Programme IOPS Winter Conference - Utrecht 2025

Thursday, 4 December, Booth zaal

10:30 – 11:00 **Reception with coffee and tea**

11:00 – 11:15 **Official opening** by Joris Mulder and welcome by Mirjam Moerbeek

11:15 – 11:45 **Networking event**

11:45 – 12:15 **Presentation Nikola Sekulovski** (University of Amsterdam)
LLM-Based Prior Elicitation for Bayesian Graphical Modeling
Discussants: Vira Dvoriak & Hadi Mohammadi

12:15 – 13:15 **Lunch**

13:15 – 13:45 **Presentation Zhiwei Dou** (KU Leuven)
Post-selection inference in multiple linear regressions: the double dipping dilemma in psychological research
Discussants: Tra Le & Iris Willigers

13:45 – 14:15 **Presentation Wong Tsz Keung** (Tilburg University)
BayesPower: A General Application of Power and Sample Size Calculation for the Bayes Factor
Discussants: Cas Goos & Camila Barragan Ibañez

14:15 – 14:45 **Presentation Hadi Mohammadi** (Utrecht University)
EvalMORAAL: Interpretable Chain-of-Thought and LLM-as-Judge Evaluation for Moral Alignment in Large Language Models
Discussants: Tijn van Hoesel & Şeyma Nur Ertekin

14:45 – 15:00 **Break**

15:00 – 15:45 **Keynote by Ellen Hamaker** (Utrecht University)
Bridging disciplinary divides in statistics and causal inference: How to & why

15:45 – 16:45 **Poster presentations & drinks**

Mohammad Behbahani (Utrecht University)
Exploring hidden patterns in relational event history data: an extension of Hidden Markov Model (HMM) for the Relational Event Model (REM)
Thijs Carrière (Utrecht University)
The relation between participant agency and willingness to participate in app-based data collection
Luke Korthals (University of Amsterdam)
AI Grading & Feedback Insights from a programming course
Diana Kolesnikova (Tilburg University)
How difficult is 2+2? Simulating expert behavior with LLMs for difficulty estimation. Angoff-like vs Pairwise procedures
Jason Nak (University of Amsterdam)
Phenomenal Foundations of Excellent Explanations
Franziska Rüffer (Tilburg University)
Correcting moderator effects in meta-analysis for publication bias
Pepijn Vink (Utrecht University)
Evaluating the Subject-Specific Effects in the Multilevel Hidden Markov Model: a Simulation Study

16:45 – 17:45 **Networking event with IOPS alumni**

18:30 – 22:00 **Conference diner – Grand Café Living, Weg tot de Wetenschap 400, Utrecht**

Friday, 5 December, Booth zaal

10:00 – 10:30 **Reception with coffee and tea**

10:30 – 11:00 **Presentation Tzu-Yao Lin** (KU Leuven)
Single-item Reliability for Intensive Longitudinal Data
Discussants: Daniëlle Remmerswaal & Pepijn Vink

11:00 – 11:30 **Presentation Yuqi Liu** (Leiden University)
Predicting distal outcomes from latent classes in the growth mixture models: robustness over various coding of time
Discussants: Karel Veldkamp & Santiago Gómez-Echeverry

11:30 – 12:00 **Presentation Daniëlle Remmerswaal** (Utrecht University)
Do respondents show higher activity and engagement in app-based diaries compared to web-based diaries? A case study using Statistics Netherlands' Household Budget Diary
Discussants: Marcel van Assen & Yuqi Liu

12:00 – 13:15 **Lunch**

12:45 – 13:30 **PhD Meeting**

13:30 – 14:30 **Workshop R packages by Hanne Oberman**

14:30 – 15:00 **Closing and Best Poster/Presentation Awards 2025**

Venue: Booth zaal, University Library Utrecht Science Park, Heidelberglaan 3, Utrecht

Bridging disciplinary divides in statistics and causal inference: How to & why

Ellen Hamaker (Utrecht University)

Empirical sciences make use of data to answer substantive questions. For this they rely on statistical methods, which are at times also combined with causal reasoning. Because researchers in different disciplines are working on similar inferential problems, you might expect them to use similar techniques, and that we could easily borrow techniques from other disciplines. But the disciplinary divides in statistics and causal inference are surprisingly deep and seem insurmountable at times. In this presentation, I provide three examples based on my own experience with cross-, multi-, and interdisciplinary work in statistics and causal inference. I discuss how to do this kind of work, emphasize why such work is needed, and what makes this work so hard.

LLM-Based Prior Elicitation for Bayesian Graphical Modeling

Nikola Sekulovski (University of Amsterdam)

In the Bayesian Graphical Modeling framework, priors on network structure encode theoretical assumptions and uncertainty about the topology of psychological constructs under study. For instance, the Bernoulli prior specifies the probability of each pairwise interaction, the Beta–Bernoulli prior governs expected network density, and the Stochastic Block prior models clustering. In practice, however, specifying informed hyperparameters is challenging: theoretical guidance is limited, and default choices can be overly simplistic or restrictive. To address this, we introduce an LLM-based prior elicitation framework in which a large language model provides inclusion judgments foreach variable pair. These judgments are converted into edge-specific prior probabilities for the Bernoulli prior and used to derive hyperparameters for the Beta–Bernoulli and Stochastic Block priors. To make the approach accessible, we provide an R package, bgmElicit, with a Shiny app implementing the methodology. We illustrate the framework in two examples. First, a validation on a subset of a PTSD network from a meta-analysis compares OpenAI GPT models across several conditions. Second, an empirical analysis of 17 PTSD symptoms shows that elicited priors can modestly strengthen evidence regarding edge presence and absence. Taken together, this work is a proof of concept, complementary to expert judgment and prior sensitivity checks.

Disussants: Vira Dvoriak & Hadi Mohammadi

Post-selection inference in multiple linear regressions: the double dipping dilemma in psychological research

Zhiwei Dou (KU Leuven)

Current psychological research often combines exploratory data analysis (EDA) with confirmatory data analysis (CDA). This leads to a practice where the same dataset is used to select variables and conduct inference. Known as invalid post-selection inference (PoSI), this practice violates key statistical assumptions and can result in misleading conclusions. This presentation introduces the issue of invalid PoSI in the context of variable selection within multiple linear regression models and discuss potential solutions. We explain how data-driven parameter selection can bias estimates, distort p-values, and inflate Type I error rates, ultimately undermining the validity of reported findings. To illustrate these points, we present results from simulation studies conducted under different model specifications. Additionally, we discuss several statistical methods to address this problem. Our goal is to didactically explain this overlooked issue and to provide practical recommendations for psychological researchers.

Discussants: Tra Le & Iris Willigers

BayesPower: A General Application of Power and Sample Size Calculation for the Bayes Factor

Wong Tsz Keung (Tilburg University)

BayesPower is an R package with a user-friendly Shiny interface for conducting sample size determination, power calculation, and Bayes factor calculation in the context of Bayesian hypothesis testing without the use of simulation. The app supports a wide range of commonly encountered statistical tests in the social, behavioral, and biomedical sciences, including standardized mean differences, correlations, linear regression and ANOVA, as well as tests for one and two proportions. In addition to traditional point-null hypothesis versus composite alternative hypothesis, BayesPower supports sample size planning for interval hypothesis Bayes factors (i.e., equivalence testing). Moreover, the Shiny app provides command-line-based R code for the reproducibility of results. For most testing problems, the sample size and power are returned almost instantly, enabling researchers to rapidly design informative and efficient studies. Preprint at : https://osf.io/preprints/psyarxiv/pgdac

Discussants: Cas Goos & Camila Barragan Ibañez

EvalMORAAL: Interpretable Chain-of-Thought and LLM-as-Judge Evaluation for Moral Alignment in Large Language Models

Hadi Mohammadi (Utrecht University)

We present EvalMORAAL, a transparent framework to evaluate moral alignment in 20 large language models (like GPT-4, Claude, Gemini). We assess models using the World Values Survey (55 countries, 19 topics) and the PEW Global Attitudes Survey (39 countries, 8 topics). While top models align closely with survey responses, we find a clear regional bias: Western regions show much stronger alignment compared to non-Western regions (a 21-point gap). Our framework combines structured reasoning and automated peer review to identify conflicts and assess quality across different cultures.
Paper link: https://arxiv.org/abs/2510.05942

Discussants: Tijn van Hoesel & Şeyma Nur Ertekin

Single-item Reliability for Intensive Longitudinal Data

Tzu-Yao Lin (KU Leuven)

With the advancement of technology, measures are more and more often collected in real-time, resulting in intensive longitudinal data (ILD). ILD allow us to study dynamic properties of one or more variables, such as moment-to-moment fluctuations in affect or changes in heartbeat rates. Despite the widespread use of ILD, the issues of measurement error and reliability are often inadequately addressed, particularly in the presence of serial correlation and complex multilevel structures. This presentation introduces a three-level linear mixed-effects model that explicitly accounts for the nested data structure and the serial correlation (i.e., moments within days, days within persons). The model incorporates two distinct autoregressive processes to capture serial correlations at both the moment and day levels, while simultaneously partitioning measurement error from within-person variability. Building on this model, we delineate two approaches to defining reliability: a variance-partitioning approach and a correlation approach. This framework yields a multifaceted set of reliability coefficients, enabling researchers to distinguish between between-person from within-person reliability, and to assess measurement consistency across different time scales (e.g., moment-to-moment and day-to-day). We further discuss the relationship between the proposed reliability measures and those derived from simpler models. We apply our model to study the reliability of negative affect in the Exam Result Study (Kalokerinos et al., 2019), where 101 students were measured multiple times a day for several days before and after receiving their exam results.

Discussants: Daniëlle Remmerswaal & Pepijn Vink

Predicting distal outcomes from latent classes in the growth mixture models: robustness over various coding of time

Yuqi Liu (Leiden University)

Growth mixture models (GMMs) with distal outcomes are used for assessing the downstream consequences of heterogeneity of growth trajectory (predicting distal outcomes from latent classes) and understanding the background mechanism (predicting from class-specific growth factors). Prior work on the latent curve models with distal outcomes highlights the substantial implications of changes in the time coding on the unique effects of growth factors in predicting the distal outcome. However, its implications on the performance of GMMs and class-based prediction of the distal outcome remains unclear. Here, we combine two real-data illustrations with a simulation study to assess the impact of changes in time coding approaches on the performance of GMMs with distal outcomes, under various manipulated factors, by examining the recovery of latent classes and the prediction of distal outcomes. We conclude that time coding is not a benign modelling choice for GMMs with distal outcomes. We recommend reporting the time coding approaches, conducting sensitivity analyses across several time codings, and interpreting distal outcomes prediction results cautiously when class separation is low or many classes are identified. Recommendations for applied research using GMMs with distal outcomes are provided.

Discussants: Karel Veldkamp & Santiago Gómez-Echeverry

Do respondents show higher activity and engagement in app-based diaries compared to web-based diaries? A case study using Statistics Netherlands' Household Budget Diary

Daniëlle Remmerswaal (Utrecht University)

Smartphones offer opportunities for official statistics, promising improved user experience, reduced response burden, and higher data quality. We investigate whether respondents show higher activity and engagement in app-based diaries compared to traditional web-based diaries. We use Statistics Netherlands' Household Budget Survey (HBS) as a case study. The HBS is a diary survey conducted every five years to capture household expenditure on goods and services. In 2020, Statistics Netherlands conducted a 4-week web-based survey. In 2021, they conducted a 2-week app-based survey on a smaller sample. We compare participation behavior of respondents in the two modes. We hypothesize that app respondents will show greater retention and engagement, driven primarily by the smartphone's constant availability, the smart features that simplify data collection, and the user-friendly design of the smartphone app. Our first research question (RQ1) examines whether retention rates are higher among app respondents, measured through stages of nonresponse and survival analysis of last activity. The second question (RQ2) assesses objective burden by comparing average time spent per purchase between app and web respondents, with a focus on the time-saving potential of the app's scanning feature. The third question (RQ3) explores whether app respondents are more active and show more spread-out reporting patterns. App-based surveys are currently transitioning from the pilot phase to full implementation in panel surveys and official statistics. Our goal is to evaluate the expectation that app-based data collection enhances activity and user engagement in diary studies.

Discussants: Marcel van Assen & Yuqi Liu