riverside Insights)

In multistage testing, groups of items (called ‘modules’) are selected on an examinee-by-examinee basis, providing a test that targets each examinee’s ability individually rather than providing a one-size-fits-all fixed form for all examinees. In MST by Shaping (MST-S; Han & Guo, 2013), modules are constructed on-the-fly using a module-shaping algorithm which iteratively selects and replaces a module’s items to minimize the difference between the information that module provides (the module information function; MIF) and a predefined target information function (TIF). This ensures that examinees receive modules with different content but similar MIFs, promoting fair and consistent measurements.

Loss (i.e., error) between the MIF and TIF is often measured using mean squared error (MSE) or mean absolute error (MAE). However, these loss functions weigh positive and negative differences between the MIF and TIF equally, despite these differences being meaningfully different. Negative differences correspond to information deficits, whereas positive differences correspond to information surpluses. Because information surpluses are generally less problematic than information deficits, weighting information deficits more heavily in loss functions may improve estimation outcomes, as uninformative item modules can result in larger biases and standard errors of ability estimates.

In this study, directionally-weighted variations of MAE/MSE were developed by implementing a directionality weight parameter to increase the loss contributed by information deficits. A simulation study was conducted to compare these directionally-weighted loss functions to their unweighted counterparts (regarding module information, reliability, and error) under different conditions by varying the number of stages, the number of items, and the TIFs.

Baseline scores and testing mode influence responder thresholds: A comparison of coefficients of repeatability between PROMIS computer adaptive tests and short forms

Thursday, 17th July - 14:15: Computer Adaptive Testing III (GH: Think 4) - Oral

Dr. Minji Lee (Mayo Clinic), Dr. David Cella (Northwestern University), Ms. Veronica Grzegorczyk (Mayo Clinic), Dr. Kathryn Ruddy (Mayo Clinic), Dr. Andrea Cheville (Mayo Clinic)

Objectives

The study aimed to identify necessary score changes for reliable and likely change in PROMIS computer adaptive testing (CAT) and to compare these with selected short forms (SFs-4a or 8a).

Methods

In a large health system symptom management intervention project, 5,823 completed PROMIS CATs for anxiety (v1.0), depression (v1.0), pain interference (v1.1), and physical function (v2.0) at baseline and at 4 months. Reliable changes were determined using the Reliable Change (RCI, 90% CI) and likely change (LCI, 68% CI) were calculated using IRT methods. RCI and LCI for short forms (SFs) were estimated for comparison using established item parameters.

Results

Reliable (likely) change required CAT and SF-4a score changes of 7(4) for CAT Anxiety, 6(4) for Depression, 5(3) for Pain Interference, and 6(4) for Physical Function when baseline scores reflected mild to slightly severe symptoms. SF-8a's required smaller changes: 5(3) for Anxiety 8a, 4-5(3) for Depression 8a, 4(2) for Pain Interference 8a, and 4-5(3) in Physical Function 8b. For baseline scores within the normal range, CAT was more accurate in detecting change compared to SFs. Typically, CAT administration ceased at 4 items once the SE fell below 3, whereas SF-8a's maintained SEs around 2 at mid-score ranges.

Conclusions

The reliable/likely change thresholds from this study apply across populations as they depend on scale characteristics, not sample distribution. The choice between LCI and RCI should consider testing mode, baseline scores, and context, helping researchers and clinicians identify responders and decliners in a way that suits their use of the scores.

Computer adaptive testing for ecological momentary assessment: Considerations and evaluations

Thursday, 17th July - 14:30: Computer Adaptive Testing III (GH: Think 4) - Oral

Dr. Teague Henry (University of Virginia)

Ecological momentary assessment studies (EMA) are an increasingly common protocol for collecting intensive longitudinal data on a variety of psychological phenomena, from psychopathology to social behaviors. A key consideration of designing EMA studies is minimizing participant burden. Put simply, a participant will be unlikely to respond to lengthy questionnaires multiple times a day. However, psychological measurement relies on longer assessments to improve measurement quality. Striking a balance between minimizing burden and optimizing measurement is vital for the use of EMA protocols.In this project, we develop and evaluate several approaches for optimizing this balance.

First, we evaluate the use of computer adaptive testing (CAT) methods for reducing the number of items administered. CAT methods allow for high quality measurement with small numbers of items, but have only been evaluated in light of the unique properties of intensive longitudinal data in a small number of studies. Here, we evaluate two different types of CAT, a within-timepoint CAT where the adaptation does not use information from previous timepoints, and a between-timepoint CAT, where estimates of the construct from previous timepoints inform the adaptation in the current timepoint. Additionally, we develop and evaluate an adaptive test *selection* approach, which uses methods from reinforcement learning to determine which constructs to collect from which participants. The overarching goal of this project is to determine EMA protocols can be modified to use these methods to reduce participant burden while maintaining rigorous measurement.

Toward a psychologist's guide to computational modeling: An interdisciplinary scoping review of modeling roadmaps

Thursday, 17th July - 13:30: Computational Modeling of Psychological Systems (GH: Think 5) - Oral

Ms. Jill de Ron (University of Amsterdam), Prof. Denny Borsboom (University of Amsterdam), Dr. Donald Robinaugh (Harvard Medical School), Dr. Olga Perski (Tampere University)

Psychological systems are inherently complex, characterized by nonlinear behavior, emergence, adaptation, and feedback loops. To better capture this complexity, researchers are increasingly using mechanistic computational models (i.e., computer code designed to represent explanatory principles or rules of a target system). However, there is little guidance on how to most effectively develop and use computational models in psychological research. Other scientific disciplines, such as ecology and chemistry, have used computational modeling extensively for decades and may offer valuable insights from which psychologists can learn. To address this gap, we conducted a scoping review containing 53 computational modeling frameworks from diverse scientific disciplines. Our narrative synthesis focuses on four key questions: (i) the general purpose of computational modeling, (ii) proposed modeling steps, (iii) best practices for progressing through these steps, and (iv) the role of empirical data in the modeling process. Drawing on these insights, I will propose an initial how-to guide for psychologists aiming to integrate computational modeling into their research.

Mapping the dynamics of idiographic network models to the network theory of psychopathology using stability landscapes

Thursday, 17th July - 13:30: Computational Modeling of Psychological Systems (GH: Think 5) - Oral

Dr. Ria Hoekstra (University of Amsterdam), Ms. Jill de Ron (University of Amsterdam), Dr. Sacha Epskamp (National University of Singapore), Dr. Donald Robinaugh (Harvard Medical School), Prof. Denny Borsboom (University of Amsterdam)

Theories in psychology and the statistical models we use to study them are intimately connected, but how well do these models truly reflect the underlying theoretical dynamics?

The network theory of psychopathology conceptualizes mental disorders as emergent from causally interconnected symptoms, predicting that individuals with more strongly connected symptom networks are at higher risk of developing a disorder. While idiographic network models are frequently used to test this hypothesis, it remains unclear whether their dynamics truly align with the network theory.

This talk presents a systematic investigation of this alignment using stability landscapes. Focusing on the Ising model (with different encodings) and the graphical vector autoregressive (GVAR) model, we show how symptom severity and symptom variability map onto theoretical predictions. Our results demonstrate that only the dynamics of the 0,1 encoded Ising model aligns with the network theory. In contrast, the −1,1 encoded Ising model yields mixed results and the GVAR model does not align with the core theoretical claims.

The findings underscore the importance of careful theory-model alignment, and we view stability landscapes as a valuable tool for assessing such correspondences. By systematically interrogating the connections between theory and model, this work highlights the need for careful theoretical validation of statistical network models before drawing substantive psychological conclusions.

Unraveling symptom dynamics: A mechanistic approach to feedback loops and causal discovery

Thursday, 17th July - 13:30: Computational Modeling of Psychological Systems (GH: Think 5) - Oral

Ms. Kyuri Park (University of Amsterdam)

Understanding the progression and persistence of mental disorder symptoms requires a mechanistic approach beyond traditional statistical models. To address this, we develop a computational model that simulates symptom intensity dynamics, with a particular focus on feedback loops as key drivers of symptom amplification and persistence. Our model captures bistability between healthy and depressed states, revealing tipping points and hysteresis effects that govern transitions.

Using this model, we systematically simulate diverse symptom network configurations to explore the role of feedback loops. Our findings indicate that increasing feedback loops generally heightens symptom levels, but this effect plateaus due to overlapping loops reducing their marginal impact. Additionally, networks with evenly distributed connectivity sustain higher symptom levels, requiring multiple simultaneous disruptions to weaken the system's cohesion and reduce persistence.

To validate these findings, we apply causal discovery methods to real-world clinical data. The results show that frequently observed network edges in empirical symptom structures closely align with simulated configurations associated with severe symptoms, highlighting the critical role of specific connectivity patterns in symptom persistence.

Our work demonstrates the utility of computational modeling in simulating diverse symptom scenarios, providing a structured framework to examine how certain connectivity influences symptom dynamics. By integrating mechanistic simulations with causal discovery, we establish a strong link between theoretical models and empirical data. This approach opens new avenues for understanding mental health progression and developing targeted interventions aimed at disrupting pathological connectivity structures more effectively.

sdbuildR: Building system dynamics models in R

Thursday, 17th July - 13:30: Computational Modeling of Psychological Systems (GH: Think 5) - Oral

Ms. Kyra Evers (University of Amsterdam), Prof. Denny Borsboom (University of Amsterdam), Dr. Eiko Fried (Leiden University), Dr. Fred Hasselman (Radboud University), Dr. Lourens Waldorp (University of amsterdam)

Despite the excitement over formal models to improve psychological theories, creating one can be a daunting task. System Dynamics offers tools and principles to help translate theories into computational models. Broadly speaking, System Dynamics aims to understand and intervene on complex systems by taking a feedback-centered perspective, where the interactions, nonlinearities, and delays within a system are central to addressing its problems. These systems can be mathematically described with differential equations as stock-and-flow models. However, software to build and simulate stock-and-flow models has historically been licensed and proprietary, such as Vensim and Stella. Despite the rise of open-access software in the past decade, in practice, System Dynamics remains underutilized if software is not easily incorporated into the platforms most researchers are used to and skilled in. Moreover, domain-specific languages limit the integration of emerging methodologies such as artificial intelligence, restricting the potential of this powerful technique. This talk introduces sdbuildR, a package to build and simulate stock-and-flow models in R. System Dynamics modelling in R offers five main benefits: accessibility, compatibility, flexibility, scalability, and reproducibility. sdbuildR allows users to easily modify and compile models, lowering the barrier to entry to modelling. In addition, sdbuildR can translate stock-and-flow models created in Insight Maker, an open-access online System Dynamics tool. This talk will focus on how sdbuildR can be used practically to develop formal models of psychopathology. Utilizing R's flexibility and accessibility, this presentation will demonstrate how sdbuildR facilitates parameter exploration, parallel simulation, and visualization - key components of iterative model development.

Bayesian evaluation of latent variable models: A practical tutorial with the R package bleval

Thursday, 17th July - 13:30: Tools and Techniques for Quantative methods and Psychometrics (GH: Think 3) - Oral

Ms. Xiaohui Luo (Beijing Normal University), Mr. Jieyuan Dong (Beijing Normal University), Prof. Hongyun Liu (Beijing Normal University), Prof. Yang Liu (University of Maryland)

Model evaluation is crucial in psychological methodology, particularly in the context of Bayesian latent variable models. Bayesian evaluation methods require integrating out latent variables to compute likelihoods for information criteria, and integrating both latent variables and model parameters to compute marginal likelihoods for Bayes factors. These processes can be computationally demanding, as likelihoods in many latent variable models are intractable. Moreover, the role of latent variables in model evaluation has been insufficiently addressed in applied research, likely due to a lack of practical guidance. This study fills these gaps by (a) offering step-by-step instructions on approximating model likelihoods using an efficient numerical method (i.e., adaptive Gauss-Hermite quadrature), (b) developing a user-friendly R package, *bleval*, designed specifically for computing information criteria and marginal likelihoods in Bayesian latent variable models, and (c) demonstrating the application of this package in various empirical scenarios, including structural equation models (SEMs), item response theory (IRT) models, and multilevel models (MLMs). The empirical examples also investigate practical issues such as the sensitivity of model evaluation results to the number of quadrature nodes, the performance of the numerical quadrature method in high-dimensional latent variable models, and the handling of latent class variables in Bayesian mixture models.

ILSAmerge and ILSAstats: Two new R packages for international large-scale assessments

Thursday, 17th July - 13:45: Tools and Techniques for Quantative methods and Psychometrics (GH: Think 3) - Oral

Dr. Andrés Christiansen (IEA Hamburg)

International Large-Scale Assessments (ILSA) in education, such as TIMSS, PIRLS, ICILS, and ICCS, provide valuable insights into educational systems worldwide. However, analyzing ILSA data poses unique challenges, including how the data is structured and the use of plausible values and replicate weights for estimating the standard errors. We introduce two R packages to address these challenges: ILSAmerge and ILSAstats.

ILSAmerge simplifies the process of downloading and merging ILSA data across multiple countries into a single dataset. By automatizing the integration of country-specific files, this package eliminates the need for manual data handling or the requirement of multiple steps.

ILSAstats is designed to analyze ILSA data. The package calculates means, proportions, quantiles, correlations, and single-level and multi-level regression models while accounting for plausible values and replicate weights, adhering to each ILSA's methodological guidelines.

Together, ILSAmerge and ILSAstats provide a comprehensive toolkit for researchers and analysts working with ILSA data, streamlining data preparation and statistical analysis.

ShinyFORC: A Shiny app for Bayesian probablistic forecasting

Thursday, 17th July - 14:00: Tools and Techniques for Quantative methods and Psychometrics (GH: Think 3) - Oral

Ms. Kjorte Harra (University of Wisconsin - Madison), Prof. David Kaplan (University of Wisconsin - Madison)

ShinyFORC conducts Bayesian probabilistic forecasting (BPF) with longitudinal data. The app implements our BPF framework that utilizes methodologies explicitly designed to obtain optimally predictive measures of rate of change as the foundation for projecting future trends. ShinyFORC can be applied to a variety of longitudinal data; including country-level large-scale assessments, repeated measures for individuals, and more. ShinyFORC integrates latent growth estimation, Bayesian model averaging (BMA), and Bayesian stacking into one interactive and accessible Shiny software application. The app uses the R program *blavaan* which runs *Stan* in the background, to provide latent growth estimates of units of analysis. These estimates serve as inputs to the *BMS*, *rstanarm*, and *loo* programs, allowing users to conduct and compare results from BMA and Bayesian stacking on the predictors of growth under various parameter and model prior specifications available. The stacking procedures can use ELPD_{loo} model weights, along with options for pseudo-BMA and pseudo-BMA+ weighting. ShinyFORC also produces in-sample and pseudo-out-of-sample predictions, performance measures, and interactive visualizations. Finally, the app can conduct true out-of-sample forecasts and provide forecasting visualizations. We finally present a comprehensive walkthrough using PISA data to forecast future trends. ShinyFORC combines multiple advanced Bayesian methods into a single user-friendly application that is useful for technical and non-technical researchers working with longitudinal data.

Sequential rank aggregation: An optimal active estimation approach

Thursday, 17th July - 14:15: Tools and Techniques for Quantative methods and Psychometrics (GH: Think 3) - Oral

Prof. Xiaoou Li (University of Minnesota)

Ranking from partial comparisons is a fundamental problem in applications such as recommendation systems, tournament design, and crowdsourced ranking. Selecting which comparisons to make adaptively can significantly enhance ranking accuracy while reducing the number of queries needed. This talk presents an active sequential estimation approach to rank aggregation, leveraging greedy information-based experiment selection rules to optimize information gain at each step. We establish theoretical guarantees, proving that maximum likelihood estimation paired with these selection strategies achieves consistency, asymptotic normality, and optimal risk performance. Numerical studies on synthetic and real-world ranking tasks illustrate the efficiency of the proposed methods in recovering accurate global rankings.a

Uncovering cognitive strategies in tower of london using n-gram analysis

Thursday, 17th July - 15:00: Natural Language Processing (GH: Meridian 1-2) - Oral

Mr. Matheus Rodrigues (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University), Mr. Gabriel Lopes (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University), Ms. Amanda Cardoso (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University), Ms. Isadora Martins (Albert Einstein Israelite Hospital), Dr. Alexandre Serpa (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University)

Research in cognitive assessment has increasingly emphasized the importance of multimodal analysis to explore problem-solving strategies beyond evaluating only final outcomes. This study investigates problem-solving strategies in a computerized tower of London (TOL-BR) task, a widely utilized neuropsychological assessment of planning abilities that requires participants to rearrange colored spheres on pegs to match target configurations within a limited number of moves. A total of 647 participants, with mean age of 20 years and 71.9% female, completed TOL-BR. Five items varying in difficulty were selected for analysis based on Item Response Theory parameters and available moves: Item 08 (“easy”, 5 moves), Items 01 and 19 (“medium easy,” 3 and 9 moves, respectively), Item 11 (“medium,” 6 moves), and Item 09 (“hard,” 5 moves). A n-gram frequency analysis was conducted to identify common action sequences, providing evidence of optimal solution patterns for item resolutions. The alignment between individual responses and optimal strategies decreased as item difficulty increased, except for item 19. This alignment was approximately 70% for item 08 and two optimal solutions identified, 63% for item 01 and one optimal solution, 25% for item 19 and three optimal solutions, 40% for item 11 and one optimal solution, and 20% for item 09 and two optimal solutions. The decreased endorsement for item 19 may be explained by the number of moves available for completing the item. These findings demonstrate the utility of multimodal analytical approaches in cognitive assessment, offering detailed insights into problem-solving strategies and emphasizing the importance of examining planning efficiency.

Predicting reading passage grades: text features vs. contextual embeddings

Thursday, 17th July - 15:15: Natural Language Processing (GH: Meridian 1-2) - Oral

Dr. Ann Hu (HMH/NWEA), Prof. Hong Jiao (University of Maryland)

Providing grade-appropriate reading passages is critical for test validity. Typically, passage grades are determined by subject matter experts (SMEs). This study explores using text analyses to predict passage grades. We investigate the efficacy of feature-based machine learning methods and BERT (Bidirectional Encoder Representations from Transformers). Text features such as average sentence length were analyzed using spaCy (Honnibal & Montani, 2017). Fourteen features were incorporated into nine machine learning models: linear, ridge, lasso, elasticNet, decision tree, random forest, gradient boosting regressor, support vector regression (SVR), and stacking.

Feature-based models often provide more interpretable results but may not capture the full context of passages, leading to less accurate predictions. Given BERT's ability to understand context and semantics, it is expected to result in higher prediction accuracy than feature-based methods. The performance of each method is evaluated using R-squared and root mean squared error (RMSE) metrics. For feature-based models, about 80% of the variance in text grades could be explained. The aggregated top five features contributing to prediction accuracy included average sentence length, syllable count, clause count, complex word percentage, and word count. BERT will also be utilized to analyze the passages, and its prediction accuracy will be compared with the feature-based models.

The findings will provide insights into the strengths and weaknesses of traditional feature-based approaches versus advanced contextual embeddings in predicting passage grades. This comparison aims to inform researchers about effective techniques for passage grade assignments, contributing to the development of more accurate and reliable assessment tools.

Documents are people and words are items: A psychometric approach to textual data with contextual embeddings

Thursday, 17th July - 15:30: Natural Language Processing (GH: Meridian 1-2) - Oral

Prof. Jinsong Chen (The University of Hong Kong)

The overarching theme of this paper is the application of psychometric modeling to textual data using contextual embeddings derived from large language models (LLMs). The study introduces a novel method where documents are treated as individuals, words as items, and contextual scores as item responses. These contextual scores, computed through dot products between word-level conditional contextual embeddings and document embeddings, serve as input for factor analysis.

The methodology involves two main stages: (1) obtaining contextual scores using transformer-based models such as BERT and (2) applying psychometric analysis, including exploratory and bifactor factor analysis. The approach was tested on the Wiki STEM corpus, a dataset rich in science, technology, engineering, and mathematics content.

Results indicate that contextual scores exhibit semantic relatedness, approximate normality, linear relationships, and stability across documents, making them suitable for psychometric analysis. Exploratory factor analysis extracted a large number of first-order factors, with second-order factor analysis reducing these to a more interpretable structure. Bifactor analysis was found to be preferable, as it effectively separated general and minor factors, with general factors aligning well with distinct STEM-related knowledge domains.

The study concludes that this method can enhance psychometric modeling in textual data analysis, enabling applications such as measuring knowledge representation in documents, analyzing latent structures in textual corpora, and extending classical psychometric methods to textual environments. However, challenges remain, including high correlations among word pairs and computational complexity, which future research should address to improve model reliability and confirmatory applications.

A joint factor-topic model for multimodal survey data analysis

Thursday, 17th July - 15:45: Natural Language Processing (GH: Meridian 1-2) - Oral

Mr. Yuxiao Zhang (Purdue University), Dr. David Arthur (University of Washington Tacoma), Dr. Yukiko Maeda (Purdue University), Dr. Hua-Hua Chang (Purdue University)

Multimodal data integration is a critical frontier in social science research (Mu et al., 2020). However, existing methods that combine numerical and textual data typically assume numeric covariates are observed and error-free (e.g., structural topic model). This assumption limits their applicability in psychological and educational research, where latent constructs measured by rating-scales are common (Brown, 2015). To address this gap, we propose a novel joint factor-topic model that integrates confirmatory factor analysis with topic modeling under a Bayesian hierarchical framework.

In this model, latent constructs (factors) derived from rating-scale data are linked to themes (topics) identified from open-ended responses. The topic proportions become a function of the factor scores, reflecting how latent traits may correlate with the themes individuals express in their open-ended responses. This joint model extends the traditional structural equation model to incorporate textual data while simultaneously expanding the topic model to include factor-based covariates, offering a more psychometrically robust and methodologically versatile tool for multimodal survey data analysis.

We estimate the model using Hamiltonian Monte Carlo in Stan and evaluate it via a systematic simulation study, examining parameter recovery and topic identification under varying sample sizes, text lengths, and factor-topic association strengths. Initial results show that the integration improves both topic discovery and the precision of trait estimates, especially when factor-topic links are moderate to strong. Finally, the model will be applied to real data to show that incorporating text can reveals deeper insights into latent traits, exemplified by linking scale-measured academic engagement with self-expressed learning experiences.

Damped linear oscillators in emotion dynamics: Influence of fundamental parameters on dynamic EGA structures

Thursday, 17th July - 15:00: Bridging Exploratory Graph Analysis and Complexity Science: Advancing the Understanding of Psychological Structures and Dynamics (GH: Meridian 3-4) - Oral

Dr. Aleksandar Tomašević (University of Novi Sad)

Emotions are inherently dynamic phenomena, fluctuating over time in patterns that vary across individuals. Within the framework of dynamical systems, the Damped Linear Oscillator (DLO) model has emerged as a theoretical approach to represent affective dynamics. The DLO characterizes how emotional states oscillate around an equilibrium with parameters capturing core features of emotion regulation (interpreted in terms of emotional lability, resilience, and vulnerability). While this model provides a foundation for modeling the temporal evolution of affect, questions remain about how its parameters influence the outcomes of psychometric analyses of emotional data.

In this talk, we bridge dynamical systems theory and network psychometrics by examining how variations in DLO parameters impact the structure of dynamic Exploratory Graph Analysis (dynEGA) models of emotion. DynEGA is a network psychometric method that identifies latent dimensions or communities among multivariate time-series data, making it well-suited to capture the evolving structure of emotions. We systematically vary key DLO parameters – such as the damping coefficient and equilibrium stability – to simulate different affective regimes. For each simulated regime, dynEGA is applied to estimate the network of interrelated emotions (revealing how emotion variables cluster into dimensions over time). Preliminary analyses indicate that different parameter settings yield distinct network structures, highlighting that variations in underlying affective dynamics can lead to differences in the estimated dimensional structure of emotion.

By exploring the theoretical principles of the DLO and leveraging dynEGA's ability to uncover dynamic latent structures, this work offers insight into the interplay between emotional dynamics and psychometric outcomes.

Taxonomic graph analysis

Thursday, 17th July - 15:00: Bridging Exploratory Graph Analysis and Complexity Science: Advancing the Understanding of Psychological Structures and Dynamics (GH: Meridian 3-4) - Oral

Dr. Alexander Christensen (Vanderbilt University), Dr. Andrew Samo (Bowling Green State University), Dr. Luis Garrido (Pontificia Universidad Catolica Madre y Maestra), Dr. Francisco Abad (Universidad Autónoma de Madrid), Dr. Hudson Golino (University of Virginia), Dr. Sam McAbee (Bowling Green State University)

Taxonomic Graph Analysis (TGA) is a comprehensive network psychometrics framework aimed at identifying hierarchical structures in psychology from the bottom-up. This framework addresses key methodological challenges that have impeded accurate recovery of hierarchical structures in large item pools, including local independence violations, wording effects, and structural stability. This talk discusses how TGA can be used to assess complex constructs such as personality and psychopathology by building up from most basic elements. Simulation support for each step of TGA will be briefly discussed. As an empirical example, TGA was applied to an open-source 300-item IPIP-NEO dataset (N = 149,337). TGA revealed a three-level structure composed of 28 first-level dimensions (facets), 6 second-level dimensions (traits), and 3 third-level dimensions (meta-traits). The overarching theme of these findings was a hierarchical structure that integrated empirical and theoretical findings that have been scattered across the personality literature, renewing structural considerations of the IPIP-NEO and demonstrating TGA's value to investigate hierarchical psychological constructs.

The Ergodicity Information Index: Bridging dynamic exploratory graph analysis with complexity science

Thursday, 17th July - 15:00: Bridging Exploratory Graph Analysis and Complexity Science: Advancing the Understanding of Psychological Structures and Dynamics (GH: Meridian 3-4) - Oral

Dr. Hudson Golino (University of Virginia)

In the last half of the 20th century, psychology and neuroscience have experienced a renewed interest in intraindividual variation. To date, there are few quantitative methods to evaluate whether a population (between-person) structure is likely to hold for individual people, often referred to as ergodicity. We introduce a new network information theoretic metric, the ergodicity information index (EII), that quantifies the amount of information lost by representing all individuals with a between-person structure. A Monte Carlo simulation demonstrated that EII can effectively delineate between ergodic and nonergodic systems. A bootstrap test is derived to statistically determine whether the empirical data is likely generated from an ergodic process. When a process is identified as nonergodic, then it's possible that a mixture of groups exist. Finally, two empirical examples are presented using intensive longitudinal data from personality and neuroscience domains. Both datasets were found to be nonergodic, and meaningful groupings were identified in each dataset. Subsequent analysis showed that some of these groups are ergodic, meaning that the individuals can be represented with a single population structure without significant loss of information. Notably, in the neuroscience data, we could correctly identify two clusters of individuals (young vs. older adults) measured by a pattern separation task that were related to hippocampal connectivity to the default mode network.

Finite mixture models for an underlying zero-one inflated Beta distribution

Thursday, 17th July - 15:00: Latent Class and Mixture Models (MAC: Johnson) - Oral

Prof. Jang Schiltz (University of Luxembourg), Dr. Cédric Noel (University of Lorraine)

Current versions of finite mixture models with underlying Beta distributions have the problem that the data y have to be contained strictly between 0 and 1. We resolve that problem by using the zero-one inflated Beta distribution instead and present our R package trajeR which allows to calibrate the model.

We illustrate some of the possibilities of trajeR by means of an example with simulated data and finish by comparing the results of the new model with the classical model for an underlying Beta distribution on a real data set.

Measuring effort with a multi-level non-hierarchical Gaussian mixture model

Thursday, 17th July - 15:15: Latent Class and Mixture Models (MAC: Johnson) - Oral

Dr. Kirk Vanacore (University of Pennsylvania), Dr. Adam Sales (Worcester Polytechnic Institute), Dr. Ryan Baker (University of Pennsylvania)

Effort measurement is underexplored in psychometrics, yet it underlies many psychological constructs (e.g., knowledge, ability, conscientiousness). During learning activities, effort is often inferred from response times. For IRT models, low-effort responses (i.e., guesses) are often identified using fixed time thresholds (Wise et al., 2004; Wise & Ma, 2012). To avoid reliance on arbitrary thresholds, Gurung et al. (2021) model response times after students viewed hints in a digital learning platform to identify high-effort (longer response times) and low-effort (rapid hint skipping) using a Gaussian mixture model. However, this approach does not account for factors that may cause individual differences in response time that are not associated with engagement, like varying hint lengths or reading speed, and does not provide student-level parameters. We propose a Multi-level Non-hierarchical Gaussian Mixture Model (MNGMM) to account for the nested structure of students responding to multiple problems without imposing a hierarchical cluster structure. Our model adds random components to cluster probabilities and means. Specifically, the student-level random intercept for high response time probability captures individual effort across problems. We apply this model to data from ASSISTments, a digital learning platform where middle school students access on-demand hints while solving math problems. The model was fit using MCMC sampling in STAN. We evaluate the relationship between student-level effort parameters and learning gains using a validated learning measure (Star et el., 2015). This study demonstrates the utility of MNGMMs for clustering nested data in psychometric research, offering a flexible approach to modeling effort during learning activities.

A mixture multidimensional nominal response model to account for different faking strategies

Thursday, 17th July - 15:30: Latent Class and Mixture Models (MAC: Johnson) - Oral

Mr. Timo Seitz (University of Mannheim), Mr. Ömer Emre Can Alagöz (University of Mannheim), Prof. Thorsten Meiser (University of Mannheim), Prof. Esther Ulitzsch (University of Oslo)

High-stakes personality assessments are often compromised by faking, where test-takers distort their responses according to social desirability. Many previous models have accounted for faking by modeling an additional latent dimension that quantifies each test-taker's degree of faking. Such models assume a homogeneous response strategy among all test-takers, where substantive traits and faking jointly influence item responses. However, such a model will be misspecified if, for some test-takers or some items, responses are only a function of substantive traits or only a function of faking. To address this limitation, we propose a mixture modeling extension of the multidimensional nominal response model. This model allows accounting for qualitatively different response strategies across subgroups of test-takers or over the course of the questionnaire by specifying latent classes on the person or person-by-item level. To complement the model, external information such as covariates of class membership or item response times can be implemented. Using simulation studies, we found the mixture modeling extension to accurately recover model parameters and outperform models not accounting for different faking strategies. The model also proved successful in applications to empirical data from personnel selection.

Bayesian and frequentist model evaluation for growth mixture modeling

Thursday, 17th July - 15:45: Latent Class and Mixture Models (MAC: Johnson) - Oral

Ms. Xingyao Xiao (University of California, Berkeley)

This study investigates the application of frequentist and Bayesian information criteria in growth mixture models (GMMs), addressing key challenges and proposing solutions to improve model selection. Critical decisions in defining information criteria include choosing between marginal and conditional likelihoods, calculating the effective number of parameters, and defining sample size. Marginal likelihoods, which integrate out random effects, are appropriate for population-level inferences, while conditional likelihoods, which condition on random effects, are suitable only for within-cluster predictions. These choices must align with the researcher's inferential goals.

For frequentist approaches, we identify limitations in the default settings of the flexmix package in R, such as insufficient random initializations (nrep = 5), which often lead to convergence to local maxima. Even when convergence is achieved, the deviance of the k-class model may decrease relative to the k−1-class model, yet the solution may still represent a suboptimal local maximum. We demonstrate that frequentist criteria perform significantly better when local maxima are validated against the minimum deviance from a Bayesian solution.

On the Bayesian side, we examine the pitfalls of the Deviance Information Criterion, particularly the occurrence of negative penalty terms due to its reliance on plug-in deviance. We advocate for DIC_pV, an alternative estimator defined as half the posterior variance of the deviance (Gelman et al., 2013), which is invariant to reparameterization and guarantees non-negative penalties. Our simulation study shows that DIC_pV, along with other Bayesian criteria like WAIC and LOO-CV, consistently outperforms standard DIC in classification accuracy across various simulation conditions.

From the roots: Likelihood ratio DIF testing for IRTrees

Thursday, 17th July - 15:00: Differential Item Functioning and Measurement Invariance III (MAC: Thomas Swain) - Oral

Dr. Anne Thissen-Roe (Harver), Ms. Amy Li (Harver)

Assessments of personality and similar constructs often use items in a rating scale format, such as Likert items, with flexible or inferred scale anchors. When we measure a construct using this type of item, responses are affected by both the intended construct and also individual differences in scale use: response sets. In recent years, our field has developed intrinsically multidimensional polytomous IRT models for this purpose. These include IRTrees, which model multiple sequential processes with latent construct and response set traits (De Boeck & Partchev, 2012; Böckenholt, 2012). These models can better fit the data, and allow better modeling of the constructs' relationships, from basic behavioral research to the valid prediction of employment outcomes. DIF testing methods have not kept up with the new item models, limiting our options in evaluating the content these models fit. If a set of items is affected by response sets, but we use a DIF test based on a construct-only model, we are left unsure if our DIF findings are really DIF, or a symptom of model misfit. Jin, Wu & Chen (2018) evaluated non-IRT DIF tests (Logistic Regression and Odds Ratio) applied to binary pseudo-items that constitute an IRTree. Here, we propose a set of general principles for applying IRT likelihood ratio DIF testing to whole IRTrees, even those that are not binary-decomposable (TDM; Thissen-Roe & Thissen, 2013), and demonstrate a real application to an operational employee selection personality instrument using six-point dual-anchored response scales.

Testing and correcting for differential test functioning

Thursday, 17th July - 15:15: Differential Item Functioning and Measurement Invariance III (MAC: Thomas Swain) - Oral

Prof. Peter Halpin (University of North Carolina at Chapel Hill)

This paper reviews some procedures for robust scaling in IRT and applies them to the problem of estimating group differences on the latent trait (impact) in the presence of differential item functioning (DIF). The robust scaling procedures do not require pre-specification of anchor items and are shown to produce consistent estimates of impact so long as fewer than 50% of items exhibit DIF. Comparing the robust estimate to a naïve estimate that ignores DIF leads to a test of whether the naïve estimate is consistent for the true impact. If it is inferred that the naive estimate is inconsistent, it can, in some settings, be useful to produce test scores that "correct for" this inconsistency. Several methods for addressing this problem are compared using simulated data and an empirical illustration.

Detecting uniform differential item functioning with the permutation test

Thursday, 17th July - 15:30: Differential Item Functioning and Measurement Invariance III (MAC: Thomas Swain) - Oral

Mr. Walton Ferguson (University of Notre Dame), Dr. Ying Cheng (University of Notre Dame)

Detecting differential item functioning (DIF) remains a pertinent issue in psychometric testing. Current methods for assessing DIF include the Mantel–Haenszel (MH) test, logistic regression approaches (LR), the likelihood-ratio test, and others. In this study, we propose a general framework of using permutation tests (PT) to detect items with DIF. It can be coupled with any DIF test statistic and can accommodate situations when no explicit group membership exists (e.g., when each individual associates with a latent class with a certain probability). In this paper, we illustrate the application of this framework when it is coupled with the MH odds-ratio statistic to detect DIF and compare the results to the generic MH test. Data was generated for a 20-item assessment from a 2-parameter logistic model. Power and type-1 error rates were calculated for the three methods under 48 simulation conditions with varying sample sizes (N = 1000, 500, 250 100, 50, and 30), effect sizes (Δ = 1.0, 0.8, 0.6, and 0.4), and reference-to-focal group balancing ratios (50/50 and 80/20). The PT was equivalent to the MH test at large sample sizes (i.e., > 100) but showed reduced type-1 error rates and equivalent or larger power at small sample sizes (i.e., 30 and 50), especially when the ratio between focal and reference group is imbalanced. With these gains in performance, the average PT computation time was between 27 and 68 times larger than the average MH test computation time. This study demonstrates the advantages and limitations of the proposed approach.

Guidelines for the interpretation of NCDIF as an effect size measure

Thursday, 17th July - 15:45: Differential Item Functioning and Measurement Invariance III (MAC: Thomas Swain) - Oral

Dr. Víctor H Cervantes (University of Illinois Urbana-Champaign), Mr. TRUNG LE (University of Illinois Urbana-Champaign)

Several statistics have been proposed to detect Differential Item Functioning (DIF) and quantify its magnitude. However, not all DIF statistics quantify the same population parameters and, therefore, cannot be interpreted interchangeably. We focus on deriving principled guidelines for the interpretation of the non-compensatory differential item functioning (NCDIF) parameter as a DIF effect size measure. This parameter quantifies the expected (average) effect of the possible differences between item parameters on the (squared differences of the) item score for the focal population of examinees. We first investigate in which situations the Delta Mantel-Haenszel (Δ_{MH}) is comparable to NCDIF, so that the ETS cutoff points can serve meaningfully as a benchmark. Then, we examine the parameter's behavior under various conditions of uniform and non-uniform DIF, as well as the effect that the distribution of the focal group exerts on its magnitude. Lastly, using one of the estimators of the NCDIF parameter, we evaluate the accuracy of the derived classification rules for NCDIF and identify an approximate bias correction for this estimator. Overall, these results provide useful guidelines for interpreting the magnitude of NCDIF that are consistent with its specific nature and improve the alignment of DIF classifications with the magnitude of the NCDIF parameter.

Rethinking discrimination: A marginal effects approach to IRT

Thursday, 17th July - 15:00: Item Response Theory III (GH: Think 4) - Oral

Dr. Brooke Magnus (Boston College), Dr. Trenton Mize (Purdue University)

Logistic regression models the log-odds of an outcome using a logit link function, making regression coefficients the expected change in the log-odds of the outcome for a one-unit increase in the predictor. However, applied researchers are typically more interested in how predictors influence outcome probabilities. Originating in econometrics, marginal effects improve the interpretability of categorical/nonlinear models by translating changes in log-odds into probability-based metrics. Despite the conceptual parallels between logistic regression and item response theory (IRT), marginal effects have received little attention in IRT (Mize, 2024).

Like logistic regression, the two-parameter logistic (2PL) model uses a logistic function to model a binary outcome (the item). The discrimination parameter, analogous to a regression coefficient, quantifies how strongly an item differentiates individuals along the latent variable. Also similar to logistic regression, the effect of the predictor — in IRT, the latent variable —is not constant across its range; it is strongest at the item's location parameter and weaker at the extremes. Traditionally, IRT users rely on item information functions to quantify how well an item measures individual differences across the latent continuum. While information functions provide more detail than a single discrimination parameter, their scale is less intuitive than probability. Marginal effects offer a probability-based alternative that may be more accessible to applied researchers. Mize (2024) recently introduced Stata commands for computing marginal effects in IRT models, yet their use in psychometrics (and psychology writ large) remains rare. This presentation will explore the potential benefits of incorporating marginal effects into IRT applications.

Item response models for rating relational data

Thursday, 17th July - 15:15: Item Response Theory III (GH: Think 4) - Oral

Dr. Chih-Han Leng (National Taiwan University), Prof. Ulf Böckenholt (Northwestern University), Prof. Hsuan-Wei Lee (Lehigh University), Prof. Grace Yao (National Taiwan University)

This paper introduces item response models for rating relational data. Relational data are collected through ratings of senders and receivers within a directed network. The proposed models enable comparisons of senders and receivers on a one-dimensional latent scale while accounting for underlying homophilic relationships. Our approach effectively captures reciprocity and clustering phenomena in relational data. Model parameters are estimated using a Bayesian framework with Markov Chain Monte Carlo methods to approximate full conditional posterior distributions. Simulation studies confirm the accuracy of parameter recovery, even with small network sizes. Finally, we present an extensive empirical application to illustrate the utility of our models for both complete and incomplete networks.

A generalization of multidimensional item response theory parameters

Thursday, 17th July - 15:30: Item Response Theory III (GH: Think 4) - Oral

Dr. Daniel Morillo (Univerisdad Nacional de Educación a Distancia (UNED)), Dr. Mario Luzardo-Verde (Universidad de la República (UdelaR))

The Multidimensional 2-Parameter Logistic (M2PL, McKinley & Reckase, 1983) model is one of the most important and widely used models in Multidimensional Item Response Theory (MIRT). Reckase (1985) and Reckase & McKinley (1991) gave a precise definition of a Multidimensional Item Difficulty (MID) and Multidimensional Discrimination (MDISC) parameter, respectively, and derived their expressions for the M2PL model. However, we argue that their derivations (1) only consider the case of items with conditional monotonously increasing probability of a positive response, and (2) implicitly assume that the items are represented in an orthogonal basis.

We argue that these assumptions, which restrict the application of the model to the assessment of non-cognitive traits, can be relaxed. We thus extend the expressions to the general case of unchanging monotony items in any non-orthonormal space, which allows to properly represent disparate covariance structures. We then use these definitions in two applications: to represent M2PL items in non-orthonormal bases, and to compute the multidimensional parameters in various applications, showing how they provide better properties compared with their "covariance-agnostic" counterparts.

These results become increasingly relevant as the modern measurement theory in social and behavioral sciences experiences as a convergence between MIRT and the common factor model. Importantly, our derivations show that the space where the items are represented is necessarily distinct from the latent space. Nevertheless, we warn against generalizing our results to other MIRT models (i.e. beyond the M2PL), without carefully considering each particular case.

An informative index for evaluating equiprecision in IRT-based assessments

Thursday, 17th July - 15:45: Item Response Theory III (GH: Think 4) - Oral

Mr. Jesus Delgado (University of Minnesota), Dr. Nana Kim (University of Minnesota), Prof. David Weiss (University of Minnesota)

Equiprecision refers to a test's ability to measure respondents of varying latent trait levels with similar precision (papersWeiss & Sahin, 2024). While this concept is commonly applied in the context of computerized adaptive testing (CAT), understanding how both adaptive and nonadaptive IRT-based assessments afford precision across the latent trait continuum remains valuable. A novel informative index is introduced to quantify equiprecision, providing a flexible approach to evaluating measurement precision across different regions of the trait distribution. This index allows users to specify focal regions of interest while incorporating a penalty for imprecision, ensuring that scenarios of "low-quality equiprecision"—where measurement is uniformly poor across all levels—are appropriately accounted for. The proposed index provides a practical tool for test developers and psychometricians seeking to optimize test design, evaluate information distribution, and compare assessments in terms of their ability to provide fair and efficient measurement across diverse populations. Example scenarios using real data illustrate the utility of this index in evaluating test precision across multiple IRT models. Potential applications in test construction, evaluation of fixed-form and adaptive tests, and practical considerations for implementation in applied settings will be discussed.

Adapting Fisher information-based difficulty and discrimination IRT measures to handle multimodality

Thursday, 17th July - 16:00: Item Response Theory III (GH: Think 4) - Oral

Mr. Peter Johnson (City University of New York), Prof. Jay Verkuilen (City University of New York)

Johnson and Verkuilen (2024) proposed Fisher information-based measures for effective difficulty and effective discrimination in binary item response theory (IRT) models, termed FIDiff and FIDisc, respectively. These measures have the appealing and intuitive meanings of the difficulty and discrimination parameters held in the two-parameter logistic (2PL) model lost outside the 2PL model and are based on computations that can be entirely numerical, not requiring complicated symbolic algebra. These metrics show promise for model comparison and parameter interpretation in models that stray from the 2PL, such as when adding parameters like in the four-parameter logistic (4PL) family, when using different link functions like the complementary log-log, and other modifications that might be made to the base 2PL model. Their proposal was limited because it required unimodal item information functions (IIFs). The current research extends FIDiff and FIDisc outside the binary, unimodal IIF case to address binary models with multimodal information (e.g., the monopoly model), polytomous IRT models, and ideal point models. We approximate the true IIF using a penalized smoother that generates a unimodal approximating IIF that is closest to the true IIF. This allows computation of FIDiff and FIDisc along with guaranteeing that level sets of the approximating IIF are convex. We illustrate using empirical examples from the SAPA 16 item intelligence battery and the 32 item SAT12 data originally from the TESTFACT manual.

Item-level heterogeneity in value added models: Implications for reliability, cross-study comparability, and effect sizes

Thursday, 17th July - 15:00: Reliability (GH: Think 5) - Oral

Mr. Joshua Gilbert (Harvard University), Mr. Zachary Himmelsbach (Harvard University), Dr. Luke Miratrix (Harvard University), Dr. Andrew Ho (Harvard University), Dr. Benjamin Domingue (Stanford University)

Value added models (VAMs) attempt to estimate the causal effects of teachers and schools on student test scores. We use Generalizability Theory to show how estimated VA effects depend upon the selection of test items. Standard VAMs estimate causal effects on the items that are included on the test. Generalizability demands consideration of how estimates would differ had the test included alternative items. We introduce a model that estimates the magnitude of item-by-teacher/school variance accurately, revealing how standard VAMs overstate reliability and precision. Using 35 datasets with item-level outcome data, we show how standard VAMs overstate reliability by an average of .12 on the 0-1 reliability scale (SD = .12). We discuss how imprecision due to heterogeneous VA effects across items attenuates effect sizes, obfuscates comparisons across studies, and causes instability over time. This suggests that accurate estimation and interpretation of VAMs require item-level data, including qualitative data about how items represent the content domain.

Discretization error in psychological science

Thursday, 17th July - 15:15: Reliability (GH: Think 5) - Oral

Dr. Mathias Berggren (Uppsala University), Dr. Jessica Flake (University of British Columbia)

When a continuous variable is measured with a few discrete categories, it will tend to attenuate its correlation with other variables. Such attenuation can complicate conclusions about the relations between variables. It can, for example, lead to an incorrect ordering for the most important predictors of an outcome, and can even lead to spurious partial correlations between other variables, or spurious differences between groups. Here, we provide an overview and review of discretization errors. We also provide a mathematical notiation for understanding different types of errors, not usually considered together, as different types of discretization errors. This notation also helps separate discretization error from random and unbiased errors considered in classical test theory. Next, we use simulations to illustrate the potential consequences of discretization errors on effect sizes, and discuss how this may influence various results in psychology. Finally, we provide some recomendations for researchers in thinking about, handling, and preferably designing studies to avoid discretization error.

Reliability of unidimensional ordinal scores: Insights from two simulation studies

Thursday, 17th July - 15:30: Reliability (GH: Think 5) - Oral

Prof. Sebastien Beland (Université de Montréal), Prof. Eunseong Cho (Kwangwoon University), Prof. Carl Falk (McGill University, Montreal, Canada)

Most research on score reliability has focused on unidimensional items with continuous responses. However, researchers in education and psychology frequently use items with ordinal responses (e.g., Likert-type or dichotomous). Yet, little is known about how reliability coefficients behave with this type of response.

To gain further insight into this topic, we made recommendations based on two simulation studies. In the first study about dichotomous responses, our analyses showed that Dimitrov's (2003) strategy, based on the two-parameter Item Response Theory (IRT) model, McDonald's total omega (based on the bifactor model), and a coefficient derived from confirmatory factor analysis, was the most effective.

In a second study, we analyzed ordinal variables of the Likert type. Preliminary analyses indicate that a little-known coefficient presented by Kaiser and Caffrey (1965) and a coefficient based on IRT are the most promising. These results lead to two key recommendations favoring reliability indices based on a psychometric model rather than on covariance/correlation. First, Cronbach's alpha is never the best reliability coefficient in our simulations. Second, coefficients designed for continuous response items perform very well when applied to ordinal responses.

Best treatment from a set of options: An optimal sequential method for principled experimentation

Thursday, 17th July - 15:00: Advances in Experimental Design, Measurement, and Predictive Inference (GH: Think 3) - Oral

Dr. Ken Kelley (University of Notre Dame), Mr. William Stamey (University of Notre Dame), Dr. Bhargab Chattopadhyay (Indian Institute of Technology Jodhpur), Dr. Tathagata Bandyopadhyay (Indian Institute of Management Ahmedabad)

Often, individuals are assigned across multiple experiments simultaneously, a practice commonly done in digital experimentation. Because of the substantial demand on the participant pool, whether from a selected group with certain characteristics or in an online environment, we propose a sequential method for selecting the best treatment among multiple alternatives, where the procedures stops with statistically sound properties as early as possible for a fixed level of statistical power and specified false positive rate. The method leverages simultaneous comparison to minimize sampling of inferior treatments and focus sampling on more promising ones. Our method can be immediately implemented in an R package we provide, and is especially attractive in digital experimentation contexts of behavioral studies (from digital firms or researchers using digital environments). Identifying the best method among a set of alternatives as compared to some benchmark value, minimum effect size of interest, or the current value (e.g., business as usual) provides a powerful way of framing trials. We justify the method as methodologically rigorous and show how it leads to efficient digital experimentation, meaning that fewer participants (e.g., customers, members of a special population) are needed

The psychometrics of uncertainty elicitation for real world forecasting problems

Thursday, 17th July - 15:15: Advances in Experimental Design, Measurement, and Predictive Inference (GH: Think 3) - Oral

Ms. Sophie Ma Zhu (University of British Columbia), Dr. David Budescu (Fordham University), Mr. Nikolay Petrov (University of Cambridge), Dr. Ezra Karger (Federal Reserve Bank of Chicago), Dr. Mark Himmelstein (Georgia Institute of Technology)

For some types of problems, it can be appropriate to elicit not just a subject's best estimate, but also a probability distribution that represents their uncertainty across the possible outcomes. This applies to forecasting judgments, such as those in forecasting tournaments or the Survey of Professional Forecasters. Predictions about continuous outcomes, like the future U.S. unemployment rate, are typically divided into mutually exclusive and exhaustive bins for which a simplex vector of probabilities is elicited. An alternative approach inverts the elicitation, requiring forecasts to be reported as a series of quantiles in the variable's original units at several fixed probabilities. Using data from 1,147 participants and 47 experienced forecasters, we compared the psychometric properties of each format. We elicited forecasts about a set of 36 problems as quantiles for five fixed probability values (.05, .25, .50, .75, .95) and probabilities across five pre-determined item-specific bins. After scoring forecasts using proper scoring rules, we found a bifactor model with one general and two method factors showed better fit than a two-factor model, suggesting a shared latent trait drives accuracy across formats. However, quantile forecasts showed stronger internal consistency, achieving a suitable reliability level with fewer items. A simulation study, where scores from quantile forecasts were more efficient at recovering simulated forecasters' latent skill parameters than scores from probability forecasts, provides theoretical justification for our results. We also consider issues related to scaling accuracy scores, as well as indicators of possible comprehension difficulties, such as violations of monotonicity.

Application of conformal prediction in language sample analysis

Thursday, 17th July - 15:30: Advances in Experimental Design, Measurement, and Predictive Inference (GH: Think 3) - Oral

Mrs. Youmin Hong (University of Maryland), Prof. Ji Seung Yang (University of Maryland), Prof. Nan Bernstein Ratner (University of Maryland), Prof. Yang Liu (University of Maryland)

Language Sample Analysis (LSA) is a method for evaluating language skills by collecting and analyzing speech or writing. It is particularly useful for young children who are not yet ready for structured writing tasks or formal testing. Researchers and clinicians use LSA to derive various linguistic indices that assess different aspects of language development, aiding in the evaluation of developmental appropriateness or the diagnosis of clinical conditions in children. While previous studies have examined the effectiveness of various language measures (e.g., Bernstein Ratner et al., 2024; Yang et al., 2022a; Yang et al., 2022b), they have often focused on a single construct of language development. In this study, we integrate multiple language measures and employ conformal prediction (CP, Vovk et al., 2005) to quantify uncertainty across different predictive methods, including traditional linear and non-linear regression-based approaches as well as machine learning algorithms. Since these predictive methods rely on different assumptions, CP is particularly suitable as a model-free inferential framework that constructs prediction intervals without relying on correctly specified models. We apply CP to both continuous (chronological ages) and binary (late talker status) outcomes, evaluating various prediction methods based on their practical utility and internal estimates. By demonstrating the effectiveness of CP in predictive validation, this study enhances methodological rigor in LSA research and underscores the importance of uncertainty quantification in language assessment.

Optimal design and analysis strategies for equivalence testing

Thursday, 17th July - 15:45: Advances in Experimental Design, Measurement, and Predictive Inference (GH: Think 3) - Oral

Dr. Zuchao Shen (University of Georgia)

Failing to reject the null hypothesis of a difference test (e.g., H_0: two group means are equal) does not allow us to infer that the two means are equal because it may be that the study is underpowered (Lakens, 2017). The equivalence test is a statistical technique that formally tests the equivalence of two-group means through null hypothesis testing (Lakens et al., 2018; Schuirmann, 1987). The null hypothesis is H_0: $|d_2\text{-}d_1| \geq \Delta$ with d_i as the mean of group i (i = 1, 2) and Δ as a trivial positive number to quantify the equivalence bound. Rejecting the null hypothesis through the two-sided t-tests can draw the inference that the two-group means are equivalent ($|d_2\text{-}d_1| < \Delta$). Literature has been limited in design strategies to detect statistical equivalence, such as failing to incorporate covariate adjustment and the lack of correct statistical power formulas.

The present study advances equivalence testing by (a) extending the method to designs with covariates to increase statistical power, (b) proposing a Monte Carlo confidence interval method testing equivalence, and (c) developing/validating the statistical power formula for equivalence tests. The methods can be extended to testing the equivalence of mediation effects. The proposed methods have been implemented in the R package *anomo* (version 1.0.0). The power analysis function can calculate statistical power, required sample size, and the minimum detectable difference between equivalence bounds and the group-mean difference. The package also includes a function to identify optimal sample allocation to achieve maximum statistical power under a fixed budget.

Don’t let your Likert scales grow up to be visual analog scales: Understanding the relationship between number of response categories and measurement error

Thursday, 17th July - 16:00: Advances in Experimental Design, Measurement, and Predictive Inference (GH: Think 3) - Oral

Ms. Siqi Sun (University of Virginia), Dr. Karen Schmidt (University of Virginia), Dr. Teague Henry (University of Virginia)

The use of Visual Analog Scales (VAS), which can be broadly conceptualized as items where the response scale is 0-100, has surged recently due to the convenience of digital assessments. However, there is no consensus as to whether the use of VAS scales is optimal in a measurement sense. Put differently, in the 90+ years since Likert introduced his eponymous scale, the field does not know how to determine the optimal number of response options for a given item. In the current work, we investigate the optimal number of response categories using a series of simulations. We find that when the measurement error of an item is not dependent on the number of response categories, there is no true optimum; rather, reliability increases with number of response options and then plateaus. However, under the more realistic assumption that the measurement error of an item increases with the number of response categories, we find a clear optimum that depends on the rate of that increase. If measurement error increases with the number of response categories, then conversion of any Likert scale item to VAS will result in a drastic decrease in reliability. Finally, if researchers do want to change the response scale of a validated measure, they must re-validate the new measure as the measurement error of the scale is likely to change.

Automated generation of creativity test items using large language models

Friday, 18th July - 09:00: Automated Item Generation (GH: Meridian 1-2) - Oral

Dr. Antonio Laverghetta Jr. (Pennsylvania State University), Mr. Simone Luchini (Pennsylvania State University), Ms. Averie Linnell (University of Nebraska-Omaha), Prof. Roni Reiter-Palmon (University of Nebraska-Omaha), Prof. Roger Beaty (Pennsylvania State University)

Creativity is considered a primary factor in individual and organizational success in modern economies (Tsegaye et al., 2019). Yet developing psychometrically valid creativity measures is also a complex task requiring many hours of trial and error to create suitable items. Automated item generation (AIG) using large language models (LLMs) has seen widespread success in creating high-quality items for a variety of constructs (Lee et al., 2023). However, approaches for AIG are often unsuitable for the constructed response formats employed in creativity tests, as they assume the use of measurement models whose parameters cannot easily be estimated for constructed responses. We developed a new AIG method for generating creativity test items. Our framework, the creative psychometric item generator (CPIG), uses a mixture of LLM-based item generators and evaluators to iteratively develop creativity items, such that items from later iterations demonstrate stronger validity evidence. CPIG combines item generation with item parameter estimation (Laverghetta et al., 2021): LLMs generate and respond to the items such that the prompt used for item generation can be refined iteratively based on LLM feedback. Across two creativity assessments, CPIG generates valid and reliable items with psychometric properties similar to human-written items. This effect is also not attributable to known biases in the evaluation process. Our findings demonstrate how AI may facilitate creativity testing and could serve as the basis for future item-generation efforts in constructed response assessments.

Automatic item generation for figure reasoning tests using generative AI

Friday, 18th July - 09:15: Automated Item Generation (GH: Meridian 1-2) - Oral

Ms. Jing Huang (Purdue University), Dr. Hua-Hua Chang (Purdue University)

The efficiency of computerized adaptive testing (CAT) relies heavily on the quality of its item bank. However, developing and calibrating high-quality items is costly and labor-intensive, making it impractical when large-scale item banks are required. Automatic item generation (AIG) offers a promising solution, but traditional AIG methods are often constrained to content areas that are easy to model and require highly skilled experts to design item templates. With advancements in generative AI (Gen-AI), researchers have explored AI-driven approaches, such as using structured prompt templates to maintain consistency in item construction or leveraging machine learning models like GPT-2 to generate items based on a few initial examples. However, these methods lack inherent difficulty classification, making it challenging to control item quality and interpret test results. Moreover, most Gen-AI-generated items are limited to multiple-choice or text-based formats currently, which may not fully engage examinees or optimize assessment accuracy. To address these challenges, this study aims to propose an innovative framework for generating Figure Reasoning CAT items by integrating Gen-AI. We design a structured prompt template with predefined difficulty levels based on the transformation methods, the number of elements and the degree of distraction. Then, we utilize text-to-image generators to create corresponding visual stimuli for both item content and answer choices. This approach enhances item diversity and test validity while advancing the integration of Gen-AI in educational assessment.

Generating quantitatively grounded free-text using large language models

Friday, 18th July - 09:30: Automated Item Generation (GH: Meridian 1-2) - Oral

Ms. Lindley Slipetz (University of Virginia), Ms. Jiaxing Qiu (University of Virginia), Dr. Teague Henry (University of Virginia)

Traditionally, psychological measurement relies on numerical scales to quantify constructs, ordering participants along their continua, though often describing a single conceptualization of a construct, potentially missing heterogeneity in how it might manifest. Free text (i.e., asking participants to respond to a given prompt however they like) is commonly used in qualitative research; but its complexity, quantitatively analyzing large amounts of free text data, creates difficulties for researchers. Large language modeling (LLM) as a tool of analyzing free-text data (via embeddings as well as generative tools) helps overcome this. To evaluate LLMs' capability to understand/quantify free text data, we need the ability to simulate quantitatively-grounded free text data.

In this study, we evaluate LLMs' ability to generate free text data when provided with a construct definition, demographic traits, and the percentile on the continuum (e.g., a free text response to "Describe the impact of your depression symptoms in the last 24 hours" from a 20-year-old Asian-American woman in the 80th percentile of the Beck Depression Inventory). Overall, the goal is to provide a simulation platform for future methodological development.

Here, we develop and test prompts, and using the embedding space, evaluate the semantic structure of the generated text. Of particular interest is the sensitivity of the generation to relevant demographic differences that might impact how real participants could respond. Future directions of this project include developing construct-specific models to convert free text responses to quantitative percentile estimates and using this in a reinforcement learning loop, fine-tuning the simulation approach.

Evaluating cut score consistency in standard setting procedures for automatic item generation testing

Friday, 18th July - 09:45: Automated Item Generation (GH: Meridian 1-2) - Oral

Dr. Stella Kim (University of North Carolina at Charlotte), Dr. Tong Wu (Riverside Insights), Dr. Kristin Villanueva (University of North Carolina at Charlotte), Ms. Qiao Liu (University of North Carolina at Charlotte), Dr. Won-Chan Lee (University of Iowa)

Many testing programs, like licensure exams or college admissions tests, require one or multiple cut scores to determine minimum qualifications. The process of establishing these cut scores is known as “standard setting.” This standard setting process involves a panel of content experts (typically 5-15 panelists) who review each item in a test form and estimate the percentage of minimally competent examinees who would correctly answer each item (the modified Angoff method). The cut score is typically determined by summing the percentages across items in the test form.

This seemingly simply process encounters a challenge, however, when each examinee receives a test form different from the one used in the standard setting procedure. This is particularly relevant in automatic item generation (AIG) testing, where the test form differs for each examinee (Gierl & Lai, 2013). To address this issue, the study collected data from 4th-grade math teachers to evaluate the consistency of cut scores across different standard-setting procedures. In the first panel, 7 teachers reviewed a “parent form.” In the second panel, 8 teachers reviewed a child form generated using ChatGPT, based on an item model developed from the parent form. In the final panel, 7 teachers reviewed distinct child forms generated for each participant.

This study applies generalizability theory to analyze the data collected from each panel. The main focus of the analysis is the standard errors of cut scores, specifically the variability in group means. Both items and teachers are considered as facets that contribute to measurement errors.

AI-driven item generation for PIRLS

Friday, 18th July - 10:00: Automated Item Generation (GH: Meridian 1-2) - Oral

Dr. Ummugul Bezirhan (Boston College), Mrs. Erin Wry (Boston College), Prof. Matthias von Davier (Boston College)

This study explores the application of Automated Item Generation (AIG) within the PIRLS framework, leveraging GPT-4o to develop reading comprehension items while reducing reliance on human subject-matter expertise. Developing PIRLS achievement items is a labor-intensive, collaborative process involving subject-matter experts and National Research Coordinators from over 50 education systems. Reading passage selection and subsequent item-writing efforts are carefully conducted, with materials reviewed to ensure cultural relevance and clarity across diverse student populations. By integrating AI with PIRLS item-writing guidelines, this study examines the feasibility of AIG for four key comprehension processes: focus on and retrieve, make inferences, interpret and integrate, and evaluate and critique.

Using two released PIRLS 2021 passages, one informational and one literary, over 150 AI-generated questions per passage were produced. After removing duplicates, a filtering strategy was applied based on empirically determined BLEU, ROUGE, METEOR, and cosine similarity thresholds to optimize quality and recall. A weighted ranking system (Cosine similarity 40%, ROUGE-L 30%, METEOR 30%) identified the top four matches per human-written question, ensuring alignment. When direct matches were not found for each question, relaxed thresholds were introduced to maintain coverage without sacrificing overall quality. This process narrowed the pool to 30 AI-generated items, which were further evaluated by reading experts.

Preliminary results indicate strong alignment between AI- and human-authored questions. Additional validation was done through psychometric comparison of fourth graders' responses to AI-generated and human-written items. Findings suggest AI based AIG can reduce human labor associated with assessment development while maintaining quality.

Perfect timing: An algorithm for leveraging optimal temporal design to enhance statistical power

Friday, 18th July - 09:00: Longitudinal Data Analysis II (GH: Meridian 3-4) - Oral

Ms. Anne-Charlotte Belloeil (Vanderbilt University), Dr. Kristopher Preacher (Vanderbilt University)

Longitudinal studies are instrumental for capturing behavioral changes over time, but they often require substantial time and financial resources. Temporal design—decisions regarding when, how often, and at what intervals to measure—is a component of research planning uniquely germane to longitudinal studies. Though often overlooked, suboptimal temporal design can lead to reduced statistical power, biased correlation coefficients and regression coefficients, misrepresentation of mediation effects, concealment of functional forms, inability to detect effects given complex growth curves, and failed replication attempts across studies with different measurement schedules. Despite methodological appeals for better temporal planning, researchers lack practical tools to optimize their study designs during the planning stage. We introduce a novel iterative grid-search algorithm, implemented in R, that provides a flexible tool for temporal design optimization. By optimizing the spacing of their measurement occasions with respect to various optimal design criteria, the algorithm enables researchers to dramatically enhance the power of their longitudinal studies without increasing the number of measurement occasions or the sample size. Multiple optimization criteria are available to meet the unique research needs of users (such as maximizing the efficiency of a focal predictor, enhancing the collective efficiency of all predictors, or improving the efficiency of the least efficient predictor, etc.) to achieve higher statistical power without additional costs, to empower scientists to leverage scarce research funds more efficiently, and ultimately to better capture the dynamics of change phenomena over time.

Person-specific updating of EWMA control limits in sparse in-control data scenarios

Friday, 18th July - 09:15: Longitudinal Data Analysis II (GH: Meridian 3-4) - Oral

Dr. Evelien Schat (KU Leuven, University of Leuven), Dr. Francis Tuerlinckx (Katholieke Universiteit Leuven), Prof. Eva Ceulemans (KU Leuven, University of Leuven)

Retrospective analyses of experience sampling (ESM) data suggest that changes in mean levels may serve as early warning signals of impending depressive episodes. Detecting these signals in real time could enable timely intervention and prevention. The exponentially weighted moving average (EWMA) procedure is a promising method to scan ESM data for the presence of mean changes in real time. First, this procedure captures the natural variation present in a set of in-control data, used to establish control limits. Afterwards, incoming data are compared to the in-control distribution, to detect and test whether and when the incoming data go out of control (i.e., when the data go beyond the control limits). However, a major challenge is the amount of in-control data needed for optimal performance (i.e., at least 50 scores). Collecting such an amount of data from a single person (in a healthy state) is often challenging, if not impossible. To address this, we explore whether we can use the person's incoming data to update the control limits over time, thereby circumventing the need for the large initial in-control dataset. This approach, already showing promising results in pilot simulations, uses a decision rule to determine whether incoming data fall within the same range as the in-control data. If so, the incoming score(s) are added to the in-control data set, yielding more data for deriving updated and more reliable control limits. To further improve performance, we propose refinements that incorporate characteristics of ESM data and evaluate their effectiveness through simulations.

Lowering participant burden in long-term ESM studies through variable sample size EWMA

Friday, 18th July - 09:30: Longitudinal Data Analysis II (GH: Meridian 3-4) - Oral

Ms. Fien De Pauw (KU Leuven, University of Leuven), Dr. Evelien Schat (KU Leuven, University of Leuven), Dr. Francis Tuerlinckx (Katholieke Universiteit Leuven), Prof. Eva Ceulemans (KU Leuven, University of Leuven)

The exponentially weighted moving average (EWMA) procedure is a promising tool for detecting changes in experience sampling (ESM) data in real time. This allows timely interventions when mental health is worsening. However, asking participants to complete multiple ESM questionnaires per day over extended periods increases their perceived burden. We, therefore, investigate whether we can lower this perceived burden by adapting the number of ESM questionnaires per day (i.e., sample size) based on the participant's current ESM scores.

The EWMA procedure determines whether a process has gone out of control by comparing incoming data to previously established control limits, which reflect the natural variation of an in-control process (e.g., affect in times of well-being). Traditionally, the EWMA procedure uses a fixed number of measurement occasions each day (i.e., a fixed sample size). We investigate whether we can use variable sample size (VSS) EWMA, in which the sample size can vary per day. VSS EWMA allows for sampling at a higher rate only when there is evidence of a change in the individual's ESM data, while keeping the sample size low—and thus minimizing burden—when the individual is doing well. As VSS EWMA is commonly used in industrial applications, we tailored this approach to the use of ESM data. Based on our simulation results, we provide insight into the potential and challenges of this VSS EWMA approach for decreasing participant burden in ESM studies.

Nonparametric estimation of latent growth parameters and heterogeneity

Friday, 18th July - 09:45: Longitudinal Data Analysis II (GH: Meridian 3-4) - Oral

Mr. Graham Buhrman (University of Wisconsin - Madison), Prof. Jee-Seon Kim (University of Wisconsin - Madison), Dr. Weicong Lyu (University of Macau)

Questions that concern multilevel data are commonplace in the social sciences. Whether they are students nested within schools, or repeated measurements on children learning to read, one need not look far to find situations where observations are clustered according to some real-world or design-induced structure. Latent growth curve models (LGCMs) with random growth parameters are widely used to analyze longitudinal data, capturing individual differences in trends and rates of change. These models are used to examine heterogeneity over time, including time-varying and time-constant effects, factors that drive growth variability, cross-level interactions, and groupwise differences. Despite their interpretability and flexibility, LGCMs rely on strong assumptions of multivariate normality and specific model structures, particularly linear relationships between growth parameters and predictors. In this study, we propose an approach to growth curve modeling which relaxes the normality and linearity assumptions while retaining the intuitive framework of LGCMs. Our approach integrates multilevel modeling with nonparametric estimation, particularly Bayesian Additive Regression Trees (BART), to identify key growth determinants and sources of variability by flexibly estimating latent growth parameters and capturing relationships with predictors. By relaxing these assumptions, we can more flexibly characterize heterogeneity in repeated measures data. We present the logic and mechanics of how we might practically relax these assumptions using nonparametric estimation, and we demonstrate an application of our approach in a small simulation study.

Estimating non-normal random effects in nonlinear random effects models

Friday, 18th July - 10:00: Longitudinal Data Analysis II (GH: Meridian 3-4) - Oral

Ms. Yue Zhao (University of Minnesota), Prof. Nidhi Kohli (University of Minnesota)

Nonlinear random effects models (NREMs) are particularly useful for modeling growth or decline over repeated measured (longitudinal) data. The standard model assumes normal distributions for random effects and random errors. However, there is a significant need in psychological and educational research to relax this assumption and allow for non-normal specifications. This study addresses this critical gap in the current literature by developing the NEfit.R algorithm, which allows for non-normal distribution for random effects and random errors within NREMs. This software algorithm uses the marginal likelihood framework and a two-stage Newton-Raphson algorithm to estimate model parameters. In the marginal likelihood approximation, we use a probability integral transformation (PIT) method to handle non-normal densities for the Gauss-Hermite quadrature method. To demonstrate the utility of the NEfit.R function, we present a real data example and conduct a comprehensive Monte Carlo simulation study to evaluate its performance. The results indicate that correctly specifying the random effects and random errors distributions significantly improved the overall model estimation, particularly in variance estimation for the intrinsically nonlinear parameter. We conclude that the correct distributional specification for random effects and random errors is essential for making accurate and valid statistical inferences from NREMs.

Confidence intervals based on scaled difference tests in SEM

Friday, 18th July - 09:00: Structural Equation Modeling II (MAC: Johnson) - Oral

Prof. Carl Falk (McGill University, Montreal, Canada), Mr. Lihan Chen (McGill University, Montreal, Canada)

Although likelihood-based confidence intervals sometimes have better coverage rates than Wald-based intervals in structural equation modeling, some challenges and unknown details remain regarding their performance. One issue concerns their robustness to distributional assumption violations. While past research has implemented a variant of likelihood-based intervals based inverting a scaled difference test by Satorra (2000), multiple scaled difference tests could have been used (e.g., Satorra & Bentler, 2001; 2010). Furthermore, these approaches have apparently not yet been compared to Wald-based intervals based on a sandwich covariance matrix with observed information (Huber-White or "MLR"). Finally, to our knowledge *lavaan*-based software implementations are quite challenging to get working properly and several solutions, including the new *semlbci* package (Cheung & Pesigan, 2023), have not yet been compared. In this research, we report the results of two simulation studies that partially address these shortcomings, evaluating three different scaled difference tests versus each other, Huber-White standard errors, and two different software implementations to obtain intervals based on inverting Satorra's (2000) difference test. As context, we examine a classic ACE behavioral genetics model and a cross-lagged panel model with an indirect effect—both under several nonnormality generation techniques. In general, we find that it is easiest to get Satorra's (2000) difference test to work well, and it can sometimes outperform Huber-White standard errors. Furthermore, we document challenges in estimation of these intervals if *lavaan* (Rosseel, 2012) is used as the underlying software.

Bayesian fit measures in detecting misspecified multilevel structural equation modeling

Friday, 18th July - 09:15: Structural Equation Modeling II (MAC: Johnson) - Oral

Dr. Chunhua Cao (The University of Alabama), Dr. Xinya Liang (University of Arkansas)

Multilevel structural equation modeling (MSEM) is frequently estimated from frequentist framework using maximum likelihood methods. Previous research has shown that common fit indices such as RMSEA, CFI, and TLI could only detect the within-group model misspecification, and SRMR-B was the sole fit index sensitive to the between-group model misspecification (Hsu et al., 2015). An increasingly popular alternative for estimating MSEM is the Bayesian framework using Markov chain Monte Carlo methods. Simulation studies have showcased the superiority of Bayesian MSEM in terms of model convergence and parameter estimation, especially when accurate priors were specified (Depaoli & Clifton, 2015). However, no study has examined the performance of Bayesian fit indices in detecting model misspecification in MSEM. The Bayesian version of approximate fit indices, including Bayesian CFI (BCFI) and the Bayesian RMSEA (BRMSEA), have been developed in Bayesian SEM analogous to frequentist fit indices (Garnier-Villarreal & Jorgensen, 2020). In addition, posterior predictive p-value (PPP) and information criteria (ICs) are valuable tools to detect model misspecification. Thus, this study compares the performance of various Bayesian fit measures, including PPP, BCFI, BTLI, BRMSEA, BIC, and DIC, in detecting model misspecification in MSEM. The design factors include the number of clusters, cluster size, intercorrelation coefficient (ICC), and the location of model misspecification (i.e., at the between-level, or at the within-level). The model misspecification can be latent factor covariance or cross-loadings misspecification. Also, the impact of different types of priors is to be examined. The implications for applied researchers will be discussed.

A priori distributions in Bayesian structural equation modeling: A scoping review protocol

Friday, 18th July - 09:30: Structural Equation Modeling II (MAC: Johnson) - Oral

Dr. Jorge Sinval (Nanyang Technological University), Dr. Sonja Winter (University of Missouri), Prof. Joseph B. Kadane (Carnegie Mellon University), Prof. Edgar C. Merkle (University of Missouri)

Bayesian Structural Equation Modeling (SEM) has received increasing interest due to its capacity to address challenges in the frequentist approach, such as nonconvergence, Heywood cases, small sample sizes, and inadmissible solutions. A defining feature of Bayesian SEM is the use of *a priori* distributions, which play a fundamental role in parameter estimation, model interpretation, and uncertainty quantification. While researchers may express skepticism regarding the subjectivity of priors, they represent a substantial advantage of Bayesian methods, enabling the integration of prior knowledge about parameters before observing the data.

However, the choice and specification of *a priori* distributions remain an underexplored aspect of Bayesian SEM. This protocol outlines the goal of a scoping review designed to explore the application of *a priori* distributions in Bayesian SEM, with a particular focus on confirmatory factor analysis and full SEM models.

Key aspects of *a priori* distribution usage will be examined, including their application across different dimensionality structures and model parameters. Special attention will be given to priors for variances, loadings, regression coefficients, and covariances, both in terms of distribution families and hyperparameter values, highlighting their impact on posterior distributions, model estimation, and performance. By synthesizing recent literature, this review will identify trends, challenges, and gaps in the use of *a priori* distributions within the Bayesian SEM framework. The findings aim to promote informed decision-making regarding prior elicitation and enhance the robustness of Bayesian SEM applications.

The influence of informative priors in the estimation of MIMIC model parameters with small sample sizes and outliers

Friday, 18th July - 09:45: Structural Equation Modeling II (MAC: Johnson) - Oral

Ms. Nancy Alila (University of Georgia), Prof. Zhenqiu Lu (University of Georgia)

With advancements in software developments, Bayesian Structural Equation Modeling (SEM) has become a widely used approach for estimating model parameters in data analysis. The effectiveness of Bayesian estimation often depends on the choice of prior distributions. In small samples, researchers have suggested that informative priors may be beneficial. Considering data in social and behavioral research are rarely normally distributed, this study examines whether selecting appropriate priors can help regularize estimation and mitigate bias when estimates are sensitive to deviations from normality, such as heavy-tailed distributions or skewed errors. Specifically, we conduct simulation studies to exam the influence of Bayesian prior distributions on parameter estimation in the Multiple Indicator Multiple Cause (MIMIC) model when errors deviate from normality due to outliers and small sample sizes. Various prior distributions are explored, with the expectation that weakly informative priors will allow data to drive the estimates while still providing regularization, helping to reduce overfitting in the presence of extreme observations, and that strongly informative priors may shift estimates toward prior knowledge, which can be particularly useful when data are contaminated with outliers. Additionally, when measurement errors follow a heavy-tailed distribution, such as in the presence of outliers, heavy-tailed priors like the Student's t-distribution are tested as an alternative to normal priors to evaluate whether they can prevent extreme values from disproportionately influencing parameter estimates in small samples. The study's findings will provide insights into practices for Bayesian estimation in non-normal conditions.

Advancing contingent paradigms evaluating model fit in structural equation modeling

Friday, 18th July - 10:00: Structural Equation Modeling II (MAC: Johnson) - Oral

Prof. Thomas Niemand (TU Clausthal), Prof. Nadine Schröder (Bielefeld School of Business, Hochschule Bielefeld – University of Applied Sciences and Arts), Dr. Andreas Falke (University of Regensburg), Prof. Robert Mai (Grenoble Ecole de Management)

Structural equation modeling (SEM) is a cornerstone in social and behavioral research. A central aspect of SEM is assessing global model fit, which determines how well a model represents the data. While traditional approaches proposing rigid cutoffs (most importantly, Hu & Bentler, 1999) are widely used for decades, recent advancements have introduced more flexible and dynamic alternatives that simulate effective cutoffs. Two prominent approaches are Flexible Cutoffs (FCO; Niemand & Mai, 2018) and Dynamic Cutoffs (DFI; McNeish & Wolf, 2023), which aim to provide more accurate assessments of model fit by generating simulated cutoffs from a given model and data. Yet, they rest on notable conceptual differences for determining the cutoffs by emphasizing either Type I or Type II errors, with specific strengths but also limitations. While FCO only controls for Type I error, DFI corrects for both error types, yet does not provide cutoffs when the overlap between the simulated correct and misspecified models is substantial. The present study suggests incorporating both approaches through revising the decision models (FCO2) or minimizing Type I and II errors (cut point, CP). With a comprehensive simulation study, we conclude that FCO2 (mean = .837) and CP (mean = .866) outperform DFI (mean = .579) in overall accuracy, largely due to DFI not considering many sources of misfit. These findings highlight the importance of carefully considering the underlying conditions and a cutoff's generalizability, ultimately guiding researchers in making informed decisions about their analyses.

An approach that can validate both Q-Matrices and attribute hierarchies in cognitive diagnosis models

Friday, 18th July - 09:00: Cognitive Diagnostic Models III (MAC: Thomas Swain) - Oral

Dr. Lingling Wang (Shenyang Normal University)

Cognitive diagnostic models (CDMs) are developed to diagnostically evaluate subjects' cognitive strengths and weaknesses based on the Q-matrix mappings of their items and attributes. Due to the subjective process of the traditional Q-matrix construction, there inevitably are misspecifications in the Q-matrix, which, if left unchecked, may result in a serious negative impact on CDMs. From another important perspective, in the empirical applications of CDMs, cognitive attributes generally do not operate independently, and a certain hierarchical relationship may be present among the cognitive attributes. The correctness of both the Q-matrix and the attribute hierarchy significantly impacts the parameter estimation ability of a CDM and the accuracy of the examinee's classification result. Recently, considerable studies have developed approaches for validating Q-matrices or testing attribute hierarchies respectively. However, no method that can validate both a Q-matrix and an attribute hierarchy simultaneously. An approach based on Bayesian networks (BN) for validating both Q-matrices and attribute hierarchies simultaneously is proposed in this research. To explore the performance of the BN method, this research conducted two simulation studies and empirical data analysis to theoretically and practically evaluate the accuracy of the Q-matrix validation and attribute hierarchy correction processes. In conclusion, the initial specified Q-matrix and attribute hierarchy can be simultaneously validated via the BN method. Then the corrected Q-matrix and the refined attribute hierarchy obtained from the data-driven BN method can again be combined with the theoretical judgments of experts to obtain a more optimized model, finally achieving more accurate diagnostic outcomes in CDA practice.

The influence of Q matrix mis-specified on the classification of nonparametric cognitive diagnosis based on Hamming distance

Friday, 18th July - 09:15: Cognitive Diagnostic Models III (MAC: Thomas Swain) - Oral

Dr. Sun Rui (Beijing Normal University), Dr. Jiahui Zhang (Beijing Normal University)

Cognitive Diagnosis Assessment (CDA) evaluates learners' knowledge based on their response patterns, with the Q matrix playing a key role in the diagnosis accuracy. However, Q matrix mis-specification can affect this accuracy. Few studies have explored the robustness of non-parametric cognitive diagnostic methods under such mis-specification. This study investigates the robustness of three methods—Hamming distance, weighted Hamming distance, and General Nonparametric Parameterized Classification(GNPC)—under Q matrix mis-specification using self-written R code. The study first examines Q matrix mis-specification at 10% and 20% error rates, testing the methods under two conditions: 3 or 5 attributes and 4 hierarchical relationships. It then compares the accuracy of Hamming distance, weighted Hamming distance, and GNPC under three types of mis-specification (missing, redundancy, and mixed) at a 10% mis-specification rate. Finally, compare the research results with the true values of the correct Q matrix and evaluate the robustness of the method using the PMR(Pattern Match Ration, PMR).The main conclusions were: (1) The sample size does not impact the PMR of the three methods under Q matrix mis-specification. (2) The Q matrix hierarchy type affects method robustness, with GNPC performing best under complex hierarchies. (3) A 20% mis-specification rate causes a greater PMR decrease than a 10% rate. (4) With larger item numbers, the decrease of PMR follows the order: redundancy > mixed > missing. (5) PMR degradation increases as hierarchical looseness grows.

Bayesian estimation of the Q-matrix and attribute hierarchy in DINA model

Friday, 18th July - 09:30: Cognitive Diagnostic Models III (MAC: Thomas Swain) - Oral

Dr. Xue Wang (Northeast Normal University), Dr. Shiyu Wang (University of Georgia), Dr. Yinghan Chen (University of Nevada, Reno)

Cognitive Diagnosis Models (CDMs), which provide a finer-grained evaluation of examinees' attribute master pattern, are widely used in educational and psychological measurement. In many applications of CDMs, it is commonly assumed that a hierarchical structure is present to delineate the relationships among attributes which plays an important role in CDMs. Furthermore, another crucial component in CDMs is the Q-matrix, which precisely specifies the attributes measured by each item. In many current studies, the hierarchical structure among attributes and the Q-matrix are commonly presupposed. However, in practice, the presupposed hierarchical structure and Q-matrix may be incorrectly specified. Previously, Chen and Wang (2023) introduced a Bayesian method for estimating the attribute hierarchy under the assumption of a pre-specified Q-matrix. Therefore, in this study, based on deterministic inputs, noisy-and gate (DINA) model, we transcend the constraint of a pre-specified Q-matrix and propose a Markov chain Monte Carlo method that achieves simultaneous estimation of both the attribute hierarchy and Q-matrix. The proposed method performs well in recovering the real attribute hierarchy and Q-matrix. A simulation study is conducted to demonstrate the validity of our method. Additionally, we employ a mini-batch method to enhance computational efficiency. At last, we demonstrated the utility of the proposed algorithm based on the real data.

A new reliability framework for cognitive diagnosis models

Friday, 18th July - 09:45: Cognitive Diagnostic Models III (MAC: Thomas Swain) - Oral

Prof. Youn Seon Lim (University of Cincinnati)

Reliability assessment in cognitive diagnosis models (CDMs) presents unique challenges due to their categorical classification structure. Traditional reliability indices, such as Cronbach's alpha, assume continuous latent traits and fail to capture classification consistency, a core aspect of CDMs. This study proposes a new reliability measure designed specifically for CDMs, focusing on the stability of attribute mastery classifications across repeated assessments. The measure quantifies the probability that examinees receive the same skill profile under repeated testing conditions, addressing the limitations of existing methods. Through extensive simulation studies, we evaluate its performance across different CDM structures, test lengths, and sample sizes, benchmarking it against existing reliability indices. Results indicate that the proposed measure provides more precise and interpretable reliability estimates, particularly for diagnostic assessments where classification accuracy is critical. Additionally, we demonstrate its practical utility using large-scale educational assessment data, highlighting its potential for improving test design and score interpretation. This work advances the methodological foundations of CDMs by introducing a reliability framework that aligns with their classification-based nature, ultimately enhancing the validity of diagnostic decisions in educational and psychological measurement.

Measuring forecasting proficiency: An item response theory approach

Friday, 18th July - 09:00: IRT applications (GH: Think 4) - Oral

Mr. Fabio Setti (Fordham University), Dr. Leah Feuerstahler (Fordham University), Ms. Sophie Ma Zhu (University of British Columbia), Mr. Nikolay Petrov (University of Cambridge), Dr. Ezra Karger (Federal Reserve Bank of Chicago), Dr. Mark Himmelstein (Georgia Institute of Technology)

Although recent computational advances have allowed statistical models to provide increasingly accurate forecasts, the human component remains a cornerstone of making decisions about uncertain events. The Forecasting Proficiency Test (FPT) is a test developed to measure individuals forecasting ability by posing realistic forecasting questions. The FPT implements a quantile forecast format, which requires individuals to make five forecasts regarding a continuous future outcome at increasing levels of certainty. In this paper, we analyze FPT items using a customized item response theory (IRT) approach. Quantile forecast items differ from items suitable for IRT because responses represent accuracy, the signed distance between the forecast and the resolution, which is a continuous measure. We lean on the IRT theoretical framework to model accuracy of forecasts at each quantile while estimating properties of the FPT items, such as item difficulty and uncertainty, as well as person-level forecasting proficiency. We posit a probabilistic model that describes the observed accuracy at each quantile and explore multiple distributions to model the idiosyncrasies of the data. Then, we interpret estimated item parameters and describe how estimated person forecasting proficiency relates to more conventional scoring rules for quantile forecast tasks. We also discuss how to tune the model to incorporate different scoring criteria. The advantages the IRT approach offers in understanding how individuals interact with forecasting items and possible benefits in test construction and person scoring for items with a quantile forecast format are discussed.

Demographic influences on spatial ability: A 2PL and Rasch tree analysis

Friday, 18th July - 09:15: IRT applications (GH: Think 4) - Oral

Mr. Justice Dadzie (The University of Alabama), Mr. Daniel Oyeniran (The University of Alabama), Ms. Patricia Quaye (The University of Alabama), Prof. Joni M. Lakin (The University of Alabama)

This study examines the influence of demographic factors on spatial ability using a combination of the Two-Parameter Logistic (2PL) model and Rasch Tree Analysis. The 2PL model will be used to estimate group differences (i.e., race, gender) in three spatial assessments: Mental Rotation Test (MRT), Object Assembly (OA), and Surface Development (SD). The Rasch Tree model provides a means to test for differential item functioning (DIF) across demographic groups concurrently. This means that DIF can be identified for intersectional groups rather than being restricted to dichotomous comparisons. By applying Rasch tree analysis, significant subgroup differences emerged, revealing gender as the primary variable affecting measurement properties, followed by test type among male participants.

The 2PL model was employed to evaluate item difficulty and discrimination, highlighting variations across test items. Results demonstrated significant gender differences, with female participants consistently outperforming males in all spatial tasks (counter to prior research). Group differences based on race and ethnicity also revealed disparities, particularly in MRT and SD performance, suggesting the potential influence of cultural and educational factors on spatial cognition. The combination of Rasch Tree and 2PL analyses provided deeper insights into how demographic variables interact with spatial ability measures, allowing for the detection of measurement invariance violations and subgroup-specific performance patterns. We recruited a sample of 123 students from grades 3 to 8, representing an age range typically between 8 and 14 years. Future research should expand these methods to further explore demographic influences on spatial reasoning and refine measurement models for diverse populations.

Desirability-matched Thurstonian IRT scale construction leveraging sentiment analysis

Friday, 18th July - 09:30: IRT applications (GH: Think 4) - Oral

Dr. Kensuke Okada (The University of Tokyo), Mr. Yoshito Tan (The University of Tokyo), Mr. Keishi Nomura (Department of Media and Communication, Toyo University), Mr. Kotaro Tsuchida (The University of Tokyo), Dr. Kyosuke Bunji (Kobe University)

Forced-choice measurements have gained attention owing to their ability to mitigate social desirability bias, which refers to the tendency of respondents to answer as if they were desirable people. To effectively reduce this bias, the desirability of the paired items must be matched. This necessitates a preliminary process of determining the desirability for each item, which generally requires collecting response data from respondents and is therefore cost- and resource-intensive. To address this issue, we propose an approach that incorporates two key elements. First, as an effective proxy for social desirability, we propose using the valence of the item, which can be quantified using sentiment analysis techniques. Second, we propose constructing a set of Thurstonian IRT items using combinatorial optimization such that the difference in valence scores within each item pair is below a pre-specified threshold. We developed a forced-choice scale for the Big Five personality traits as an empirical application of the proposed framework. Following preregistration, we collected responses on social desirability and valence for 500 personality trait words and observed that the two were highly correlated (> 0.99). Subsequently, using a blueprint that specified the factors to be measured as well as the forward/reverse directions for each item, we performed combinatorial optimization to finalize the scale. The proposed approach enables the construction of a psychometric scale that is both resilient to social desirability bias and capable of leveraging IRT advantages—such as equating and adaptive testing—in a cost-effective manner, making it particularly well-suited for high-stakes assessments.

Resource usage and its influence on ability estimation – exploring different IRTree models in the context of PISA-LDW data

Friday, 18th July - 09:45: IRT applications (GH: Think 4) - Oral

Dr. Leonard Tetzlaff (DIPF | Leibniz Institute for Research and Information in Education), Mr. Lothar Persic-Beck (DIPF | Leibniz Institute for Research and Information in Education), Dr. Ulf Kröhne (DIPF | Leibniz Institute for Research and Information in Education), Prof. Carolin Hahnel (Ruhr University Bochum), Prof. Frank Goldhammer (DIPF | Leibniz Institute for Research and Information in Education)

Digital assessment environments sometimes provide learning resourcesthat can be accessed by test takers on demand. To measure the ability underlying students' performance in the assessment, it is important to know whether the use of thoseresources influences the estimation of item difficulty and/or interindividual differences in ability.

To explore these questions, different models that decompose the response process into ordered subprocesses (IRTree models, DeBoeck & Partchev, 2012), including the decision to use offered resources or not, were considered. One group of models assumes that learners make the decision on whether to use resources before engaging with the task; the other group assumes that learners only decide on whether to use the offered resources after unsuccessfully engaging with the task.

We use data from the German piloting study (N = 737) on the innovative domain in PISA 2025 "Learning in the Digital World" (LDW). This LDW ability is demonstrated by solving computational and scientific inquiry problems in an environment that offers worked examples as learning resources to be accessed on demand by test-takers.

We compare and contrast alternative IRTree models based on the plausibility of their assumptions, the resulting parameter estimates and their relations to other variables. Three out of four models conclude that items that were solved after accessing resources measure a different ability than items that were solved without resources. Substantive inferences drawn from the different models differ qualitatively in respect to the key research question, leading to more general implications for the use of IRTrees to model sequential behavior.

Measuring print exposure using a bifactor unipolar IRT model

Friday, 18th July - 10:00: IRT applications (GH: Think 4) - Oral

Prof. Qi (Helen) Huang (Purdue University), Prof. Daniel Bolt (University of Wisconsin - Madison)

Unipolar IRT has been suggested to provide a better theoretical representation of certain educational (e.g., reading frequency, vocabulary) and psychological (e.g., depression, addictive behaviors, extreme beliefs) constructs than the more traditional bipolar IRT approaches. However, potential multidimensional extensions of unipolar IRT applications remain underdeveloped, limiting its broader application and the ability to capture more complex construct representations. In this study, we extend the unipolar IRT framework to incorporate a bifactor structure and demonstrate its usefulness using an Author Recognition Test (ART; Wimmer & Ferguson, 2023), a measure of print exposure. A bifactor representation is argued to be particularly appropriate in this setting as it can accommodate the increased tendency for reading to become focused in particular genres as reading frequency increases. The value of the bifactor unipolar IRT model is seen not only through its superior fit to the data, but also by its ability to explain and reduce the tendency for item-specific hazard functions to rapidly decline as reading frequency increases. The paper thus furthers our understanding of how survival modeling techniques can be of relevance in educational and psychological measurement.

Fitting propensity analysis in R

Friday, 18th July - 09:00: Advances in the Evaluation of Statistical Model Complexity (GH: Think 5) - Oral

Dr. Sonja Winter (University of Missouri), Mr. Tive Khumalo (University of Missouri)

Conducting a fitting propensity (FP) analysis has recently become more accessible through the introduction of easy-to-use R packages. However, guidance on how best to design and execute an FP analysis for accurate results is still lacking. Thus, in this talk, we will present the results of an extensive investigation in which we systematically varied characteristics of the data space (e.g., number of variables in data space, permissible correlations) and choices surrounding model estimation and evaluation (e.g., number of ML estimator iterations, fit index selection) to evaluate their potential impact on FP analysis performance. In this examination, we focused on latent variable models with continuous indicators and the *ockhamSEM* (Falk & Muthukrishna, 2023) R package. Results indicate that several design choices affect FP analysis findings. For example, we found that conclusions regarding a model's FP depend on whether the data space consists of all positive definite covariance matrices or only those matrices in the positive manifold. We will discuss all findings and provide practical recommendations for researchers interested in conducting FP analyses, including guidance on selecting a sufficient data space size for the number of variables included. Broader implications for the study of fitting propensity and extensions to other variable types and software packages will be discussed.

The fitting propensity of 1-parameter item response theory models

Friday, 18th July - 09:00: Advances in the Evaluation of Statistical Model Complexity (GH: Think 5) - Oral

Dr. Hyejin Shim (University of Missouri), Dr. Wes Bonifay (University of Missouri), Dr. Yon Soo Suh (Houghton Mifflin Harcourt)

In item response theory (IRT), certain one-parameter models based on alternative link functions—specifically, the complementary log-log (CLL), negative log-log (NLL), and cauchit models—often outperform more complex traditional link models such as the three- and four-parameter logistic models (Shim et al., 2023a; 2023b). These findings suggest that a model's capacity to capture diverse data patterns depends not only on its number of parameters but also on the choice of its underlying link function.

In this study, we investigate how link function selection influences IRT model complexity, even when such models include just one item parameter. We use Preacher's (2006) fitting propensity analysis (a simulation-based method aligned with the principle of minimum description length) to evaluate the complexity of the (logit) Rasch, CLL, NLL, and cauchit models. Specifically, we fit each model to 1,000 datasets that were randomly and uniformly sampled from the complete categorical data space and compared their relative propensities to achieve goodness-of-fit via the Y2/N statistic.

Our results indicate that these one-parameter alternative link IRT models: (a) vary in complexity; (b) capture distinct data patterns; and (c) preferentially align with response patterns that reflect behaviors such as guessing and slipping. These findings reinforce the notion that IRT model complexity cannot be quantified simply by counting the number of free parameters in the model; psychometrics researchers must also consider that a model's functional form affects its propensity to fit well, both in general and to specific meaningful data patterns.

The fitting propensity of multilevel models

Friday, 18th July - 09:00: Advances in the Evaluation of Statistical Model Complexity (GH: Think 5) - Oral

Ms. Yun-Kyung Kim (University of California, Los Angeles)

This study applies fitting propensity (FP) analysis to quantify the complexity of multilevel models, using multilevel confirmatory factor analysis (ML-CFA) models as an example. ML-CFA models are commonly used to develop and validate measurement tools using datasets with a nested structure. In the process of model selection, practitioners often compare the goodness-of-fit (GOF) of competing ML-CFA models that differ in latent factor specifications and/or cross-level measurement invariance constraints. While model fit should be evaluated based on a balance between GOF and model parsimony, the latter—or model complexity, as its complement—has only been approximated by the number of freely estimated parameters. This metric, however, inaccurately represents model complexity and overlooks the distinct complexities at different levels of ML-CFA models. To address the gap, this study presents level-specific FPs for ML-CFA models that represent how well the models fit the within- and between-group level covariances within the complete data space. Results indicate that the interplay of level-specific FPs gives rise to model complexity beyond what is captured by the number of freely estimated parameters. For example, ML-CFA models tend to prioritize fitting within-group covariances over between-group covariances. Moreover, cross-level measurement invariance constraints are found to disproportionately impact the between-group level FP. These findings provide context for the GOF that practitioners encounter in their applications of these models, facilitating a more nuanced evaluation of the trade-off between GOF and model parsimony, while serving as empirical evidence that underscores the need for level-specific model fit evaluation and estimation.

The fitting propensity of factor analysis models

Friday, 18th July - 09:00: Advances in the Evaluation of Statistical Model Complexity (GH: Think 5) - Oral

Dr. Wes Bonifay (University of Missouri), Dr. Li Cai (University of California, Los Angeles), Prof. Carl Falk (McGill University, Montreal, Canada), Dr. Kristopher Preacher (Vanderbilt University)

Model complexity is a critical consideration when evaluating a statistical model. To quantify complexity, one can examine fitting propensity (FP), or the ability of the model to fit well to diverse patterns of data. The scant foundational research on FP has focused primarily on proof-of-concept rather than practical application. To address this oversight, the present work joins a recently published study in examining the FP of models that are commonly applied in factor analysis. We begin with a historical account of statistical model evaluation, which refutes the notion that complexity can be fully understood by counting the number of free parameters in the model. We then present three sets of analytic examples to better understand the FP of factor analysis models that are widely used in applied research. We show that a) exploratory and confirmatory models with the same number of parameters differ in FP; b) bifactor models with the same number of parameters but different specific factor structures differ in FP; and c) models that are equivalent in terms of overall FP may fit well to distinct patterns. We characterize these findings relative to previously disseminated claims about factor model FP. Finally, we provide some recommendations for future research on FP in latent variable modeling.

Characterisations of phenomena

Friday, 18th July - 11:45: Theory Construction Methodology (GH: Meridian 1-2) - Oral

Mr. Jason Nak (University of Amsterdam)

Psychological phenomena have been proposed as a more appropriate form of explananda in psychology. Phenomena are robust, general trends across datasets and less prone to the idiosyncrasies of data. Datasets are formed by a set of decisions regarding instruments, samples, etc. Phenomena are what remains when the influence of specific decisions are lessened through, for instance, conceptual replication. An example phenomenon is the response-time delay in incongruent trials, evidenced by data from both the Stroop-task and the Flanker-task. This leads to a triptych of theory-phenomena-data in which data evidence phenomena, while theories explain phenomena.

Phenomena are heterogenous in many ways, one of which is the ease with which they can be formally represented. Take for instance the difference between the positive manifold of cognitive skill and something like 'maturing out' of substance abuse. While the former has a clear representation, the latter may be captured by a correlational pattern, an interaction-effect, or a linear trend over time. As all these patterns could represent the phenomenon, all data that show these patterns can evidence the phenomenon. This is ideal from a robustness perspective. However, which one you choose to represent your phenomenon, for instance when testing whether a formal theory can recreate the pattern, becomes more difficult.

In this talk I will discuss the different levels on which phenomena can be described, ending with the representational level which concerns the statistical patterns that represent and evidence them. I will then discuss the implications this has for formal theory building and testing.

The boundaries between assumptions and phenomena in formal modeling: A case study of a computational model of addiction

Friday, 18th July - 11:45: Theory Construction Methodology (GH: Meridian 1-2) - Oral

Mr. Jesse Boot (University of Amsterdam)

An important criterion for evaluating the explanatory power of a formal model or theory is whether it can explain relevant phenomena: stable features in the world. In the context of formal modeling, phenomena are often characterized as statistical patterns. To test whether a formal model can explain a phenomenon, one can then run simulation studies and see whether patterns emerge that correspond to important phenomena that are reproduced in the simulation study. However, when a formal model produces a phenomenon, it must always mean that the phenomenon is to some extent “built in”. A model will not produce behavior that is not somehow derived from its assumptions. Using a formal model of addiction theory as an example, this talk will explore the boundaries between “building in” a phenomenon as an assumption, phenomena as emergent properties of interactions of modeling assumptions, and unexpected insights from formal models.

How cause-effect reasoning impedes theory development: Why we should focus on mechanisms and functions instead

Friday, 18th July - 11:45: Theory Construction Methodology (GH: Meridian 1-2) - Oral

Dr. Noah van Dongen (Erasmus University Rotterdam)

Scientific theories aim to explain and predict phenomena. In our field, our theories address phenomena of behavior, social relations, cognition, emotion, psychopathology, and so on. Few would disagree that (human) psychology and our social organization are systems with the number and type of processes that merit the adjective “complex.” This makes cause-effect reasoning problematic. And even if our systems of interest were merely “large” or “difficult to parse”, our standard causal reasoning may still present problems.

I will argue that our typical cause-effect reasoning often impedes theoretical progress. The appeal of causal explanations is intuitive: they provide clear, testable links between variables. However, this approach risks oversimplification, missing the forest for the trees, and (rein)forcing linear explanation. Exemplary problems of cause-effect reasoning are: many causes to one effect, irrelevant prime movers, differences in the value of “cause” in well-functioning and dysfunctional systems.

I will contrast cause-effect reasoning with mechanistic and functional (modeling) perspectives. Mechanistic reasoning focuses on the system’s particular underlying structures–consisting of components, their properties and their relations–that jointly generate observed phenomena. Functional reasoning takes a more abstract perspective of capacities, affordances, and processes across these particular instantiations.

Using examples from meteorology, automotive engineering, mammalian physiology, and psychopathology, I will illustrate the shortcomings of cause-effect reasoning and the advantages of mechanistic/functional reasoning for theory development. In conclusion, I will speculate on the methodological implications and requirements for the next steps toward comprehensive, explanatory theories of (human) psychology.

AI-assisted theory construction with theoraizer: Do LLMs and humans agree on causal relationships?

Friday, 18th July - 11:45: Theory Construction Methodology (GH: Meridian 1-2) - Oral

Ms. Meike Waaijers (University of Amsterdam)

The Causal Loop Diagram (CLD) method is a theory construction technique in which a group of domain experts identifies causal relationships between variables. However, as the number of variables increases, the process becomes more labor intensive, limiting its scalability. Large Language Models (LLMs), with their advanced text processing capabilities and extensive knowledge bases, offer a way to streamline this process.

We have developed theoraizer, an R package designed to assist researchers in the early stages of CLD development by integrating LLMs as a digital extension of the expert group. Researchers can use theoraizer to define a list of potential variables and query an LLM to provide candidate causal relationships between them. Rather than replacing expert judgment, this approach aims to increase efficiency by generating an initial set of possible relationships, allowing researchers to focus on evaluation and refinement. By easing the workload of constructing a candidate CLD, theoraizer supports a more efficient theory building process.

To assess the validity of LLM-generated causal judgments, we conducted a study comparing human and LLM causal judgments. We curated a dataset of 36 variable pairs with unidirectional, bidirectional, and non-causal relationships that varied in complexity (easy vs. hard) and sign (positive vs. negative). University researchers judged the presence, direction, and sign of these causal relationships, and we compared their judgments with each other and those generated by an LLM.

In this talk, I will present our findings and discuss how LLMs can support causal inference and contribute to theory development in scientific research.

Graduate student Datathon competition

Friday, 18th July - 11:45: Graduate Student Datathon Competition (GH: Meridian 3-4) - Symposia

Prof. Silvia Cagnone (University of Bologna)

In this session, the three finalist teams of the IMPS 2025 Student Datathon will present their work. The Datathon is a collaborative competition designed to engage participants in the analysis of an innovative dataset using state-of-the-art psychometric methods

An homage to John R. Nesselroade: The Poet Laureate of quantitative psychology

Friday, 18th July - 11:45: Symposium: Remembering John Nesselroade (MAC: Johnson) - Symposium Overview

Dr. Hudson Golino (University of Virginia), Dr. Sy-Miin Chow (Pennsylvania State University)

“In the future, if life should give you lemons, make lemonade. Then, find someone to whom life has given Jack Daniels- blend your assets and have a party!”. In “An Homage to John R. Nesselroade: The Poet Laureate of Quantitative Psychology,” we celebrate the wit, wisdom, and enduring influence of one of our field’s most beloved figures. Through a series of brief, personal reflections, former students, colleagues, and collaborators will share the creative insights they’ve gleaned from John’s pioneering work in dynamic measurement, multivariate modeling, and idiographic methods. Expect lighthearted “poem-jokes,” surprise guest appearances, and the signature blend of intellectual rigor and congeniality that John has always championed. By weaving anecdote with academic perspective, this symposium will illuminate how John R. Nesselroade’s mentorship and scholarship have shaped careers, and inspired us all to find joy in the quantitative quest.

Factor score indeterminacy of sum score and common factor score

Friday, 18th July - 11:45: Factor Analysis (MAC: Thomas Swain) - Oral

Mr. Hoang Nguyen (University of Minnesota)

Social scientists routinely develop psychological instruments to measure individual differences using factor analysis without addressing the factor score indeterminacy (FSI) issue. The FSI, quantified as the correlation between latent traits and their estimates, is highly relevant for test development as it can affect the measurement and interpretation of latent traits. To better understand in what data-model conditions and to what extent FSI can affect trait measurement, I conducted a comprehensive simulation study that examined the FSI when different factor score estimation methods were used. These methods included the unit-weight sum score (UWSS), Thurstone's, Bartlett's, and ten Berge's scores. In total, I simulated 25,920 data sets by varying (a) factor loading size, (b) number of factor indicators, (c) factor cross-loadings, (d) factor correlation size, (e) model approximation error, (f) sample size, and (g) item response distribution. My results showed that when trait measurements included many items with high factor loadings, the effect of FSI was minimized. In these models, UWSS and other factor score estimators could recover latent traits equally well. However, in factor models with few, small-loading items, FSI became larger and Thurstone's score typically provided the best estimation of latent traits.

WORD COUNT: 193

Keywords: Factor score indeterminacy, sum score, factor score, Monte Carlo

Integration of latent space and confirmatory factor analysis to explain unexplained person-item interactions

Friday, 18th July - 12:00: Factor Analysis (MAC: Thomas Swain) - Oral

Dr. Inhan Kang (Yonsei university), Prof. Minjeong Jeon (School of Education & Information Studies, University of California Los Angeles)

As with many other latent variable models, the confirmatory factor analysis (CFA) model is built upon the conditional independence (CI) assumption, which states that latent variables and item parameters can fully explain covariations between item responses. However, growing evidence in psychological and educational measurement research challenges this assumption, raising concerns regarding conditional dependence (CD). As the main model parameters correspond to the main person and item effects, CD implies the presence of unexplained person-item interactions. To leverage this information from non-binary item responses, we propose integrating a latent space model with CFA. The resulting model assumes that persons and items have coordinates on a shared metric space called an interaction map, where distances between persons and items reflect CD and their interactions. With this approach, the model quantifies and visualizes person-item interactions, leading to further practical analyses of CD. We present a series of simulation studies to examine various statistical properties of the proposed model. We also provide empirical examples to demonstrate the utilities and advantages of the proposed model, such as (1) deriving personalized diagnoses and evaluations for respondents, (2) quantifying individual differences in perceived item properties, and (3) facilitating investigations of CD with external variables and method effects.

A GLLAMM approach for measuring child-teacher interaction quality

Friday, 18th July - 12:15: Factor Analysis (MAC: Thomas Swain) - Oral

Dr. JoonHo Lee (The University of Alabama)

We develop a Generalized Linear Latent and Mixed Models (GLLAMMs) approach for measuring classroom process quality in early childhood education settings, using data from the 2018 Early Head Start Family and Child Experiences Study (Baby FACES). Typical two-level confirmatory factor analysis can handle hierarchical data when all responses are continuous and uniformly assessed, but it struggles when items have mixed distributions (e.g., continuous or binary) and vary by subgroup (e.g., infant vs. toddler classrooms). GLLAMMs provide a single, integrated model that accommodates these complexities without separate partial analyses.

First, GLLAMMs seamlessly include both binary and continuous items for three latent domains—support for social-emotional, cognitive, and language development—in one framework. Second, they naturally handle missing-by-design assessments by letting each classroom's subset of items contribute only what is observed. Third, by specifying random effects at the classroom and center levels, GLLAMMs decompose each factor's variance across organizational tiers, producing intraclass correlation–type measures that reveal whether variability in measured quality stems more from differences among teachers within a center or across centers. We demonstrate how the Laplace approximation used in GLLAMMs can lead to near-zero boundary estimates of factor variances and present a fully Bayesian solution that addresses such estimation challenges. Together, these features make GLLAMMs a powerful alternative to standard two-level CFA for large-scale early childhood data, yielding richer, more robust insight into child–teacher interaction domains.

A new representation of factor score and its theoretical properties

Friday, 18th July - 12:30: Factor Analysis (MAC: Thomas Swain) - Oral

Dr. Naoto Yamashita (Kansai University), Dr. Shin-ichi Mayekawa (National Center for University Entrance Examinations)

Factor score indeterminacy is known as a fundamental property in factor analysis model, indicating that factor scores are not uniquely determined under the same structural parameters. This study reformulates factor scores as solutions to simultaneous equations derived from the factor analysis model. The research proposes an analytical expression that explicitly separates determinate and indeterminate parts of factor scores. The derived factor scores are shown to be equivalent to regression-based scores and those obtained through matrix decomposition factor analysis. These findings offer new insights into factor score estimation and specification. Further, approaches for factor score specification is also discussed.

Development and validation of a Renyuan measurement scale: Exploring social self-perception in Chinese cultural contexts

Friday, 18th July - 12:45: Factor Analysis (MAC: Thomas Swain) - Oral

Ms. Ching Yi Chiang (National Chengchi University)

Popularity refers to the extent to which an individual is valued or liked within a group and has long been a significant research topic in social and developmental psychology. However, studies on popularity often rely on peer ratings as the primary measurement method, with some also examining the discrepancy between self- and peer ratings as an indicator of social self-perception. *Renyuan* is a concept in Chinese cultural contexts that parallels popularity yet differs in its emphasis. Unlike the individualistic notion of popularity, which highlights an individual's central position, *renyuan* focuses on social integration. Based on this perspective, this study aims to develop a measurement tool for *renyuan* and examine the self–peer rating discrepancy as an indicator of social self-perception in Chinese culture.

The current study collected approximately 200 adjectives describing *renyuan* from literature and open-ended questionnaires, using a sample of 230 college students and 131 non-student adults. Exploratory factor analysis identified four primary factors: interpersonal relationships, intrapersonal relationships, group contribution, and personal characteristics. A second-order exploratory factor analysis confirmed a common underlying factor explaining these four dimensions. Content analysis of open-ended questionnaire responses suggested that individuals described as having high *renyuan* are perceived to be sensitive to social acceptance and rejection cues in interpersonal interactions, indicating a potential association with social self-perception. Accordingly, this study developed a self-report scale incorporating these adjectives—an approach rarely used in previous Western studies on popularity—to examine its reliability, validity, and relationship with social desirability.

Keywords: Renyuan, Popularity, Social Self-Perception

Generalized structured component analysis accommodating convex components: A knowledge-based multivariate method with interpretable composite indexes

Friday, 18th July - 11:45: Multivariate Analysis (GH: Think 4) - Oral

Dr. Gyeongcheol Cho (The Ohio State University), Dr. Heungsun Hwang (McGill University)

Generalized structured component analysis (GSCA) is a multivariate method for examining theory-driven relationships between variables including components. GSCA can provide the deterministic component score for each individual once model parameters are estimated. As the traditional GSCA always standardizes all indicators and components, however, it could not utilize information on the indicators' scale in parameter estimation. Consequently, its component scores could just show the relative standing of each individual for a component, rather than the individual's absolute standing in terms of the original indicators' measurement scales. In the paper, we propose a new version of GSCA, named *convex GSCA*, which can produce a new type of unstandardized components, termed convex components, which can be intuitively interpreted in terms of the original indicators' scales. We investigate the empirical performance of the proposed method through the analyses of simulated and real data.

A multivariate generalization of the Glass delta effect size

Friday, 18th July - 12:00: Multivariate Analysis (GH: Think 4) - Oral

Prof. Jay Verkuilen (City University of New York), Dr. Christopher Siefert (Sandia National Laboratory)

Effect size measures are important for interpreting data analysis as well as in contexts such as meta-analysis. While there are many effect size measures for univariate data, only correlation family effect sizes and Mahalanobis distance exist for multivariate data. Correlation family effect sizes are primarily useful for power analysis while M-distances require homoscedasticity across groups. In a univariate analysis, Glass's Delta uses a reference group, e.g., the control, standard deviation to standardize comparison groups, e.g., treatment groups. We explore the linear algebra of standardizations of this type, which are related to matrix whitening of one group. We consider the notion of relative principal components to perform data reduction to allow visualization. This has a close connection to Kullback-Leibler divergence and the generalized eigenvalue problem.

Comparison of missing data techniques in generalized structured component analysis

Friday, 18th July - 12:15: Multivariate Analysis (GH: Think 4) - Oral

Ms. Luqi He (McGill University), Dr. Gyeongcheol Cho (Department of Psychology, The Ohio State University), Dr. Heungsun Hwang (McGill University)

Generalized structured component analysis (GSCA) is a comprehensive method for component-based structural equation modeling (SEM), where constructs that are not directly measurable are represented as components or weighted composites of indicators. In practice, missing data are often unavoidable due to various reasons such as nonresponse and attrition. In SEM, missing data can introduce bias and reduce efficiency, affecting the accuracy and stability of parameter estimates. GSCA currently offers three approaches for handling missing data: listwise deletion, mean substitution, and least-squares (LS) imputation. The first two are simple, model-free techniques applied before conducting GSCA, whereas LS imputation is a model-based approach that iteratively estimates missing values using model parameter estimates. LS imputation has been implemented into free software programs. However, a systematic evaluation of its performance compared to conventional missing data techniques in GSCA has not been conducted. To address this gap, we conducted a simulation study to examine the parameter recovery capabilities of LS imputation and traditional missing data techniques under various experimental factors, including sample size, missing data mechanisms, and the proportion of missing data. Our study aims to provide insights into the effectiveness of different missing data approaches under controlled conditions and to offer practical recommendations for selecting appropriate methods in GSCA.

Evaluating assumption violations in latent APIM: Implications for effect estimation

Friday, 18th July - 12:30: Multivariate Analysis (GH: Think 4) - Oral

Ms. Shiyao Wang (KU Leuven, University of Leuven), Prof. Eva Ceulemans (KU Leuven, University of Leuven)

The Actor-Partner Interdependence Model (APIM; Cook & Kenny, 2005) is commonly used to study interpersonal relationships within dyads. The APIM acknowledges the effect of individual's trait scores on their own outcome (i.e., actor effect) and their partner's outcome (i.e., partner effect), as well as the correlation between both outcomes. Although the APIM is primarily developed for cases where both the predictor and the outcome are observed variables, in interpersonal relationship research, variables of interest are often latent and measured by multiple indicators jointly.

To accommodate latent variables within the APIM, researchers have proposed extending the APIM with a latent model (e.g., Kim & Kim, 2022), where the same common factor model is built for each partner. However, the assumptions underlying this practice are often violated in interpersonal relationship research. For example, when investigating mother-child dyads, the temperament of the mother and child might be measured using different questionnaires, leading to different latent models for their variables.

In the current research, we investigate how these violations influence the implementation of the latent model, the estimation of actor and partner effects, and the interpretation of these effects.

Evaluating statistical power in generalized structured component analysis

Friday, 18th July - 12:45: Multivariate Analysis (GH: Think 4) - Oral

Ms. Zhiyuan Shen (McGill University, Montreal, Canada), Dr. Gyeongcheol Cho (Department of Psychology, The Ohio State University), Dr. Heungsun Hwang (McGill University)

Generalized structured component analysis (GSCA) is a well-established approach to component-based structural equation modeling, where constructs are statistically represented as weighted composites of observed indicators. Statistical power—the probability of correctly rejecting the null hypothesis when a true effect exists—is critical for designing efficient studies that reliably identify meaningful relationships among variables. Despite its importance, statistical power in GSCA has not been systematically investigated. To address this gap, we conducted a Monte Carlo simulation study, manipulating various experimental factors such as sample size, target effect size, number of indicators per component, indicator reliability, and correlation among exogenous components. Our study aims to provide practical guidelines for conducting power analysis in GSCA, enabling researchers to optimize study designs while maintaining sufficient power to detect meaningful effects.

Examining heterogeneity in causal mediation effects

Friday, 18th July - 11:45: Causal Inference II (GH: Think 5) - Oral

Prof. Hanna Kim (University of California, Santa Cruz), Prof. Jee-Seon Kim (University of Wisconsin - Madison)

Recent advances in causal mediation analysis have clarified the conditions necessary for identifying direct and indirect effects in observational studies. However, these effects are often assumed to be uniform across subpopulations, overlooking potential variation in how mediators transmit exposure effects to outcomes. Addressing this gap, this study extends causal mediation analysis by examining heterogeneity in natural direct and indirect effects across both observed group membership and potentially unobserved (latent) subgroups.

Heterogeneity may arise from a well-defined characteristic or an observed variable, but it may also reflect the existence of participant subgroups that are not directly observable, despite substantial variability in effects or distinctive patterns. To capture this complexity, we incorporate latent moderators into the causal mediation framework, providing a methodological tool to account for variation in mediational processes and analyze distinct causal pathways. When heterogeneity is driven by an observed discrete moderator, the proposed latent variable framework simplifies to groupwise heterogeneity in mediation effects.

By integrating causal mediation analysis with latent variable modeling, this study provides a flexible framework for examining heterogeneity in mediating processes. We conclude by discussing how this approach advances the study of developmental processes in longitudinal intervention research and its broader implications for causal inference in contexts where treatment effects and mediational pathways vary across subpopulations.

A taxonomy of heterogeneity in causal effects

Friday, 18th July - 12:00: Causal Inference II (GH: Think 5) - Oral

Prof. Jee-Seon Kim (University of Wisconsin - Madison), Mr. Graham Buhrman (University of Wisconsin - Madison), Prof. Xiangyi Liao (University of British Columbia), Prof. Wen Wei Loh (Maastricht University)

Addressing heterogeneity in treatment effects is essential for developing interventions that accommodate various needs and strategies across subpopulations. This paper introduces a taxonomy of treatment effect heterogeneity, classifying effects based on key dimensions such as observability and the type of heterogeneity. This classification provides a comprehensive framework for capturing the diverse characteristics of treatment effects. Furthermore, this study positions the taxonomy as the centerpiece of an analytical workflow for causal research, guiding analysis from initial theory and data exploration to estimation, sensitivity analysis, and inference. This iterative process establishes a feedback loop that refines theories and generates new insights. Grounded in causal inference theory, the framework integrates latent variable modeling with machine learning-based causal algorithms to enhance the identification, estimation, and interpretation of treatment effects. By capturing both observed and unobserved heterogeneity, as well as the various forms of its manifestation, the taxonomy and analytical workflow support nuanced and contextually informed interventions. The workflow is relevant to experimental, quasi-experimental, and observational studies, ensuring broad applicability across research designs.

Item-level heterogeneous treatment effects in instrumental variables regression: Fixed- and random-item approaches

Friday, 18th July - 12:15: Causal Inference II (GH: Think 5) - Oral

Dr. Sanford Student (University of Delaware), Mr. Joshua Gilbert (Harvard University), Mr. Jesse Eze (University of Delaware), Dr. Benjamin Domingue (Stanford University)

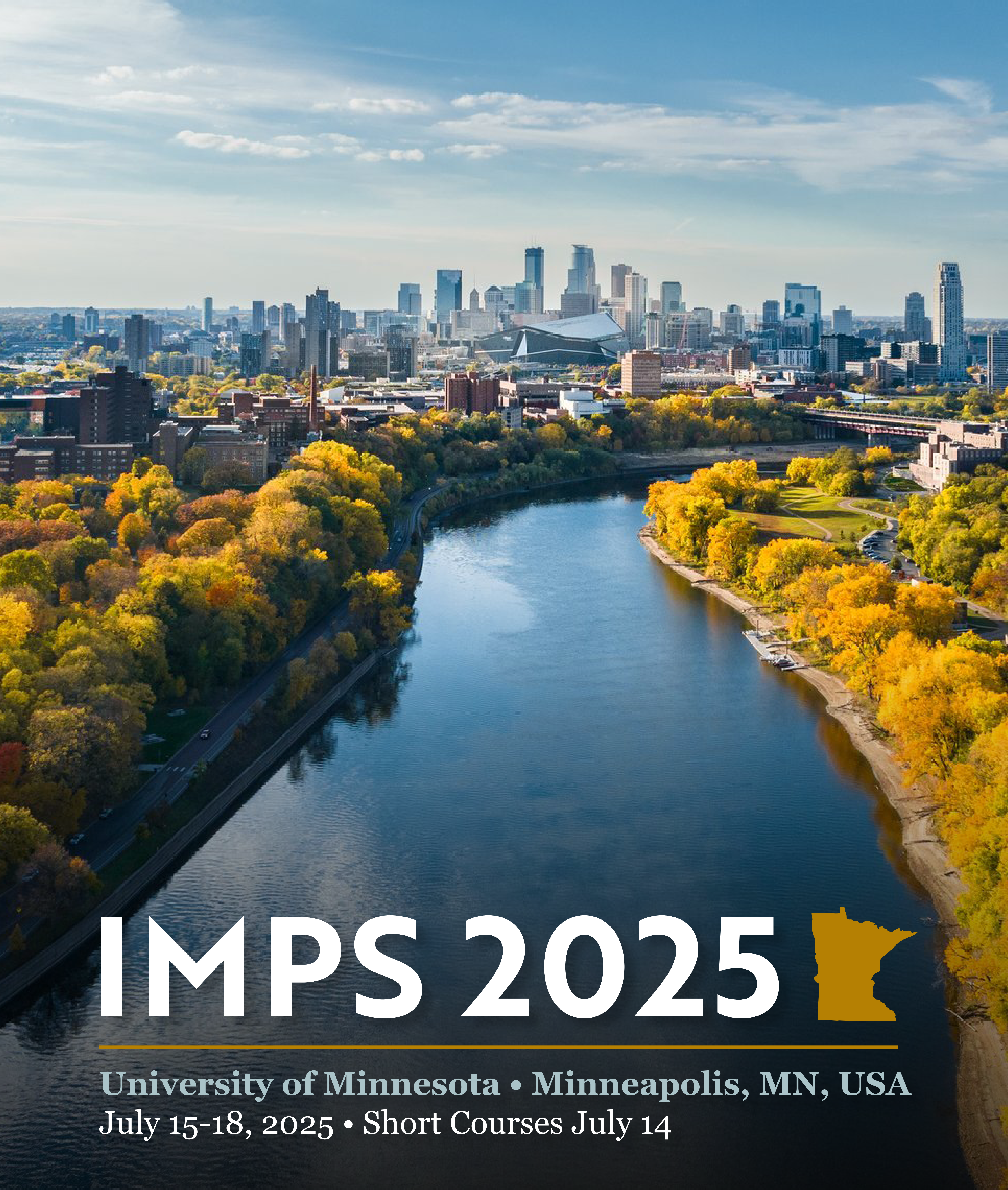
Recent research has highlighted that modeling of heterogeneity of treatment effects across the individual items on an outcome measure (item level heterogeneous treatment effects; IL-HTE, e.g. Gilbert et al., 2023, Ahmed et al., 2024) can yield additional insights into interventions' effects beyond what is available from comparisons of average scores. So far, this work has generally focused on randomized controlled trials. However, in social science research, randomization often does not hold perfectly, or is not possible. This can lead to treatment endogeneity – residual correlation between the predictor and the outcome. Instrumental variables regression (IVR) is a popular method for correcting the resulting bias by introducing an exogenous variable that drives change in the predictor but is uncorrelated with the outcome. Though typically estimated via two-stage least squares, IVR can also be represented and estimated as a latent variable model (Maydeu-Olivares et al., 2019). Extending IL-HTE analysis to IVR is not entirely straightforward for reasons we outline in this paper. We develop models for estimating IL-HTE with instrumental variables in the general latent variable modeling framework, supporting more valid causal inferences about impacts of nonrandom treatments/dosages on test items. We present methods for modeling items as fixed or random, and discuss the affordances and tradeoffs of the two approaches. We illustrate the use of these models with a real data example analyzing potential IL-HTE in a study of the impact of time spent on homework on performance on a large-scale international assessment of high school mathematics.

Estimation of individual treatment effect with transfer learning

Friday, 18th July - 12:30: Causal Inference II (GH: Think 5) - Oral

Ms. Şeyda Aydin (Eberhard Karls Universität Tübingen), Prof. Holger Brandt (Eberhard Karls Universität Tübingen)

Generalizing causal knowledge across diverse environments is a major challenge in scientific research, particularly when insights from large scale datasets must be applied to smaller or distinct contexts, where external validity becomes critical. Models – particularly from machine learning - for estimating individual treatment effects (ITE) typically require large sample sizes, limiting their applicability in domains such as social or behavioral sciences that rely on smaller datasets. To address this issue, we demonstrate how transfer learning using Treatment Agnostic Representation Networks (TARNet; Shalit et al., 2017) can substantially improve ITE estimation by leveraging knowledge from extensive source datasets and adapting it to new settings. We consider four scenarios illustrating various degrees of overlap between source and target data, including a partial transfer case where only control group data is available in the source dataset (e.g., from large scale assessments with observational data). Additionally, we incorporate inverse probability weighting (IPW) to account for confounding and reduce bias which is an important step for smaller or datasets with non-randomized interventions. We employ the integral probability metric to assess potential distributional discrepancies between source and target datasets to determine whether transfer learning is appropriate. Empirical results across these scenarios show that combining IPW with TARNet based transfer learning significantly enhances ITE estimates, providing a robust framework for causal inference in behavioral studies. Beyond improving statistical power, this approach broadens the applicability of transfer learning, ensuring that causal knowledge derived from large scale datasets can be responsibly generalized to more limited or heterogeneous contexts.



IMPS 2025

University of Minnesota • Minneapolis, MN, USA
July 15-18, 2025 • Short Courses July 14

ABSTRACT BOOK: POSTERS

Comparing traditional and AI-based IRT estimation: An empirical study

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Olukayode Apata (Texas A&M University), Mr. Segun Ajose (University of Kentucky)

The increasing role of Artificial intelligence (AI) is reshaping psychometric modeling, yet its accuracy in item response theory (IRT) parameter estimation remains uncertain. This study compares traditional maximum likelihood estimation (MLE) using Stata 16.0 with an AI-based (ChatGPT 4o) estimation in Python to analyze the *Dataset for 21st-Century Character Exploratory Factor Analysis in Vocational High School Students* (Effiyanti, 2024). The dataset from vocational high school students (N = 340) measures greeting behaviors, conflict resolution, integrity, academic honesty, adaptability, and information literacy using 18 Likert-scale items. We evaluated the one-parameter logistic model (1PL), two-parameter logistic model (2PL), and three-parameter logistic model (3PL) and estimated their item difficulty, discrimination, and guessing parameters.

Our results indicate that difficulty estimates in the 1PL model were similar, with Stata estimating Q1 = -1.49 and AI estimating Q1 = -1.14. However, we noticed a slight deviation in the 2PL model estimated discrimination parameters (Q1: Stata = 0.95; AI = 1.21). In the 3PL model, we observed a notable variation in the guessing parameter (Q1: Stata = 0.043, AI = 0.323). The Stata analysis produced stable parameter estimates, while the AI analysis exhibited inconsistencies, particularly in discrimination and guessing parameters.

These findings highlight the strengths and limitations of AI-driven IRT estimation. While AI models can efficiently estimate item parameters, the system should be improved to ensure alignment with traditional MLE-based methods. We encourage future research to look at ways to refine AI estimation methods especially for complex IRT models and large-scale educational assessments.

Network psychometrics: Node redundancy

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Abraham Iniguez (The University of California, Davis), Dr. Mijke Rhemtulla (The University of California, Davis), Dr. Jonathan Park (The University of California, Davis)

Over the past decade, network analysis has increasingly become a popular tool for modeling psychological systems. Estimating the correct population network is critical for making inferences to real-world data. However, choosing which variables or symptoms to include in a network is often challenging because the set of variables should reflect the elements of the system that the network is meant to represent. When choosing variables, it is recommended to avoid including redundant nodes. Node redundancy has previously been defined as two strongly correlated variables that share the same weighted associations with other variables (Christensen et al., 2023) with other possible definitions as well. Given the limited research on node redundancy, how it affects estimated networks remains unclear. Plausible effects could include biasing centrality estimates, altering a network's density, or affecting its structure. Therefore, we aim to clearly define the problem of node redundancy and investigate the extent to which this problem affects network model interpretations.

To address this aim, we simulated several forms of redundancy from network models of varying size. The effects of redundant nodes on the network metrics were evaluated with and without these nodes. Additionally, we assessed various forms of addressing node redundancy to see whether networks could be returned to their original topology. Our findings provide evidence that redundant models change the structure of estimated networks, leading to different interpretations than if redundancy were not present.

Comparing the Rasch equating method and the d scoring equating method

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Ahmed Haddadi (Education and Training Evaluation Commission), Dr. Mohammed Alqabbaa (Education and Training Evaluation Commission)

This study aimed to compare the test equating under two measurement frameworks: Item response theory (IRT) and the classical d scoring method (DSM-C). Specifically, the study compared the equating under the Rasch model and the DSM-C. Test forms were equated using the non-equivalent groups with anchor test (NEAT) and mean/sigma approach to compute the rescaling coefficients. Participants were 15,707 3rd grade students who took six equivalent forms of a mathematics test. The results showed that both approaches are highly comparable in terms of the rescaling coefficients, rescaled item parameters, and equated person scores. The study results recommend adopting the equating under the DSM-C given its practicality and simplicity and call for more studies examining the DSM equating under various conditions.

Assessing cross-group and cross-time measurement invariance

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Alexander Miles (University of Southern California), Ms. Meltem Ozcan (University of Southern California), Dr. Mark Lai (University of Southern California)

To ensure that valid and meaningful comparisons can be made, a construct must have equivalent measurement qualities both between and within groups. Measurement invariance testing assesses whether the psychometric properties of a measurement instruction are equivalent across groups or time points, but applied research examining both cross-group and cross-time invariance is relatively scarce. We address this gap by rigorously testing cross-group and cross-time measurement invariance, employing large datasets and robust statistical methods, including 10-fold cross-validation. We examined measurement invariance of a self-efficacy in math scale, across two large datasets: the Educational Longitudinal Survey (ELS; N = 16,197; 2010, 2012, 9th to 11th grade) and the High School Longitudinal Survey (HSLS; N = 23,503; 2002, 2004; 10th to 12th grade), with five items in ELS and four items in HSLS. We tested for both between group invariance (comparing psychometric properties between the surveys) and within group (comparing psychometric properties over time) using the lavaan software package, and ML-Robust estimation. We tested invariance at the weak, strong, and strict levels. Due to the large sample size, we used a combination of changes in CFI, RMSEA, and SRMR to determine if measures were equivalent. The data supported strong factorial invariance across time for both groups; however, it was found that strong invariance was violated between groups. As such, a partial invariance model was constructed using traditional measurement invariance techniques. Additionally, the large sample size enabled us to utilize 10-fold cross validation, broadly confirming our results.

Impacts of Q-matrix design on learning DCM linear attribute hierarchies

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Olasunkanmi Kehinde (Elizabeth City State University), Dr. Alfonso Martinez (Fordham University), Mr. Mubarak Mojoyinola (University of Iowa)

This study examines how different Q-matrix designs, for linearly structured attributes, impact the accuracy of item parameters and attribute profile estimations in diagnostic classification models (DCMs). While DCMs effectively assess student competencies and identify skill gaps, modeling learning progressions within them remains challenging due to the hierarchical nature of learning. In particular, this study focuses on a linear attribute hierarchal structure where attribute A is a prerequisite for B, and attribute B is a prerequisite for attribute C. Using the G-DINA model, we investigate the performance in recovering item parameters and attribute profiles across three Q-matrix representations that encode information about the linear hierarchy: linear independent (LI), linear adjacent (LA), and linear reachable (LR). Item response data was generated using the GDINA package in R, with a fully-crossed simulation design incorporating three sample sizes (200, 500, 1000), two item quality levels (low, high), and two marginal attribute base rates (e.g., equal base-rates or declining base-rates to reflect the hypothesis that attributes further in the hierarchy are more difficult to master), totaling 18 conditions replicated 200 times each. Preliminary findings from a small-scale pilot study suggest that the LI Q-matrix representation yields better classification rates than the LA and LR Q-matrices. The full simulation study is currently in progress and is expected to be completed by April 2025. Findings from this study will provide valuable insights into how linear hierarchy Q-matrix designs impacts inferences made from DCMs and may provide insights into optimal ways of representing linear learning progressions.

Evaluating sample size requirements for context questionnaire scales in international large-scale assessments

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Andrés Christiansen (IEA Hamburg), Dr. Ana María Mejía-Rodríguez (IEA Hamburg), Dr. Rolf Strietholt (IEA Hamburg)

Besides assessing student achievement, International Large-Scale Assessments collect valuable information through context questionnaires. These questionnaires are evaluated during a field trial (FT), intended to test the functioning of context scales in each country and provide empirical evidence to revise material. However, the usefulness of the FT depends on the quality of evidence, and a key issue is the sample size per country.

While previous studies have explored minimum sample size for scale construction, such studies are theoretical or use simulated data. Whether these findings apply to actual data remains unclear. To investigate, we used existing student and teacher questionnaire data from TIMSS 2019 and ICILS 2018, with data for up to 65 educational systems and 70 scales.

We analyzed how different sample size minimums influence the distribution and precision of indicators from confirmatory factor analysis. We focused on how goodness-of-fit indices (CFI, TLI, RMSEA, SRMR) change between sample sizes and how this may impact scale adequacy decisions based on standard thresholds.

Our analysis confirms bias decreases as sample size increases and supports the N≥200 guideline, where bias remains minimal. However, samples of N≥100 may still provide reliable evidence, with slightly higher bias compared to 200 observations. Therefore, while N≥200 is preferred, sample sizes within 100-200 may be a practical alternative in specific contexts.

Early assurance admission program outcomes for matriculation years 2019 - 2022

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Anita Wilson (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Aaron Douglas (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Alisa LoSasso (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Katherine Berg (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Kathleen Day (Sidney Kimmel Medical College at Thomas Jefferson University), Ms. Bonnie Emilius (Sidney Kimmel Medical College at Thomas Jefferson University), Ms. Michelle Calderoni (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Steven Herrine (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Wayne Bond Lau (Sidney Kimmel Medical College at Thomas Jefferson University), Dr. Charles Pohl (Sidney Kimmel Medical College at Thomas Jefferson University)

The goal of the early assurance admission programs is to find and recruit students with an interest in attending medical school and becoming physicians. Sidney Kimmel Medical College (SKMC) grants three types of early admission to students who meet the eligibility criteria from secondary and post-secondary institutions. This paper summarizes and compares the performance of early admission students to the performance of a matched sample of traditional admission students.

The data for this study comes from SKMC students of the graduating classes of 2023 and 2024 and expected classes of 2025 and 2026. Given the admission requirements for early assurance students, students are matched based on the common data among all participants: cohort, dual degree status, and overall pre-clerkship performance. Welch's ANOVA is used to compare the average student performance of early admission students and traditional students. Fisher's Exact test for pairwise comparisons was used to make two sets of pairwise comparisons: 1) the USMLE Step 1 and Step 2 CK passing rates between the groups and 2) the proportion of early assurance students who graduated within 4 years to all SKMC students who graduated within 4 years.

Results from Welch's ANOVA indicate no statistically significant difference in the average overall clerkship performance between groups of students. When compared to all SKMC students, the Fisher's exact test shows there is no difference in the proportion of conditional early assurance, 3+4 early assurance, and traditional students who graduate within 4 years.

Supporting student progression: The relationship between the LASSI, preclinical course exams, and USMLE Step 1

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Anita Wilson (Sidney Kimmel Medical College at Thomas Jefferson University), Mr. James Dyksen (Sidney Kimmel Medical College at Thomas Jefferson University)

Academic support faculty consistently seek new strategies and opportunities in which to provide students with the tools they need to be successful in their UME program. This study explores the relationship between student performance on Learning and Study Strategies Inventory (LASSI), preclinical course exams, Comprehensive Basic Science Self-Assessment (CBSSA) and the United States Medical Licensing Exam (USMLE) Step 1.

The participants in this study represent students from a large urban medical college for the entering classes of 2025 and 2026 who took the (LASSI), preclinical course exams, Comprehensive Basic Science Self-Assessment (CBSSA) and the United States Medical Licensing Exam (USMLE) Step 1.

A paired t-test was used to identify any statistical differences in student performance across the LASSI time points. A multinomial logistic regression (MLR) model was generated to predict student performance on the USMLE Step 1 exam.

The results from the paired t-test show that student percentile rank was significantly higher at time point 1 than at time point 2 on Attitude, Concentration, and Motivation for the 2025 cohort. The student percentile rank was significantly higher at time point 1 than time point 2 for Attitude, Motivation, and Using Academic Resources for the 2026 cohort. There was a significant decrease from time point 1 to time point 2 on the Information Processing subscale. For the MLR model, the Test Taking Strategies subscale (time 1 and time 2) and Medicine 201 block exams were the only significant predictor of student success on Step 1.

Comparing logit, probit, and glogit link functions: A multinomial logistic regression model with unbalanced data

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Anita Wilson (Sidney Kimmel Medical College at Thomas Jefferson University)

Undergraduate medical education students desire to perform well on the first licensing exam to prevent a reduction of choice in residency specialty. Unprepared students have the option to delay taking the exam. This delay creates four time/performance paths for students: 1) on-time/pass (P) 2) delay/pass (DP), 3) delay/fail (DF), or 4) on-time/fail (F). Most students are in the P class, yet there are a small number of students who are in the DP, DF, and F classes thus leading to unbalanced data when implementing predictive modeling. The purpose of this study is to compare the prediction of the logit, probit, and glogit link functions for the multinomial logistic regression model (MLRM) when the data is significantly unbalanced.

Using data from two cohorts of students, a MLRM with 87% accuracy is fitted using the original data. The accuracy was improved to 91% with oversampling the minority classes. Data are simulated for 1,000 samples each with 270 observations. The simulated data sets mimic the target cohort size and time/performance class distributions from the real data.

The results suggest the root mean square error of prediction (RMSE) and bias was below 2.6 for all glogit models and above 2.75 for all logit and probit models. The standard error of prediction (SEP) was above 0.69 for the glogit models and below 0.65 for the logit and probit models. The performance of the logit and probit models were similar, yet the logit and probit models underperformed when compared to the glogit models.

Avoiding conflated random slopes in multilevel analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Bladimir Padilla (University of Iowa), Dr. Lesa Hoffman (University of Iowa)

The literature on multilevel analysis commonly warns about the potential for level 1 predictors to have conflated fixed slopes: an uninterpretable weighted average of a predictor's fixed effect at the within level (e.g., level 1 individuals) and the between level (e.g., level 2 clusters). Said differently, a conflated fixed slope assumes no between-level contextual fixed slope (i.e., no incremental contribution of the cluster beyond that of the individual). However, less attention has been paid to the problem of conflated *random* slopes, which are a different kind of uninterpretable blend—of slope heterogeneity (i.e., between-level variation around the within-level fixed slope) and intercept heteroscedasticity (i.e., between-level variation of the cluster intercepts across the cluster means of the within-level predictor). Conflated random slopes can bias the estimation, interpretation, and inferences of the between-level variances in the model, as well as the standard errors and Type I error rates of their corresponding fixed effects. The purpose of this simulation study is to examine the performance of three multilevel models that avoid random slope conflation by manipulating several relevant factors: sample size at level 1 and level 2, intraclass correlation of the level 1 predictor, and reliability and magnitude of the random slope variances. To help guide future best practices in models using random slopes, results will be presented for model convergence, as well as bias, power, Type 1 error rate, empirical standard error, and coverage of the relevant parameters of each model.

Time series and multilevel modeling for longitudinal item response theory data

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Caio Azevedo (Department of Statistics, Universidade Estadual de Campinas), Dr. Dalton Andrade (Department of Informatic and Statistics, Universidade Federal de Santa Catarina), Dr. Jean-Paul Fox (Department of Research Methodology, Measurement and Data Analysis, University of Twente)

In this work we consider some stationary and nonstationary time series and multilevel models to represent longitudinal Item Response Theory (IRT) data. We developed a Bayesian inference framework, which includes parameter estimation, model fit assessment and model comparison tools, through MCMC algorithms. Simulation studies are conducted in order to measure the parameter recovery. All computational implementations are made through the Stan program, using the rstan package, from the R program. A real data analysis, concerning the Amsterdam Growth and Health Longitudinal Study, is properly analyzed and usefull conclusions are drawn from the obtained results.

Causal mediation in clustered data with machine learning

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Cameron McCann (The University of Texas at Austin), Dr. Xiao Liu (The University of Texas at Austin)

Mediation analyses in educational and behavioral research often involve clustered data. However, causal mediation methods addressing unmeasured cluster-level confounders remain limited. This study extends a multiply robust estimation method to assess causal mediation effects for a single mediator in clustered data. Multiply robust estimators remain consistent if at least one of multiple combinations of nuisance models is consistently estimated. Furthermore, the multiply robust method can incorporate machine learning approaches to relax modeling assumptions while allowing uncertainty quantification via asymptotic variance and confidence intervals. To control for unmeasured cluster-level confounders, we include cluster means and cluster dummies in the nuisance model estimation, comparing this approach to a specification that omits cluster dummies. Through simulations, we evaluate the proposed method's performance in estimating individual- and cluster-average effects under both continuous and binary mediators and outcomes, varying the number and size of clusters and the nonlinearity (linear or quadratic) in data generation. We also compare parametric and machine learning estimation of nuisance models, with the latter employing the Super Learner ensemble of multiple machine learning prediction models. Simulation results indicate that the proposed method performs satisfactorily in the presence of unmeasured cluster-level confounding, with the machine learning-based nuisance model estimation improving performance when quadratic relations exist. Finally, we demonstrate the method using data from the National Longitudinal Study of Adolescent to Adult Health.

Beta copula diagnostic classification models for attribute estimation using testlet-based visual analogue scaling

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Prof. Chen-Wei Liu (National Taiwan Normal University)

This study introduces beta copula diagnostic classification models (BCDCMs) for inferring and classifying latent attributes from testlet-based visual analogue scaling (VAS) data. BCDCMs incorporate copula-based dependency structures to enhance the accuracy of classification and parameter estimation. Model properties, estimation procedures, goodness-of-fit assessment, and visualization techniques are examined. An empirical analysis using the Symptom Checklist-90 dataset demonstrates the model's effectiveness in estimating latent psychiatric symptoms, with simulation studies confirming parameter stability. Findings suggest that BCDCMs provide a robust and precise alternative to traditional cutoff-score methods for VAS data, offering valuable insights for psychometric assessment.

Investigating stopping rules for variable-length multistage adaptive testing

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Chia-Ling Hsu (Hong Kong Examinations and Assessment Authority)

This study conducted a series of simulation studies to compare various stopping rules with respect to the recovery of true abilities and mean test length to terminate the test for variable-length multistage adaptive testing (MST). The stopping rules are: *maximum standard error (SE) rule* (or the *minimum information rule*), *absolute change in theta (CT) rule*, *minimum information rule*, and *joint rule*. The number of items for assembling test, the maximum number of items administered, and the distribution of the item characteristics for each stage of MST are also manipulated. The simulation results showed that a stringent precision criterion leads to higher measurement precision and lower efficiency. CT rule requires a longer average test length than SE rule to obtain equivalent precision. CT rule enhances the efficiency of SE-CT rule when SE rule is dominant. A less stringent precision criterion in the earlier stage(s) is sufficient when a strict precision criterion is utilized in the later stage(s). In sum, the simulation findings provide insights on the utility of various stopping rules for difference scenarios of variable-length MST.

Demarginalizing the intersection of DIF: Traditional vs intersectional approach

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Christopher Green (Morgan State University), Dr. Damon Bryant (Morgan State University)

Differential item functioning (DIF) is a statistical method used to detect potential biases in test items by evaluating whether individuals from different demographic groups (e.g., by gender, socioeconomic status, or race) with the same underlying ability perform differently on specific items (Albano, 2024). The purpose of this proposed study is to investigate the detection rates of differential item functioning (DIF) using an intersectional modeling approach. Specifically, we seek to compare it to the traditional DIF detection approach. Are intersectional and traditional DIF detection modeling approaches the same or different in the detection of DIF? Here we intend to compare the intersection of race and sex in a computer simulated environment. DIF detection using open source software and widely available techniques, such as R or Python and logistic regression will be demonstrated in this study. The independent variables are sample size [50,150, 250, 350 per group], number of questions [20, 40, 60], amount of DIF [small, medium, and large], ability group differences [0, .25, .50, 1 SD], and modeling approach traditional DIF detection (race and sex independently) or intersectional (race and sex in combination).

Psychometric modeling of performance and response strategies in forecasting judgements

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Clifford Hauenstein (Johns Hopkins School of Medicine), Dr. Eunbee Kim (Georgia Institute of Technology), Dr. Ricky Thomas (Georgia Institute of Technology), Dr. Michael Dougherty (University of Maryland)

Using data from a geopolitical forecasting tournament and a traditional Brier score metric, Mellers et al. (2014) concluded that forecasting ability was improved by allowing participants to work in teams and providing them with probability training. However, few constraints were imposed on response behavior; participants were given much discretion in terms of which items to respond to and when to respond during the testing window. Previous concerns have been raised about the potential for such testing procedures to produce missing responses not at random, and measurement models have been developed to address this issue by considering the relationship between latent forecasting ability and item selection decisions. We have extended these measurement models by additionally accounting for the influence of latent forecasting ability on response times. We then applied this measurement model to the geopolitical forecasting tournament data and reevaluated Mellers et al.'s conclusions. Across the first two years of the tournament, our model demonstrated substantially improved fit over other measurement models which did not take into account the influence of forecasting ability on item selection and response time. Furthermore, the adjusted forecasting ability estimates from our model either eliminated, reduced, or reversed the originally observed effects of teaming and training on forecasting ability. Lastly, we embed our measurement model into a larger structural analysis that evaluates the relationships of other stable latent traits (intelligence, open-mindedness) on forecasting ability.

Application of explanatory Rasch model: Standardizing African certificate examination

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Daniel Oyeniran (University of Alabama), Mr. Justice Dadzie (University of Alabama), Mr. Christopher Ocheni (University of Alabama), Mr. Yusuf Isah (University of Alabama)

This study examines student performance on the Basic Education Certificate Examination (BECE) mathematics test using a Many-Facet Rasch Model (MFRM) to analyze student ability, item difficulty, gender, and school type effects. A subset of responses (male: 598, female: 570; public: 584, private: 584) from the 2022 administration was analyzed. Additionally, a Dichotomous Rasch Model was employed to assess differential item functioning (DIF) across gender and school type. Results indicate that while some items favor specific student demographics, DIF is not substantial. Wright Maps reveal a normally distributed ability estimate centered around 0.5, with items more suited to average and lower-achieving students. Logit-scale calibration shows that students have a higher average location (M=0.38, SD=0.32) compared to school type, class, and items, all centered at zero logits. Fit statistics (Infit/Outfit MSE ≈1.00) confirm model validity. Separation reliability indicates poor differentiation among students (-0.06) and class groups but moderate variation for school type (0.48) and high item reliability (0.97). Group comparisons show negligible differences in ability estimates, with private schools having a slightly lower location (δ =-0.01) than public schools (δ =0.01), while male and female students share the same location (δ =0.0). Overall, results suggest minimal measurement bias, strong model fit, and appropriate scale functioning across subgroups.

Impact of item drift on student scores in pre-equated assessments

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Daniel Edi (North Carolina Department of Public Instruction), Mr. Eric Asare (Old Dominion University)

In high-stakes summative assessments, equating is crucial to ensure meaningful comparisons of student performance across time. While post-equating provides precise estimates of student ability, it occurs after test administration and is time-consuming. Pre-equating, on the other hand, allows pre-constructed scoring tables and automatic score reporting but does not account for changes in item or test function between administrations, such as item preknowledge. This simulation study assesses the impact of item drift due to item preknowledge on student scores. Using Monte Carlo simulations, student responses to a 50-item test with dichotomous response categories were generated. The sample size ranged from 10,000 to 50,000, and the percentage of items and students affected varied from 0% to 50%. Fixed item parameter calibration was performed by anchoring non-drifted items. Student scores were estimated using both original and updated score tables. Results indicated that 42% to 61% of students would be classified at different proficiency levels even after applying fixed item parameter calibration. Additionally, the robust z procedure's ability to detect drifted items was explored. These findings highlight the significant impact of item preknowledge on score classification and highlight the need for robust methods to detect and mitigate item drift in pre-equated assessments.

Modeling interaction terms among level 1 predictors in multilevel models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Erica Dorman (University of Iowa), Dr. Lesa Hoffman (University of Iowa)

Multilevel modeling is a methodological tool commonly used in social sciences to account for dependency in data and address questions at multiple levels of analysis simultaneously. While the use of centering for level 1 predictors in multilevel models to avoid conflating between- and within-level main effects is well understood, how centering should be used for interaction terms among level 1 predictors has inherent ambiguity. As highlighted by Loeys et al. (2018), different orders of operations in creating level 1 interaction terms can yield different results. The purpose of this study is to understand the consequences of different choices for modeling interaction terms between two level 1 predictors in a Monte Carlo simulation. Manipulated factors include the centering method for level 1 predictors (cluster-mean centering or grand-mean centering), the order of operations in creating the interaction terms (multiply first or center first), and the correlation between the level 2 component of the level 1 predictor and a separate level 2 predictor. The potential influence of these factors will help us to understand the impacts of misspecified interaction terms among level 1 predictors, as well as when alternative models can or cannot be made equivalent by including additional cross-level and between-level interaction terms. Results from the simulation study will be evaluated using bias, empirical standard error, and coverage to help guide future best practice.

Evaluating aberrant response patterns in reading using mixture models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Fathima Jaffari (Qiyas,ETEC,Saudi Arabia), Dr. Georgios sideridis (Boston Children's Hospital, Harvard Medical School, Boston, MA, United States,)

The purpose of the present study was to employ mixture modeling as a means to identify homogeneous subgroups that display specific aspects of aberrant responding. The convergent validity of the proposed method was tested using person-fit indicators, namely the U3 index and NCI, commonly employed in Item Response Theory (IRT) as a means to identify specific instances of aberrant responding. Participants were 21,900 3rd graders from the Kingdom of Saudi Arabia whose behavioral response patterns were examined in a reading comprehension test. Using mixture modeling results pointed to the presence of four distinct groups with class 4 being a low-achieving group termed "aberrant." This group of students had low levels of achievement, irrespective of item difficulty levels, thus, pointing to aberrance in the form of inattention and/or carelessness. The two person-fit indices corroborated with the conclusion from the latent class analysis as they flagged approximately 70% of the class 4 students using strict distributional cutoff criteria. Further evidence showed null effects due to gender but significantly more aberrant responders who took the test in the English language. It is concluded that the combination of mixture modeling with person-fit indicators may provide cross-validation evidence on the presence of aberrant responses at school.

Mediation analysis with misclassification in binary variables

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Fei Gao (University of California, Merced), Mrs. Haiyan Liu (University of California, Merced)

Traditional methods for analyzing binary data—such as logistic regression and contingency table analysis—rely on the critical assumption that the binary variables are measured without error. However, in social and behavioral research, this assumption is often violated. The observed category may differ from the underlying true category, leading to measurement error in categorical data, a phenomenon known as misclassification. The misclassification in a binary outcome variable is well known and studied. However, the misclassification in a binary predictor has not yet drawn much attention. To deal with this issue, we propose a new modeling approach that explicitely account for misclassification by introducing false positive and false negative parameters. This new model not only estimates the probability of each type of misclassification but also produces unbiased estimates of the regression coefficients. Given the widespread use of mediation analysis in psychology, sociology, and related disciplines, we extend our framework to address misclassification in simple mediation analysis. Specifically, we examine three scenarios:(1) misclassified independent variable, (2) misclassified mediator, and (3) misclassified dependent variable. The model is estimated using a Bayesian method and a simulation study is conducted to evaluate its performance. To demonstrate the usefulness of the new models, an empirical example is included.

Keywords: Mediation analysis, Binary variable, Misclassification, Probit regression

A two-stage path analysis approach to model interaction effects for categorical indicators

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Gengrui Zhang (University of Southern California), Dr. Mark Lai (University of Southern California)

Latent variable interactions are an important focus in psychological research. While traditional methods such as latent moderated structural equations (LMS) and unconstrained product indicator (UPI) have been studied extensively with continuous indicators, their performance with categorical indicators has been evaluated in fewer studies. Prior research suggests that categorical indicators can introduce estimation challenges, particularly under conditions of non-normality or low reliability.

This study introduces an extension of the two-stage path analysis framework (2S-PA-Int) to model latent interactions with categorical indicators. The proposed approach first estimates factor scores from categorical indicators based on a measurement model and then applies path analysis with measurement error corrections. Unlike continuous indicators, categorical items have varying error variance across latent trait levels. 2S-PA-Int addresses this by modeling individual-specific measurement error variance using definition variables, allowing for more precise estimation while maintaining model simplicity and computational efficiency. In addition, 2S-PA-Int Allows to use different estimation methods for measurement and structural models (e.g., EAP scores using the IRT framework).

A Monte Carlo simulation will compare 2S-PA-Int, LMS, UPI, and SAM across different conditions, including sample size, number of response categories, indicator reliability, and predictor correlations. These methods will also be applied to an empirical dataset from the Panel Study of Income Dynamics (PSID) to evaluate their performance in practice.

By systematically assessing these approaches, this study will provide insights into their strengths and limitations when applied to categorical indicators. The findings will help researchers make informed decisions about modeling latent interactions in psychological research.

Do large language models replicate human psychometric bias? Investigating DIF in simulated CES-D responses

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Gengrui Zhang (University of Southern California)

Psychometric scales are fundamental tools in psychological and clinical research, yet they are often susceptible to measurement bias, particularly across demographic groups. Large language models (LLMs) offer a novel approach to addressing these challenges by simulating diverse persona responses, potentially aiding in survey validation and scale refinement. This study investigates whether LLM-generated responses to the Center for Epidemiologic Studies Depression Scale (CES-D) exhibit differential item functioning (DIF) across gender groups, and how these biases compare to those observed in human data.

We generated 4,000 responses using four proprietary LLMs (GPT-4o, GPT-3.5, Google Gemini, and Claude) by prompting models to adopt simulated demographic characteristics. We then applied the generalized partial credit models (GPCM) to assess DIF and employed item response curves (IRCs) to visualize response patterns. Additionally, item similarity matrices and textual analyses were conducted to examine potential sources of bias. Findings reveal that while LLMs exhibited measurement bias, the pattern of DIF differed significantly from human data, underscoring the challenge of interpreting LLM-based psychometric responses. Item content characteristics and word frequency analysis provided limited explanatory power for these biases, suggesting that response variability may be model-dependent rather than content-driven.

Our results highlight both the potential and limitations of LLMs in psychometric research. While these models can efficiently simulate survey responses, their biases and response distributions diverge from human patterns, warranting further methodological refinement. This study contributes to the ongoing discourse on AI-assisted psychometrics, emphasizing the need for rigorous validation of LLM-generated data in psychological assessment.

Integrating response time into stopping rules in computerized adaptive testing

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Hahyeong Kim (University of Illinois Urbana-Champaign), Dr. Justin Kern (University of Illinois Urbana-Champaign)

Computerized adaptive testing (CAT) selects test items based on an examinee's responses, optimizing measurement precision while reducing test length. However, practical constraints such as testing time and speededness remain critical concerns. Existing stopping rules in CAT, such as the standard error (SE) and minimum information (MI) rules, prioritize measurement accuracy but do not account for response time (RT). This omission can lead to excessively long tests or unfair disadvantages for test-takers provided with more time-consuming items. While RT has been incorporated into item selection models, its potential in stopping rules has received little attention.

This study explores the integration of RT into CAT stopping rules to balance measurement precision and efficiency while mitigating speededness-related biases. A review of existing approaches highlights how prior research has managed RT constraints in CAT item selection and classification testing. Applying to Sie et al.'s (2015) RT-based stopping rule in computerized classification testing (CCT), this study proposes an adaptive stopping rule that incorporates RT within a variable-length CAT framework. By limiting overly time-consuming items relative to remaining test time, this approach aims to enhance fairness and efficiency without compromising precision.

A simulation study will evaluate the impact of this combined stopping rule across various testing conditions, including different speededness thresholds and item exposure constraints. The findings will inform best practices for balancing test length, accuracy, and fairness in CAT applications. This research contributes to improving adaptive testing methodologies by integrating RT as a key factor in termination decisions.

Comparison of RDSEM approaches in continuous-time and discrete-time

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Hosung Ryu (Chonnam National University), Prof. Soyoung Kim (Chonnam National University)

This study compares the Residual Dynamic Structural Equation Model (RDSEM) and the Continuous-Time Residual Dynamic Structural Equation Model (CT-RDSEM) in the context of intensive longitudinal data under both discrete-time and continuous-time frameworks. RDSEM, which is useful for analyzing intensive longitudinal data, assumes that observations occur at discrete time intervals. However, when missing values occur, the inclusion of imputed values can lead to slower convergence rates or inaccurate estimation. In contrast, CT-RDSEM can continuously process randomly observed time intervals, making it applicable even to data collected at discrete intervals with missing values, thereby improving estimation accuracy. The results of this study identified that CT-RDSEM provides relatively accurate estimates compared to RDSEM under conditions of smaller sample sizes (N), fewer time points (T), and higher missing data rates.

Results are discussed along with limitations and suggestions for future CT-RDSEM methodological research, as well as implications for researchers applying the model in practice.

Evaluating detection and correction methods for dependent effect sizes in meta-analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Prof. Hsiu-Ting Yu (National Chengchi University), Mr. Chi-Yun Deng (National Chengchi University)

This study evaluates the performance of various detection and correction methods for publication bias in meta-analysis. Common approaches include graphical methods and selection models. The graphical methods examined are funnel plots, the trim-and-fill method, and Egger's linear regression, while the selection models include weight function models, Iyengar and Greenhouse's generalization, and p-methods. A two-condition, between-subjects design is employed in Monte Carlo simulations, assuming that observations within conditions are normally distributed with a common but unknown variance. The study evaluates several factors, including (1) population average effect size, (2) between-study heterogeneity, (3) number of primary studies, (4) sample size per study, and (5) publication mechanisms. Additionally, varying degrees of dependency among effect sizes within sub-studies are incorporated. The number of sub-studies and the level of dependency within the same study are systematically manipulated. The evaluation criteria include bias, root mean square error (RMSE), coverage percentage, and coverage width. This study also compares the advantages and limitations of graphical methods and selection models in detecting and correcting publication bias. The findings will summarize and assess the effectiveness of these methods, providing practical guidelines for addressing publication bias in meta-analyses. Special attention is given to scenarios where sub-studies exhibit dependent effect sizes.

Exploring attribute hierarchies in cognitive diagnosis from a graph-theoretic perspective

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Hyunjoo Kim (University of Illinois Urbana-Champaign), Dr. Hans Friedrich Koehn (University of Illinois Urbana-Champaign)

Attribute hierarchy models (AHMs) in cognitive diagnosis (CD) assume that cognitive skills ("attributes") are interrelated and hierarchically structured. However, this structure is not known a priori and must be estimated either by content experts or directly from item responses ("data-driven"). Although AHMs offer a realistic and flexible approach to modeling assessment data, they have been underused in CD because imposing a hierarchy on the attributes causes complications. Existing data-driven estimation methods typically conceptualize attribute hierarchies as directed networks, with edges identified using variable selection techniques via a penalized EM algorithm or Bayesian methods using the Metropolis-Hastings algorithm within Gibbs sampling. These approaches often face limitations, including difficulty specifying penalty terms, heavy parameterization, and potential convergence issues.

To meet this challenge, the current study explores graph-theoretic approaches to reconstruct item hierarchies directly from item responses, which are assumed to point at the hierarchical structure of the underlying attributes. Tatsuoka (2009) introduced the idea that items designed to measure specific skills may reflect possible relations among attributes through directed acyclic graphs, thereby formalizing both item and attribute hierarchies. This idea is in line with Gagne's (1965) suggestion that an attribute hierarchy underlying test items should manifest itself in the hierarchical organization of the items. A small-scale experiment using the asymmetric component from the dominance matrix—applied with various scaling procedures and hierarchical clustering—provides promising preliminary evidence for graph-theoretic approach.

Equating coefficients and their standard errors under skew-normal ability distribution

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Ikko Kawahashi (Meiji Gakuin University), Dr. Saori Kubo (Tohoku University)

In item response theory, when test equating is conducted using a common examinee equating design, the ability parameter distribution of the common examinee group often exhibits negative skewness. To address this issue, Kawahashi (2023) proposed asymptotic standard errors for the estimators of equating coefficients that do not rely on specific assumptions about the ability distribution. However, one limitation of Kawahashi's estimators is that a large sample size (N □ 5,000) is required to ensure their asymptotic properties. In this study, we derive estimators for equating coefficients and their asymptotic standard errors that can provide accurate estimates with a smaller sample size. We adopted the skew-normal distribution for the probability density function of the ability distribution to allow the estimators to exhibit asymptotic properties with reduced sample sizes. We developed an expectation-maximization algorithm for estimating the equating coefficients based on the skew-normal distribution and then derived the corresponding estimators for the asymptotic standard errors. A simulation demonstrated that the proposed estimators can estimate the asymptotic standard errors with greater accuracy than those of Kawahashi (2023), even when the ability distribution is highly skewed and a significant discrepancy in scales exists between the two test forms, all with a smaller sample size (i.e., N = 1000).

Second-order structural equation modeling with both factors and components

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. In-Hyun Baek (McGill University, Montreal, Canada), Dr. Gyeongcheol Cho (Department of Psychology, The Ohio State University), Dr. Heungsun Hwang (McGill University)

Higher-order measurement models enable researchers to model constructs at multiple hierarchical levels, with higher-order constructs representing broader traits and lower-order constructs capturing specific facets. Structural equation modeling (SEM) is a general statistical framework for specifying and estimating these models. However, existing SEM methods are limited to handling higher-order models with either factors or components, rendering them unsuitable for those that involve both factors and components. To address this limitation, we propose a method termed second-order integrated generalized structured component analysis (IGSCA), which can accommodate four distinct cases of second-order relationships involving factors or components at different levels: (1) first-order factors paired with second-order factors, (2) first-order components paired with second-order components, (3) first-order factors paired with second-order components, and (4) first-order components paired with second-order factors. We evaluate the performance of second-order IGSCA through a simulation study, using partial least squares path modeling as a benchmark. Additionally, we demonstrate the practical applicability of the proposed method with an analysis of real-world data.

Comparison of Cronbach’s Alpha and McDonald’s Omega for nationally collected cross-sectional panel data

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Jae-A Lim (Sogang University), Prof. Hye Won Suk (Sogang University)

In social sciences and related fields, reporting the reliability of measured variables is crucial, as reliability indices reflect the internal consistency of measurements. A widely accepted definition of reliability describes it as the ratio of true score variance to the sum of true score variance and error variance. Cronbach’s alpha is a commonly used estimator for reporting reliability; however, it can be biased if its underlying assumptions—such as equal variance among items, a sufficient number of items, essential tau-equivalence, unidimensionality, normally distributed errors, and independent errors—are violated. National panel data collected cross-sectionally is typically obtained using stratified, clustered, and systematic sampling at the city/province level. Due to the diverse demographic and socioeconomic characteristics of respondents in panel data, there is a high likelihood of violating the tau-equivalence assumption.

McDonald’s omega is a reliability estimator based on a latent variable model. It offers several advantages over Cronbach’s alpha, as it does not assume tau-equivalence or unidimensionality. Until now, various approaches to measuring reliability have been developed and refined; however, there remains a need to expand the literature on estimating reliability in cross-sectionally collected national panel data.

This study addressed how to compare Cronbach’s alpha and McDonald’s omega under different panel data conditions. We examine the reliability of a 6-item scale under varying conditions, including the number of clusters, the number of factors, item intraclass correlation, and the reliability levels of the indicators (i.e., within- and/or between-level). Some considerations are presented along with simulations and an empirical case.

Structural bias in causal effects: Two-wave nonequivalent control group designs

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Jaehyun Shin (Virginia Tech), Dr. Yasuo Miyazaki (Virginia Tech)

a two-wave nonequivalent control group design is a quasi-experimental design frequently employed in educational and psychological research when randomization is not feasible. traditional analytical approaches, such as ANCOVA and change-score analysis, often yield contradictory results in this design, as illustrated by Lord's Paradox, thereby underscoring the structural biases in the treatment effects and highlighting the need for explicit causal assumptions among the variables involved. despite extensive discussions on the assumptions required to obtain unbiased treatment effects from a graphical causal model perspective, comprehensive and consistent understandings of these assumptions under various causal relationships remain unexplored. accordingly, the study demonstrates how tools such as the backdoor criterion and path-tracing rules developed in the graphical causal model framework can address such challenges. various plausible scenarios within the two-wave design are examined to illustrate how these tools can be utilized, followed by a discussion of how different forms of bias may arise and which specific assumptions are needed to obtain unbiased treatment effects under different scenarios. finally, a theoretical rationale explains why the backdoor criterion and path-tracing rules can identify either the entire or partial structural bias under specific analytic approaches, demonstrating how this method is more accessible than the algebraic OLS estimation method under both approaches for identifying causal assumptions required to obtain unbiased treatment effects, thereby underscoring its contributions to the field.

Reconciling reliability across measurement frameworks for Adverse Childhood Experiences Scales

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Jay Jeffries (University of Nebraska-Lincoln), Ms. Amanda Barrett (University of Nebraska-Lincoln), Ms. Melanie Willis (University of Nebraska-Lincoln), Dr. James Bovaird (University of Nebraska-Lincoln), Ms. Whitney Thomas (University of Nebraska-Lincoln)

The psychometric evaluation of Adverse Childhood Experiences (ACE) scales has predominantly relied on reflective measurement, often overlooking formative measurement principles. This gap fosters misunderstandings in ACE assessment reliability and validity across research and practice. Although traditional notions of internal consistency are widely used to assess ACE scale reliability, researchers adopting formative measurement practices often exclude such metrics due to conceptual inconsistencies. This omission hinders the comparability of scale psychometrics across the ACE research corpus. This investigation examines the relationship between traditional internal consistency metrics and a formative measurement alternative, the proportion of variance explained (PVE) by a unidimensional principal component analysis, to determine whether these measures can serve as interchangeable psychometric evidence for ACE measurement. Coefficient alpha, McDonald's omega, and PVE were computed for ACE scales from two large, nationally representative datasets (N = 124,764). Pearson correlations evaluated the associations between these psychometric measures over full-sample and state-level data. Internal consistency estimates generally met conventional thresholds of adequacy, while PVE ranged from 44.25% to 48.17%, on average. The PVE strongly correlated with traditional reliability estimates ($r \geq .88$). These results suggest that the one-component PVE may serve as alternative psychometric evidence for ACE scale internal consistency, particularly for researchers adhering to formative measurement. These findings bridge reflective and formative measurement perspectives, offering a viable method for researchers to report psychometric properties regardless of their measurement framework. This approach enhances methodological transparency, facilitates cross-study comparisons, and supports the integration of ACE measurement in clinical applications.

The role of the mean structure in models of co-development

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Jennifer Traver (University of North Carolina at Chapel Hill), Dr. Patrick Curran (University of North Carolina at Chapel Hill)

The Random Intercept Cross-Lagged Panel Model (RI-CLPM) and Latent Curve Model with Structured Residuals (LCM-SR) are popular methods for modeling co-development. Though they arise from disparate frameworks (i.e., the CLPM and LCM respectively), they have many similarities. The primary difference between the two involves how each model captures change over time. The mean structure is saturated for the RI-CLPM but restricted for the LCM-SR. To examine the impacts of this difference, we conducted a Monte Carlo simulation study. We compared performance (i.e., model fit and parameter estimation) between the LCM-SR and RI-CLPM as most typically parameterized in the literature. Various conditions were examined, including the presence or absence of a trend and the magnitude of the slope variance, cross-lagged effects, and covariance between slopes. Mixed-effect ANOVAS were fit to statistically examine results, and effects with a partial eta-squared greater than 5% were deemed meaningful

We found that both models had similar performance when a trend was absent. However, when a trend was present, the typical parameterization of the RI-CLPM (i.e., with no slope factor and a saturated mean structure) had worse model fit and more biased parameter estimates than the typical parameterization of the LCM-SR. Performance of the RI-CLPM deteriorated as slope variance increased. Autoregressive effects were particularly biased, with an average relative bias of 132% when the slope variance was large. These findings highlight the importance of the mean structure in models of co-development and suggest using a model with a restricted mean structure if a trend is expected.

Comparing generalized linear mixed models and generalized estimating equations with longitudinal count data: A demonstration

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Jichuan Wu (University of Minnesota), Dr. Haoran Li (University of Minnesota)

Longitudinal count data posits unique challenges for applies researchers to choose appropriate models and provide correct interpretations. This study compares Generalized Linear Mixed Models (GLMM) and Generalized Estimating Equations (GEE) in analyzing longitudinal count data, which provides conditional (subject-specific) and marginal (population-average) estimates, respectively. Using data from the National Longitudinal Survey of Youth (NLSY97), we fitted GLMM and GEE models to examine the trend and sex differences regarding the number of drinks consumed. For GLMMs, a negative binomial model with piece-wise age effects and cross-level interactions demonstrated optimal fit, while for GEEs, a Poisson model with piece-wise age effects and AR(1) correlation structure performed best. Findings from both models underscored similar and significant trend change in drinking behavior, with a critical turning point at age 18, and confirm that gender significantly influences drinking patterns. The similarities and differences in the results between GLMMs and GEEs were graphically illustrated and discussed. These findings provide practical guidance for researchers in selecting appropriate analytical approaches for longitudinal count data based on their inferential goals.

An innovative approach to latent underlying structure by employing Bayesian network analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Jingyang Li (University of Georgia), Prof. Zhenqiu Lu (University of Georgia), Prof. Pengsheng Ji (University of Georgia)

Traditional psychometric methods, such as exploratory factor analysis (EFA), are widely used in social and behavioral sciences to identify the underlying relationships among a set of observed variables by grouping them into a smaller number of latent factors. These methods, assuming linear models, often rely on assumptions of linearity and simple dependence that may not adequately reflect the complexity of trait interdependencies. Probabilistic modeling approaches provide new tools to explore latent structures. Among these, Bayesian Network Analysis (BNA), a probabilistic graphical modeling approach, stands out for its ability to model conditional relationships and capture probabilistic dependencies among items. In this study, we propose an innovative approach to uncovering the underlying latent structure by employing BNA. Unlike linear models, BNA constructs a network where nodes represent variables and edges denote probabilistic influences, offering a dynamic framework for understanding how observed variables interact (Litvinenko et al., 2017). To illustrate this approach, we conduct a real data analysis by applying BNA to the Big Five Inventory-2 (BFI-2). Specifically, we reclassify the 60 items of the BFI-2, utilizing algorithms like hill-climbing and Louvain to explore item groupings. We further evaluate the resulting structures against metrics like Purity and Adjusted Rand Index (ARI) to assess their alignment with the BFI-2's theoretical framework. Findings highlight BNA's potential to complement traditional methods, providing deeper insights into the BFI-2's structure, while addressing limitations of linear models. This study contributes to advancing psychometric methodologies and highlights the value of probabilistic approaches in personality assessment.

Performances of collapsed Gibbs sampling, posterior mode estimation and joint maximum likelihood estimation on diagnostic classification models for boundary problems

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Jingyang Li (University of Georgia), Prof. Zhenqiu Lu (University of Georgia), Prof. Matthew J. Madison (University of Georgia), Prof. Zuchao Shen (University of Georgia)

Diagnostic Classification Models (DCMs) offer a structured approach to assessing latent cognitive attributes, but their estimation procedures are prone to boundary problems, where slipping and guessing parameters converge to extreme values, reducing classification accuracy. This study evaluates the performance of Posterior Mode Estimation (PME), Collapsed Gibbs Sampling (CGS), and Joint Maximum Likelihood Estimation (JMLE) in addressing these challenges through a systematic Monte Carlo simulation. By varying sample sizes, attribute dimensions, item numbers, and item quality (slipping and guessing parameters), the study examines classification accuracy and boundary problem occurrence rates across different estimation methods. The findings indicate in many conditions JMLE achieves the highest classification accuracy while maintaining the lowest boundary problem occurrence rate. However, its performance declines in cases of high Q-matrix complexity, suggesting limitations when the number of measured attributes exceeds available item information. In most simulation conditions, the computation efficiency (CE) of JMLE is worse than CGS and PME. Its boundary problem occurrence rate (BPOR) is better than PME. When item quality is sufficient, CGS can effectively mitigate boundary problems and outperforms JMLE in classification accuracy. Additionally, CGS exhibits greater robustness under extreme noise conditions, maintaining higher classification accuracy than PME and JMLE when slipping and guessing values are set at 0.6. These findings provide insights into the trade-offs between Bayesian estimation methods and frequentists methods in DCMs, show the strengths and limitations of each method in optimizing classification accuracy and parameter stability.

Measuring reflective higher-order consumption worldviews

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Julien Geissmar (TU Clausthal), Prof. Thomas Niemand (TU Clausthal)

Worldviews nowadays are widespread attitudes ranging from brand communities (e.g., Apple, Tesla) over health-related concerns (e.g., COVID-19 vaccination) to security beliefs (e.g., Windows and Viruses). Yet, their measurement is complex and hence sparsely investigated. This research examines consumer worldviews, which are formed reflectively via three hierarchical levels of multi-dimensional consumption context-specific attitudes towards products, brands and environmental influences. We apply repeated measures nested in two-group SEM, weighted by attitude strengths and higher-order measurement models for the example of brand-, product- and privacy-related sub-worldviews regarding smartphone operating systems (Apple and iOS, Google and Android). We demonstrate that this reliable (ω_{pre} = .75, ω_{post} = .82) and validated worldview works as a mediator between reactance and decision variables like trust and word-of mouth. Further, we reveal that this worldview helps to explain a (reversed) boomerang effect manipulating valence in a mixed-design experiment.

Modeling engagement dynamics: Integrating deep learning with generalized additive models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Jung Yeon Park (George Mason University), Dr. Linghan Zhang (Eindhoven University of Technology), Dr. Nirup Menon (George Mason University)

Higher education institutions increasingly offer online courses, yet real-time measurement of student engagement remains a methodological challenge. Traditional self-report measures provide subjective and static assessments, limiting their utility in capturing engagement dynamics. This study proposes a modeling framework to analyze engagement trajectories in online learning environments. A Deep Neural Network (DNN) was trained on the Dataset for Affective States in E-Environments (DAiSEE), which includes over 10,000 expert-annotated video clips. Engagement was classified into four discrete levels—very low, low, high, and very high—based on facial features, eye gaze, and head movement. Model generalizability was assessed through a lab-based experiment, where students viewed online lectures, completed quizzes, and provided self-reported engagement ratings. Given the non-linearity and temporal dependence of engagement processes, a Generalized Additive Model (GAM) was implemented to estimate individual engagement trajectories. The model incorporated time-varying predictors, self-reports, and quiz performance while adjusting for classification uncertainty in the DNN-derived engagement labels. The GAM framework was selected due to its capacity to accommodate smooth functions, providing flexibility in modeling engagement trends without restrictive parametric assumptions. This study integrates machine learning and statistical modeling to refine engagement measurement in online education. The proposed approach addresses construct-irrelevant variance in classification outputs and enhances the validity of engagement estimation in digital learning environments.

A new frontier in psychometrics: Leveraging large language model for psychological construct measurement

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Kentaro Suzuki (York University), Dr. Xijuan Zhang (York University)

Recent advancements in Artificial Intelligence (AI), particularly through Large Language Models (LLMs), offer a faster, more scalable, and cost-effective approach to developing psychological scales. However, research on using LLMs for generating items for these scales is still limited. Our study aims to address this gap by focusing on three psychological constructs: global self-esteem, mindfulness, and cognitive fluency. We manipulated three independent variables (IVs) for the prompts: types of LLMs, simplicity, and creativity. The LLMs included ChatGPT, Gemini, and DeepSeek. For "simplicity," we tested: 1) a very-simple condition with a one-sentence prompt; 2) a simple condition with 2-3 sentences describing the construct and item-writing guidelines; 3) a detailed condition with comprehensive paragraphs on the construct and item-writing guidelines. The "creativity" IV had two levels: 1) no mention of creativity; 2) a prompt requesting creativity. Preliminary analyses show that the "creativity" IV significantly impacts item generation. For instance, in the self-esteem construct, items from the very-simple and non-creative condition differed notably from those in the simple-creative condition. ChatGPT, in the simple-noncreative setting, produced items similar to the Rosenberg Self-Esteem Scale (RSES), while in the simple-creative condition, items diverged to blend self-esteem with resilience constructs. The type of LLM used also had a significant effect. For example, Gemini's simple-creative condition included more elements of self-compassion compared to ChatGPT. We plan to validate these LLM-generated items through real-world data collection and by examining their factor structure, reliability, and validity.

The use of AI tools to develop and validate Q-matrices

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Kevin Fan (Fulton County Schools), Dr. Jacquelyn Bialo (Georgia State University), Dr. Hongli Li (Georgia State University)

Cognitive diagnostic modeling (CDM) is a valuable approach that allows assessments to provide rich diagnostic information to inform instruction and learning. A crucial step in this process is identifying the specific skills (or attributes) required for each test item, a process known as Q-matrix construction. However, developing a Q-matrix—particularly for an existing assessment—can be challenging, as the cognitive processes involved in responding to test items are often not fully understood. For instance, Li and Suen (2013) constructed an initial Q-matrix for a reading comprehension test by synthesizing evidence from relevant literature, consulting content experts, and analyzing students' think-aloud protocols and then validated it with empirical data. This process was both time-intensive and complex.

Recent research has demonstrated the potential of AI in educational measurement. The purpose of this study is to explore the application of AI in Q-matrix development and validation. Specifically, we will compare the performance and capabilities of various AI tools —such as ChatGPT, OpenAI API, LLaMA, and DeepSeek—in generating Q-matrices for a reading comprehension test. The AI-generated Q-matrices will be compared to the final Q-matrix constructed by Li and Suen (2013). Using AI for this purpose has the potential to streamline the Q-matrix construction process, facilitating the use of CDMs to provide detailed diagnostic insights into student test performance.

Extending biclustering models to polytomous psychological scales

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Koji Kosugi (Senshu University), Dr. Kojiro Shojima (National Center for Universtity Entrance Examinations)

Biclustering is a method that simultaneously clusters data in both row and column directions. Shojima (2022) applied this method to test data, proposing a way to classify both examinees and test items simultaneously and further extended it to rankclustering which arranges subject classes by achievement levels. This study extends these models to polytomous response data. Specifically, we adapted it to nominal scale data by changing the Bernoulli distribution in the original model to a categorical distribution and the beta prior distribution to a Dirichlet distribution. Psychological scales, typically using 5-7 point Likert scales, are commonly standardized through factor analysis. However, it is often overlooked that factor analysis is a model assuming latent variables and merely extracts common dimensions from the item response space. This can lead to inappropriate interpretations suggesting the existence of psychological entities within the human mind. In contrast, our proposed method classifies based on response pattern similarities, allowing classification of subjects and items without referring to item response mechanisms, making interpretation straightforward and empirically grounded. Furthermore, we implemented these models in R and developed an analysis package with visualization functions that help researchers interpret the results. This enables researchers to easily apply biclustering to polytomous response data and gain insights into the structure of psychological scales. Future work will explore extensions to ordinal scale data, which will further enhance the applicability of this approach.

Differential test functioning procedures in small and disparate group samples

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Laura Jamison (Naval Research Laboratory), Dr. Cyrus Foroughi (Naval Research Laboratory)

Much of the research in differential functioning both at the test (DTF) and item (DIF) level has focused on model accuracy in large samples. This is because most latent variable modeling applications require large sample sizes for stable parameter estimates. However, in practical applications, small sample sizes and disparate group sample sizes are common data conditions. One testing procedure identified for DIF in small sample conditions is to compare the results of a likelihood ratio test and a logistic regression (Lai et al. 2005; Belzak, 2019). However, these studies didn't investigate the behavior of DTF models under these same conditions. The recommended testing procedure is to first test DIF, then test DTF. If no DTF is found, it could be because the DIF is not severe enough to impact test functioning, or because there is balanced DIF within the items. In this simulation study, we investigate testing procedures for DIF and DTF in both small and disparate sample sizes varying several key data conditions including sample size and sample size ratios, ratio of DIF, underlying ability mean and variance, balanced and unbalanced DIF, and uniform and nonuniform DIF. We will compare the performance of DTF testing procedures including signed/unsigned DTF tests and SIBTEST. Specifically, we are interested in providing guidelines for testing DTF based on DIF results from the outlined testing procedures in small and disparate group sample size conditions. Initial simulation results indicate a higher Type II error rate in DTF when sample sizes fall below 100.

Evaluating the performance of SEM fit indices in MASEM and MAGNA models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Laura Maria Fetz (University of Amsterdam)

Meta-analytic structural equation modeling (MASEM) and meta-analytic gaussian network aggregation (MAGNA) are powerful methods for synthesizing results across multiple studies. In our experience variations in software implementations—such as differences in data pooling, optimization routines, and fit index calculations—have made direct comparisons between factor and network models, hence model selection, challenging. In this study, the aim is to enhance model selection by integrating both approaches within a unified framework using the R package OpenMx. Next, we aim to evaluate the performance of fit indices—such as the chi-square test of exact fit, RMSEA, TLI, CFI, AIC and BIC—through a simulation study with saturated, factor, and network data-generating models. We test if the fit indices correctly identify the underlying data-generating model under each condition. If so, the approach taken makes direct model comparisons possible, supporting better informed decisions in meta-analysis, ultimately enhancing the rigor of meta-analytic research.

The impact of configural invariance violations on the performance of alignment optimization

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Lindsay Alley (McGill University), Dr. Jessica Flake (University of British Columbia)

Measurement invariance is a prerequisite for making valid comparisons of observed scores across groups or pooling such data from different groups together. Using the alignment method, researchers can obtain comparable scores even when invariance does not hold for some items. However, this method assumes that the baseline model is correctly specified and the same across groups, also referred to as configural invariance. For many scales, this assumption is not met. While Asparouhov and Muthen (2014) advise researchers to modify the configural model until they achieve acceptable fit, it is unclear what constitutes 'acceptable'. This is important to understand, as iterative model modification is challenging, especially with many groups, and often does not result in a correct model. However, in addition to different amounts, configural invariance can be violated in different ways, representing conceptually distinct patterns in the data and presenting practical challenges. Using Monte Carlo simulation, we examine the implications of different types (e.g., crossloadings, correlated residuals) and amounts of configural non-invariance for the performance of alignment optimization. To operationalize amount of configural non-invariance comparably across non-invariance sources, we use the mean absolute discrepancy between the data generation correlation matrices of the reference and focal groups. We report bias and coverage for estimation of group factor means and variances, and item loadings and intercepts, as well as power and Type I error for the detection of DIF. We provide advice to researchers on how to identify problematic violations of configural invariance for alignment optimization.

Investigating multidimensionality in symptom presence/severity

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Lionel Meng (University of Wisconsin - Madison), Prof. Daniel Bolt (University of Wisconsin - Madison)

Methodological studies of self-report rating scale data in psychological assessments frequently identify a dimensional distinction between ratings of the presence of symptoms versus the severity of symptoms. We consider both sequential (hurdle) models and a multidimensional nominal response model to illustrate the potential for item difficulty/dimensionality correlations to manifest as rating scale (presence/severity) multidimensionality. We use both simulations and an empirical dataset measuring symptoms of nicotine withdrawal to demonstrate the potential to distinguish these causes of multidimensionality.

Uncovering dynamic cognitive signatures in spatial problem solving

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Lucy Okam (Morgan State University)

cognitive signatures—unique and consistent patterns of cognitive processing—provide a valuable framework for analyzing individual differences in problem-solving strategies. this study investigates cognitive signatures in the context of spatial problem solving by examining response time (RT) dynamics, strategy flexibility, and accuracy. specifically, it examines whether cognitive strategy processes remain consistent across tasks or change depending on item complexity and total time commitment. response time slopes and variability were calculated for each participant to determine cognitive signatures. Shannon entropy (H), which measures the predictability of RT distributions, was used to estimate strategy flexibility. the entropy of a variable is defined as the "amount of information" it holds (Vajapeyam, 2014). it is based on a sequence that results in some uncertainty (Chen, 2016). despite large entropy differences among clusters (F(9,72) = 9.057, p <.001), entropy was not a strong predictor of accuracy (r = 0.004, p =.768), implying that higher flexibility does not always translate to superior performance and that certain task may require stability rather than flexibility for maximum performance. instead, analysis of overall task time revealed that moderated processing (as assessed by total time) was connected to improved accuracy, whereas exceptionally fast or slow response times suggested ineffective techniques. however, task design appears to alter cognitive approach tendencies. these findings emphasize the potential of cognitive signatures in psychometric modeling, adaptive testing and learning, and real-time cognitive assessment, paving the path for more personalized testing and training interventions.

Latent profile analysis of health and well-being in COVID-19 patients

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Marco Viola (University of Turin), Prof. Rosalba Rosato (University of Turin)

Latent Profile Analysis (LPA) is a statistical method used to identify latent subgroups within heterogeneous populations based on observable continuous variables. Unlike other classification techniques, LPA allows for the identification of homogeneous groups of individuals without requiring predefined thresholds, thereby improving the understanding of interindividual variability. In the medical field, LPA has been widely applied to analyze symptom patterns and psychophysical well-being, offering a more detailed representation of patient variability.

In the context of COVID-19, LPA proves particularly useful for examining the physical and mental health conditions of patients. This study analyses a sample of 601 individuals who were hospitalized for COVID-19 at the Molinette Hospital in Turin, Italy, using data on quality of life and physical and mental health collected in the post-discharge period. The aim is to identify latent profiles based on health indicators such as quality of life, anxiety, depression, stress, sleep quality, and fatigue. Identifying distinct subgroups based on physical and psychological characteristics will allow for the assessment of the extent to which disease severity influences membership in specific latent profiles. Furthermore, this study seeks to identify clinical variables contributing to this classification to optimize targeted therapeutic interventions based on patients' profile membership.

Sequence mining of cognitive strategies in a computerized neuropsychological task

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Matheus Rodrigues (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University), Mr. Gabriel Lopes (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University), Ms. Amanda Cardoso (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University), Ms. Isadora Martins (Albert Einstein Israelite Hospital), Dr. Alexandre Serpa (Computational Intelligence in Psychometrics and Epidemiological Research, Mackenzie Presbyterian University)

The study examines cognitive strategies during problem-solving using the Tower of London (TOL-BR) task. A sample of 647 participants aged between 12 and 29 years completed a computerized ToL test, which required arranging colored spheres on pegs to match a target configuration with the fewest moves possible. Longest Common Subsequence (LCS) method, a sequence mining technique, was applied to compare each participant's sequence of moves with an optimal solution. Two main metrics were derived: similarity (the degree of match with the optimal sequence) and efficiency (the number of extra moves made beyond the minimum required). Results revealed some age-related differences. Young adults (18–24 and 25–29 years) consistently showed higher similarity and efficiency scores compared to teenagers (12–17 years), particularly on more challenging items, like item 11. Statistical analyses, including non-parametric tests and Bayesian ANOVA, confirmed these developmental differences, indicating a clear improvement in planning strategies with age. In contrast, differences between male and female participants were minimal, with only a slight efficiency disadvantage observed in males for one item. Overall, the findings highlight the effectiveness of sequence mining in uncovering subtle cognitive processes and emphasize that executive function assessments, particularly those measuring planning and problem-solving, should account for age-related variability,

Assessing metacognitive calibration in biological model evaluation using Rasch analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Mei Grace Behrendt (University of Nebraska-Lincoln), Dr. Jordan Wheeler (University of Nebraska-Lincoln), Dr. Joe Dauer (University of Nebraska-Lincoln), Dr. Carrie Clark (University of Nebraska-Lincoln)

Metacognitive calibration—the ability to align confidence with accuracy (Andrade, 2019; Stone, 2000)—is a key factor in STEM learning, particularly in biological modeling (Tanner, 2012). Understanding how students monitor their reasoning during model evaluation tasks can inform instructional strategies that enhance learning outcomes (Dauer et al., 2024, Schraw & Dennison, 1994; Tobias & Everson, 2009). The purpose of this study is to describe the development and validation of a 17-item Biology Metacognitive Scale (BMS) that assesses students' metacognitive calibration on biology model evaluation tasks. Each item on the BMS displayed a biological model (e.g., carbon cycle) and asked students to determine whether an error existed while also rating their confidence. Metacognitive calibration for an item was determined based on the alignment between task accuracy and confidence ratings. Fifty undergraduate life sciences students participated in this study and completed the BMS while undergoing fMRI scanning. The psychometric properties of this scale were assessed through Rasch measurement theory. Exploratory factor analysis supported a unidimensional structure and the variance explained the Rasch model exceeded 20% (Reckase, 1979). Analyses also showed that moderate internal consistency with acceptable infit and outfit mean-squared statistics (0.87 to 1.25). Additionally, results demonstrated that students with higher metacognitive calibration also had higher performance on the biological model task (r = .88). Results from this study suggest the BMS could serve as an effective tool for assessing and improving metacognitive calibration in STEM education, with potential applications for developing targeted interventions to enhance student learning and reasoning.

Detecting differential item functioning using varying coefficient models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Michaela Cichrová (Institute of Computer Science, Czech Academy of Sciences; Faculty of Mathematics and Physics, Charles University), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University)

Testing for Differential Item Functioning (DIF) is an important step in the analysis of multi-item measurements. Traditionally, DIF has been investigated using a categorical grouping variable, which can be either dichotomous (e.g., gender, where groups are male and female) or polytomous (e.g., education level, with multiple categories such as high school, undergraduate, and graduate). This approach works well in many situations, but there are circumstances where the variable of interest is continuous, such as age or income. In such cases, treating the variable as categorical by arbitrarily dividing it into discrete groups may lead to a significant loss of valuable information and potentially obscure complex relationships within the data.

We propose the use of varying coefficient models to better handle the continuity of covariates in DIF detection. These models allow the effect of the covariate to vary smoothly, capturing more intricate patterns that simpler methods may overlook. To demonstrate their effectiveness, we compare several model forms through a comprehensive simulation study. The simplest model is a direct extension of the logistic regression model, commonly used for DIF detection with dichotomous grouping variables.

We compare the models in terms of their accuracy while also considering computational challenges, interpretability, and the identifiability issues that arise in more complex models.

Detection of aberrant response behavior in context of robust logistic regression methods

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Michaela Varejkova (Institute of Computer Science, Czech Academy of Sciences; Faculty of Mathematics and Physics, Charles University), Dr. Patrícia Martinková (Institute of Computer Science, Czech Academy of Sciences; Faculty of Education, Charles University)

Accurately modelling item responses is a key objective in psychometrics. Generalized linear and nonlinear models provide a useful framework for this task. However, traditional logistic regression is highly sensitive to outliers, which can strongly influence parameter estimates and compromise assessment validity. These outliers often arise due to measurement errors, misclassifications, or aberrant responses. To address this issue, we explore robust logistic regression using M-estimators within the generalized linear model framework. This approach incorporates robust weighting functions to reduce the influence of extreme values in both the response and predictor spaces. Our simulation study demonstrates that while classical logistic regression is heavily affected by misclassified outliers, robust logistic regression remains stable and provides more reliable estimates. This method presents a practical alternative to traditional item response modeling approaches, effectively mitigating the impact of anomalies without introducing additional item parameters.

Bridging psychometrics and reality: Mitigating selection bias in factor analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Miguel Bosch Francisco (Universidad Autónoma de Madrid)

Factor analysis, widely used in psychological measurement, often assumes a reflective model where observed variables manifest underlying latent constructs. However, standard factor analytic procedures often violate key reflectivity assumptions due to item selection biases, imposing range restrictions on factor loadings. This study demonstrates how these restrictions inflate factor variance estimates, distort latent correlations, and bias fit indices, challenging the assumption that psychometric models provide unbiased representations of psychological constructs.

A critical issue arises from the conflation of psychological constructs with latent variables. While constructs are theoretical entities, latent variables are statistical abstractions. This conflation leads to misinterpretations of factor analytic results, potentially undermining construct validity.

Through theoretical derivation and Monte Carlo simulations, we show that selecting items based on observed loadings alters the empirical distribution of factor loadings, leading to overestimation of factor strength and underestimation of error variance. This bias propagates into structural analyses, affecting latent correlations and goodness-of-fit statistics.

We propose a novel correction method adjusting for the distributional distortion of factor loadings, preserving the reflective measurement model's properties. Our approach is compared with existing corrections across various sample sizes and item selection criteria.

This research bridges psychometrics and psychological reality, addressing the gap between latent variable modeling and the complex nature of psychological constructs. Our findings have significant implications for construct validity and the interpretation of factor analytic results in psychological research, contributing to more faithful representations of psychological phenomena in measurement practices.

Identifying core symptoms of depression for Hawaii farmers: A network analysis for PHQ-2 screening tool

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Min Liu (University of Hawaii), Dr. Thao Le (University of Hawaii)

Introduction

Network analysis shows that depression measures may function differently across populations. Hawai‘i farmers, a high-risk, underserved group with the 4th highest depression and suicide rates nationally (CDC, 2021), may experience depression uniquely due to cultural factors like respect, responsibility, and endurance. This study hypothesizes that the PHQ-9 depression network and central symptoms among Hawai‘i farmers differ from mainland populations.

Methods

A cross-sectional survey collected PHQ-9 responses from 375 Hawai‘i farmers; 144 with mild to severe depressive symptoms were included in the network analysis. A Gaussian Graphical Model, based on a Spearman correlation matrix, was used to estimate the depressive symptom network in RStudio. Model stability and accuracy were assessed via nonparametric bootstrapping. ROC analysis compared the sensitivity and specificity of a new PHQ-2 (guilt & fatigue items) against existing PHQ-2 versions. AUC and Gini index values assessed the new PHQ-2's accuracy.

Results

The analysis identified three symptom clusters: Guilt-Mood-Anhedonia, Sleep-Fatigue-Appetite, and Suicide-Motor-Focus, with stronger associations within clusters than between them. Guilt and fatigue emerged as central symptoms, showing the highest values across strength, closeness, betweenness, and expected influence indices. The new PHQ-2 version (guilt & fatigue) demonstrated good criterion validity (AUC = 0.883) and acceptable reliability (Cronbach's alpha = 0.632).

Discussion

Guilt and fatigue were key symptoms of depression in Hawai‘i farmers, reflecting cultural influences. The new PHQ-2 offers a practical screening tool with potential applications for other underserved populations. These findings demonstrate the utility of network analysis in designing culturally tailored mental health interventions.

CDMs in language assessments: A systematic review

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Mirai Nagasawa (University of Alabama), Mr. Benjamin Lugu (University of Alabama), Prof. Wenchao Ma (University of Minnesota)

Cognitive Diagnostic Models (CDMs) have garnered growing attention in language assessments due to their potential to identify learners' specific strengths and weaknesses in specific skills. Despite the increasing number of studies on CDMs, they often differ substantively in terms of model selection, model fit evaluation, and reliability and validity investigations. In addition to the existing literature reviews of CDMs in language assessments (e.g., Lee & Sawaki, 2009; Mei & Chen, 2022), which identified certain limitations, our work provides more detailed information regarding the psychometric properties of CDMs, such as model selection and validity arguments, which require further investigation. This review included 56 articles, revealing that 78.57% retrofitted CDMs to existing tests, while only 5.36% created new assessments. Additionally, 51.79% of studies employed pre-selected CDMs, with the Fusion Model being the most popular among these, followed by G-DINA and LCDM. This review also examines detailed information such as sample size, item characteristics, model selection, attribute specification, or the detailed procedure for constructing Q-matrices. This study contributes to research on language assessment by providing an overview of CDM usages and revealing areas where current practices may be lacking. Applying CDMs is a complex process that requires extensive knowledge of both statistics and psychometrics (Javidanmehr & Anani Sarab, 2017). Evaluating examinees' language skills from the perspectives of both language assessment and psychometrics can enhance the accuracy and effectiveness of diagnosing language proficiency. Further details are provided in the full paper.

Testing Russell’s affect circumplex in corpora: A latent semantic scaling approach

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Mizuki Kobayashi (Department of Psychology, Senshu University), Mr. Keishi Nomura (Department of Media and Communication, Toyo University), Dr. Koji Kosugi (Department of Psychology, Senshu University)

Russell’s circumplex model explains emotional structure using valence and arousal dimensions, with emotions distributed circularly. While this model is widely referenced in psychological research, computational approaches typically rely on unidimensional scales. This study aims to test Russell’s model using large language corpora through latent semantic scaling (LSS), addressing the gap between theoretical emotional structures and their manifestation in large-scale linguistic data. LSS calculates polarity values (i.e., numeric scores indicating the position of words along bipolar dimensions such as positive-negative or arousal-sleepiness) based on the cosine similarity between seed words and target words, offering interpretability through theoretically informed seed selection. We extended this approach by independently applying unidimensional scaling twice—once for valence and once for arousal dimensions—by selecting specific seed words for each dimension, calculating similarity scores using pre-trained language models, and mapping emotion terms onto a two-dimensional space. Results show that while some emotion terms aligned with the circular arrangement predicted by Russell’s model, the overall distribution differed from the circular pattern, particularly along the arousal dimension where emotion terms showed a limited distributional range. This finding suggests that appropriate seed word selection might potentially help acquire emotion structures similar to the circumplex model from language corpora. Future research should focus on developing inherently multidimensional extensions of LSS, rather than applying unidimensional scaling twice as in our current approach. Furthermore, it will be important to compare the structure of emotion words across different languages and cultural backgrounds, as well as to optimize seed word selection.

Developing OSCE examinee score reports using the latent space item response model

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Nai-En Tang (National Board of Chiropractic Examiners), Dr. Igor Himelfarb (National Board of Chiropractic Examiners)

This study explores the use of the Latent Space Item Response Model (LSIRM) to analyze Objective Structured Clinical Examinations (OSCEs). It aims to improve score reporting by identifying examinees' strengths and weaknesses. OSCEs are key for evaluating clinical skills and knowledge, and providing diagnostic score reports is essential for guiding remedial efforts, particularly for candidates who fail. The LSIRM represents the interactions between examinees and test items in a low-dimensional latent space, with distances between them revealing insights into their performance.

The study applied LSIRM to an OSCE exam for chiropractic case management, consisting of 10 stations with 20 items. Using Bayesian estimation, the study analyzed data from 902 examinees, applying the MCMC algorithm. The results showed item easiness parameters, with some items being easier than others. They also displayed clustering of items from the same station in latent space. This clustering indicated a testlet effect, where items from the same scenario were more closely related.

An interaction map visualized the positions of both items and examinees in latent space, highlighting the relationship between examinee ability and item difficulty. Examinees closer to certain items performed better, while those farther away struggled. Additionally, individual examinee score profiles were created, showing probabilities of correct responses with a focus on the shorter distance component, indicating areas for remedial efforts.

The study demonstrates that LSIRM and interaction maps can enhance OSCE score reports. By identifying specific areas of strength and weakness, they can provide more detailed insights into examinees' medical competence.

The creativity conundrum: Examining the gap between concept and measurement

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Navrose Bajwa (University of Houston), Ms. Anna Snyder (University of Houston), Dr. Kaylee Litson (University of Houston)

Creativity is broadly recognized as a cornerstone of human innovation and development and is widely studied within sub-fields of psychology. Despite broad relevance, defining and measuring creativity remains a challenge, in part because creativity has historically been described as something both “novel and useful” (Guilford, 1950), with some scholars advocating that creativity also involves “surprise” (Boden, 2004; Simonton, 2018; Thaygaurd, 2012; Tsao et al., 2019). Recently, a universal definition of creativity has been proposed as “a multiplicative function of originality, utility, and surprise” (Simonton, 2023). Despite past work done to theorize and measure creativity, the literature is littered with creativity scales that measure distinct concepts. We report an in-progress systematic review of meta-analyses and review articles relevant to creativity measurement (published after 2010) that examine how creativity is defined and measured, with particular attention to the 4P model: Person, Process, Product, and Press (Rhodes, 1961). Drawing from an initial pool of N=463 and a final selection of N=36 articles, we cataloged various definitions studies adopted and mapped them to the measurement instruments used (e.g., divergent thinking tasks, self-report scales, product assessments). Preliminary qualitative results revealed that researchers often invoke broad or multifaceted definitions of creativity but then rely on measurement tools (often a single divergent thinking task or narrow self-report) that only capture a small slice of what they define as “creativity”. We will provide an overview of coded articles and discuss ways to address mismatch between creativity as a concept and its respective measurement tools.

Comparing traditional and AI-based IRT estimation: An empirical study

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Olukayode Apata (Texas A&M University), Mr. Segun Ajose (University of Kentucky)

The increasing role of Artificial intelligence (AI) is reshaping psychometric modeling, yet its accuracy in item response theory (IRT) parameter estimation remains uncertain. This study compares traditional maximum likelihood estimation (MLE) using Stata 16.0 with an AI-based (ChatGPT 4o) estimation in Python to analyze the *Dataset for 21st-Century Character Exploratory Factor Analysis in Vocational High School Students* (Effiyanti, 2024). The dataset from vocational high school students (N = 340) measures greeting behaviors, conflict resolution, integrity, academic honesty, adaptability, and information literacy using 18 Likert-scale items. We evaluated the one-parameter logistic model (1PL), two-parameter logistic model (2PL), and three-parameter logistic model (3PL) and estimated their item difficulty, discrimination, and guessing parameters.

Our results indicate that difficulty estimates in the 1PL model were similar, with Stata estimating Q1 = -1.49 and AI estimating Q1 = -1.14. However, we noticed a slight deviation in the 2PL model estimated discrimination parameters (Q1: Stata = 0.95; AI = 1.21). In the 3PL model, we observed a notable variation in the guessing parameter (Q1: Stata = 0.043, AI = 0.323). The Stata analysis produced stable parameter estimates, while the AI analysis exhibited inconsistencies, particularly in discrimination and guessing parameters.

These findings highlight the strengths and limitations of AI-driven IRT estimation. While AI models can efficiently estimate item parameters, the system should be improved to ensure alignment with traditional MLE-based methods. We encourage future research to look at ways to refine AI estimation methods especially for complex IRT models and large-scale educational assessments.

Classification-based DIF analysis of the National Benchmark Tests (NBT)

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mrs. Precious Mudavanhu (University of Cape Town), Dr. Fernando Austria (Assessments Systems Corporation (ASC))

The NBT tests are national high-stakes tests that assess high school students' proficiency in Academic Literacy (AL), Quantitative Literacy (QL), and Mathematics (MAT) domains. These tests help identify students needing academic support, determine appropriate programme placements, and inform curriculum development in South African higher education. This study explores the classification-related differential item functioning (DIF) of NBT AL, QL, and MAT test items using logistic regression methods with the R difR package. DIF analysis is used to detect and address potentially biased test items that may unfairly advantage or disadvantage certain groups of students, (Zumbo,1999). Effect sizes are calculated using common odds ratios and interpreted on the ETS Delta scale (negligible, moderate and large) to assess the magnitude of DIF. The analysis was conducted on three test forms for each domain, with each form written by over 950 test-takers. Classification was based on self-declared categories: Black, Coloured, White, Indian/Asian, and Other. The results show that no items were flagged for DIF as the effect sizes were negligible despite significant chi-square values. These findings suggest that the NBT tests do not exhibit substantial classification-related DIF. The results reinforce the fairness and validity of the tests across self-declared classification groups, affirming their suitability for high-stakes decision-making in higher education. Further analyses incorporating additional variables, such as province and home language, will be presented as a quality check process on test items to assess the presence of DIF bias.

NLP-based evaluation of LLMs' response alignment in couples counseling

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Qian Shen (University of Florida), Ms. Paige Carter (University of Florida), Ms. Mufan Sun (University of Florida), Dr. Walter Leite (University of Florida)

The rise of powerful large language models (LLMs) has led more people to seek counseling advice through them, though they cannot replace qualified therapists. Researchers have used qualitative analysis to assess satisfaction with LLM-generated counseling responses and develop specialized LLM agents, but few studies have quantitatively evaluated their alignment with people's counseling needs.

In this study, we collected two sets of common couples counseling questions with similar meanings using different prompting strategies and obtained responses from two powerful LLMs, OpenAI o3-mini and Deepseek R1. Using multiple NLP-based features, we have quantified alignment between LLMs' responses and people's needs and applied statistical tests to examine differences due to model choice and prompting strategies.

Results show that LLMs' responses closely match people's phrasing in common counseling questions, effectively addressing core needs with strong empathy, safety, and appropriate vocabulary diversity. However, they still fall short in readability and capturing question-related emotions. OpenAI o3-mini outperforms Deepseek R1 in matching phrasing, core meaning, and vocabulary diversity. Responses to more detailed questions demonstrate better empathy, alignment with core needs, readability, and vocabulary diversity than those for less detailed questions. Model choice and prompting strategies also show a significant interaction effect on the alignment between LLMs' responses and the core needs of the questions.

In the future, we will refine alignment features and compare more LLMs' responses to couples counseling questions, aiming to contribute to the integration of LLMs into psychological counseling.

Detecting differential item functioning in response time using regularized regression

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Qizhou Duan (University of Notre Dame), Dr. Ying Cheng (University of Notre Dame)

This study proposes a regularized regression approach to detect Differential Item Functioning (DIF) in response time data. DIF in response time is important to examine because response time or reaction time is of primary interest in certain tests, speed plays an important role in test administration and scoring, and differential cost in mental effort or cognitive load is relevant to test fairness. We extend the LASSO-based DIF detection method from item response models to response time data to address several drawbacks of traditional approaches, including multiple testing issues and the assumption that non-tested items are DIF-free.

The proposed method uses penalized regression to automatically identify items showing significant group differences in response time patterns while accounting for individual working speed. A simulation study with 500 test-takers and 20 items was conducted to evaluate the method's performance. Results showed that the method maintains a conservative Type I error rate at approximately 2% and demonstrates increasing power rates as the difference in time-intensity parameters between focal and reference groups grows larger.

We discuss various approaches for selecting tuning parameters, including cross-validation, Mallows Cp, Adjusted R^2, AIC, and BIC, as well as methods for determining variable importance through visual inspection, resampling methods, and covariance tests. Future work will expand the simulation conditions to include different group ratios (250/250, 50/450) and test lengths (10 and 30 items), and compare the performance with existing least square DIF detection methods in both simulation and real data settings.

Detection of cognitive diagnostic model misspecification using Lancaster-Chesher information matrix tests

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Oral

Prof. Richard M. Golden (University of Texas at Dallas), Ms. Reyhaneh Hosseinpourkhoshkbari (University of Texas at Dallas)

Existing Pearson Chi-Squared and Likelihood Ratio Goodness-of-Fit (GOF) tests use Cognitive Diagnostic Model (CDM) GOF statistics whose variance increases exponentially as a function of the number of exam questions d. The more recently proposed M2 statistic has the advantage that its misspecification test statistic variance increases only as a quadratic function of the number of exam questions d. Still, the large variance of all of these misspecification test statistics imply poor statistical power for large d.

In this talk, we extend prior work by Hosseinpourkhoshkbari and Golden (2024) and derive entirely new mathematical formulas for seven information matrix misspecification tests. These new specification tests only have 1 or 2 degrees of freedom and are not dependent upon d. In a preliminary simulation study, we estimated the parameters of a five skill CDM with 15 questions using the Fraction-Subtraction data set. Then we generated 500 simulated data sets from the estimated model where each data set consists of 2000 samples. We then checked the p-value calculated using our new formulas against a range of significance levels for each simulated data set. For example, our mathematical theory predicts that about 25 out of the 500 bootstrap data sets would be incorrectly classified as misspecified for a significance level of 0.05. We found good quantitative agreement between the theoretically expected p-values and the simulation results for two out of seven of our new misspecification tests. Further work is planned to investigate performance for different numbers of bootstrap data sets and sample sizes.

Evaluating Thomson Bonds model as an alternative data generating model

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Robert Chapman (University of Minnesota)

Sir Godfrey Thomson introduced the “Bonds” measurement model in 1916 as an alternative to Spearman’s general factor model of intelligence. The Bonds model uses numerous independent random samples to approximate a traditional factor structure. With the growth of high-powered computing, Thomson’s model is now a researcher-accessible data generating model for quantitative inquiry in psychology. The lack of alternative data generating models is a problem in quantitative inquiry in psychology, where data is often simulated with the same model being evaluated. The Bonds model could be used as a neutral data generating model for the evaluation of model fit or comparative model fit of measurement models.

In this study, data was simulated using the Bonds model across multiple conditions and data generation parameters, including number of people, number of items, type of items, number of scales, number of bonds, and proportion of shared and specific variance within and between scales. Factor analysis and unidimensional and multidimensional Item Response Theory (IRT) models were fit to the data. Dimensions predicted by factor analysis were compared to true generating Thomson bonds model dimensions for factor and factor-item alignment. Estimated person and item parameters and domain-item structure were evaluated according to true generating Thomson Bonds model parameters.

The results evaluate the use of Thomson’s Bonds model as an alternative data generating model in the context of conducting quantitative inquiry with psychological measurement models. Results highlight the impact of Bonds model parameters on data generation, IRT model parameter estimation and factor analytic detection of dimensionality.

Linear measurement of parent and family Involvement in K-12 education

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Robert Anderson (Morgan State University)

The purpose of this study is to examine the benefits of using linear measurement models to analyze the results of large-scale surveys. While the use of latent trait theory in the analysis of these datasets is growing, this paper argues much richer analysis of these datasets could be realized if the results were analyzed using linear measurement models.

Utilizing the 2019 Parent and Family Involvement in Education survey, this study found the Rasch model is an appropriate tool for analyzing the survey results. The data largely fit the model. Given model fit, the analysis further analyzed both person and item measures with statistical and graphical techniques to produce an analysis much more informative than the dense tables of numbers used in the official reports. Among other findings, the results showed patterns of parental involvement in their children's K-12 education largely consistent among racial and ethnic lines. The findings also showed higher levels of parental involvement in the public schools as opposed to various private schools. A somewhat surprising and counter-intuitive finding was parents with a less than high school education were more involved in their children's education than were more educated parents.

The ultimate goal of this research is to encourage more researchers, including those in the survey community, to consider the use of linear measurement tools in the analysis of these large datasets. The richness of the linear measurement results make the effort to use these tools worthwhile.

What do web-based memory assessments measure? A convergent validity analysis.

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Romeo Penheiro (University of Houston), Dr. Kaylee Litson (University of Houston)

Objective. Large-scale datasets measuring cognition, like memory, are becoming commonplace using web-based assessments. The underlying properties of the assessments must be considered to use such data in a meaningful way at scale. One such property is the convergence of different assessments aiming to measure a focal construct (i.e., convergent validity; Campbell & Fiske, 1959). Applied to large-scale data, confirmatory factor analysis for multimethod data can demonstrate which assessments within a memory battery achieve high convergence and low method-specificity, but few studies have directly evaluated this question using web-based cognitive assessments. ***Method.*** The current study used an open-access web-based dataset of multiple cognitive assessment batteries in the NeuroCognitive Performance Test (Jaffe et al., 2022). Data for N=738,230 adults (mean age=46.74, SD=16.46, range=18-90) who completed five memory tests (Two Target Search, Forward Memory Span, Reverse Memory Span, Verbal List Learning, Delayed Verbal List Learning) were used. Two models were fit to the five memory tests: a model including only a trait-factor, and a model including both trait- and method-factors (Eid, 2000). The "Memory Span" and "List Learning" method-factors grouped their respective tests. ***Results.*** Results showed that including the method-factors improved model fit (X^2=1008.31, df=3, p<.001). The "List Learning" method-factor showed a high variance proportion of reliable method-specificity (88.78-95.25%), suggesting that these tests do not share much variability with other tests of memory. Sensitivity and covariate analyses will be described. ***Conclusion.*** Findings show a lack of convergent validity in the memory tests. Construing these measures as a single "memory" construct may be incorrect.

Evaluating factor retention methods in large exploratory factor analysis models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Ruoqian Wu (University of Illinois Urbana-Champaign), Dr. Yan Xia (University of Illinois Urbana-Champaign)

Exploratory factor analysis (EFA) is widely used to identify latent structures in observed variables. Nevertheless, determining the number of factors remains a methodological challenge, especially when the size of the exploratory factor analysis model is large (e.g., > 10 items > 5 factors). This study evaluates the performance of various factor extraction methods under large factor analysis models, including parallel analysis (PA) using the 50th and 95th percentiles, exploratory graph analysis with LASSO and TMFG estimations, sequential χ^2 model tests, fit indices (CFI, TLI, RMSEA), and the Kaiser-Guttman criterion. Additionally, we examine the effectiveness of the very simple structure method with varimax and oblimin rotations, the comparison data approach, the minimum average partial test, and the Hull method using both the comparative fit index and Cattell's acceleration factor criteria. The simulation design leads to 432,000 datasets under 432 unique conditions. Specifically, we manipulated the number of factors (5, 6, and 7), items per factor (5, 10, and 15), sample sizes (100, 200, 500, 800, 1,000, 1,500, 2,000, and 4,000), inter-factor correlations (.30, .50, and .70), and factor loadings (.40 and .70). Findings provide empirical guidance on selecting robust factor extraction techniques in complex data settings with large models in psychological assessments and offer recommendations for researchers applying EFA to high-dimensional measures.

Using dual scaling to understand rater cognition

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Prof. Ruth Childs (University of Toronto), Dr. Amanda Brijmohan (University of Toronto), Dr. Ryan Hargraves (University of Toronto)

Dual scaling is a flexible and powerful approach to modelling categorical data that was developed by Shizuhiko Nishisato. Drawing from four research studies, we illustrate how combining dual scaling analyses of ratings with interview data can help us understand rater cognition. The studies investigated: (1) why university faculty disagree when rating applicants' essays, (2) how types of evidence are valued by school leaders, (3) what university students consider when they rate peers' performance on written tasks, and (4) what characteristics university faculty leaders consider when forming admission committees. We discuss the insights on rater differences gained from this approach when combined with qualitative data.

Linking across grades using calibrated projections with censored populations

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Ryan Lerch (University of California, Los Angeles)

This project explores a bivariate scoring prior that is suitable in cases of non-normality as a result of censoring. Growth projections are often an important concern when handling English language assessments. To this end, Hansen & Monroe (2018) demonstrated how Thissen's (2011) calibrated projection method can be used to link adjacent grades using multidimensional IRT, even in the absence of a traditional vertical scale. Hansen & Monroe (2018) used a bivariate normal scoring prior for the two grades, which did not represent the true structure of their language assessment data. Given that the most proficient students from one year are not subject to assessment in subsequent years, the population of test-takers across two grades ought to be right-censored in the upper grade.

The current study overcomes this limitation by using a multivariate Gaussian copula (Nelsen, 2006) to construct a censored bivariate distribution with the five necessary parameters (two means, two variances, and one covariance). Through simulation and real data from a large-scale language assessment, this work explores the improvements in accuracy and precision obtained by using the copula-generated censored bivariate scoring prior as opposed to assuming bivariate normality. Specifically, we explore population mean, variance, and covariance estimates for the projected grade, as well as standard errors of measurement. Thus, the intent of this method remains consistent with the intent of a traditional vertical scale: Growth between grades at the population level is of primary concern, as opposed to growth at the individual level.

Strengthening cross-grade comparisons on a vertical scale using moderated nonlinear factor analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Sanford Student (University of Delaware)

Vertical scales are intended to establish a common metric for scores on test forms targeting different levels of development in a specified domain. They are often constructed using common item, nonequivalent group designs that implicitly rely on the linking items being effectively free from differential item functioning (DIF) or the DIF being symmetric to produce unbiased linking constants. Moderated Nonlinear Factor Analysis (MNLFA) is a measurement model that can be used to understand both the presence of DIF among vertical scale common items and the extent to which the presence of DIF may bias grade-to-grade score distributions. Specifically, this study considers two approaches to MNLFA with empirical anchor item identification to investigate grade-wise DIF in the vertical scaling context. The approaches considered are penalized maximum likelihood (i.e., regularized) estimation of MNLFA (Belzak and Bauer, 2024) and DIF detection based on robust statistics (Halpin, 2024). The performance and affordances of both approaches are explored. Simulation and real data applications show how models that do and do not account for DIF in vertical scale common items can produce very different answers to the fundamental question of how much students grow from one grade to the next, but that when DIF is not present, MNLFA provides effectively identical growth estimates to traditional concurrent and characteristic curve approaches to vertical linking.

A network approach to procrastination: Modeling dynamic interactions with the Irrational Procrastination Scale

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Shirin Rezvanifar (Ph.D), Prof. Denny Borsboom (University of Amsterdam), Dr. Riet Van Bork (University of Amsterdam)

Procrastination, a prevalent self-regulatory failure, affects 20% of adults and 70-95% of students, leading to stress, reduced academic success, and psychological distress (Steel, 2007; Ellis & Knaus, 1977). Traditionally viewed as irrational delay with adverse outcomes (Klingsieck, 2013), it encompasses passive (traditional) and active (rational, time-pressured) forms (Chu & Choi, 2005). Existing models—psychoanalytic (death anxiety), rational-emotional (perfectionism, guilt), and motivational (task avoidance)—highlight diverse causes, while factor analyses identify predictors like fear of failure and task aversiveness (Solomon & Rothblum, 1984; Steel, 2007). Coping strategies range from problem-focused (e.g., goal-setting) to emotion-focused (e.g., self-forgiveness) (Folkman & Lazarus, 1980). Despite extensive research, these factors lack a coherent, dynamic framework. This study applies network analysis to the Irrational Procrastination Scale (IPS; Steel, 2002) with 414 students, who completed the IPS for personalized feedback. By modeling procrastination as a network of interacting items (e.g., delay, excuses), we aim to uncover dynamic relationships among task-related (e.g., self-efficacy, aversiveness) and personality-related (e.g., impulsiveness, fear of failure) factors. Preliminary findings suggest central nodes driving procrastination cycles, offering a multidimensional alternative to unidimensional models. This approach informs tailored interventions—e.g., enhancing self-regulation for passive procrastinators or leveraging time pressure for active ones—potentially improving academic and psychological outcomes. These insights advance procrastination's conceptualization and its measurement in psychometric research.

Keywords: procrastination, network analysis, Irrational Procrastination Scale, self-regulation, dynamic modeling

Simulation-based sensitivity analysis for power of Rasch family models

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Shuhei Hanadate (University of Tsukuba), Dr. Kazuhiro Yamaguchi (University of Tsukuba)

This study presents a simulation-based sensitivity analysis for sample size planning in extended Rasch models, treating them as generalized linear mixed models (GLMMs). The research addresses the challenge of determining an appropriate sample size for reliable parameter estimation and sufficient statistical power in psychometric modeling. Existing analytical solutions are limited, making simulation-based power analysis a practical alternative.

Using the R packages lme4 and mixedpower, we conducted sensitivity analyses on four Rasch family models: the Rasch model (RM), the linear logistic test model (LLTM), the rating scale model (RSM), and the linear rating scale model (LRSM). We examined how sample size, item difficulty, discrimination, and item-design matrices affect power estimates. Our results show that while increasing sample size improves power, model complexity and item-design matrices also play a significant role. LLTM and LRSM, which incorporate item-design matrices, exhibit higher power under complex specifications.

Replication of eRm functions using lme4::glmer demonstrated comparable coefficients estimates, highlighting the feasibility of using widely available R packages for power analysis. However, this approach faces challenges related to computational costs, data dependency, and the inability to apply to partial credit models. We discuss potential solutions, including surrogate modeling and sequential analysis, to improve efficiency.

This study provides insights into sample size determination for extended Rasch models and offers a practical framework for researchers conducting power analyses in psychometric studies. Future research should explore power estimation for more complex item response models, such as the linear partial credit model.

Distinguishing bifactor models from alternative models: Insights from loading patterns

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Sijia Li (University of British Columbia), Dr. Victoria Savalei (University of British Columbia)

In psychology research, confirmatory factor analysis (CFA) models such as bifactor model (BFM), higher-order model (HFM), and correlated factors model (CFM) are popular models for multidimensional data with several correlated domains. However, it is not always easy to differentiate between these models. Simulation studies have found that BFM tends to fit better by traditional model fit indices (e.g., CFI, TLI, RMSEA, SRMR). Although BFMs fit well, review studies found that they tend to produce problematic solutions with irregular loading patterns. Given these properties, methodologists emphasized the importance of examining the interpretability of loading patterns when evaluating BFMs. However, previous simulation studies mainly focused on model fit, and not on the interpretability of loading patterns. To fill this research gap, in this study, we explored whether alternative models (i.e, CFM, HFM) can produce more interpretable solutions when a BFM solution is problematic. In our study, all three models were fit to population covariance matrices conforming to the BFM structure. We varied three design factors: number of group factors, factor loading sizes, and solution interpretability (i.e., presence of negative or very small loadings). We found that when the data-generating BFM structure is interpretable, CFM and HFM also produce interpretable solutions. However, for almost half of the conditions when the BFM data-generating structure was of low interpretability, CFM and HFM produced more interpretable solutions. We discuss the implications of these results.

Beyond cross-sectional calibration: Evaluating within-person variability in multilevel item response theory

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Siqi Sun (University of Virginia), Dr. Teague Henry (University of Virginia)

Psychological constructs such as mood often fluctuate from day to day, yet traditional item calibration procedures typically rely on cross-sectional data that neglect these within-person variations. This challenge is further compounded by sampling issues related to individual differences in trait levels. This study examines whether using cross-sectional data alone underestimates the true variability of dynamic states and thereby compromises the validity and precision of psychological measures. To address this gap, we apply a multilevel item response theory (IRT) model that incorporates both person-level and time-level variability. We present a simulation study employing a multilevel item response theory (IRT) model to examine (a) how ignoring temporal dynamics and individual differences in trait-level stability biases estimated item parameters, and (b) the required sample size for intensive longitudinal designs to mitigate such bias. We expect that growing temporal dependencies and increasing individual differences will amplify bias in both item parameters and factor scores. Ultimately, this project aims to determine when this bias becomes a problem for accurate measurement of latent constructs and to clarify under what conditions scales should be recalibrated using intensive longitudinal data.

Evaluating the performance of repeated items in longitudinal exams

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Siyu Wan (American Board of Internal Medicine)

In operational testing, item reuse is necessary, but most organizations strive to prevent the same examinee from encountering the same item more than once to avoid potential advantages from recall. The literature on repeated items shows mixed results. Studies in academic and organizational psychology suggest performance improves with repeated exposure (Hausknecht et al., 2007). However, certification and licensure exam research indicate that repeated exposure does not always confer an advantage (Boulet et al., 2003; Raymond et al., 2009; Wood, 2009; Swygert et al., 2010; O'Neil et al., 2015; Fienberg et al., 2015).

This study investigates how item difficulty changes when items are repeated to the same examinees in medical certification exams. Unlike prior studies, examinees in our study did not retake the exam after an initial failure but encountered repeated items due to the exam's longitudinal nature of administration— a format that is becoming increasingly common in modern assessment designs. Moreover, examinees received the correct answers and explanations of most repeated items after their initial attempt.

The study explored: (1) Does item difficulty change significantly when items are repeated to the same examinees? (2) Do changes in difficulty depend on the length of time between administrations? We found that item difficulty significantly decreased when examinees encountered the same items again, regardless of the time between administrations. These results emphasize the impact of repeated exposure in longitudinal assessments and support maintaining designs that limit repeated exposure to ensure assessment fairness.

Exploring suicidal ideation risk in first-year students with zero-inflated model

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Sohee Kim (University of South Alabama), Dr. Yulim Kang (Yonsei university), Dr. katie koo (University of Georgia)

This study investigates the risk factors influencing both the probability and severity of suicidal ideation among first-year college students using a zero-inflated model. Data were collected from 957 first-year students at a large public research university in the Southwestern United States through an 18-item self-report questionnaire assessing suicidal ideation. Additional variables include demographic factors (age, gender, ethnicity, and college GPA), socioeconomic status (SES), educational aspirations, quality of life, self-esteem, spirituality, career satisfaction, social support, and major satisfaction. To analyze these factors, zero-inflated Poisson and zero-inflated negative binomial regression models will be employed, effectively addressing excess zero responses and potential overdispersion in the data. Preliminary expectations suggest that quality of life, self-esteem, and social support may serve as protective factors against suicidal ideation. The educational significance of this study lies in its potential to deepen understanding of the mental health challenges faced by university freshmen, guiding the development of targeted early intervention strategies for at-risk students. Moreover, the findings could provide empirical evidence to inform university-level prevention and intervention programs, ultimately strengthening student support policies and enhancing the effectiveness of mental health education initiatives.

Cross-classified mediation with a between-person predictor

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Prof. Soyoung Kim (Chonnam National University)

Mediation models have received increasing attention in analyzing intensive longitudinal data(ILD). While these models offer valuable insights, certain methodological aspects, such as when the independent variable is at the between-person level and when accounting for autocorrelation in longitudinal data, have received relatively less attention. This study explores approaches for testing mediation, focusing on cross-classified mediation with ILD. This study assumes that the mediator explains the relationship between the independent and dependent variables, where the independent variable is measured at the between-person level, while the mediator and the dependent variable are measured at the within-person level. Different assumptions were considered regarding the within-person level b-path, which represents the effect of the mediator on the outcome variable. Specifically, three models were analyzed. The first model assumes that the b-path is a fixed parameter. The second model allows the b-path to vary over time while remaining constant across individuals. The third model accounts for individual differences by allowing the b-path to vary across individuals while remaining constant over time. By examining these models, this study would contribute to the understanding of various cases about within-person level coefficients when testing mediation in dynamic mediation models.

Bullying Victimization patterns and school climate among middle school students in East Asian cultural circles: Based on latent profile analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Sun Rui (Beijing Normal University)

This study selected the Bullying items (BEINGBULLIED, ST038) in PISA2022 and analyzed the bullying victimization patterns of 15-year-old students in Singapore, Hong Kong, Macau, Taipei, Japan, and South Korea. Latent profile analysis (LPA) was conducted on victimization by verbal, physical and relational bullying. This study also explored the relationship between the different bullying victimization patterns and school atmosphere. Results: (1) The bullying patterns of 15-year-old students can be divided into four patterns: “high physical victimization group” (3.5%), “high verbal victimization group” (4.5%), “high bullying involvement group” (10.8%), and “low victimization group” (81.3%). (2) Different bullying patterns have certain commonality in physical bullying, verbal bullying, and relational bullying. Except for the “low victimization group”, the “high physical victimization group” suffered more severe physical bullying, the “high verbal victimization group” was mainly verbal bullying, and the “high bullying involvement group” suffered more severe bullying in physical, verbal, and relational bullying. (3) Different bullying patterns showed significant differences in gender, grade and country(or region). Boys, Hong Kong and Japanese students were more likely to be involved in cross-bullying and bullying that was mainly physical aggression. (4) Individuals with different patterns of bullying showed heterogeneity in school atmosphere: students in the low victimization group scored higher in school safety and belong, while students in the high verbal victimization group and high bullying involvement group scored higher in school risk. The study provides empirical evidence for school bullying governance in East Asia and recommends the implement the targeted intervention measures.

Sample size imbalance and its impact on measurement invariance

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Teddy Chen (The University of California, Davis), Dr. Mijke Rhemtulla (The University of California, Davis)

Measurement invariance testing is essential for ensuring that constructs are measured equivalently across groups. Researchers typically use goodness-of-fit indices and chi-square difference tests to evaluate invariance, but these methods are known to be sensitive to sample size and group ratio. As the group ratio increases, both chi-square tests and fit indices tend to suggest better model fit, which may lead to misleading conclusions. Previous studies have not systematically disentangled total sample size from group ratio. In this study we investigate the impact of these factors on measurement invariance assessment using Monte Carlo simulations, comparing power for the chi-square test and RMSEA test of not-close fit, as well as average fit index values under both invariant and non-invariant conditions.

Additionally, we examine the effectiveness of a subsampling method proposed by Yoon and Lai (2018) as a potential solution to mitigate the biases of group size imbalance on fit indices. This method involves randomly subsampling from the larger group to match the smaller group's size, creating a balanced group ratio to reduce the bias when invariance does not hold. We compare the subsampling approach to an alternative approach in which fit index calculations are adjusted for the sample size imbalance. We discuss the implications of our findings for researchers conducting measurement invariance tests with unbalanced group sizes.

Mediation analysis using saturated and restricted structures

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Valerie Polad (University of North Carolina at Chapel Hill)

Mediation analysis is widely used to evaluate causal mechanisms, yet the impact of saturated versus restricted model specifications remains understudied. Saturated models freely estimate all possible paths, while restricted models impose theoretically driven parameter constraints. Despite the flexibility of both regression and Structural Equation Modeling (SEM) to incorporate restrictions, applied researchers often default to saturated models—even when theoretical hypotheses do not justify the estimation of all paths. This approach may serve as a safeguard against bias due to model misspecification but also introduces potential drawbacks, including reduced statistical efficiency and lower statistical power for detecting specific indirect effects. This study examines the implications of model specification choices for a common serial mediation structure through a Monte Carlo simulation, assessing parameter estimation, standard errors, confidence intervals, and statistical power across varying sample sizes and effect size magnitudes. Preliminary findings suggest that model specification meaningfully impacts the accuracy and efficiency of indirect effect estimates, underscoring potential tradeoffs in common analytic practices. Theoretical and practical considerations for aligning model specification with hypothesized mechanisms are discussed.

Assessing the utility of metacognitive wrapper data in early grade prediction for student assessment

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Walton Ferguson (University of Notre Dame), Dr. Ying Cheng (University of Notre Dame), Mr. Thomas Joyce (University of North Carolina at Chapel Hill), Dr. Alex Ambrose (University of Notre Dame), Dr. Shawn Miller (University of Notre Dame)

Predictive analytics in educational settings is often concerned with making accurate predictions of a student's competence as quickly as possible to enable timely interventions. Traditionally, researchers use graded assessment data when making such predictions. However, metacognitive data–a student's individual appraisal of their performance or related aspects–may also help predict student learning outcomes. Using a large sample from an organic chemistry course at a private midwestern university, this study aims to use both course assessment and metacognitive data to predict students' final grade in a large, 15-week organic chemistry course (N = 423) at weeks 4, 6, and 7. Methods include six different classification algorithms each with three different types of data balancing (ROS, SMOTE, and no balancing). To compare models, we collected accuracy rates and quadratic weighted kappa metrics (a type of balanced accuracy; QWK) Support Vector Machines, Proportional Odds, and Ordinal Forest typically showed the best performance of the six tested models. Results indicate that course data alone was able to predict final course grades with up to 80% accuracy at week 4 (QWK = 0.77), which increased to 81% by week 7 (QWK = 0.78). However, including metacognitive data as predictors does not appear to increase performance accuracy or model QWK. This study extends previous research by demonstrating the possibility of accurate early grade prediction using real classroom data. Although the metacognitive predictors did not aid in prediction accuracy, they provide diagnostic value for instructors looking to intervene in students that are predicted to perform poorly.

Youth's perspective of facility climate in restrictive education settings: A latent class analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Prof. Wenchao Ma (University of Minnesota), Dr. Kristine Jolivette (University of Alabama), Dr. Olivia Hester (University of Alabama), Dr. Sara Sanders (University of Alabama), Dr. Stephanie Shelton (University of North Carolina at Chapel Hill), Dr. June Preast (University of Alabama), Dr. Allyson Pitzel (University of Alabama), Dr. Kimberly Odom (University of Alabama), Dr. Nicole Prewitt (University of Alabama)

Some restrictive education settings (e.g., community school programs, residential treatment facilities, juvenile justice facilities) have increasingly shifted away from reactive, punitive approaches toward multi-tiered systems of support to enhance facility climate. Despite this transition, the perspectives of youth—key stakeholders in these settings—remain less heard. This study uses latent class analysis, a person-centered method, to investigate how youth perceive their facility climate based on survey data.

The analysis draws on responses from 668 youth across 25 restrictive education settings nationwide and examines their views on five critical tenets of facility climate: behavioral expectations, acknowledgment, discipline, safety, and respect. For each tenet, one or two focal survey questions were selected for analysis. Latent class models ranging from two to five classes were tested. The four-class model is shown to fit data best based on the bootstrapped chi-squared statistics. These four classes highlight distinct youth perspectives. For instance, approximately 45% of youth in Class 1 reported clear understanding of expectations, received praise from staff, and perceived the facility as safe with fair discipline. In contrast, about 20% of youth in another class understood expectations but reported little praise or reinforcement as part of the facility climate. Additional analyses will be conducted to explore how demographic factors influence class membership. By identifying these distinct youth perceptions and their association with demographic factors, this study offers deeper insights into group-level variations in climate efforts and has implications for tailoring interventions to better align with youth perspectives and improving their overall facility climate.

Integrating anchor selection strategies with dual-scale purification for DIF detection

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Ya-Hui Su (National Chung Cheng University), Ms. Wei-Tzu Wu (National Chung Cheng University)

Differential item functioning (DIF) refers that examinees with the same ability level perform differently on a given item due to belonging to different groups, thereby compromising test fairness. Before conducting a DIF analysis, it is necessary to place the groups on a common scale for comparison. Wang (2004) reviewed methods for establishing a common scale, and Wang (2008) proposed the DIF-free-then-DIF (DFTD) procedure, which did not guarantee the selected anchor items were DIF-free. To address this issue, Chen and Hwu (2018) introduced the dual-scale purification (DSP) procedure, adding a step to detect the anchor items for DIF during the DFTD procedure. In practice, various anchor strategies were available for researchers. Kopf et al. (2015a) proposed anchor strategies, such as the mean test statistic threshold (MTT) and mean p-value threshold (MPT), which could be combined with the DFTD procedure for DIF detection. Their findings indicated that these strategies generally controlled Type I error rates well and improved statistical power. However, when DIF entirely favored the reference group and the DIF proportion reached 40%, the first and second-ranked anchor item candidates were only 80% and 90% likely, respectively, to be truly DIF-free. This limitation increased the risk of including DIF items in the anchor set, potentially leading to failure in establishing a common scale and inflating Type I error rates. Therefore, this study aimed to investigate the effectiveness of various anchor strategies combined with the DSP procedure in DIF detection.

Evaluating model fit in confirmatory factor analysis for cross-classified data

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Yen Lee (Uniformed Service University)

Multilevel data structures are common in social and behavioral research (Bryk & Raudenbush, 1992). These structures may be hierarchical, where lower-level units belong to a single higher-level unit, or cross-classified, where lower-level units belong to multiple higher-level units simultaneously. Hierarchical linear models (HLMs) address dependencies in hierarchical data, while generalized linear mixed models (GLMMs) handle dependencies in cross-classified data (CCd).

In applied research, cross-classification is often overlooked, leading to biased variance estimates and standard errors (Ye & Daniel, 2017). To account for latent structures in cross-classified data, GLMMs have been extended through two R packages: **PLMixed** (Jeon & Rockwood, 2019) and **galamm** (Sørensen, 2024), which implement profile maximum likelihood and maximum marginal likelihood estimation, respectively. Although these methods can be applied in confirmatory factor analysis (CFA), model fit evaluation for cross-classified CFA models remains underexplored.

This study investigates the performance of traditional fit indices and explores appropriate cut-off criteria for CFA when estimated using **PLMixed** and **galamm**. Simulations vary factor loadings (0.3, 0.6, 0.8), model structures (two factors with 6 or 12 items; one factor with 3 or 6 items), random effect variances (1, 3), and sample sizes (100, 200, 500, 1000). Results will offer practical guidelines for evaluating model fit in CCREMs with latent structures, supporting more accurate model specification in applied research.

Detecting multidimensionality that causes differential item functioning: A simulation study

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Yijun Cheng (University of Washington), Dr. Chun Wang (University of Washington)

Although it has long been known that differential item functioning (DIF) can be caused by multidimensionality, and it is the recommended to conduct dimensionality check before DIF evaluation. However, no study has explored whether the state-of-art dimensionality analysis methods (e.g., parallel analysis, exploratory item factor analysis) can really detect multidimensionality causing DIF. This study addresses this gap by providing evidence-based guidelines on when dimensionality checks are effective. We constructed scenarios where the focal and reference groups have equal ability on the primary dimension (i.e., target construct of interest) but differ on one or multiple secondary dimensions (benign or nuisance). DIF items loaded on secondary factors to simulate the impact of multidimensionality. Then we use parallel analysis and model fit indices from exploratory item factor analysis (IFA) such as AIC and BIC to evaluate (1) whether more than one dimension is detected and (2) whether the correct number and loading structure can be discovered. Results demonstrate that IFA can effectively identify one or multiple secondary dimensions. Identification requires at least two (with equal loadings) or three indicators per factor for the factor to be identified. Additionally, we found that the number of indicators, sample size, correlation between the latent factors, and proportion of students in the focal group to those in the reference group all important influenced factors discovery. In conclusion, this study provides comprehensive simulation evidence to support a robust guideline on whether, and if so when, one should use dimensionality checks as the first step in DIF analysis.

The effects of within- and between-person reliability on the performance of single-indicator multilevel MEAR(1)

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. You Kyoung Hwang (Sogang University), Mrs. Young Soo Lee (Sogang University), Prof. Hye Won Suk (Sogang University)

The single-indicator multilevel MEAR(1) model, based on the Dynamic Structural Equation Modeling (DSEM) framework, extends first-order autoregressive (AR(1)) modeling by incorporating measurement error in intensive longitudinal data (ILD) collected with a single indicator. Traditional DSEM typically assumes no measurement error, implying perfect reliability. However, this assumption is often unrealistic in psychological research, and ignoring measurement error can lead to biased parameter estimates and misleading conclusions.

While several approaches exist for calculating reliability in ILD, a consensus has emerged that within-person reliability and between-person reliability should be evaluated separately. Despite this, no systematic simulation studies have examined the separate or combined effects of within-person reliability and between-person reliability on the performance of DSEM. Moreover, few simulation studies have investigated single-item ILD reliability, even though single-item measures are often used in ILD to reduce participant burden.

This study aims to systematically examine the effects of within-person and between-person reliability on the estimation accuracy of autoregressive effect within the single-indicator multilevel MEAR(1) model. Using simulations, we evaluate how varying reliability levels affect key performance metrics, including bias, mean squared error (MSE), coverage rates, and statistical power (or Type I error rates). Furthermore, we compare the performance of the MEAR(1) model to that of a traditional AR(1) model under different reliability conditions.

Penalized Gaussian maximum likelihood of latent structural vector autoregressive models for high-dimensional discrete-valued time series

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Younghoon Kim (Cornell Univesity)

In psychology and neuroscience, discrete-valued, both count and ordinal, intensive longitudinal data is widely used to examine time-dependent relationships between variables. Despite the abundance of such datasets, most modeling approaches rely on dynamic structural equation models, state-space models, or point process models. However, the former is challenging to extend to high-dimensional settings, while the latter is often not able to capture interpretable network structures. Furthermore, both approaches lack flexibility in modeling the marginal distributions of discrete-valued observations. We propose a high-dimensional discrete-valued network time series model, where the reduced form latent structural vector autoregressive (SVAR) model is assumed to have sparse transition matrices and a sparse inverse covariance matrix. The estimators of these matrices, obtained through simultaneous multivariate linear regressions, capture directed temporal dependence by examining Granger causality and undirected contemporaneous dependence by computing partial correlations. The second-order properties of this latent SVAR model and the observed series with any marginal distributions are connected through the link functions. We conduct simulation studies to assess estimation and forecasting performance against renowned benchmarks. We apply this model to multimodal behavioral data in developmental psychology to improve forecasting and to spike train data in brain connectomes to investigate neural network structures.

A new IRT model for correcting social desirability bias in VAS

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Youxiang Jiang (Beijing Normal University), Prof. Hongbo Wen (Beijing Normal University)

As self-reported measurements become increasingly digitalized, the Visual Analogue Scale (VAS), originally used to measure pain, has been applied more widely in evaluating psychological constructs such as attitudes, personality, and anxiety. Social desirability bias, a response bias influenced by the participant's tendency to present themselves in a favorable light, is prevalent in self-reported measurements and undermines the reliability and validity of such measures. However, previous research has primarily focused on correcting response bias in discrete data (e.g., Likert scale responses) while overlooking bias correction in continuous bounded response data (e.g., VAS responses). This study develops a new Item Response Theory (IRT) model capable of correcting social desirability bias in continuous bounded response data, thereby offering more accurate estimates of individual characteristics. The study employs empirical data to test the fit of the new model and the validity of its estimation of participants' social desirability tendencies. The results indicate that the new model fits the data better than models that do not account for social desirability bias and can effectively identify participants' social desirability tendencies. A simulation study confirmed the accuracy of parameter estimation across varying sample sizes and test lengths. In summary, this study is the first to achieve statistical control of social desirability bias in continuous bounded response data from VAS (or similar measures, such as slider bars) within the IRT framework.

A new method for detecting DTF using sum scores

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Dr. Yutaro Sakamoto (Recruit Management Solutions Co., Ltd.), Dr. Ryuichi Kumagai (Tohoku University)

When conducting differential item functioning (DIF) analyses in test development, item-level differences and differential test functioning (DTF) must be examined. With the renewed focus on the practical utility and interpretability of sum scores (Sijtsma, Ellis, & Borsboom, 2024), the need for effective DTF detection using these scores has increased. Current DTF detection methods rely on item response theory (IRT), leaving a gap in techniques that utilize sum scores.

This study presents a new method for detecting DTF through sum scores. When two examinee groups are to be compared for DTF, first, a DIF analysis is conducted to identify non-DIF items designated as anchor items. Second, the sum score for the anchor items is calculated, and examinees are stratified by the sum scores. Within each stratum, the mean difference in overall test sum scores between the two groups is calculated, considering signed and unsigned values. This difference is weighted according to the proportion of examinees in each stratum, and the final DTF index, interpreted as an effect size, is derived by summing these weighted values across all strata.

The simulation studies demonstrated that the proposed DTF index correlates approximately 0.8 with an existing DTF index (Meade, 2010), affirming its validity. These findings suggest that DTF can be detected without the reliance on complex psychometric models like IRT. Furthermore, the proposed method is straightforward and requires only basic arithmetic operations, making it a practical tool for test developers and practitioners.

The effect of item difficulty on credit scores: Integrating speed and accuracy in a cognitive performance measure

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Zarah Chaudhary (Multi-Health Systems Inc.)

Psychometric models often assume a uniform relationship between speed and accuracy. In performance-based assessments where scoring integrates both speed and accuracy components, understanding how test-level information, such as item difficulty, interacts with test-taker performance can offer insights into test-taker strategies and cognitive processes. This study seeks to model the systematic relationship between item difficulty and composite credit scores that reward higher speed and accuracy.

A representative sample of 938 participants (F:50%, M:50%) from the U.S. completed a 30-item Flexibility of Closure task that measures the ability to recognize familiar patterns within a complex background. Hierarchical mixed-effects statistical modeling will be employed to investigate whether difficulty significantly predicts Flexibility of Closure credit scores accounting for both within-person and between-item variability. Specifically, whether i) higher difficulty items lead to systematically lower credit scores, ii) whether test-takers with different ability levels have different patterns of score changes with increased difficulty and iii) whether high and low-performing individuals show evidence of employing different strategies (e.g. high performers adjusting by slowing down to prioritize accuracy). To capture individual differences in response adaptation, random effects for individuals and items will be incorporated where appropriate, to understand speed-accuracy trade-offs under varying difficulty levels.

The findings on how item complexity interacts with test-taker behaviour have important implications for adaptive test design and the evaluation of cognitive performance measures with time-sensitive scoring protocols. A clearer understanding of the speed-accuracy relationship in cognitive tasks is essential for developing scoring algorithms that enhance score reliability and support valid score interpretations.

Understanding spatial cognitive processes on a digital numeracy item: An n-gram based machine learning approach

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Yiming Chen (Georgetown University), Dr. Qiwei (Britt) He (Georgetown University)

Technological advances have transformed educational assessments, particularly in how we collect and analyze data about individuals' cognitive processes and interactions with assessment items. The rich data recorded in log files throughout human-machine interactions, often referred to as process data, offers new insights into problem solving behavior. This study focuses on exploring how adult respondents transform geographical knowledge into spatial recognition on an interactive numerical item ("Map") from the 2012 PIAAC by using sequential process data. The objective of this study is two-fold: (1) to explore different behavioral patterns that distinguish success or failure on a spatial numerical item, and (2) to identify spatial cognitive process features across high and low numeracy performance levels. Using a sample of 596 U.S. adult respondents, we employed the time-embedding n-grams method and two machine learning methods, random forest and XGBoost, to predict respondents' numeracy performance and to identify robust classifiers. Results indicate that time-related features are the most predictive of adults' numeracy skills. The findings highlight the potential of process data to support analyses of latent numeracy skills and cognitive processes in educational contexts.

Applying CoxNet onto educational sequential process data

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Binhui Chen (Georgetown University), Dr. Qiwei (Britt) He (Georgetown University)

Machine learning (ML) methods are increasingly applied across disciplines but remain underutilized in the social sciences, where traditional parametric models often fall short in handling high-dimensional, censored, and temporally structured data. A novel approach is proposed in this study to leverage CoxNet, a deep learning survival analysis framework, to model sequential behavioral data from the Problem Solving in Technology-Rich Environments (PSTRE) component of the PIAAC assessment. Using time-stamped user interaction sequences from 14 digital tasks, we predict binary task outcomes (correct vs. incorrect). Our approach yields exceptionally high predictive accuracy. Moreover, we evaluate three input strategies for representing behavioral sequences: (1) action embeddings with time interval features, (2) a weighted sparse matrix incorporating time-on-task, and (3) an unweighted sparse matrix with elapsed time. Among these, the action embedding representation yields the highest predictive accuracy, highlighting the importance of input design in sequence modeling. This work advances the application of survival models in process data analysis and sets the stage for future improvements in model architecture and interpretability, particularly for domains requiring insight into behavioral sequences.

The impact of latent confounder distribution statistics on direction dependence analysis

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Guyin Zhang (University of South Carolina), Dr. Dexin Shi (University of South Carolina)

Determining the causal direction between two variables X and Y—whether X causes Y (x → y) or vice versa (y → x)—is a fundamental question in psychological research. Direction Dependence Analysis (DDA) is a statistical tool that leverages higher-moment information to detect directional model misspecifications in observational data. Previous research has primarily focused on distinguishing the correct causal model from a mis-specified alternative under the assumption that no hidden confounders influence the relationship. However, in real-world data, latent common causes are prevalent and can distort causal inferences. This study examines how the presence of hidden confounders affects the performance of DDA and identifies the distributional assumptions necessary for maintaining its validity. Specifically, we investigate the conditions under which causal model selection remains reliable when using higher than second moments (skewness and kurtosis) DDA measures of observed variables and residual terms. Monte Carlo simulations systematically evaluate the impact of different confounder distributions on causal inference accuracy. Our findings highlight the strengths and limitations of DDA in the presence of latent confounders and provide recommendations for its use in psychological research.

A multi-perspective analysis of retracted education publications

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Mr. Qian Shen (University of Florida), Mr. Yilin Zhang (University of Florida), Ms. Xinyi Tai (University of Florida)

As a prominent example of academic misconduct, the increasing number of retracted publications has sparked widespread concern and reflection both within academia and beyond. Current studies on retracted publications have primarily focused on retractions about biomedicine, while retractions in educational research have received relatively little attention.

Using the publicly available publication retraction database from Watch Retraction, we selected 4,759 retracted publications related to education, all published in the 21st century and retracted as of February 4, 2025. We applied topic modeling using pre-trained models such as RoBERTa and Gemini 2.0, along with the Apriori algorithm and other analytical methods, to examine the data associated with these publications.

Based on multiple topic modeling results, we found that studies related to arts education, ideological education, fundamental subject education, and educational technology accounted for the majority of retracted publications in educational research. The number of retracted publications in educational research peaked in 2010 and 2011, significantly exceeding the surrounding years, and has been steadily increasing since 2020. Retracted publications from IEEE and Hindawi far outnumbered those from other publishers. Investigations by journals or publishers, unreliable research results, and scrutiny from peer reviewers and third-party institutions were the primary reasons for these retractions. The results of the Apriori algorithm indicate that many retractions in educational research were triggered by third-party investigations uncovering unreliable findings, which subsequently led to further investigations by journals and publishers, ultimately resulting in the retraction of these academic publications.

Unravelling the pattern complexity between youth depression and risk behaviors

Tuesday, 15th July - 17:15: Welcome Remarks and Poster Session (Memorial Hall of McNamara Center) - Poster

Ms. Jessica Sun (Children's Mercy), Dr. Ayanda Chakawa (Children's Mercy)

This study examines the multifaceted relationships between depression and a broad spectrum of 9 risk behaviors among high school students, utilizing psychometric measurement and advanced statistics to examine a national sample of nearly 1.5 million observations. The risk behavior categories include unsafe driving, weapon-related behaviors, physical fights, suicidality, smoking, alcohol consumption, substance abuse, unhealthy eating, and physical inactivity – based on 50 individual indicators. Statistically, we employed normalized association tests to ensure comparability across diverse demographic groups by adjusting for differences in sample sizes. Complementing these tests, logistic regression models were used to further validate the strength and direction of the associations. In addition, a bootstrapping approach was applied to assess the stability of the results across subgroups defined by gender, grade level, and ethnicity. We also implemented multidimensional mapping techniques based on unsupervised machine learning to visualize the intersectional patterns among student subgroups, revealing how depression differentially influences risk behaviors in diverse populations. The results of our novel multidimensional framework underscore the heterogeneity of depression's impact, with certain risk behaviors showing stronger associations in specific demographic segments. These insights are critical for informing targeted mental health strategies and interventions. Actionable implications from our findings will be discussed on how educators, policymakers, and mental health practitioners can implement data-driven programs that mitigate risk behaviors and improve student well-being across diverse student populations. Recommendations will also be discussed on generalizing the statistical framework from this study to improve knowledge and practice implications beyond depression indicators to other areas of psychological study.

Authors Index

IMPS 2025

University of Minnesota • Minneapolis, MN, USA
July 15-18, 2025 • Short Courses July 14