Universiteit Utrecht

14th International Multilevel Conference
March 12 & 13, 2024

Conference Program

Organizing committee
Emmeke Aarts
Beth Grandfield
Sara van Erp

Utrecht University
Department Methodology & Statistics

Conference program

Day 1 (March 12)

Room: Kerkzaal

9:00	**Registration**
9:25	Opening
9:30	***Keynote 1: Dr. Mirjam Moerbeek*** *(Utrecht University, The Netherlands)* Sample size calculations.
10:30	*Camila Barragan Ibanez (Utrecht University, The Netherlands)* Method for sample size determination in cluster randomised trials using the Bayes factor.
10:55	*Ulrich Lösener (Utrecht University, The Netherlands)* Hypothesis Evaluation in Multilevel Models with the Approximate Adjusted Fractional Bayes Factor.
11:20	**Coffee and Tea Break**
11:40	*Prof. dr. Bill Browne (Centre for Multilevel Modelling and School of Education, United Kingdom)* Optimal simulation-based sample size calculations for complex multilevel models.
12:10	*Ana Carolina Franco Castiblanco (University of Bremen, Germany)* Sample size determination for multilevel trials with heterogeneous within cluster variance.
12:35	*Dominiek Vollbracht (RPTU Kaiserslautern-Landau, Germany)* Slider Scales vs. Radio Buttons: A Comparison of Psychometric Properties in Experience Sampling Methods.
13:00	**Lunch**
14:00	*Prof. dr. George Leckie (Centre for Multilevel Modelling and School of Education, United Kingdom)* Intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA): A Review, Critique, and New Extensions.
14:30	*Dr. Elif Çoker (Mimar Sinan Fine Arts University, Türkiye)* A Comparative Study on Health Behaviour in Turkish School-Aged Children: Multilevel Regression Models vs. Multilevel Path Models.
14:55	*Dr. Ethan McCormick (Leiden University, The Netherlands)* Deriving mixed-effects models of change with interpretable parameters: linear estimation with nonlinear inference.
15:20	*Prof. dr. Mirka Henninger (University of Basel, Switzerland)* Tree-based machine learning methods for multilevel data: opportunities, pitfalls, and potential solutions.
15:45	**Short Break**
16:00	Poster Session: *David Most* - How might multilevel models might be used to characterize the temporal dimensions of doctoral student outcomes? *Joost Meekes* - Modeling left-censored concentrations of many chemical compounds with true zeros. *Salome Li Keintzel* - Are Larger Distractor Effects associated with Slower Reaction Times in Subsequent Trials? *Cristian Marquez Romo* - Does rising inequality increase perceived social conflict? *Daniel Ventus* - How can we analyze measurement invariance of composites?

17:00	End of Day 1
18:00	**Conference Dinner** *(for those who registered)*

Day 2 (March 13)

Room: Kerkzaal

9:00	**Doors open**
9:25	Opening
9:30	*Dr. Terrence Jorgensen (University of Amsterdam, The Netherlands)* Can Bayesian methods yield more robust estimates of summary statistics for two-stage maximum likelihood estimation of multilevel structural equation models?
10:00	*Javier Aguilar (TU Dortmund, Germany)* Intuitive Joint Priors for Bayesian Linear Multilevel Models: The R2D2M2 prior.
10:25	*Aditi Bhangale (University of Amsterdam, The Netherlands)* Hyperparameters of Prior Distributions for MCMC Estimation of the Multivariate Social Relations Model.
10:50	**Coffee and Tea Break**
11:10	*Dr. Xynthia Kavelaars (Open Universiteit, The Netherlands)* Bayesian analysis of multilevel data from multiple correlated binary outcome variables.
11:35	*Hanne Oberman (Utrecht University, The Netherlands)* Imputation of Incomplete Multilevel Data with R.
12:00	*Dr. Christian Röver (University Medical Center Göttingen, Germany)* How trace plots help illustrating hierarchical models.
12:25	**Lunch**
13:30	*Dr. Chiara di Maria (University of Palermo, Italy)* Structural multilevel models for longitudinal mediation analysis: a definition variable approach.
13:55	*Dr. Joran Jongerling (Tilburg University, The Netherlands)* Robust Autoregressive Modeling: Protecting Against Bias Caused by Omitted Lags Using Random Residual Variances.
14:20	*Dr. Leonie Vogelsmeier (Tilburg University, The Netherlands)* Disentangling changes in careless responding from changes in substantive item interpretation in ecological momentary assessment.
14:45	*Sebastian Mildiner Moraga (Utrecht University, The Netherlands)* A Bayesian multilevel hidden Markov model with Poisson-lognormal emissions for longitudinal count data.
15:10	**Short Break**
15:25	PhD-award ceremony
15:30	***Keynote 2: Prof. Dr. Dan McNeish*** *(Arizona State University, USA)* Measurement in Intensive Longitudinal Data.
16:30	Closing remarks and End of Day 2

Keynote 1: Sample size calculations.
Dr. Mirjam Moerbeek (Utrecht University, The Netherlands)

One of the main steps to be taken in the design of a study is the calculation of sample size. In this presentation I will give a summary of my past, present and future research on this topic, with a focus on cluster randomized trials. With cluster randomized trials, complete clusters such as schools, general practices or neighborhoods are randomized to treatment conditions and all subjects in the same cluster receive the same condition.
The first part of this presentation focused on sample size calculations from a frequentist point of view. It will be shown how to calculate how many clusters and how many subjects per cluster should be included in the trial. These sample sizes can be shown to depend on the intra-class correlation coefficient. An a priori estimate of this model parameter is not always available and various approaches to deal with this will be discussed.
The second part of this presentation focuses on Bayesian sample size calculation. It will be shown how the Bayes factor is used to evaluate informative hypotheses and a criterion for a priori sample size determination is introduced. Furthermore, Bayesian sequential designs are discussed. With such designs, additional subjects are recruited during the course of the study until sufficient support for either informative hypothesis is achieved.

Method for sample size determination in cluster randomised trials using the Bayes factor.
Camila Barragan Ibanez (Utrecht University, The Netherlands)

In the initial phases of designing a research study, a key step is determining the sample size. Employing small sample sizes may lead to underpowered studies, while it is unrealistic to expect from researchers to use a large number of participants considering the limitations in resources and that it may be unethical to involve more participants than necessary. To ensure that a study possesses a sufficient number of participants to secure statistical power, researchers can employ a prior power analysis. Determining the sample size in complex research designs such as cluster randomised trials becomes intricate due to the hierarchical structure of the data, meaning that the sample size needs to be determined at each level. Conventionally, the sample size for this design is based on null hypothesis significance testing, an approach known for its numerous pitfalls. These drawbacks can be avoided by using the Bayes factor instead. While previous studies have proposed methods for determining sample size when using the Bayes factor, 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 the approximated adjusted fractional Bayes factor 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 increases with decreasing effect sizes and with increasing intraclass correlation and Bayes factors. We encourage researchers to use our method when planning a cluster randomised trial where the Bayes factor is used for hypothesis testing.

Hypothesis Evaluation in Multilevel Models with the Approximate Adjusted Fractional Bayes Factor.
Ulrich Lösener (Utrecht University, The Netherlands)

Evaluating hypotheses about multilevel model (MLM) parameters using *Bayes Factors* is a viable, often more straightforward alternative to the frequentist method using *p*-values. However, applied researchers have to make a number of choices in the process which have non-trivial consequences for the inferential process. This raises common questions such as *Which prior distribution should I choose?, How should I calculate the Bayes Factor(s)?,* and *What level of evidence do I consider convincing?*. In this presentation, I elaborate on these queries by means of using the Approximate Adjusted Fractional Bayes Factor as an example to evaluate hypotheses within a simple MLM. For illustrative purposes, only two competing hypotheses about a single model parameter are considered here, though this method can be generalized to more complex cases. Because this talk is part of a larger symposium on Sample Size Determination using Monte Carlo simulation, computational efficiency in hypothesis evaluation is of high priority. This is achieved by normally

approximating the posterior rather than repeatedly sampling from it, significantly reducing the required computation time and resources.
A prominent theme throughout the presentation is the attempt to strike a balance between methodological soundness and accessibility to the applied researcher. This relationship can in many cases be looked upon as a trade-off. Opinions and ideas about this balance from the audience is highly appreciated.

Optimal simulation-based sample size calculations for complex multilevel models.
Prof. dr. Bill Browne (Centre for Multilevel Modelling and School of Education, United Kingdom)

Deciding on how large a study should be is one of the most asked questions of statisticians in collaborative research. For simpler statistical models there are formulae that will produce suggestions for sample size requirements to satisfy criteria like specific hypothesis test power.
As the data to be collected becomes more complex, for example when the data exhibit clustering, both the (multilevel) models used to fit the data and the equivalent sample size calculations also get more complicated.
For simple multilevel models the sample size theory has been extended and packages like PINT can be used however it is often now more common to use simulation-based approaches, either by using specialist packages like MLPowSim or simply writing R code to perform the simulation. When running simulations there are various decisions to make in terms of what designs to simulate, how many simulations to perform and how to use the results of the simulation to calculate the required sample sizes. Often people use a 0/1 type approach where for each of a series of different sample size scenarios, a large number of simulated datasets are generated and the model of interest is fitted with the proportion of simulated datasets that show a significant effect for a test giving an estimate of the power of that test. Then some form of interpolation is used to find the specific sample size that satisfies the desired power.
In this talk we look at approaches that can improve on the simple 0/1 approach by:
(i) capturing more information than simply the significance or not for each simulation
and (ii) utilising transformations to best share information across different scenarios.
We show that by considering these two improvements we can construct simulation-based sample size calculations that are considerably faster and more accurate.

Sample size determination for multilevel trials with heterogeneous within cluster variance.
Ana Carolina Franco Castiblanco (University of Bremen, Germany)

Multilevel data, also known as hierarchical or nested data, refers to data that is organized in multiple levels or layers of observations. The determination of appropriate cluster and sample sizes are an important element in the planning of a multilevel study. Like for classical randomized trials, the goal of the cluster and sample size determination is to achieve a specific target power under a specific predetermined effect. For multilevel trials this step is far more complex than for individually randomized trials for several reasons. One fact that complicates the issue is that the data cannot be regarded as independent and the correlation among observations within the same cluster needs to be accounted for in the sample size calculation. This level of dependency is commonly measured using the intraclass correlation coefficient (ICC). The ICC decompose the total variance in two, the between cluster variance which captures how much the clusters differ from each other with respect to the outcome, and the within cluster variance which measure how dissimilar the individual observations are within the same cluster. The standard sample size formulae assumed that the within cluster variability is homogeneous among clusters. In practice, however, the within cluster variability may not be constant and the standard formula may be biased. We propose a sample size formula for multilevel trials (with constant treatment effect) when the within cluster variability is heterogeneous, for both constant and variable cluster-wise sample sizes, and its simplification for two and three level trials. Additionally, we illustrate how the variance components can be estimated based on conditional means and variances. Furthermore, we conduct a simulation study to investigate the behavior of the proposed sample size formula and variance components estimation and we compare it with the standard sample size formulae and the estimation of the variance components via multilevel linear models.

Slider Scales vs. Radio Buttons: A Comparison of Psychometric Properties in Experience Sampling Methods.
Dominiek Vollbracht (RPTU Kaiserslautern-Landau, Germany)

Slider scales, a type of visual analogue scale, are commonly used as a response format in smartphone-based experience sampling methods. This may be due to several advantages of slider scales, including the use of a metric scale instead of a categorical (ordinal) scale (as in radio buttons), and the ease with which participants can respond to a slider scale on their smartphone touchscreen. However, only a few studies have compared the psychometric properties of slider scales with those of classic radio buttons (e.g. Likert scales), and a limitation of the existing studies is that they (mainly) relied on cross-sectional data and used manifest variable models. In this research, our goal was to scrutinize the psychometric properties of the two response formats (slider scale vs. radio buttons) in experience sampling methods using latent variable models. In an ongoing experience sampling study, we manipulated the response scale format between persons experimentally (21 days, 4 measurement occasions per day, planned N of individuals = 320) by presenting multiple items measuring different latent constructs using a slider scale in one group and a radio button scale in the other group. We will use a Multigroup Multilevel Structural Equation Modeling framework to compare the experimental groups in terms of measurement invariance, within-person variability, reliability (two-level omega), and validity (correlations with other constructs). The project was preregistered before the start of data collection, which is almost complete. Preliminary results show negligible differences between response formats. Final results will be presented at the conference.

Intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA): A Review, Critique, and New Extensions.

Prof. dr. George Leckie (Centre for Multilevel Modelling and School of Education, United Kingdom)

Intersectional Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA) is a recently proposed approach to investigate sociodemographic intersectional inequalities in individual outcomes.

The approach proves to be an unusual application of two-level individuals-within-clusters random-intercept linear regression. Specifically, the clusters are not organizations or areas, but population strata defined by combinations of individuals’ sociodemographic characteristics, such as age, gender, ethnicity, and socioeconomic status.

Intersectional MAIHDA is argued to efficiently estimate mean outcomes across strata, capture the extent to which differences in individual outcomes are explained by sociodemographic characteristics versus unmodelled individual-level factors, and measure the extent to which sociodemographic characteristic effects are additive or multiplicative. Particular interest is associated with whether the approach can be used to reliably identify higher-order interactions between sociodemographic characteristics.

In this presentation, we first review the intersectional MAIHDA approach. We then draw attention to some unusual assumptions and results implied by modelling strata as clusters. We then discuss potential extensions to this approach and their implications for studying intersectionality, including allowing the regression slopes on an additional predictor and the outcome variance to vary across strata.

A Comparative Study on Health Behaviour in Turkish School-Aged Children: Multilevel Regression Models vs. Multilevel Path Models.

Dr. Elif Çoker (Mimar Sinan Fine Arts University, Türkiye)

A person's health behaviours include their habits, choices, and practices to protect their health. These behaviours can affect an individual's physical, mental, and emotional health. In this context, the health of our children, who are the representatives of the future adult generation, holds significant importance for our society. Investing in children's health is a fundamental and crucial step to ensure the development of a healthy and successful generation, enhance community welfare, and create a sustainable future.

This study aims to analyze the current health behaviors of school-aged children in Türkiye by utilizing data from the "Health Behaviour in School Aged Children: A World Health Organization Collaborative Cross-National Study" (HBSC) survey, currently conducted in 51 different countries and regions worldwide.

The dataset used in this study is taken from the latest available HBSC survey which is from 2018 and includes the information about the 2017/2018 period for Türkiye. In total, there are 5848 students who are taken from 288 classes, which are nested within 97 schools and 29 regions. The main interest is on the Body Mass Index (BMI), which is an indicator of students’ health measure.

Since there are so many variables to predict BMI, firstly factor analysis is performed for the purpose of data reduction. Following the identification of factors, both multilevel regression models and multilevel path models are employed to investigate BMI. The results from both models are compared and discussed.
And, finally, the results obtained from this study will be compared with those of the 2006 survey, which was conducted before the previous one in 2010.

Deriving mixed-effects models of change with interpretable parameters: linear estimation with nonlinear inference.
Dr. Ethan McCormick (Leiden University, The Netherlands)

Linear multilevel models are a mainstay of the psychological and behavioral sciences, able to model hierarchical and longitudinal data (Raudenbush & Bryk, 2002). While amazingly useful models, widely implemented in available software, and with attractive estimation properties, linear parameter models nevertheless impose many restrictions on the ability to ask specific theoretical questions because of the need to formulate a model into a linear equation. To address these issues, prior work has derived alternative nonlinear expressions. Unfortunately, these models do not appear as standard options in major software packages, and many applied researchers remain unaware of their potential utility. Additionally, nonlinear expressions present additional estimation challenges – especially in growth modeling contexts with random effects – and for these reasons, largely remain the provenance of researchers with training in and access to more advanced statistical methods and software options.

Here, I address several extant issues for formulating multilevel models with interpretable and meaningful parameters, with an eye for expanding the utility and accessibility of these approaches. First, I review a history of and motivation for interpretable parameter models and walk through a general approach for deriving new parameters of interest (Cudeck & du Toit, 2002; McNeish et al., 2021). I then extend these principles and derive two alternative forms of a cubic polynomial with meaningful parameters and show how this new model is related to the standard linear parameter version. I also consider a multiphase version of this model which can serve as an approximation of S-shaped nonlinear models (e.g., logistics). To address the common estimation issues with nonlinear versions of these alternative models, I lay out an approach of linear estimation with nonlinear inference (LENI), where the standard linear parameter model is estimated, and then results are transformed *post hoc* into the parameters of interest from the nonlinear alternative models. I derive transformation equations for the point estimates and standard errors of fixed, random, and conditional effects, allowing inferences to be made on the meaningful parameters as if we had directly estimated the nonlinear equation. Finally, consider extensions of the LENI framework to fit multilevel models of any known nonlinear equation.

I highlight the utility of the LENI framework for addressing substantive questions of interest by considering sex-specific trajectories of learning during adolescence and young adulthood.

Tree-based machine learning methods for multilevel data: opportunities, pitfalls, and potential solutions.
Prof. dr. Mirka Henninger (University of Basel, Switzerland)

Machine learning methods, such as decision trees or random forests, are robust, yet powerful methods to capture and interpret complex dynamics and non-linear effects of predictor variables on outcomes. As a result, they have gained popularity in psychological research in recent years. Initial attempts have been made to adopt these machine learning methods for multilevel data, with a focus on integrating random effects structures. However, these adaptations have not yet addressed the role of predictor variables, namely whether they are assessed on Level-1 or Level-2. First, we demonstrate through simulation studies that existing decision tree and random forest extensions for multilevel data exhibit an increased type-1 error rate for Level-2 predictor variables. This risk of inaccurately selecting Level-2 predictor variables becomes more pronounced with higher intra-class correlations. In contrast, the level of the predictor variable appears to have a negligible impact on prediction accuracy, as seen in variable importance measures.
Second, we illustrate a possible application scenario of the multilevel decision tree using Level-1 predictor variables. Specifically, we employ the multilevel decision tree to predict moments of emotional similarity in romantic couples using time-varying covariates measured on Level-1. We conclude by highlighting the need for future developments and propose a potential solution by

incorporating an alternative version of the score-based test statistic into the multilevel decision tree.

Can Bayesian methods yield more robust estimates of summary statistics for two-stage maximum likelihood estimation of multilevel structural equation models?
Dr. Terrence Jorgensen (University of Amsterdam, The Netherlands)

The multilevel structural equation model (ML-SEM) is a generalized latent variable model that includes as special cases the well-known univariate multilevel model (MLM) and structural equation model (SEM) for single-level multivariate data. An advantage of SEM is that for normally distributed complete data, software accepts summary statistics (a covariance matrix of modeled variables) as input, rather than raw data. In contrast, software for fitting a ML-SEM (e.g., M*plus* and the R package lavaan) offer full-information maximum likelihood (FIML) estimation, which therefore require raw data as input. However, two-stage limited-information estimators have been proposed for ML-SEM, which estimate level-specific covariance matrices in Stage 1, followed by standard SEM in Stage 2. The advantage of two-stage approaches is that an ML-SEM can be estimated using any standard SEM software, where each level of analysis is treated as a group in a multigroup SEM. For balanced cluster sizes, Muthén (1994) showed how it was possible to obtain maximum likelihood estimates from level-specific summary statistics, by placing scaling constraints on the Level-2 SEM that are proportional to the common cluster size. The disadvantage of this so-called “MuML” estimator is that in practice, cluster sizes typically vary, so the necessary constraints (which are tedious to specify) yield biased estimates to the degree that they are not realistic. Yuan & Bentler (2007) later proposed fitting a separate SEM for each level of analysis, to enable evaluating each level’s model unambiguously. This simplifies model evaluation relative to Ryu & West’s (2009) method of saturating all levels except the one whose model is evaluated. Unfortunately, Yuan & Bentler only provided a SAS program to implement the estimation of level-specific covariance matrices and the corrected standard errors and test statistics for Stage-2 SEM. The limited availability of that highly expensive software has prevented wide-scale adoption of their method, but open-source alternatives are now quite prevalent, especially in the R environment. In this talk, I will revisit Yuan & Bentler’s idea for two-stage maximum likelihood estimation of level-specific SEMs, employing Markov chain Monte Carlo (MCMC) estimation for Stage-1 estimation of level-specific summary statistics. A benefit of MCMC estimation (and Bayesian methods in general) is the ability to incorporate prior beliefs and evidence by specifying informative priors, which make it possible to supplement a lack of information from small samples. This is particularly problematic for ML-SEM, which can provide unstable estimation of Level-2 model parameters even when there are 100 clusters, depending on the complexity of the model. I demonstrate a way to use the R package blavaan to obtain posterior estimates of level-specific covariance matrices, as well as uncertain about those point estimates. This Stage-1 information can then be passed to lavaan to obtain Stage-2 estimates of SEM parameters. Although the example application is rather arbitrary (because blavaan can simply be used to fit the SEM of interest using MCMC in one step), it is a proof of concept that has promising implications for more complex scenarios (e.g., 3-level or cross-classified models, or ML-SEM for ordinal data).

Intuitive Joint Priors for Bayesian Linear Multilevel Models: The R2D2M2 prior.
Javier Aguilar (TU Dortmund, Germany)

The training of high-dimensional regression models on comparably sparse data is an important yet complicated topic, especially when there are many more model parameters than observations in the data. From a Bayesian perspective, inference in such cases can be achieved with the help of shrinkage prior distributions, at least for generalized linear models. However, real-world data usually possess multilevel structures, such as repeated measurements or natural groupings of individuals, which existing shrinkage priors are not built to deal with.

We generalize and extend one of these priors, the R2D2 prior by Zhang et al. (2020), to linear multilevel models leading to what we call the R2D2M2 prior. The proposed prior enables both local and global shrinkage of the model parameters. It comes with interpretable hyperparameters, which we show to be intrinsically related to vital properties of the prior, such as rates of concentration around the origin, tail behavior, and amount of shrinkage the prior exerts.

We offer guidelines on how to select the prior's hyperparameters by deriving shrinkage factors and measuring the effective number of non-zero model coefficients. Hence, the user can readily evaluate and interpret the amount of shrinkage implied by a specific choice of hyperparameters.

Finally, we perform extensive experiments on simulated and real data, showing that our inference procedure for the prior is well calibrated, has desirable global and local regularization properties and enables the reliable and interpretable estimation of much more complex Bayesian multilevel models than was previously possible.

Hyperparameters of Prior Distributions for MCMC Estimation of the Multivariate Social Relations Model.

Aditi Bhangale (University of Amsterdam, The Netherlands)

The social relations model (SRM) is a linear random-effects model applied to examine multivariate dyadic data (e.g., round-robin data) within social networks. Such data have a unique multilevel structure in that dyads are cross-classified within individuals who may be nested within different social networks. The SRM decomposes perceptual measures into multiple components at two levels: individual-level random effects (incoming and outgoing effects) and dyad-level residuals (relationship effects), the associations among which are often of substantive interest. Nestler (2018) proposed maximum likelihood (ML) estimation to estimate multivariate SRM. We propose a Bayesian estimator, specifically Markov chain Monte Carlo (MCMC), for multivariate SRMs. MCMC provides some practical advantages to estimating complex or analytically intractable models, but its accuracy may vary depending on the priors specified. In this study, we compare ML estimation to MCMC using various prior distributions. In a Monte Carlo experiment, we manipulate the accuracy of the location of prior distributions, as well as their precision, when estimating a trivariate SRM using MCMC. We compare the accuracy and efficiency of ML and MCMC point (and interval) estimates when round-robin data are normally distributed.

Bayesian analysis of multilevel data from multiple correlated binary outcome variables.

Dr. Xynthia Kavelaars (Open Universiteit, The Netherlands)

In social, and behavioral research, datasets with a multilevel structure and multiple correlated dependent (binary) variables are common. These data are frequently collected from a study population that distinguishes several subpopulations with different (i.e., heterogeneous) effects of an intervention. Despite the frequent occurrence of data with such a multilevel, multivariate, and heterogeneous structure, methods to analyze these aspects together are less common. Researchers therefore sometimes resort to either ignoring the multilevel and/or heterogeneous structure, analyzing only a single dependent variable, or a combination of these. These analysis strategies are suboptimal: Ignoring multilevel structures inflates Type I error rates, while neglecting the multivariate or heterogeneous structure might mask detailed insights and subtle nuances in the data.

To analyze such data comprehensively, a recent Bayesian framework for statistical decision-making regarding (treatment) superiority can be used. During this talk, the three elements of the framework will be discussed in the context of multilevel data. First, a Bayesian multilevel multivariate logistic regression model is introduced. The analysis model is suitable to analyze data while taking their clustered, heterogeneous, and multivariate nature into account. Second, a transformation procedure to facilitate interpretation is discussed. This transformation procedure aims to express results in terms posterior success probabilities and differences between them rather than the less intuitive multivariate logistic regression parameters. Third, the complementary decision procedure allows for posterior inferences regarding (treatment) superiority with accurate Type I error rates.

Together, these three elements can be used to predict treatment effects and to make superiority decisions within subpopulations, while taking advantage of the size of the entire study sample and while properly incorporating the multilevel structure of the data.

Imputation of Incomplete Multilevel Data with R.

Hanne Oberman (Utrecht University, The Netherlands)

Incomplete multilevel data requires careful consideration of the missing data problem and analysis strategy. In this tutorial, we focus on a popular strategy for accommodating missingness in multilevel data: replacing the missing data with plausible values, i.e., imputation.

Imputation separates the missing data problem from the analysis of scientific interest. Consequently, the completed data can be analyzed as if it had been fully observed, without added complexity in the analysis of scientific interest.
This tutorial illustrates the imputation of incomplete multilevel data with the statistical programming language *R*. We aim to show how imputation can yield less biased estimates from incomplete clustered data. We provide practical guidelines and code snippets for different missing data situations, including non-ignorable missingness mechanisms. For reasons of brevity, we focus primarily on multilevel imputation using chained equations with the popular *R* package *mice*, in combination with other *R* packages which are used for applications and visualizations.
The case study datasets cover typical data structures from the social and biomedical sciences. These include an example of clustering in individual patient data meta-analyses and a 'missing not at random' missingness mechanism.

How trace plots help illustrating hierarchical models.
Dr. Christian Röver (University Medical Center Göttingen, Germany)

The trace plot is a helpful tool for interpreting results of a meta-analysis; it shows the dependence of (conditional) estimates on the magnitude of the between-study variance component (the *heterogeneity*), and it may serve as a sensitivity check or as an illustration of the "inner workings" of the analysis. We introduce the trace plot in the context of the simple hierarchical model commonly used for meta-analysis, we discuss relevant extensions as well as Bayesian and frequentist variations. We will also touch upon possible applications in more general hierarchical models.

Structural multilevel models for longitudinal mediation analysis: a definition variable approach.
Dr. Chiara di Maria (University of Palermo, Italy)

Mediation analysis is used to assess the direct effect of an exposure on an outcome, and the indirect effect transmitted by a third intermediate variable. Longitudinal data are the most suited to address mediation, since they allow mediational effects to manifest over time. There exist several approaches to deal with longitudinal mediation analysis, and one of the most widely spread, especially in social and behavioural sciences, consists of using multilevel models. However, when applied to mediational settings, these models present some limitations, for example the difficulties in estimating the covariance between random effects belonging to different models (the mediator and the outcome model), and the fact that it is impossible to model a relationship where a level-2 variable depends on a level-1 variable. All these shortcomings can be overcome moving to a structural perspective. We propose a new formalisation of multilevel models within a structural framework combining the reticular action model notation and the definition variable approach. We reconsider two multilevel mediation designs very frequent in longitudinal settings from this structural perspective, discuss the advantages and limitations of such an approach and provide an empirical example.

Robust Autoregressive Modeling: Protecting Against Bias Caused by Omitted Lags Using Random Residual Variances.
Dr. Joran Jongerling (Tilburg University, The Netherlands)

Experience sampling data is often analyzed using multilevel first-order autoregressive (AR(1)) models in which individuals' current scores are regressed on the immediate preceding one (i.e., a so-called lag-1 or first-order effect). These models assume that a person's scores from more than one measurement ago can be ignored when modeling the current one; that is, the current score does not have to be predicted, for example, from the preceding two or three observations (i.e., there is no lag-2 or lag-3 effect). There is little theoretical or empirical work to support this assumption, however. In fact, analyses of real data suggest that multiple preceding measurements are often significantly related to the current one and that the number of lags that need to be considered varies from person to person. Therefore, analyzing experience sampling data with a multilevel AR(1) model (that assumes the same lag-1 structure for everyone) can lead to biased results.
This problem is not easy to solve since there is no consensus on what methods work best for detecting the number of lags one should include in the model, and as mentioned above, the number of relevant lags might differ between individuals, which is hard to incorporate into one overarching multilevel model for all participants. In this study, we therefore check if we can at least make analyses robust against the erroneous omission of lagged effects by including random

residual variances in the model. Residual variances reflect all factors influencing the process under study not explicitly included in the model. As such, they should also “capture” the influence of higher-order lagged effects that are not explicitly included in the model. If this indeed the case, we should still be able to get unbiased estimates of the lag-1 effect even if we don’t include all the lagged effects that are truly present in the data. However, since the scores on the outcome variable vary across time for each individual, and individuals will also have different mean scores on the outcome variable, the omitted higher-order lagged effects will also differ across time and individuals. To properly partial out these higher-order lagged effects, we therefore need to allow the residual variance to be random across individuals and/or time.
We therefore undertake an extensive simulation study to investigate under what circumstances a multilevel AR(1) model with residual variances that are random across time and/or person provides accurate lag-1 parameter (i.e., AR(1) parameter) estimates if the true data generating process is a higher-order (e.g., AR(2) or AR(3)) autoregressive model.

Disentangling changes in careless responding from changes in substantive item interpretation in ecological momentary assessment.

Dr. Leonie Vogelsmeier (Tilburg University, The Netherlands)

Intensive longitudinal data collected via methods like ecological momentary assessment (EMA) have great potential for studying the dynamics of psychological constructs such as well-being in daily life. However, a key challenge is detecting and accounting individual- and occasion-specific **careless and insufficient effort responding (C/IER)**, which is vital for accurate inferences about the dynamics of psychological constructs. EMA is especially vulnerable for C/IER because the intense assessment burdens participants significantly.
Although studied excessively in cross-sectional research, only a few studies have focused on C/IER in EMA. A common approach is including additional items to assess attentiveness (e.g., instructional manipulation check items where participants are asked to choose a specific response option) or monitoring aberrancies in response patterns such as multivariate outliers. Both approaches have limitations: the former lengthens the already burdensome EMA questionnaires, and the latter has to be tailored for the different forms C/IER can take. Furthermore, both approaches require the (arbitrary) choice of thresholds and the uncertainty in flagging observations as C/IER is not captured. Therefore, rather than relying solely on indicators, methods for detecting C/IER that use theory-based statistical models (specifying a mixture of factor analytic or item response theory measurement models for attentive responses and unstructured distributions from C/IER) are preferred. These methods use probabilistic assignments of responses to the attentive and C/IER components, respectively. Initial studies have shown the potential for the use of model-based approaches to detect C/IER in EMA. However, a major concern is that these approaches assume measurement invariance in the measurement model of the attentive responders, which is easily violated, for example, when item interpretation differs across situations. If violated, the invariance assumption can falsely flag answers as C/IER or fail to do so.
The recently proposed **latent Markov factor analysis (LMFA)** has the potential to disentangle changes in C/IER and substantive changes in the measurement model for attentive individuals as this exploratory method captures any changes in the measurement model, whether induced by C/IER or substantive changes in item interpretation. LMFA combines a discrete- or continuous-time latent Markov model with mixture exploratory factor analysis to cluster observations into separate latent states that differ in the measurement models underlying the responses in EMA. However, LMFA is not yet tailored to capturing C/IER, which leads to severe convergence problems, possibly because the exploratory factor analyses applied in all the latent states make distributional assumptions that are violated for C/IER (e.g., if responses stem from distributions other than the normal distribution). These convergence issues can, however, be reduced by applying specific model constraints as suggested in the cross-sectional literature.
In this talk, I will first explain how LMFA with the imposed constraints can be used to disentangle changes in C/IER and substantive item interpretation in EMA and, thus, how to flag C/IER in the presence of measurement non-invariance. Subsequently, I will show how well the adjusted LMFA works with regard to accurately determining the appropriate number of underlying measurement models and assigning observations to these models for different types of C/IER encountered in EMA.

A Bayesian multilevel hidden Markov model with Poisson-lognormal emissions for longitudinal count data.

Sebastian Mildiner Moraga (Utrecht University, The Netherlands)

Hidden Markov models (HMMs) are probabilistic methods in which observations are seen as realizations of a latent Markov process with discrete states that switch over time. Moving beyond standard statistical tests, HMMs offer a statistical environment to optimally exploit the information present in multivariate time series, uncovering the latent dynamics that rule them. Here, we extend the Poisson HMM to the multilevel framework, accommodating variability between individuals with continuously distributed individual random effects following a lognormal distribution. The multilevel HMM proposed allows a probabilistic decoding of the sequence of hidden states underlying multivariate count time-series data based on individual-specific parameters and offers a framework to measure between-individual variability formally. We assess the estimation performance of the multilevel HMM for count time-series under different conditions of between-individual heterogeneity with a Monte Carlo study, and we show that it outperforms an equivalent single-level HMM. Finally, we illustrate how to use our model to explore the latent dynamics governing complex multivariate count data in an empirical application concerning pilot whale diving behaviour in the wild, and how to identify neural states from multi-electrode recordings of motor neural cortex activity in a macaque monkey in an experimental set up. We make the multilevel HMM introduced in this study publicly available in an extension to the R-package mHMMbayes in CRAN.

Keynote 2: Measurement in Intensive Longitudinal Data.

Prof. Dr. Dan McNeish (Arizona State University, USA)

Intensive longitudinal data – where participants are measured many times over a short duration – have recently increased in popularity due to technological advances like wearables and smartphones. Ecological momentary assessment (EMA) is one such study design that has been particularly common in behavioral research when studying mood or affect. Models in EMA studies often feature outcomes and predictors that are created from sum scoring Likert-type or binary item responses at each time point. However, the behavioral literature has recently emphasized the importance of psychometrics and potential benefits of more modern psychometric approaches such as factor analysis and item response theory, especially relating to assessing measurement invariance across time and people. This talk discusses how to combine psychometric approaches with multilevel models for EMA designs to increase measurement precision and more accurately reflect the construct being studied. The proposed model is applied to motivating EMA data from a study on people with binge eating disorder to demonstrate the importance of psychometrics in intensive longitudinal designs. Specifically, statistical models are only as good as the data and variables to which they are applied – if scores on behavioral variables are imprecisely created, conclusions could be driven by inadequate measurement practices rather than the underlying dynamics of the construct of interest.