39th IOPS Winter Conference 20-21 December 2023

Conference: University of Amsterdam
Roeterseilandcampus
Building M (Amsterdam Business School)
Room M 1.03
Plantage Muidergracht 12

See website for map and directions

Dinner: **Restaurant Olijfje**
Valkenburgerstraat 223D, 1011 MJ Amsterdam

Wednesday 20 December (Room M1.03)

12.00 – 13.00 **Lunch**

13.00 – 13.05 **Official Opening by Rob Meijer and welcome by local organizer**

13.05 – 13.35 **Presentation Sophie Berkhout** (Utrecht University)
What happens during sleep? Theoretical implications of modeling night gaps in ESM data
Discussants: Ria Hoekstra, Anna Langener

13.35 – 14.05 **Presentation Martijn Schoenmakers** (Tilburg University)
A Polytomous Extension of the Clustering Approach to Uniform Differential Item Functioning
Discussants: Hadi Mohammadi, Andres Perez Alonso

14.05 – 14.35 **Presentation Jason Nak** (University of Amsterdam)
Phenomena: Bridging the gap between data and theory
Discussants: Marie Stadel, Hanne Oberman

14.35 – 14.55 **Break**

14.55 – 15.25 **Presentation Camila Natalia Barragán Ibáñez** (Utrecht University)
Sample Size Determination for Cluster Randomized Trials Using Bayes Factors
Discussants: Guiseppe Arena, Nikola Sekulovski

15.25 – 15.55 **Presentation Chuenjai Sukpan** (Utrecht University)
Selecting the correct RI-CLPM using Chi square-type tests and AIC-type criteria
Discussants: Joris Mulder

15.55 – 16.25 **Presentation Anna Langener** (Groningen University)
Challenges in Using Passive (Smartphone) Measures: Proposing a Preregistration Template to Move Forward
Discussants: Ria Hoekstra, Chuenjai Sukpan

16.25 – 16.45 **Break**

16.45 – 17.30 **Invited speaker Reinoud Stoel** (CBS, Statistics Netherlands)
Life after a PhD

17.30 – 18.25 **Poster Session with drinks and snacks**

Hadi Mohammadi (Utrecht University)
Explainable AI-Generated Text Detection
Hannah Heister (Groningen University)
Bayesian IRT Based Continuous Norming
Klazien de Vries (University of Groningen)
Adjusting for non-representativeness in normative samples: using multilevel regression and post-stratification in continuous norming.
Sara Keetelaar (University of Amsterdam)
Are psychological networks sparse or dense? A Bayesian test for sparsity.
Şeyma Nur Ertekin (University of Amsterdam)
Playing with Memory: Do Kids Remember Differently Than Adults?
Adam Finnemann
Building theories in psychology using statistical mechanics

18.30 – 19.30 **Social event (Amsterdam light tour walk)**

19.30 – 22.00 **Conference dinner**

Thursday 21 December (Room M1.03)

09.45 – 10.15 **Receipt with coffee and tea**

10.15 – 10.45 **Presentation Ria Hoekstra** (University of Amsterdam)
Testing for similarity between idiographic networks: The Individual Network Invariance Test (INIT)
Discussants: Sara Keetelaar, Hanne Oberman

10.45 – 11.15 **Presentation Karel Veldkamp** (University of Amsterdam)
Deep learning based item response theory for missing data
Discussants: Thom Volker, Kevin Kloos

11.15 – 11.45 **PhD Meeting (Room M1.03)**

11.45 – 12.45 **Lunch**

12.45 – 13.15 **Presentation Reeta Kankaanpää** (University of Turku)
Factor scores over sum scores: worth the trouble? The effect of measurement model choice on the effectiveness of an intervention
Discussants: Thom Volker, Karel Veldkamp

13.15 – 13.45 **Presentation Thom Volker** (Utrecht University)
Density ratios to evaluate and improve the utility of synthetic data
Discussants: Zeynep Bilici, Nikola Sekulovski

13.45 – 13:55 **Time to vote for the best presentations!**
14.00 **Closing and Best Presentation and Poster Awards**

Best presentations

Please use the links or scan the QR code in order to bring out your vote.

Poster Presentation

SCAN ME

Oral Presentation

SCAN ME

Abstracts

What happens during sleep? Theoretical implications of modeling night gaps in ESM data

Sophie Berkhout (Utrecht University)

The experience sampling method (ESM) is a popular technique for collecting intensive longitudinal data by measuring a person multiple times a day over several days to capture their real-time experiences. Often, researchers analyze ESM data using dynamic models that have assumptions such as stationarity and equally spaced time intervals. Unfortunately, ESM data violates the assumption of equally spaced intervals in two ways: first, the timing of the measurements follows a semi-random sampling scheme; second, the measurements rely on self-report leading to gaps in the data when participants are asleep. So far, researchers have applied several modeling approaches to deal with the problem of unequally spaced intervals. In this presentation, I focus specifically on the night gap and discuss the theoretical implications of currently applied approaches using the simple case of an AR(1) model. Furthermore, I propose an alternative model and show that the former approaches are nested within this model, which allows researchers to easily test which of the approaches is most supported by the data. Finally, I demonstrate with an empirical example how this helps researchers make well-informed modeling decisions.

A Polytomous Extension of the Clustering Approach to Uniform Differential Item Functioning

Martijn Schoenmakers (Tilburg University)

To objectively compare different groups on any latent trait using tests, the absence of differential item functioning (DIF) is crucial. While the importance of DIF has been well-established in the psychometric literature, the question of how to adequately select DIF-free items is still largely open, with many different approaches being proposed. The fact that the difficulty of an item is not identified from the observations alone may be a reason no widely agreed upon approach to DIF testing has been developed. Recently, DIF tests utilizing the differences between item difficulties across groups, which are identified, were proposed for the Rasch (Bechger & Maris, 2015) and 2-parameter logistic models (Pohl et al., 2021). The current paper aims to extend this clustering approach to the polytomous case using the partial credit model. To achieve this, the clustering approach to DIF in a polytomous item is split into two steps. First, the distances between the item thresholds within an item are compared across groups using a multivariate Wald test. When items lack equidistant thresholds, these items are classified as DIF items. The other remaining items are clustered on the differences between a single item threshold across groups. Due to previously disregarding the items lacking equidistant item thresholds, this is equivalent to clustering the remaining items on differences between all thresholds across groups. Performance of the new approach is assessed using a simulation study and practical recommendations are made.

Phenomena: Bridging the gap between data and theory

Jason Nak (University of Amsterdam)

In the aftermath of the replication crisis, many efforts have focussed on combatting QRPs, publication bias and other reproducibility initiatives with genuine methodological advancement as a result. The development of stronger theory however, has not seen many great methodological reforms, even though the weakness of theory in psychology is a definite contributor to the replication crisis. Weak theory leads to ambiguous hypotheses, dependent on unclear auxiliary hypotheses, that are easily explained away when non-significant results appear. This undermines the hypothetico-deductive method as it weakens our ability to falsify, both consciously and unconsciously. We therefore propose an addition to our conceptual repertoire in the form of phenomena, stable and general features of the world. These are transcendent over data and do not suffer from the idiosyncrasies that single samples bring with them, but they are also not explanations of data, they merely serve as known facts to be explained by theory. We believe that these phenomena will strengthen theory by separating two inferential streams, one from data to phenomena, and one from phenomena to theory. When theory no longer aims to explain single datasets but robust, general patterns found across data, this will create stronger explanatory theory from which new phenomena may be implied.

Sample Size Determination for Cluster Randomized Trials Using Bayes Factors

Camila Natalia Barragán Ibáñez (Utrecht University)

In the initial phases of designing a research study, a critical step is the determining the sample size. Employing small sample sizes may lead to underpowered studies. However, considering the limitations in resources and that it may be unethical to involve more participants than necessary, it is unrealistic to expect from researchers to use a large number of participants. To ensure that a study possesses a sufficient number of participants to secure the statistical power, researchers can employ a prior power analysis. Determining the sample size in complex research designs such as cluster randomised trials becomes considerably intricate due to the hierarchical structure of the data, meaning that the sample size at each level needs to be determined. Conventionally, the sample size for this type of design is based on null hypothesis significance testing, an approach known to have multiple pitfalls, which can be avoided by using Bayes factors instead. While methods have been proposed in past studies for determining sample size when using Bayes factors, these are limited to trials without a multilevel structure, making them unsuitable for cluster randomised trials. In this study, we present a method to determine the required sample size for one-period two-treatment parallel cluster randomised trial when using approximated adjusted fractional Bayes factors for hypothesis testing. We implement this method in an R package and provide explanation on how to use this tool for sample size determination in a one-period parallel-group design. Simulation results show that the required sample size increase with decreasing effect sizes and with increasing intraclass correlation and Bayes factors. We encourage researchers to use our methodology when planning a cluster randomised trial, utilizing Bayes factor for hypothesis testing.

Selecting the correct RI-CLPM using Chi square-type tests and AIC-type criteria

Chuenjai Sukpan (Utrecht University)

In the field of behavioral and social science, the random intercept cross-lagged panel model (RI-CLPM) is increasingly gaining popularity among researchers. However, a challenge is the selection of the appropriate RI-CLPM type, more specifically, the selection of the number of random intercepts (RIs). This study aims to address this concern by comparing four techniques: the Chi-square difference test, the Chi-bar-square difference test, Akaike's information criterion (AIC), and the new AIC-based criterion known as Generalized Order-Restricted Information Criterion Approximation (GORICA). The results demonstrate the effectiveness of each technique in selecting the correct model under various conditions. The findings indicate that the performance of GORICA surpasses AIC, Chi-square difference test, and Chi-bar-square difference test in selecting the optimal number of RIs in a RI-CLPM.

Challenges in Using Passive (Smartphone) Measures: Proposing a Preregistration Template to Move Forward

Anna Langener (Groningen University)

Passive smartphone measures hold significant potential and are increasingly used in psychological and biomedical research to capture an individual's behavior. However, utilizing these measures presents methodological challenges during both data collection and analysis. Researchers are faced with multiple decisions when working with such measures, which can result in different conclusions. In this talk, I will delve into these methodological challenges associated with working with passive measures and propose to adopt preregistration to enhance transparency and reproducibility in digital phenotyping studies. I will share preliminary results of a preregistration template developed for digital phenotyping studies that guides researchers in making informed decisions ahead of time.

Testing for similarity between idiographic networks: The Individual Network Invariance Test (INIT)

Ria Hoekstra (University of Amsterdam)

In many applied settings, the task of comparing idiographic network structures has been a challenging endeavor. Previously, researchers resorted to methods such as eyeballing the estimated network structures, computing correlations between them, or deploying techniques that make use of the multilevel structure of the data like GIMME and mlVAR. While these methods have their benefits, they fall short in one crucial aspect: they lack the capability to directly test the (in)equality of idiographic network structures. In this talk, the Individual Network Invariance Test (INIT) will be presented. INIT extends model comparison practices of Structural Equation Modeling (SEM) into the realm of idiographic network analysis. The performance of INIT on both saturated (i.e., fully connected) and pruned (i.e., some of the matrix elements have been set to zero) idiographic network structures was evaluated in a simulation study. Results indicated that INIT performs adequately when t = 100 per individual. The possibilities of this new technique will be illustrated, and practical recommendations will be provided, highlighting how INIT allows testing not just for (in)equality between idiographic network structures but also within idiographic network structures.

Deep learning based item response theory for missing data

Karel Veldkamp (University of Amsterdam)

Recently Variational Autoencoders (VAEs) have been proposed as a method to estimate high dimensional Item Response Theory (IRT) models on large datasets. Although these improve the efficiency of estimation drastically compared to traditional methods, they have no natural way to deal with missing values. In this paper, we adapt three existing methods from the VAE literature to the IRT setting, and propose one new method. We compare performance of the different VAE based methods to each other and to marginal maximum likelihood for increasing levels of missing data in a simulation study for both three- and ten-dimensional IRT models. Additionally, we demonstrate the use of the VAE based models on an existing algebra test dataset. Results confirm that VAE based methods are a time efficient alternative to marginal maximum likelihood, but that a larger number of importance weighted samples are needed when the proportion of missing values is large

Factor scores over sum scores: worth the trouble? The effect of measurement model choice on the effectiveness of an intervention

Reeta Kankaanpää (University of Turku)

Randomized control trials (RCTs) are often used in psychology to make causal claims about the effects of interventions on adolescents' psychological wellbeing. An investigation of the psychometric properties of the assessment tools precedes the effectiveness evaluation. The properties are typically investigated using factor analytic techniques modeling the theoretical constructs as latent variables. Ideally, intervention effects would be evaluated using the latent variable model that was deemed satisfactory in the psychometric evaluation. In practice, structural models involving latent variables and testing intervention effects are often too complex and have technical issues. Instead, intervention effects are estimated using sum scores where items are simply summed together. Recent studies have shown that sum scores may yield biased estimates (McNeish & Wolf 2020; Kankaanpää et al., 2023), and that latent variable -based factor scores could produce much better estimates for intervention effects (Soland 2022; Kuhfeld & Soland 2022). However, less is known about the superiority of factor scores over sum scores in the case where the measures might contain systematic error in addition to random error. This study will investigate the effect of scoring method on intervention effects using simulation and empirical analysis. We will evaluate two scenarios: with and without a possible systematic error, termed as method-related effect. First, with simulated data, we will compare the estimates from factor scores and sum scores to estimates from a latent variable model when testing the intervention effect. Second, with empirical data, we will compare the estimates from factor scores and sum scores with a real school intervention study. This study will guide researchers conducting RCTs whether to replace sum scores with factor scores in future studies. Considering the importance of the randomized controlled designs in decision-making, we need more research evidence to help us decide what is the most appropriate statistical measurement model for evaluating interventions.

Density ratios to evaluate and improve the utility of synthetic data

Thom Volker (Utrecht University)

Synthetic data is an increasingly popular solution to deal with disclosure risks. Rather than publishing the observed data, a simulated dataset that is statistically similar to the observed data can be released to protect the privacy of the respondents. However, generating high quality synthetic data is a challenging task that typically evolves in a cyclical manner, in which model adjustments and quality evaluations are performed iteratively. To ease this process, we introduce a novel framework to evaluate the utility of synthetic data based on direct density ratio estimation. We show how the estimated density ratios can be used to evaluate the utility of the synthetic data, and how these methods can be used directly to improve the utility of synthetic data.