Universiteit Utrecht

15th International Multilevel Conference
April 14 & 15, 2026

Program with Abstracts

Organizing committee
Dr. Emmeke Aarts
Dr. Beth Grandfield
Dr. Sara van Erp
Marianne Geelhoed

Utrecht University
Department Methodology & Statistics

Day 1 — Tuesday, April 14
Dynamic and Longitudinal Multilevel Modeling

Time	Activity
09:15 – 09:45	Walk-in & coffee
09:45 – 09:50	Conference opening
09:50 – 10:50	Keynote Lecture — Donald Hedeker **Two-stage mixed-effects location scale (MELS) models for intensive longitudinal data**
10:50 – 11:00	Short coffee break
11:00 – 12:20	Session 1 — Dynamic Models for Intensive Longitudinal Data
12:20 – 13:05	Lunch
13:05 – 14:15	Session 2 — Methods and Applications for Intensive Longitudinal Data
14:15 – 14:35	Refreshment break
14:35 – 16:00	Session 3 — Latent and Applied Longitudinal Multilevel Models
16:00 – 17:00	Poster session with drinks

Day 2 — Wednesday, April 15
Methodological and Bayesian Advances

Time	Activity
09:00 – 09:30	Walk-in & coffee
09:30 – 10:45	Session 4 — Bayesian Design and Power for Multilevel Studies
10:45 – 11:15	Coffee break
11:15 – 12:55	Session 5 — Computational and Bayesian Methods for Multilevel Models
12:55 – 13:45	Lunch
13:45 – 15:00	Session 6 — Missing Data, Causal Inference, and Evidence Synthesis in Multilevel Models
15:00 – 15:30	Refreshment break
15:30 – 15:40	PhD award ceremony
15:40 – 16:40	Closing Keynote Lecture — Paul Bürkner **Amortized Bayesian inference for multilevel models**
16:40 – 16:45	Closing remarks

Day 1 — Dynamic and Longitudinal Multilevel Modeling

Keynote 1: Two-stage mixed-effects location scale (MELS) models for intensive longitudinal data

Prof. dr. Donald Hedeker (University of Chicago)

Intensive longitudinal data are increasingly encountered in many research areas. For example, ecological momentary assessment (EMA), experience sampling method (ESM), and/or mobile health (mHealth) methods are often used to study subjective experiences within changing environmental contexts. In these studies, up to 30 or 40 observations are usually obtained for each subject over a period of a week or so, allowing one to characterize a subject's mean and variance and specify models for both. In this presentation, we focus on a smoking study of dual users (i.e., both combustible and electronic cigarette users) using EMA where interest is on characterizing changes in mood variation associated with these nicotine products, and whether subjects' mood response can predict future nicotine product use. At the first stage, the MELS model includes random subject effects for the mean (i.e., location), which characterize subjects' differential mood response to combustible and electronic cigarettes. A random effect for subjects' variability (i.e., scale) in mood responses is included to characterize subjects' mood consistency/erraticism. These random location and scale effects are used in a second stage regression, both linear and multinomial, model to predict future nicotine product use. Since the random effects are estimates, repeated draws from the posterior distribution of the random effects for each subject are utilized in the second stage model (i.e., plausible value replications), with results averaged across these repeated draws. A software program, MixWILD, which facilitates this two stage modeling approach, is described.

Session 1 — Dynamic Models for Intensive Longitudinal Data (11:00-12:20)

Dynamic latent factor Poisson network model for longitudinal relational data

Presenter: Rosa Gentile

Understanding the dynamics of relational networks is crucial in social contexts, where interactions are often observed as count data. This paper proposes an extension of Poisson network models to capture the temporal evolution of relationships between nodes. The model incorporates dynamic latent sender and receiver effects to account for node-level heterogeneity, as well as multiplicative dyadic effects to model higher-order dependencies, such as clustering. The temporal evolution of network nodes is modelled with additive sender and receiver effects following autoregressive processes, while multiplicative dyadic effects evolve according to a vector autoregressive (VAR) structure. To illustrate the model's validity and practical applicability, we apply a

Bayesian framework to hospital transfer data within a regional healthcare system.

Estimation Accuracy of Subject-Specific Effects in the Multilevel Hidden Markov Model

Presenter: Pepijn Vink

The last decades have seen a surge in the use of intensive longitudinal data (ILD) in behavioral sciences. These are time-series of multiple persons, obtained using, for example, ecological

momentary assessment, and allow researchers to investigate person-specific moment-to-moment dynamics, for example in mood, on a very granular level.

Naturally, as ILD are comprised of measurement moments nested within persons, it is common to model them using a multilevel model. When processes can be described as switches between stable means, these may be modeled using a multilevel hidden Markov model (MHMM). The MHMM is a multilevel multivariate mixture model for ILD that decomposes observed data into latent discrete states. The data are assumed to follow a different probability distribution in each state; for continuous data, for example, these are normal distributions with different means and variances. Duration of the states and switches between them are then modeled. The MHMM has been applied in psychology to data of patient-therapist dyads, as well as data of patients with bipolar disorder, depression, and psychosis.

Existing simulation studies that evaluated the estimation accuracy of the MHMM have focused on the group-level parameters, relying on decoding accuracy of the state sequence as a proxy for performance on the subject-level. In contrast, the subject-level estimates have great potential to provide a personalized description of the dynamics of an individual's process. Moreover, the subject-specific effects are especially affected by label switching, an estimation issue that sometimes occurs in mixture models that use MCMC.

The current study therefore has two primary aims. First, we aim to obtain insight into the estimation performance of the subject-level effects. Second, we aim to mitigate the label-switching issue in the MHMM.

To this end, we perform a simulation study, varying the number of subjects, time-series length, degree of state separation, and number of observed variables, to assess the reliability and accuracy of these person-specific models. We implement two online adaptive relabeling algorithms in the mHMMbayes R package, of which one relabels based on the parameters and one on the state sequences at each MCMC iteration.

Preliminary results indicate that subject-specific effects are estimated well with good reliability. Moreover, the implemented relabeling algorithms perform well with minimal increase in computation time.

Simultaneously capturing and characterizing states and fluctuations in intensive longitudinal data

Presenter: Dr. Jessica Schaaf

Multilevel time series models are increasingly used to investigate fluctuations in psychological processes over time. Advanced time series models such as Dynamic Structural Equation Models have proven a powerful tool to capture, for example, mood or cognitive fluctuations.

Arguably, temporal dynamics of many psychological processes are not gradual but involve states, for example, being attentive or distracted. Several models have been proposed to capture such states in intensive longitudinal data. However, these models either do not allow for detailed quantifying of transitions between states (e.g., change point models) or do not allow for quantifying the dynamics within each state (e.g., Hidden Markov Models).

Therefore, we propose an extended multilevel time series model in which we implement states and allow temporal dynamics parameters to vary across these states, the Hidden Markov Dynamic Structural Equation Model (HM-DSEM). We show how this model outperforms stateless models in capturing mood and cognitive fluctuations and how model parameters provide valuable information on psychological processes.

Three-Step State Space Mixture Modeling to Compare Dynamic Processes Across Many Individuals

Presenter: Manuel Rein

Research on within-person processes in psychological constructs often seeks to uncover how individuals differ in these dynamic processes. For instance, researchers may be interested in finding differences and similarities between individuals in the extent to which positive affect and negative affect carry over and interact with each other from one moment to the next. While some studies compare predefined groups of individuals (e.g., via multi-group modeling), in many cases such groupings are not known in advance. Instead, researchers explore heterogeneity across individuals in a data-driven manner. As researchers routinely gather data from 100 or more individuals, there is a growing need for modeling approaches that can efficiently capture and compare dynamic processes across many individuals.

State space models make it possible to capture temporal relations among latent (i.e., unobserved) variables by combining a structural model (the dynamic process) with a measurement model (which describes how the latent variables are measured by observed items). To accommodate unknown heterogeneity in these dynamics, mixture approaches such as state space mixture modeling offer one way to identify discrete subgroups of individuals who follow qualitatively distinct processes. However, jointly estimating the measurement model, the structural model, and the mixture components can be computationally demanding and vulnerable to local misspecifications (such as cross-loadings that are not included in the model). Moreover, fit indices such as the BIC reflect the combined fit of both the measurement model and the structural model. To address these issues, stepwise estimation methods that separate the estimation of the measurement model from the estimation of the structural model have been developed. This allows researchers to scrutinize (and potentially adjust) the measurement model before estimating the parameters of the structural model. It also reduces the number of parameters that need to be estimated simultaneously and may thus improve convergence and reduce the sensitivity to local maxima.

To leverage these advantages, we present Three-Step State Space Mixture Modeling, an extension of the Three-Step Latent Vector Autoregression framework to state space mixture modeling. In Step 1, solely the measurement model is estimated using factor analysis. In Step 2, estimated factor scores (i.e., estimated scores on the latent variables) are computed. In Step 3, the structural model and the mixture clustering based on it are estimated using the factor scores from Step 2 as single indicators of the respective latent variables and account for their inherent uncertainty.

We demonstrate the method's performance in obtaining correct estimates of the structural model parameters and the individuals' cluster memberships by means of a simulation study, and illustrate the implementation of the method in the R package ezLVAR.

Session 2 — Methods and Applications for Intensive Longitudinal Data (13:05-14:15)

Analyzing Skewed ESM Data

Presenter: Anastasiia Galkina

Intensive longitudinal data collected via the Experience Sampling Method (ESM) or Ecological Momentary Assessment (EMA) frequently exhibit right-skewed distributions due to floor effects arising from bounded measurement scales. When outcomes such as negative affect of psychopathological symptoms are assessed on a scale with a lower boundary, a substantial proportion of observations cluster at or near that boundary, producing distributional characteristics that violate the assumptions of standard linear mixed-effects models, which are most frequently used for the analysis of ESM data.

The current work advocates for conceptualizing bounded ESM observations as censored manifestations of an underlying latent continuous variable and proposes mixed-effects tobit models as the appropriate analytical framework. The tobit model explicitly accounts for the censoring mechanism: observations at the boundary are treated as manifestations of the true latent score, making this approach particularly useful for analyzing data with substantial skewness.

We evaluate the proposed approach using data from N=322 participants across three diagnostic groups (healthy controls, individuals with depression, and individuals with psychotic disorder), each completing mood assessments ten times daily over six consecutive days. Two complementary analyses are presented. First, an artificial censoring manipulates the degree of skewness in a near-normally distributed positive affect variable by progressively shifting scores toward the lower bound, allowing direct assessment of parameter recovery under controlled conditions. Second, a natural analysis compares standard and tobit mixed-effects models on a genuinely right-skewed negative affect outcome, examining the practical consequences of model choice for substantive conclusions.

Results demonstrated that the standard linear mixed-effects model produced slope estimates with a systematic bias toward zero as censoring increases, whereas the tobit model recovered the original effect sizes with only slight deterioration. Applied to natural negative affect data, the tobit model estimated substantially larger associations between predictor and outcome, specifically in the control group, where floor effects were most pronounced. Moreover, the mixed-effect tobit model revealed greater between-person variability in intercepts and reduced intercept-slope correlations.

These findings carry direct implications for the interpretation and replicability of ESM research: underestimation of fixed effects and overestimation of their correlation with random intercepts may have led to systematic mischaracterization of psychological processes in clinical and non-clinical populations. We suggest ESM researchers use the mixed-effects tobit model for analyzing skewed data, as it offers important theoretical and statistical advantages for analyzing bounded psychological constructs compared to conventional linear mixed-effects models.

Deep Generalised Mixed Effects Models: a Novel Neural Network Structure for Analysing Hierarchical Data

Presenter: Nina van Gerwen

Background: The Experience Sampling Method (ESM) is an intensive longitudinal research design where participants report their thoughts, emotional states and behaviours multiple times a day. ESMs have become increasingly popular to investigate individuals' daily experiences. Our work is motivated by ESM data collected by the GrowIt! app. During the COVID-19 pandemic, the app was released to investigate daily mood changes among young adults, give users insight into their emotions, and enhance users' resilience. Current procedures to analyse ESM data face various challenges. In particular, ESM data are high-dimensional and exhibit complex correlation structures. Although procedures exist to generalise multilevel models to higher dimensions, they may still not adequately capture the complex correlation structure or the models' assumption may be violated. Alternatively, machine learning procedures, such as recurrent neural networks, can be used to model ESM data and accommodate these correlations. However, these procedures face a problem with missing data. In our motivating dataset, adolescents often stopped using the app due to previous strong feelings of negative emotions. Hence, the implied missing data are of the missing-at-random type that standard machine learning procedures cannot accommodate.

Methods: We develop a novel neural network (NN) architecture that generalises mixed effects models to deep learning to overcome these challenges. Our Deep Generalised Mixed Model (DGMM) allows semi-parametric and highly flexible modelling of the data's mean and correlation structure with NNs using fixed and random effects. Classical estimation of mixed models requires integration over the random effects distribution, which is intractable when we estimate the random effects with a NN. Therefore, we use an adaptation of variational autoencoders to estimate the DGMM. By specifying a tractable variational distribution to sample from, we approximate the marginal log-likelihood as an expectation with respect to the variational distribution and the Kullback-Leibler divergence between the variational distribution and the marginal distribution of the random effects, together known as the Evidence Lower Bound. The variational distribution can also be seen as a nonlinear function of the data, which we estimate with another NN. Through this approach, the DGMM is able to accommodate longitudinal outcomes following any generic distribution, scale well to high-dimensional settings and provide valid inference when data is missing at random.

Results: In the GrowIt! app data, the DGMM showed good predictive performance for the multivariate analysis of five longitudinal outcomes following varying distributions. A simulation study of the DGMM also showed good performance in various settings, yet the model can be susceptible to small hyperparameter changes.

Conclusion: We have implemented the DGMM in Python using Keras and Tensorflow. The model can be valuable for a flexible semi-parametric analysis of the multivariate analysis of high-dimensional longitudinal data. However, this comes at the cost of reduced interpretability.

Taking a Sledgehammer to Crack a Nut: Analyzing the Dynamics of Auditory Distraction Using Multigroup Multilevel Structural Equation Modeling

Presenter: Salome Keintzel

Advanced Structural Equation Models (SEMs) have gained popularity in recent years. However, they are rarely applied in experimental psychology as most existing longitudinal SEM frameworks are designed for panel data (small number of time points, observational data in wide format) rather than the typical data structure of repeated-measures experiments (large number of time points, randomized experimental conditions, data in long format).

We demonstrate how longitudinal SEMs that were originally developed for designs with relatively few measurement occasions can be adapted for experimental settings with many repeated measures by segmenting data into shorter prototypical sequences. These trial sequences, nested in participants, can be analyzed in a multilevel SEM building on the basic idea of the autoregressive latent trajectory model to get new insights into trial-level dynamics (carryover effects) in cognitive processes.

Using data from adults and children performing an attention-control task, we illustrate how different cognitive and longitudinal phenomena (trends, autoregressive effects) map onto specific parameters within this SEM framework. Crucially, we show how the presented model can flexibly account for various interactions of these longitudinal phenomena with (multiple) experimental conditions (e.g., stimulus types, level 1) and person-characteristics (e.g., age groups, level 2) while maintaining parameter interpretations close to those known from linear mixed-effects models, that are more frequently used in the field. We provide a comprehensive account of the model specification and conclude with results from a sample size simulation study assessing the model's data requirements.

The interplay between problematic gambling and biopsychosocial risk and protective factors: a dynamic network approach

Presenter: Ladan Esmalian Khamseh

Background: Diary study designs combined with network analysis (NA) offer a powerful framework for capturing the complex, dynamic interplay between problematic gambling and biopsychosocial risk factors in individuals' natural contexts. Although NA has been widely applied in psychiatric and psychological research, its use in gambling research-particularly to examine within-person daily dynamics remains limited.

Methods: The present study employed a 40-day diary design within an ambulatory assessment framework. Participants were recruited via Prolific Academic in two phases. In Phase 1, 3,001 individuals completed the Problem Gambling Severity Index (PGSI). The 200 participants with the highest z-scores were invited to Phase 2, of whom 179 completed daily assessments. Participants reported on gambling behavior and time-varying biopsychosocial factors each evening at 20:00 (London time), using a fixed 24-hour sampling schedule. Daily questionnaires assessed experiences across four daily segments (night, morning, afternoon, evening). To reduce participant burden, single representative items from validated instruments were used to assess problematic gambling, exposure to gambling advertisements, depressive symptoms, anxiety, aggression, impulsivity, boredom, loneliness, emotion dysregulation, alcohol and nicotine use, physical activity, daily hassles, and negative interpersonal interactions. Dynamic associations were examined using multilevel vector

autoregressive (mlVAR) network models, allowing estimation of temporal, contemporaneous, and between-person networks.

Results: In the temporal network, emotion dysregulation, nicotine use, and exposure to gambling advertisements showed positive associations with subsequent problematic gambling (money spent beyond affordability), whereas problematic gambling showed a negative association with drug use. In the contemporaneous network, problematic gambling was positively associated with exposure to gambling advertisements, anxiety, depression, loneliness, boredom, emotion dysregulation, negative interpersonal interactions, daily hassles, nicotine use, and alcohol use.

Conclusions: These findings underscore the dynamic and time-dependent nature of problematic gambling, highlighting distinct temporal and within-day associations with emotional, behavioral, and environmental factors. Network-based diary approaches provide valuable insights into the daily mechanisms that may maintain or exacerbate gambling-related harm.

Session 3 — Latent and Applied Longitudinal Multilevel Models (14:35-16:00)

Predicting intersectional inequalities: MAIHDA vs. Descriptive Statistics

Presenter: Prof. dr. George Leckie

Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) is an increasingly popular method for studying intersectional inequalities in individual outcomes. A typical application might examine how mean reading scores vary across combinations of sex, ethnicity, social class, and disability.

MAIHDA allows researchers to ask: What is the mean outcome for each combination? How large are these inequalities? To what extent are they additive, and to what extent do complex interactions play a role?

Proponents of MAIHDA argue that it improves upon linear regression with many interaction terms, both by being easier to interpret and by statistically controlling the noise inherent in comparing means across many small groups.

However, MAIHDA is essentially a sequence of two-level random-intercept multilevel models-albeit with a highly unusual definition of clusters: the combinations themselves. The purported statistical control of noise stems from partitioning variance and applying shrinkage (partial pooling) when predicting cluster-specific means.

At the last conference, I highlighted various multilevel modelling conceptual critiques of MAIHDA stemming from its definition of clusters, only to be met with the pragmatic response: “If it works, it works.” At this conference, I therefore examine how well MAIHDA actually performs in practice, focusing on how much intersection means vary by method and, in turn, how much more accurate MAIHDA means are compared with simple averages.

Given the simplicity of the underlying multilevel model, I use analytic expressions to investigate the variance of each set of intersection means, their correlations with the true means, and the bias, variance, and mean squared error (MSE) associated with each method of calculating the intersection mean for a given intersection.

I conclude that, while MAIHDA means are more accurate than simple means, many MAIHDA studies ultimately need larger samples within their rarest intersections to make the desired inferences with an acceptable degree of precision. Multilevel models can only compensate so much for limited data.

Prediction of latent disease status in a multivariate longitudinal model on Duchenne Muscular Dystrophy patients

Presenter: Chiara Degan

In psychology, it is common to collect multivariate quantitative longitudinal data intended to measure a single underlying construct of interest. In such cases, a continuous and subject-specific latent variable can be postulated to describe the unobserved underlying process that generates the set of multiple observed outcomes. Similar situations might occur in medical settings, where treatment decisions or diagnostic evaluations often rely on examining subject-specific trajectories of several outcomes simultaneously. We considered an illustrative example in the context of Duchenne Muscular Dystrophy (DMD), a degenerative neuromuscular disease characterized by early loss of muscle function and reduced life expectancy. Disease progression in DMD is routinely monitored using established physical patients' assessments that clinicians rely on to guide medical decisions, e.g., treatment initiation or adjustment. These tests capture complementary aspects of motor function, each reflecting functional changes in specific body regions and stages of the disease. However, they are often time-consuming, physically demanding for patients, and logistically challenging to obtain at regular intervals. Consequently, there is growing interest in identifying serum biomarkers that could counterpart, or eventually partially substitute for these physical tests. Particularly, our aim is to identify a panel of biomarkers that,

when considered jointly, can accurately predict the hypothesized patient's underlying disease status. Ultimately, the long-term goal is to infer this latent disease status directly from biomarker data, thereby reducing or possibly eliminating the need for physical assessments. To achieve this goal, we propose a modeling framework that borrows some structural elements from both the flexible latent process approach of Proust-Lima et al. (2013), and the current-value formulation commonly employed in joint models (Rizopoulos, 2012). We introduce a model that links three key components:

- a longitudinal biomarker described by a linear mixed-effects model;

- a latent variable that captures the correlation among the physical tests and describes the underlying disease progression of each patient;

- test-specific equations for the physical assessments, each following an appropriate distribution within the exponential family.

The linear mixed-effects model for the biomarker allows to properly account for measurement error and enables the estimation of noise-free, subject-specific biomarker trajectories. The latent process serves as the bridge between the biomarker and the test outcomes. It consists of a population-level component that captures the effects of covariates (e.g. age, treatment, etc.), and a subject-specific component derived solely from the random effects of the biomarker model. Through this adaptation of the latent variable framework, future prediction of the latent process can be achieved without requiring the observation of the outcomes. Finally, the tests-specific equations link each observed longitudinal physical assessment to the latent process, enabling the model to represent how the underlying disease state manifests across heterogeneous outcome types. Model estimation has been performed using the Bayesian Integrated Nested Laplace Approximation (INLA) method. We will illustrate this alternative formulation by presenting preliminary results from a cohort of DMD patients monitored at the LUMC.

Identifying anomalously high abundance in highly seasonal populations of the common house mosquito in the Netherlands

Presenter: Lukas Sprengers

Objective

Adult mosquito abundance exhibits substantial seasonal variation driven by meteorological factors. Higher mosquito abundances could lead to negative public health outcomes through increases in transmission of vector borne diseases as well as increased mosquito nuisance. To help inform risk communication to policy makers and the public, this modelling project aimed to identify anomalously high outliers mosquito abundances using observed abundances from earlier years.

Methods

Frequentist and Bayesian multi-level generalized linear models with Gaussian, Poisson and negative binomial distributions were fitted to historical weekly mosquito counts from 12 locations ranging from 2019-2024 using year and location as random effects, and week number as a fixed effect. We evaluated the models' performances using historical mosquito counts from 2025 as novel data. We calculated the quantile rank of these novel data in simulated distributions obtained by Monte Carlo sampling from the fitted models. Observations were classified as extreme if they fell above the 95th percentile of the simulated distribution.

Results

Generalized linear models with negative binomial distributions had the best fit. Furthermore, the Bayesian model performed better than the frequentist model. The latter classified historical observations as 'extreme' more often than desired. The values classified as extreme by the Bayesian model, were also considered extreme by expert judgement.

Conclusion

Multi-level Bayesian modelling methods combined with Monte Carlo sampling function is a promising method for identifying future extreme values of mosquito abundance counts at multiple loci. Such models can help inform institutions when risk communication is warranted during the mosquito season.

Multilevel multinomial time series models with applications to population health issues in Australasia

Presenter: Prof. dr. Alice Richardson

Official statistics on health outcomes for small domains are highly prized by policymakers and researchers, for measuring and monitoring progress of communities towards healthy lifestyles. Countries use nationally representative surveys e.g. the National Health Survey (NHS) in Australia or the Demographic and Health Survey (BDHS) in Bangladesh, to monitor adult health behaviours. However, the surveys cannot be used to estimate prevalence at disaggregated levels due to lack of information.

The talk will describe three applications of multilevel multinomial time series models, estimating trends in population health outcomes at different levels of geographical detail (i.e. small areas or domains).

Direct estimates of outcomes and their smoothed standard errors for small domains are used as input for developing multilevel models, which are expressed in a hierarchical Bayesian framework and fitted by Markov Chain Monte Carlo (MCMC) simulation. The developed models provide consistent estimates at the most detailed level domains by borrowing cross-sectional and spatial strengths. The detailed level domain predictions are then aggregated to obtain estimates at higher aggregation levels. Prevalences and their standard errors are visualised in a variety of relevant ways which permit exploration of the modelled spatio-temporal changes in outcomes.

The first application will focus on trend estimation of sub-national level (state and territory) smoking prevalence in Australia by age and sex using NHS data collected over 2001-2021. The usefulness of a socio-economic index for areas (SEIFA) as a contextual variable is investigated.

The second application will focus on trend estimation of antenatal care in Bangladesh at admin-2 (64 districts) and admin-3 (544 sub-districts) levels using BDHS data collected over 2000-2018. Inclusion of remote-sensed data in the model has improved the trend prediction of stunting level, particularly in the non-survey years.

In the third application, we show how socio-economic disadvantage at the local area (i.e., SEIFA)) and the remoteness (i.e., the accessibility/remoteness index for Australia) contribute to improved prevalence estimates of child development vulnerability in statistical areas level 3 (359 SA3s) and 4 (108 SA4s) across Australia. These three examples will showcase how we can have improved and reliable estimates of disaggregated level population health outcomes (when compared to the direct survey-based results) through the model-based SAE approach.

Our small area model-based estimators describe significant social, economic, and geographical disparities, allowing us to identify regions exhibiting progress, stagnation, or decline. The model-based estimates demonstrate improved precision over direct estimates and provide valuable insights for designing targeted interventions to maximize impact. Additionally, the inclusion of spatio-temporal terms further refines estimates of disparities. Knowledge of these is crucial for localised planning and decision-making to ensure health equity.

Predicting Distal Outcomes from Latent Classes in The Growth Mixture Models: Robustness over Various Coding of Time

Presenter: Yuqi Liu

Growth mixture models (GMMs) with distal outcomes are commonly used to study heterogeneity in longitudinal trajectories and to compare downstream outcomes across latent classes. In growth models, time coding is part of the model specification that shapes the scale and interpretation of growth factors, and can affect estimation stability and substantive conclusions, including inference about distal outcomes. Prior work on latent curve models with distal outcomes, where distal outcomes are related to growth factors, shows that alternative time-coding schemes can yield substantively different distal-outcome inferences. In GMMs, distal outcomes are often related to class membership rather than to growth factors. Thus, time-coding choices may affect distal-outcome inferences indirectly by altering class enumeration and membership assignments, rather than by changing the partial effects of growth factors. However, the sensitivity of GMMs with distal outcomes to time-coding specifications has not been systematically evaluated.

We combined two real-data illustrations with a simulation study to evaluate how time-coding choices impact class enumeration, latent class recovery, and distal-outcome inference across varying data conditions. Results show that time-coding choices can meaningfully affect the selected number of classes, classification accuracy, and distal outcome conclusions, particularly under challenging conditions with many identified classes and weaker separation. For applied researcher, we therefore recommend reporting the time-coding specification, conducting sensitivity analyses across plausible codings, and interpreting distal-outcome results cautiously under high classification uncertainty.

Poster Presentation Session (16:00 – 17:00)

Do you have Open Educational Resources on Multilevel Modeling to share?

Presenter: Mariska De Moor

Conducting research according to Open Science principles has become mainstream for most scientists. However, as teachers we are still lagging behind, as not everyone is yet familiar with or works according to Open Education principles, which includes developing, sharing and using Open Educational Resources (OER). With presenting this poster, we aim to gauge how many scientists have developed learning materials on multilevel modeling and are willing to share their materials by making them open. All OER collected will be shared through a newly developed thematic page about Methods & Statistics on www.edusources.nl, a Dutch platform for OER in higher education.

Cognitive-Affective States Transitions: mHMMbayes Model Priors

Presenter: Nicole Fridling-Cook

This study examines how transitions between cognitive-affective states are influenced by prediction errors, negative affect, and positive affect. Using self-report data (0-100 scale,

multimodal) from 2795 individuals (~10.4 timepoints each), I fit a multilevel hidden Markov model in mHMMbayes (Aarts & Mildiner Moraga, 2025) with 3 hidden states, 14 observed variables, and 1 transition covariate. I am seeking guidance on selecting emission and transition priors to improve model fit.

How might multilevel models might be used to characterize the temporal dimensions of doctoral student outcomes?

Presenter: David Most

This project investigates how a variety of factors at various levels (e.g., student, department, institution, discipline) are associated with the likelihood of doctoral students completing the Ph.D. Over 5000 doctoral students are clustered in 80 departments, 16 institutions, and five disciplines. Student progress over a nine-year period is captured in the data. A survival analysis approach can be used to model the "risk" of degree completion over time as a function of various factors. As factors are measured at various levels, and as students are clustered in various ways, how might multilevel models be used to characterize relationships of interest?

Within-day changes of food craving in longitudinal experience sampling data

Presenter: Christoph Bamberg

I am analysing within-day changes of food-craving in a longitudinal experience sampling dataset with Bayesian Autoregression models in STAN. Responses are grouped within days, missings are estimated from previous and successive time-points. Currently, I model single individuals because I am interested in between-day differences.
My questions are:
I "only" have six time points per day—am I using the multilevel structure optimally, pooling information between days?
If I add a grouping-level to analyse the whole sample (N=60), how can I retain high resolution for day-to-day differences? Can I improve the posterior estimation of missings with this additional level?

Modelling intra-individual differences in variability using 4-level hierarchical dynamic structural equation models (DSEM): How to optimally trade-off fixed and random effects with respect to model complexity?

Presenter: Kevin Reniers

Within the CODEC project (Coolen et al., 2024) children's performance is measured on five different cognitive tasks in 15 sessions over an in-class burst week. We hope to estimate children's variability in reaction time at the trial, session, and day timescale, investigating individual differences and their interrelatedness. To do so, we expand proprietary, 2-level hierarchical dynamic structural equation models to open-source (R/Rstan), 4-level (and up) models. Several challenges remain in model development and parameter estimation. How can we best investigate variability at the population and individual level, accounting for nesting, the trade-off between fixed and random effects, and model convergence?

Day 2 — Methodological and Bayesian Advances

Keynote 2: Amortized Bayesian inference for multilevel models

Prof. dr. Paul Bürkner (TU Dortmund University)

Abstract: With the advent of probabilistic programming languages, speed remains the only main limiting factor of Bayesian inference. This is because current gold-standard posterior approximators, in particular MCMC, are very slow, especially compared to optimization-based approaches. In the end, it seems we have to pay this price in order to achieve principled and accurate uncertainty quantification. Or do we? What if we could have accurate and fast Bayesian inference at the same time? This question leads us to what we call neural amortized Bayesian inference, a promising new field at the intersection of Bayesian inference and deep learning. I will highlight some or our recent advances as well as existing challenges in the field. This inspires a look into a potential future of Bayesian inference, accelerated by the learning and generalization abilities of neural networks where trustworthiness and speed are no longer conflicting goals.

Session 4 — Bayesian Design and Power for Multilevel Studies (09:30-10:45)

How do I start planning my multisite project? A tutorial on multilevel simulation-based power and sensitivity analysis

Presenter: Alicia Franco-Martinez

As psychological science moves toward large-scale, multisite collaborations, researchers face a challenge: traditional power analysis tools (such as G*Power) are often insufficient for the hierarchical nature of these projects. When data are nested within laboratories, simple effect size estimates fail to account for the heterogeneity between sites. This talk will provide a practical but brief tutorial for using multilevel simulation to justify plan resources in collaborative research.

We will move beyond the "one-N-suffices" approach by demonstrating how to build a three-level generative model: trials (Level 1) nested within participants (Level 2), nested within laboratories (Level 3). This structure allows researchers to move past the (sometimes) problematic practice of aggregating trial-level data, which often discards the very variance necessary for robust inference. Participants will learn to navigate two primary strategies: First, power analysis (i.e., determining the number of laboratories, participants, and trials needed to detect a minimum effect size of interest) and, second, sensitivity analysis (i.e., identifying the minimum detectable effect size given fixed constraints, such as a maximum manageable number of labs or type of tasks). By varying the assumed between-laboratory variance in their simulations, researchers can stress-test their designs against the "unavoidable complexity" of site heterogeneity. Notably, we extend these simulations to address a common oversight in multisite designs: power analysis for detecting heterogeneity. While many-labs projects often aim to explore site-level differences, they rarely plan for the statistical power required to detect them. We demonstrate how simulations can be easily adapted to ensure a design is powered not only for fixed effects but also for measuring the variance components themselves.

A central theme of the presentation will be the alignment of multilevel models with the target of inference. We will discuss the practical implications of treating labs and participants as fixed versus random effects, and how

these choices dictate whether findings generalize to the specific samples collected or to a broader population of laboratories. Finally, we will provide an accessible R script template and an empirical illustration to help researchers transition from intuition-based planning to simulation-based design.

Bayesian Power Analysis for Multilevel Models – the BayesSSD package

Presenter: Ulrich Lösener

Sample size determination (SSD), the procedure of determining the number of subjects necessary to achieve a desired level of statistical power, is essential in planning an experiment. However, available software are mostly limited to the framework of Null Hypothesis Significance Testing. The few existing software for Bayesian SSD are either a) limited to simpler models such as ANOVA and t-test and cannot handle longitudinal data or b) unable to handle more than two treatment conditions. Current software also neglects participant attrition - a common occurrence that may substantially reduce the power of a longitudinal experiment. The present work addresses this gap by introducing the open-access R package BayesSSD, which performs simulation-based Bayesian SSD for longitudinal (and cluster randomized) experiments with two or more treatment conditions. In the presented method, various patterns of expected attrition can be specified via parametric and non-parametric survival functions and accounted for in the SSD procedure.

In this talk, I will outline the theoretical background of Bayesian hypothesis evaluation in multilevel models and elaborate on essential concepts of the package such as modelling attrition and the general mechanism of the algorithm. The effects of different attrition patterns on Bayesian power are demonstrated through the results of a simulation study.

Sample Size Determination for trials with the Bayes factor

Presenter: Camilla Barragan Ibanez

The determination of sample size is a key step in a study design to avoid underpowered studies. This step becomes particularly complex in studies with hierarchical data, such as cluster-randomised trials (CRTs) and longitudinal intervention studies. Most existing methods for sample size determination are designed for null hypothesis significance testing. Approach for hypothesis testing with numerous limitations in its use and has been subjected of considerable criticism over the past few decades. An alternative approach is the Bayes factor, which quantifies the relative evidence for competing hypotheses, providing a more direct and informative answers to the question of what the data supports. The current methods for sample size determination with the Bayes factor, however, are limited to non-hierarchical models. To address this gap, we developed a simulation-based methodology for sample size determination for hierarchical data. This methodology is implemented in the R package BayesSSD. The use of the package is demonstrated a range of study designs, from a simple two-arm CRT with binary and continuous outcomes to more complex scenarios involving multiple outcomes, longitudinal studies with participant dropout, and studies with multiple treatment conditions. We illustrate the impact of key design elements, such as the intraclass correlation coefficient, effect sizes, and variance components, on the required sample size. We conclude by providing practical recommendations for researchers designing studies with hierarchical data.

Bayesian sequential designs in studies with multilevel data

Presenter: dr. Mirjam Moerbeek

In many studies in the social and behavioral sciences, the data have a multilevel structure, with subjects nested within clusters. In the design phase of such a study, the number of clusters to achieve a desired power level has to be calculated. This requires a priori estimates of the effect size and intraclass correlation coefficient. If these estimates are incorrect, the study may be under- or overpowered. This may be overcome by using a group-sequential design, where interim tests are done at various points in time of the study. Based on interim test results, a decision is made to either

include additional clusters or to reject the null hypothesis and conclude the study. This contribution introduces Bayesian sequential designs as an alternative to group-sequential designs. This approach compares various hypotheses based on the support in the data for each of them. If neither hypothesis receives a sufficient degree of support, additional clusters are included in the study and the Bayes factor is recalculated. This procedure continues until one of the hypotheses receives sufficient support. This paper explains how the Bayes factor is used as a measure of support for a hypothesis and how a Bayesian sequential design is conducted. A simulation study in the setting of a two-group comparison was conducted to study the effects of the minimum and maximum number of clusters per group and the desired degree of support. It is concluded that Bayesian sequential designs are a flexible alternative to the group sequential design.

Session 5 — Computational and Bayesian Methods for Multilevel Models (11:15-12:55)

Variable selection via knockoffs for clustered data

Presenter: Prof. dr. Silvia Bacci

The selection of relevant predictors affecting a response is a fundamental issue in assessing a statistical model. It is particularly challenging in high-dimensional contexts, when numerous predictors are available. Indeed, different selection strategies may yield different results, with the risk of including variables with null effects in the model or, conversely, excluding variables with non-null effects.

The knockoffs approach has the advantage of explicitly controlling either the Per-Family Error Rate (PFER), which is the expected number of false discoveries, or the False Discovery Rate (FDR), which is the expected proportion of false discoveries.

To date, the available knockoff procedures are not explicitly designed for complex data structures, such as clustered (or hierarchical) data, a setting often encountered in applications, such as panel data (waves nested within subjects), repeated measures (occasions nested within subjects), or when data are collected from multiple groups (e.g., patients within hospitals).

In this contribution, we extend the knockoffs method for selecting predictors in the case of clustered data. In such a setting, variable selection is complex because some predictors are measured at the observation level (level 1), whereas others are measured at the cluster level (level 2), so their values are constant within clusters.

The solution we propose is to conduct variable selection separately at the two levels. To this end, we suggest a two-step approach: (i) decompose each level 1 predictor into level 2 and level 1 components by replacing it with the cluster mean and the deviation from the cluster mean; (ii) perform variable selection separately at the two levels, where the level 1 data matrix includes the deviations from the cluster means and the level 2 data matrix includes the cluster means of level 1 predictors and the level 2 predictors.

To evaluate the performance of the proposed approach, we conduct a simulation study comparing two knockoff approaches - the sparse sequential knockoff and the derandomized knockoff - and a standard selection variable approach, that is, the Lasso. Specifically, we address the selection of level 2 predictors (i.e., variables that are constant within clusters), focusing on both continuous and binary variables. The study shows satisfactory results in terms of false discovery rate and power. It is worth noting that all methods fail when applied to the complete data matrix, including both level 1 and level 2 predictors. In contrast, all methods perform better when applied to the level 1 and level 2 data matrices separately. Moreover, the sparse sequential knockoffs method performs substantially better than the Lasso and the derandomized knockoffs.

Our proposal to implement the knockoffs method in a clustered data framework is feasible, flexible, and effective. A key strength of the proposed approach is that it can be implemented by exploiting the

available knockoff procedures. This is especially true when the focus is on selecting level 2 variables, while further work is needed to handle binary level 1 predictors.

Post-selection inference in Linear Mixed Models

Presenter: Anna Nikolei

In psychological research, model selection is often performed on the same data that are later used for statistical inference, for example when predictors are selected via best subset selection. This practice alters the sampling distribution of parameter estimates in the selected model and can inflate type I error rates if not appropriately addressed. Although several correction methods exist for general linear models, substantially less work has examined this issue in linear mixed models. The approach proposed by Rügamer et al. (2022) provides a post-selection inference framework for mixed models by numerically approximating the distribution of test statistics after model selection, thereby enabling valid inference conditional on the selected model. Prior simulation studies have investigated type I error control, but only under idealized conditions in which the selected and data-generating models coincide. We conducted a more realistic simulation study that incorporates fixed- and random-effects misspecification and developed strategies to address issues that arise under such conditions.

Eliminating rank deficiency in Bayesian models with the ZeroSumNormal distribution

Presenter: Adrian Seyboldt

In Bayesian multilevel regression with categorical predictors, the standard "drop one level" trick resolves rank deficiency but introduces an arbitrary reference category. While the likelihood is invariant -- so that frequentist estimates don't change -- the resulting prior over identifiable contrasts is not. The choice of reference can meaningfully affect shrinkage. I show how the rank deficiency can be removed using a Normal distribution restricted to vectors with zero sample mean, which I call the ZeroSumNormal distribution. This translates ideas of sum coding from classical regression to Bayesian statistics in a clean and generative way. Using an orthogonal Householder transformation, it can be implemented at linear cost. I discuss extensions to higher-rank constraints (e.g., interaction effects), and show the newly implemented support in PyMC, NumPyro and Stan. I discuss the practical implications for whether effects should be interpreted relative to a sample mean (finite-category settings such as the set of Dutch provinces) or a population mean.

Dynamic Extended Generalized Many-Facet Rasch Model for Rater Cognitive Bias

Presenter: Dr. Giuseppe Mignemi

Rating procedures play a central role in many applied fields, including psychological and medical diagnosis, educational assessment, and peer-review and grant evaluation. A major methodological challenge in these contexts is disentangling multiple sources of variability in ratings, such as subject ability, rater behavior, and item characteristics. To address this challenge, Item Response Theory (IRT)-based models-most notably Multifacet Rasch Models and related hierarchical rater models-have been widely adopted due to their interpretability and ability to jointly estimate subject, rater, and item effects. In this work, we introduce a general and flexible class of Bayesian semiparametric models, referred to as a Dynamic Extended Generalized Many-Facet Rasch Model, that extends existing approaches in several important directions. The proposed framework specifies a measurement model in which both location and scale parameters (i.e., difficulty and discrimination) jointly depend on rater-specific features (e.g., severity and consistency) and item characteristics (e.g., item difficulty and discrimination). As a result, distinct measurement parameters are defined for each item-rater combination, allowing for fine-grained modeling of rater-item interactions.

Our framework serves as a unifying structure encompassing most multi-rater, multi-item models proposed in the literature over the past decades, including Multifacet Rasch Models and Hierarchical Rater Models. Beyond unification, the model introduces two novel rater-related components that have

not been explicitly modeled before. First, we account for rater anchoring bias by introducing a rater-specific cross-lagged effect between the latent attributes of consecutively rated subjects. This parameter captures the tendency of raters to anchor their current judgment to previous evaluations, a well-documented cognitive bias in human decision-making. Second, we model practice and workload effects by allowing rater consistency to vary dynamically over the course of the rating process. This temporal evolution is captured through a Bayesian P-spline defined over the subject order, enabling the detection of learning, fatigue, or adaptation effects at the individual rater level. Subject heterogeneity is addressed through a structural model that employs Bayesian nonparametric priors—specifically Dirichlet Process and Pitman-Yor process mixtures-to flexibly learn the latent ability distribution without restrictive parametric assumptions. We discuss theoretical properties of the model and propose computational strategies to improve Markov chain Monte Carlo efficiency. Simulation studies and real-data applications illustrate the advantages and limitations of the proposed approach. Potential extensions to multidimensional and longitudinal settings are outlined as directions for future research.

A Multilevel Psychometric Framework for Experimental Tasks: The Generalized Hierarchical Factor Model

Presenter: Ricardo Rey-Sáez

Experimental psychologists are increasingly concerned by the low correlations observed between experimental measures, even in tasks designed to measure the same underlying cognitive processes. In practice, researchers often follow a simple workflow: participants complete many trials, trial-level responses are aggregated into a single score per participant (e.g., a mean or difference score), and these scores are then correlated across tasks to draw conclusions about underlying processes. This approach implicitly treats task scores as if they were measured with little error. Yet in many experimental paradigms, trial-to-trial variability is large relative to between-participant variability, so aggregation often produces noisy (i.e., unreliable) individual scores. As the reliability of two measures decreases, their estimated correlation becomes increasingly attenuated relative to the true association. Consequently, weak correlations may be mistakenly taken as evidence for theoretical dissociations when the data are simply too noisy to recover the underlying relationship. In this talk, we present Generalized Hierarchical Factor Models (GenHFMs) as a Bayesian alternative to the "aggregate-then-correlate" approach for estimating individual differences. GenHFMs combine hierarchical modeling, which separates true individual differences from trial-level noise, with psychometric factor models that explain the covariance among these differences via shared latent processes. This framework allows researchers to (a) jointly estimate group-level effects and true individual variability; (b) recover true correlations with unprecedented precision; (c) directly test for the presence of common latent factors; (d) evaluate each task's discriminative capacity to capture these shared processes; and (e) model skewed trial-level distributions, such as the heavy-tailed patterns characteristic of response times. Beyond reliability, our simulation study shows that ignoring trial-level asymmetry (i.e., modeling trial data with a Gaussian distribution) can underestimate true correlations by up to 50%. Across two empirical illustrations, we show that GenHFMs fitted with widely used response-time distributions can yield markedly different correlation estimates and factor-loading patterns, leading to substantively different conclusions. Because the true data-generating process is unknown in empirical settings, we use leave-one-out cross-validation to compare candidate trial-level distributions and select the model with stronger empirical support. As we will show, GenHFMs achieve higher predictive accuracy than other modeling strategies commonly recommended in experimental psychology, such as hierarchical models with covariates or drift-diffusion models.

Session 6 — Missing Data, Causal Inference, and Evidence Synthesis in Multilevel Models (13:45-15:00)

Multiple Imputation of Missing Data in Longitudinal Designs: A Comparison of Different Strategies

Presenter: Mark Lustig

Longitudinal studies are commonly affected by missing data arising, for example, from attrition or wave nonresponse. Multiple imputation (MI) is often recommended to address these missing data, either through multilevel MI, which treats repeated measures as nested within participants and uses multilevel imputation models to explicitly model growth trajectories over time, or single-level MI, which treats repeated measures as separate variables. Previous research has shown that both strategies can perform well in applications of latent curve models (LCMs), but has largely focused on scenarios in which the assumptions underlying multilevel MI were met. In this talk, we first discuss the conceptual similarities and differences between common implementations of single- and multilevel MI for time-structured longitudinal designs. We argue that, in this context, multilevel imputation models often represent constrained versions of single-level models. Consequently, common implementations of multilevel MI can be significantly less flexible than single-level MI, which can, at least in principle, facilitate many different types of analyses in longitudinal designs. We then present results from two simulation studies examining (1) applications of LCMs in which the assumptions of conventional methods for multilevel MI were violated, and (2) applications with multiple-indicator designs, in which multilevel MI is difficult to implement and single-level MI may become computationally unstable due to the large number of variables. Our results indicate that multilevel MI can yield biased estimates and reduced power to detect model misspecification when its assumptions are not met. Single-level MI provided more accurate results; however, in designs with many variables, it required additional strategies for reducing the complexity of the imputation model, such as passive imputation or partial least squares regression.

Multiple imputation of multilevel data with single-level models: A fully conditional specification approach using adjusted group means

Presenter: Prof. Dr. Simon Grund

Missing data are a common challenge in multilevel designs, and multiple imputation (MI) is one of the most commonly recommended methods for handling them. Past research has shown that multilevel MI can be extremely effective at handling missing data in multilevel designs, provided that the imputation model adequately takes the multilevel structure into account. This is particularly important in multilevel analyses that involve nonlinear effects or random slopes, and many specialized methods have been developed for these applications. However, multilevel MI can be difficult to apply in practice, where the multilevel structure is often not very pronounced or not of immediate interest in the intended analyses. In these applications, existing methods can become unstable and often struggle to provide accurate results. In this talk, we present a fully conditional specification approach to multilevel MI that combines single-level imputation methods with group means (GM) or adjusted group means (AGM) to accommodate the multilevel structure. In addition, we present the results for a theoretical investigation and several simulation studies, in which we evaluated the statistical properties of these methods and their performance in different applications, including applications with balanced designs, unbalanced designs, and larger numbers of variables. Finally, we discuss the strengths and weaknesses of these methods and their implications for practice.

Mixed-effects location-scale models in meta-analysis

Presenter: Dr. Wolfgang Viechtbauer

Meta-analysis encompasses a family of statistical methods designed to synthesize the findings of related studies investigating a common phenomenon. At its core, meta-analysis rests on a simple idea: Translate the results of individual studies into a common metric (an ‘effect size’ or some other

quantitative ‘outcome measure’) and then combine these estimates to obtain more precise and generalizable conclusions than any single study can provide.

Standard meta-analytic models can be formulated as (generalized) linear mixed-effects models and hence as multilevel models (e.g., Goldstein et al., 2000; Hox & de Leeuw, 2003; Lambert & Abrams, 1995; Pastor & Lazowski, 2018; Turner et al., 2000; Van den Noortgate & Onghena, 2003) where effect size estimates are nested within studies (e.g., Konstantopoulos, 2011; Van den Noortgate et al. 2013). However, meta-analytic models also exhibit some distinctive features. The sampling variances of the effect size estimates are heteroscedastic, study-specific, and treated as known constants. In addition, the model includes an estimate-level random effect to capture ‘heterogeneity’, that is, variability in the effect sizes beyond what would be expected based on their sampling variances alone.

In this talk, I first describe the connection between meta-analysis and multilevel modeling, highlighting both shared structure and distinctive characteristics. I then discuss the extension of location-scale models to meta-analysis (Viechtbauer & López-López, 2022), which allows modeling potential differences in the degree of heterogeneity across studies. Such models can be readily fitted using the metafor package in R (Viechtbauer, 2010), although recent findings regarding their statistical performance suggest some caution (Blázquez-Rincón et al., 2025).

A key challenge in this context is the typically small number of studies (i.e., the meta-analytic sample size). Moreover, coefficients in the scale component of the model may diverge toward plus or minus infinity, similar to perfect separation in logistic regression. Potential remedies for such cases, for example via the use of Bayesian methods, will be discussed.

Meta-Analytic Pooling of Intraclass Correlation Coefficient Estimates

Presenter: Dr. Bethany Hamilton Bhat

The intraclass correlation coefficient (ICC) is a central parameter in multilevel modeling, representing the proportion of total variance attributable to clustering. ICCs are necessary for prospective power analyses for clustered data. For this, researchers rely on ICC estimates reported in prior studies or in secondary databases; however, any single estimate is subject to sampling error, particularly when drawn from small or dissimilar samples. When only a single ICC estimate is available, the upper limit of the ICC’s confidence interval is often used for power analyses although this results in an unnecessarily large sample size. Researchers with access to multiple ICC estimates use simple averages or Bayesian approaches for power analyses, however, these rely on strong assumptions and can fail to account for differences in ICCs’ precisions. A more rigorous alternative for obtaining the ICC for a power analysis is to use meta-analytic methods that weight each ICC estimate by its precision. Although ICCs have been treated as meta-analytic focal outcomes in psychology, education, and the health sciences, little methodological work has examined best practices for pooling ICC estimates. Pooling ICCs poses two challenges. First, the ICC is a ratio of multilevel model variance components with an unknown sampling distribution bounded between 0 and 1. Empirical and simulation studies show that ICC distributions are often skewed, with shape depending on the true ICC and cluster structure. Second, although multiple large-sample approximations to the ICC’s sampling variance have been proposed, it remains unclear which performs best when used as inverse-variance weights in random-effects meta-analysis of ICCs. Applied meta-analysts have used a wide range of variance formulae and estimation approaches with limited methodological justification. We examined the performance of alternative meta-analytic methods for pooling ICCs via Monte Carlo simulation. For each simulated primary study, participant-level data were generated under a two-level unconditional random effects model with normally distributed level-1 and level-2 residuals.

We manipulated the number of clusters, average cluster size, cluster imbalance, true ICC magnitude, between-study heterogeneity, and the number of studies in the ICC meta-analysis. Variance components were estimated using REML, ICCs computed, and different sampling variance formulae calculated for and compared as meta-analytic weights. Meta-analytic results were compared using REML or a distributionally agnostic method of moments (MoM) estimator. Performance was evaluated using a variety of simulation performance metrics for the pooled ICC and corresponding standard error.

Across conditions, REML and MoM performed similarly in recovering the population ICC, suggesting that both are viable for pooling independent ICCs. The choice of sampling variance formula affected bias and accuracy, with a variance estimator based on Fisher's normalizing transformation consistently producing the least biased and accurate pooled ICC estimates and standard errors. It avoided bias when the true ICC was small to moderate and performed competitively when the ICC was large or cluster sizes were modest. Larger primary-study samples improved recovery across variance formulae, whereas the number of studies pooled had little effect. These findings offer guidance for researchers pooling ICCs for power analyses for clustered designs or for other uses.