Universiteit Utrecht

11th International Multilevel Conference
April 12 & 13, 2017

Conference Program
& Abstracts

Organizing committee
Leoniek Wijngaards – de Meij (chair)
Mirjam Moerbeek
Rens van de Schoot
Dan McNeish

Utrecht University
Department Methodology & Statistics

Conference program

Day 1 (April 12) morning

	Room: Kerkzaal	Room: Huiskamer
9:00	**Registration**	
9:25	**Opening**	
9:30	**Keynote 1** Stapleton, Laura (University of Maryland, USA) *When clustering is endogenous*	
10:10	Loeys, Tom (Ghent University, Belgium) *A cautionary note on centering lower level interaction effects in multilevel models*	
10:30	Rosche, Benjamin (Utrecht University, The Netherlands) *The impact of ignoring multiple membership structures in multilevel models with survival endpoints*	
10:50	**Coffee and Tea Break**	
11:10	Doretti, Marco (University of Perugia, Italy) *Assessing the performance of nursing homes through a multilevel Latent Markov model*	Kuiper, Rebecca M. (University Utrecht, The Netherlands) *Studying multi-person time-lagged effects using real-life-self-report-data: It is time to go continuously (in multilevel modeling)*
11:30	Firth, Nick (University of Sheffield, United Kingdom) *Beyond therapist effects: Community effects and the contribution of indicators of deprivation and complexity to psychotherapy outcomes*	Karch, Julian D. (Max Planck Institute for Human Development, Berlin, Germany) *Gaussian process panel modeling – A new flexible modeling approach for longitudinal data*
11:50	Goffette, Céine (Université Paris-Saclay, France) *Does the household context matter for smoking? A European overview*	Souren, Pierre M. (Radboud University, Nijmegen, The Netherlands) *Interrater-reliability in case of missings and many raters with multilevel analysis*
12:10	Franzen, Minita (University of Groningen, The Netherlands) *Intra- and Interindividual Variability in the Affective, Behavioural, and Perceptual Effects of Alcohol Consumption in a Social Context*	Matta, Tyler H. (University of Oslo, Norway) *Partitioning measurement error and residual variance in longitudinal analysis with observed scores.*

Day 1 (April 12) afternoon

	Room: Kerkzaal	*Room: Huiskamer*
12:30	**Lunch**	
13:30	Almansa, Josue (University of Groningen, University Medical Center Groningen, The Netherlands) *Multivariate latent-class trajectories: modelling simultaneously several sets of growth classes.*	Manzi, Giancarlo (University of Milan, Italy) *Waiting for Godot: A multilevel framework for peer-review waiting times and research impact*
13:50	Vidotto, Davide (Tilburg University, The Netherlands) *Bayesian Multilevel Latent Class Models for the Multiple Imputation of Nested Categorical Data*	Bein, Christoph (Nederlands interdisciplinair Demografisch Instituut and *University of Groningen, The Netherlands)* *A multilevel analysis of the impact of religiosity on fertility intentions*
14:10	**Short break**	
14:20	Vogelsmeier, Leonie V.D.E. (Tilburg University, The Netherlands) *Latent Markov Factor Analysis for Exploring Within-Subject Measurement Model Differences in Experience Sampling Studies*	
14:40	Dudgeon, Paul (The University of Melbourne, Australia) *Improvements to Robust Confidence Intervals in Latent Growth Modelling*	
15:00	**Coffee and Tea Break**	
15:20	Leckie, George (Centre for Multilevel Modelling and Graduate School of Education, United Kingdom) *Avoiding bias when estimating the consistency and stability of value-added school effects*	
15:40	Rosseel, Yves (Ghent University, Belgium) *Multilevel SEM: history, computational approaches and software*	
16:00	**End of day 1**	
19:00	**Conference dinner**	

Day 2 (April 13) morning

	Room: Kerkzaal	*Room: Huiskamer*
9:00	**Doors open**	
	Young Researcher Award nominees	
9:30	Pouwels, J. Loes (Radboud University, Nijmegen, The Netherlands) *Stability of Peer Victimization: A Multi-Level Meta-Analysis of Longitudinal Research*	
9:50	Arroyo Resino, Delia (Complutense University of Madrid, Spain) *Adjustments of a matrix structures in multilevel growth models*	
10:10	Baumann, Petra M. (University of Graz, Austria) *The impact of fitting a hierarchical linear model to an ordinal variable with few response categories. A Monte Carlo simulation study.*	
10:30	Talloen, Wouter (Ghent University, Belgium) *Bootstrap versions of the Hausman test for cluster-level endogeneity in random slope models*	
10:50	**Coffee and Tea Break**	
11:10	Paccagnella, Omar (University of Padua, Italy) *New Insights on Students' Evaluation of Teaching in Italy*	Pillinger, Rebecca (University of Edinburgh, United Kingdom) *Understanding and exploring the results of models with random slopes on polynomial terms*
11:30	Flunger, Barbara (Utrecht University, The Netherlands) *The role of teacher differences for students' homework learning types: Applying multilevel latent profile analyses*	Teerenstra, Steven (Radboud Institute for Health Sciences, Nijmegen, The Netherlands) *Sample size and power calculations for Stepped Wedge trials with >2 levels*
11:50	Quené, Hugo (Utrecht University, the Netherlands) *Individual differences in contrasting similar [s] sounds across languages*	Moerbeek, Mirjam (Utrecht University) *The consequences of treatment non-compliance in cluster randomized trials*
12:10	Koch, Tobias (Leuphana University Lüneburg, Germany) *When the mimic approach fails - Transforming explanatory variables in g-factor models for single- and multilevel data*	Chiou, Hawjeng (National Taiwan Norman University, Taiwan, Republic of China) *The impacts of intra-class correlation (ICC) and item number on the estimation of manifest and latent contextual effect: A comparison of Bayesian and ML approach*

Day 2 (April 13) afternoon

12:30	**Lunch**
13:30	Ippel, Lianne (Tilburg University, The Netherlands) *Predicting Individual-level effects in a click-stream*
13:50	Jak, Suzanne (University of Amsterdam, The Netherlands) *Relating measurement invariance, cross-level invariance and multilevel reliability*
14:10	**Short break**
14:20	Smid, Sanne C. (Utrecht University, The Netherlands) *Bayesian vs Maximum Likelihood Estimation for Multilevel Models with Small Samples: A Systematic Review*
14:40	McNeish, Dan (University of North Carolina, Chapel Hill, USA) *Multilevel Mediation with Small Samples: A Cautionary Note on Multilevel Structural Equation Modeling*
15:00	**Coffee and Tea Break**
15:20	**PhD-award ceremony**
15:30	**Keynote 2** Hamaker, Ellen (Utrecht University, The Netherlands) *At the frontiers of dynamic multilevel modeling*
16:00	**End of day 2**

Abstracts

Multivariate latent-class trajectories: modelling simultaneously several sets of growth classes.

Almansa, J.[1]
[1] Department of Health Sciences, Division of Community and Occupational Medicine, University of Groningen, University Medical Center Groningen, The Netherlands

Summary
GMM and LCGA model distinctive trajectories of a specific outcome over time. It is possible to extend these analyses to a multivariate set of longitudinal outcomes, but in this case there is a huge number of possible ways to model their association across all outcomes (within and/or between individuals, within and/or between classes, etc). In my experience so far, the selection of the specific method has to do much more with the research question and the nature of the data than with any statistical criteria - but I'd like to open this for discussion.
I intend to show a limited number of possibilities for analyzing multiple LCGA's simultaneously. Then, I will present (at least) one example on real data: Trajectories of mental health and work outcomes (depression, anxiety, work functioning and return to work percentage) in patients with common mental disorders, who have returned to work after a (short) sick leave. The association among the four sets of trajectories was captured by means of a hierarchical 2-level latent class, which 'summarizes' the trajectory-class membership probabilities of all outcomes.

Keywords
Multivariate growth; trajectories; hierarchical latent classes.

Adjustments of a matrix structures in multilevel growth models

Arroyo Resino, Delia*
*'PhD-student', supervisor: María Castro Morera,
COMPLUTENSE UNIVERSITY OF MADRID, SPAIN

Summary
This research presents the results of a secondary analysis of data obtained in an educational assessment in Spain. The data reflect a multilevel structure with three levels (time, students and classrooms) and three measures of scientific competence (longitudinal structure). The students had evaluated at the end of the academic years 2010-2011 (4th grade) 2011-2012 (5th grade), and 2012-2013 (6th grade). The general objective is to propose a growth model based on the data of a real evaluation in Spain, with a sample of 2441 students in the second and third cycle of Primary Education (E.P) from 71 schools To measure performance in scientific competence, we used standardized tests, linked to international vertical performance scales. The data analysis was made with the SPSS statistical program, which helped estimating growth models based on Hierarchical Linear Models (HLM) and Mixed Linear Models (MLM) with repeated measures. The results show that the adjustment of both HML and MLM, a specific residual variance-covariance structure is necessary. In the case of HML, in order to obtain the adjustment to the data, the final model required a change of the intrasubjet error (time level), considered as a single value by an autoregressive matrix, a diagonal matrix for the student level and a structure matriz for the classroom level. In the case of MLM we used autorregresive matrix for the time level, an identity matrix for the student level and an unstructured matrix for the classroom level. Once the adjustment of both models has been achieved, the quality indexes, AIC and BIC, show that the MLM shows a bigger adjustement to the data of a real educational evaluation in Spain

Keywords:
multilevel, matrix structure, longitudinal

New Insights on Students’ Evaluation of Teaching in Italy

Bassi, F.[1], Grilli, L.[2], Paccagnella, O.[1*], Rampichini, C.[2], Varriale, R.[3]
[1] Department of Statistical Sciences, University of Padua, Italy
[2] Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy
[3] ISTAT-Italian National Statistical Institute, Italy
* Presenting author

Summary
Students’ opinions and judgements of teaching performances play a substantial role in higher education; the relationship between student-, teacher-, course-specific characteristics and student evaluation of teaching is the topic of a huge amount of works in the literature.
It is generally accepted that a multilevel analysis of the students’ ratings is a satisfactory approach for investigating teaching evaluations, because of the hierarchical nature of the data (i.e. university students nested into classes).
This work aims at enriching the multilevel literature on the student evaluation of teaching proposing some original analyses based on a wider set of teacher-specific characteristics, including also teachers’ opinions on their teaching activities.
This work exploits an innovative and original dataset available at the University of Padua, obtained after linkage of survey and administrative data coming from three different sources: first, the conventional survey on the student evaluation of teaching carried out among university students (for the academic year 2012-2013); second, administrative data related to the main features of the teachers and didactical activities they are involved in the academic year 2012-2013; third, a new CAWI survey carried out by means of the research project PRODID (*Teacher professional development and academic educational innovation*). The PRODID project started at the University of Padua in 2014, with the aim of developing strategies to support academic teachers and enhance their teaching competences. A specific questionnaire was then developed and addressed to all professors involved in any didactical activity during the academic year 2012-2013 (the response rate was slightly lower than 50%). This new survey collected opinions, beliefs and needs of the university professors, with regard to their teaching activities developed in the university classes.

The final dataset is then composed by about 49000 students' evaluations and allows to further separate the set of teacher-specific characteristics in objective and subjective characteristics.
On the one hand, findings of this work support the need of taking into account the hierarchical nature of these data and the heterogeneity of the different classes. On the other hand, the role of the teacher perceptions and needs on their teaching activities is deeply investigated, highlighting how these characteristics affect all other level-1 and level-2 variables.

Keywords
Record linkage; student evaluation of teaching; teacher opinions.

The impact of fitting a hierarchical linear model to an ordinal variable with few response categories. A Monte Carlo simulation study.

Baumann, Petra Martina[1]*
[1] University of Graz, Department of Sociology, Austria
* Presenting author (PhD-student, supervisor: Prof. Johanna Muckenhuber)

Summary

Fitting linear regression models to ordinal response variables is a common – though disputed – practice in sociology, which has, by analogy, been extended to hierarchical linear models (HLMs). The reasons for this are manifold, but not necessarily rooted in statistical theory—ease of interpretation of linear regression coefficients, believed robustness of linear models to violations of model assumptions, or unfamiliarity with alternative modelling options.

If the robustness belief was reasonable, it would indeed be preferable to fit the more straightforward, less computationally demanding model. From a statistical viewpoint, however, this belief does not seem justified. HLMs require a continuous response which, conditional on the covariates, is normally distributed. But ordinal variables, in general, are not. Nevertheless, papers fitting HLMs to ordinal variables with as few as three response categories are published in highly-ranked sociological journals.

The aim of this paper is therefore to check if this practice is justified and evaluate its impact on model results. The thematic background for this study is country-comparative research which is a popular field for the application of multilevel modelling in sociology.

In order to illustrate the real-life implications of fitting a linear multilevel model to an ordinal response, a published model (Lyness et al. 2010) is reproduced with real data (ISSP and additional sources for the country level) and model diagnostics are applied. The model chosen is a rather extreme example: a hierarchical linear model fitted to an ordinal variable with only three response categories. Note that the reproduction of a published study is not done to expose a single account of (possible) bad practice. It rather serves as a demonstration of the statistical problem in applied sociological research.

Additionally, the parameters of the published model act as a starting point for the second part of the paper, a Monte Carlo simulation. This part, again, focusses on an ordinal variable with three response categories. This narrow focus allows for greater variation of other sample, data and model settings.

Two different population-generating processes are applied. First, a linear model with a continuous response is used. The response variable is then split into an ordinal variable by two cut-off points (cf. Carsey/Harden 2014). This process mimics the common (and convenient) assumption that an ordinal variable stems from a latent continuous distribution. Second, a generalized linear mixed model (GLMM) is used for data generation to mimic a “truly” ordinal variable. To achieve comparable estimates, the linear model estimates are first transformed to the appropriate scale of the GLMM (cf. Bauer/Sterba 2011). In both population-generating processes the sample sizes, the distribution of the response variable as well as the size of the random effects are varied.
Model performance is assessed in terms of bias, efficiency, MSE, coverage rate of the 95% CI as well as by comparing the same model applied to the continuous response and model diagnostics. The latter two are done to illustrate the results in a way more approachable for the applied researcher whom this paper ultimately aims to serve.

Keywords
Hierarchical linear model, ordinal response, Monte Carlo simulation

A multilevel analysis of the impact of religiosity on fertility intentions

Bein, Christoph[1], Gauthier, Anne H.[2], Mynarska, Monika [3]
[1] PhD-student (presenting author), Nederlands interdisciplinair Demografisch Instituut and University of Groningen
[2] PhD supervisor, Nederlands interdisciplinair Demografisch Instituut and University of Groningen
[3] PhD supervisor, Uniwersytet Kardynała Stefana Wyszyńskiego, Warsaw, Poland

Abstract
The aim of this study is to investigate the role of religiosity in forming fertility intentions, using a cross-national perspective. Religion and its impact on fertility constitutes a reemerging field in demographic research and coincides with a general renewed interest in religion in public discourse (Hubert 2015). In the past, the focus of research lied mostly on examining differences in fertility behavior among denominations (e.g. van Poppel (1985)). In view of an increasing secularization and availability of suitable micro data, a shift occurred towards studying the influence of actual religious behavior or individual religiosity. While the relationship between religiosity and fertility outcomes has been thoroughly analyzed in many different settings (Kaufmann 2010; Hubert 2015; Skirbekk, Stonawski et al. 2015; Peri-Rotem 2016), research on religiosity's influence on intentions has been limited so far, especially using multilevel analysis.
The purpose of this paper is to model the micro and macro effects of religiosity using a cross-national design. Specifically, we are interested in the effect of religiosity on fertility intentions at the micro level and how this effect is moderated by the country context at the macro level. Drawn from the existing literature, we will be testing three hypotheses. The first hypothesis is that we expect more religious people to have higher intentions for having children. The second hypothesis concerns effects by birth order: we expect religiosity to have a stronger influence on higher birth order intentions than on lower ones. Thus, the lowest influence is expected to be on the intention to have a first child. Finally, it is expected that religiosity has a stronger impact on fertility intentions in countries
having less supportive policies towards employed mothers. In these countries, women have to actively decide between having a family or a career. More religious women therefore might be more inclined towards having a family, fulfilling traditional gender roles promoted by most religions.

The dataset used for the analyses consists of the wave 1 of the Generations and Gender Survey (GGS). This is a large cross-national survey based on nationally representative samples of the population aged 18-79 years old. In our case, we restrict the analysis to men and women of childbearing age (18-49 years old) in the 14 European countries which include data on our key variables.[1] Our dependent variable is the intention of having or not a(nother) child within the next three years. Answers ranged from “definitely yes” to “definitely no”. Our key independent variable is religiosity, measured here by the frequency of attending religious services (coded into “never”, “less than monthly”, “monthly and more often”).
Because the number of countries included in our analysis is small, and because of concerns over possible biased standard of errors (Bryan and Jenkins 2016), we will follow the two-step approach used by Koops, Liefbroer and Gauthier (2016) in first running a standard multilevel model, and then in testing its robustness by performing a meta regression.

[1] Bulgaria, Russia, Georgia, Germany, France, Italy, Netherlands, Romania, Norway, Austria, Lithuania, Poland, Czech Republic and Sweden

Keywords
Fertility intentions, religiosity, GGP

References

Bryan, M. L. and S. P. Jenkins (2016). "Multilevel Modelling of Country Effects: A Cautionary Tale." European Sociological Review **32**(1): 3-22.
Hubert, S. (2015). The impact of Religiosity on Fertility. Wiesbaden, Springer VS.
Kaufmann, E. P. (2010). Shall the religious inherit the earth? : demography and politics in the twenty-first century. London, Profile Books.
Koops, J. C., A. C. Liefbroer and A. H. Gauthier (2016). Cross-national variation in the effect of parental educational attainment on becoming a cohabiting mother: the influence of economic inequality and norms towards family formation.
Peri-Rotem, N. (2016). "Religion and Fertility in Western Europe: Trends Across Cohorts in Britain, France and the Netherlands." Eur J Popul **32**: 231-265.
van Poppel, F. W. A. (1985). "Late fertility decline in the Netherlands: The influence of religious denomination, socioeconomic group and region." European Journal of Population / Revue européenne de Démographie **1**(4): 347-373.
Skirbekk, V., M. Stonawski, S. Fukuda, T. Spoorenberg, C. Hackett and R. Muttarak (2015). "Is Buddhism the low fertility religion of Asia?" Demographic Research **32**: 1-28.

The impacts of intra-class correlation (ICC) and item number on the estimation of manifest and latent contextual effect: A comparison of Bayesian and ML approach

Dr. Chiou, Hawjeng*, and Dr. Lin, Pi-Fang
Department of Business Administration
National Taiwan Norman University, Taiwan, Republic of China

Summary

In multilevel analysis, the definition and estimation of collective constructs (e.g., organizational climate in the management research and classroom climate in the education study) involve the issues of measurement errors (related to the number of items been used) and sample errors (related to sample size) at the different levels. In general, the magnitude of intra-class correlation (ICC) reflects the influence of the sample size, while the factor analysis is available for dealing with measurement error. In order to take both issues into account, the multilevel structural equation modeling (MSEM) is proposed in the literatures. However, the assumption of normality as well as the statistical power of MSEM are based on the enough sample size on both macro (n_c) and micro level (n_j), i.e., $n_c \geq 100$ and $n_j \geq 30$. In combination with the sample size issue, the role of item number is not clear in the MSEM. The present study introduces the Bayesian estimate into the MSEM for dealing with the challenges of sample size and the measurement error. A simple Monte Carlo simulation shown that while both the sample size was large, the performance of Bayesian and ML estimates were similar if the ICC of measured variables were huge (i.e. .50). In contract, if the ICC was small (i.e. .10), Bayesian approach was superior to ML estimates in terms of lower mean square error and higher coverage rate. The impacts of item number (three vs. six items) of latent variables on both the macro and micro level shown no differences with both estimation methods. An empirical dataset contained 38 companies and 1200 employees were adopted to explore the efficiency of Bayesian MSEM in the estimation of collective constructs as well as contextual effects. Further simulation is recommended for more detail consideration on the various conditions of the measurement model of the MSEM, such as factor loadings and correlations among latent variables. The procedures of Bayesian MSEM with methodological implications were discussed in this study.

Keywords multilevel structural equation modeling, contextual variable, Bayesian inference

Assessing the performance of nursing homes through a multilevel Latent Markov model

Doretti, M.*, Montanari, G.E., Bartolucci, F.
University of Perugia, Italy

Summary

In this work, we study the temporal evolution of the health status of a number of elderly patients hosted in some nursing homes. Specifically, we consider a longitudinal dataset gathered from the Long Term Care Facilities (LTCF) Programme, a public health care plan developed in Umbria, a region of Italy. The LTCF data we focus on consist of repeatedly administered questionnaires during the years 2012-2013, in average around 4 waves per individual. These questionnaires measure several aspects of patients' general health status, which is the unobserved characteristic of main interest. Considering this and the longitudinal structure of the data, Latent Markov models represent a suitable modelling strategy in this context.

A Latent Markov model typically assumes that a set of categorical outcome variables (e.g., questionnaire items) is measured at a number of time occasions and probabilistically influenced by a latent process (e.g., health status). The latent process is modelled like a first order discrete-time Markovian process with a finite number of states (Bartolucci et. al., 2013). In this model, three sets of parameters fully describe the assumed structure: conditional response probabilities (probabilities of specific outcome categories given the latent state), initial probabilities (probabilities of latent states at the first measurement occasion) and transition probabilities (probabilities of latent states at following occasions given previous latent state). We also allow initial and transition probabilities to depend on certain individual covariates.

The data are also characterized by a multilevel structure due to patients being hosted in different nursing homes. Taking such a multilevel structure into account is in order. As a matter of fact, the goal of our analysis is to provide a ranking of nursing homes according to their capability in preserving/improving their patients' health status. In principle, this aim could be achieved by including nursing homes effects as fixed effects in the regression equations of initial and transition probabilities. However, this approach is unfeasible in practice due to the low sample sizes of some nursing homes. To overcome such a problem, a bivariate random effect is included to account for nursing home effects on initial and transition probabilities respectively, so that our model can be interpreted as a mixed Latent Markov model (Maruotti, 2011). Maximum likelihood estimates of

relevant parameters are obtained by a suitable optimization algorithm, whereas the posterior (i.e. given outcome variables) distribution of random effects is used to assess nursing homes' performances. As usual in latent variable models, sensitivity of results to the number of latent states is explored.

References

Bartolucci, F., A. Farcomeni and F. Pennoni (2013). *Latent Markov Models for Longitudinal Data*. Statistics in the Social and Behavioural Sciences. Chapman & Hall/CRC.

Maruotti, A. (2011). Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review 79(3)*, 427-454.

Keywords

Latent Markov models, multilevel analysis, nursing homes performance.

Improvements to Robust Confidence Intervals in Latent Growth Modelling

Dudgeon, P.
School of Psychological Sciences
The University of Melbourne. Victoria. 3010
AUSTRALIA

Summary
It is well-known that heteroscedastic-consistent (HC) estimators are preferable to using least squares for the calculation of standard errors and large-sample confidence intervals (CIs) in linear regression under misspecification. It is also well-established that HC3 or HC5 versions are superior in linear regression to the original Huber-White HC0 proposal (see, e.g., Long & Everitt, 2000). The HC0 estimator is sometimes referred to as a "sandwich" or "robust" estimator, and it is the same estimator used in structural equation modelling (SEM) when researchers request, e.g., MLR as the estimation method in Mplus or *lavaan*. This talk proposes a method to calculate HC3 and HC5 standard errors, and associated confidence intervals, in conditional latent growth modelling (LGM). It investigates whether these alternatives are an improvement over current robust procedures using Monte Carlo simulations in which both model misspecification and non-normality are incorporated into the design. Results indicate that these two HC estimators are notably better than the current MLR approach. An additional improvement is gained for variance estimates by employing a confidence interval transformation for bounded parameters proposed by Browne (1982).

Long, J. S., & Ervin, L. H. (2000). Using heteroskedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217-224.
Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.) *Topics in applied multivariate analysis* (pp. 72-141). London: CUP.

Keywords
Latent growth modelling; robustness; confidence intervals.

Beyond therapist effects: Community effects and the contribution of indicators of deprivation and complexity to psychotherapy outcomes

Firth, N.[1*], Saxon, D. [1*], Barkham, M.[1]
[1] University of Sheffield, United Kingdom
* Presenting author

Summary

Introduction: Evidence indicates that around 5-10% of unexplained variance in clinical psychotherapy outcome is attributable to the individual therapist delivering the intervention (known as a "therapist effect"). This study aimed to extend this body of evidence by a) determining the proportion of unexplained variance at the level of the therapy provider organisation (which we have termed a "community effect"), and b) testing demographic and process variables at each level to try to explain this effect. Although previous research supports the idea of a "neighbourhood effect" on individual physical health, we are not aware of any comparable research to date on the subject of psychotherapy outcome.

Method: The sample comprised data from 26,888 patients, seen by 462 therapists, across 30 therapy provider organisations across the United Kingdom. A three level model was constructed (patient, therapist, community). The dependent variable was patients' log-transformed post-therapy symptom severity, as measured by the Clinical Outcomes in Routine Evaluation - Outcome Measure (CORE-OM). At the patient level, explanatory variables included initial symptom severity, employment status, ethnicity, age, sessions offered, sessions attended, percentage of sessions attended. Aggregates of these variables were also included at the therapist and community levels. The type of provider organisation was also included at the community level. Before and after testing explanatory variables, Markov-Chain Monte Carlo estimation was used to produce 95% confidence intervals for the detected therapist and community effects.

Results: With no explanatory variables, a community effect of 8.2% was detected, compared with a significantly smaller therapist effect of 3.1%. However, after adding explanatory variables, the community effect was significantly reduced to 2.1%, whilst the therapist effect did not significantly change (3.4%).

Initial symptom severity explained 30% of unexplained variance at the community level, with patient employment status explaining another 15%. The type of therapy provider organisation (secondary care vs. other) explained a further 15%. Perhaps most interestingly, the proportion of ethnic minority patients accessing the provider organisation explained a further 20% of unexplained variance at the community level (over five times that explained by individual ethnicity).
Conclusions: The therapy provider organisation is important in determining clinical outcomes in psychotherapy – potentially a stronger predictor than the individual therapist. However, much of this effect can be explained by variables linked to deprivation and mental health complexity. These findings indicate that, in addition to individual factors, the patient's community context also appears to influence their therapeutic outcome. From this study alone, it is difficult to determine whether these findings implicate organisational factors (i.e. in the *clinical* community - budget, resource allocation, etc.) and/or the demographics of the neighbourhood in which the patient lives (i.e. in the *geographical* community - socioeconomic status, social support, community identity, etc.). Thus, further research is recommended to disentangle these uncertainties. In particular, work investigating specific indicators of deprivation is recommended, along with comparisons of organisations working in high ethnic minority versus high ethnic majority contexts.

Keywords
Psychotherapy, Community, Deprivation

The role of teacher differences for students' homework learning types: Applying multilevel latent profile analyses

Flunger, B.[1*], Trautwein, U.[2], Nagengast, B.[2], Lüdtke, O.[3], Niggli, A. [4], & Schnyder, I. [5]

[1] Utrecht University, Netherlands

[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen; Germany

[3] IPN, Leibniz Institute for Science and Mathematics Education, Germany

[4] College of Teacher Education Fribourg, Switzerland

[5] University of Fribourg, Switzerland

* Presenting author

Summary

The present study applied multilevel latent profile analyses (MLPA) to a longitudinal data set with 1812 eighth-grade students. MLPA offer the examination of several un-resolved questions regarding students' homework behavior. That is, recent research identified distinct learning types with regards to students' homework behavior, by considering students' differences regarding homework effort and time spent on homework (Flunger et al., 2015). These learning types were characterized by mixed patterns of high time investment (high effort and struggling learners) or low time investment (fast learners, average students, minimalists) and homework effort. However, data on students' homework behavior has an inherent multilevel structure (students are nested within teachers); the impact of teachers on students' homework behavior has been confirmed in several studies (Núñez et al., 2015; Trautwein, Lüdtke, Schnyder, & Niggli, 2006). Therefore, the current study used MLPA to investigate the dependence of the student homework learning types on their teachers at two time points. The sample consisted of Swiss students who had been surveyed twice within a school year on their homework behavior in French as a second language. Different specifications of the multilevel structure of the data were compared. Thereby, MLPA models with two to six latent student profiles and two to five latent teacher profiles were estimated. The models were compared considering a set of classification criteria (AIC3, individual-based and group-based BICs), the student class probabilities across the teacher classes and the interpretability of the solution. At both time points, the models specifying three teacher profiles and five student homework learning profiles were preferred. At the first time point, one teacher profile was characterized by low probabilities of the

two student profiles with high time investment (i.e., high effort learners and struggling learners). Regarding the probabilities of the other homework learning types, the results revealed little differences across the teacher profiles. Also at the second time point, the student class probabilities of all five student profiles did not show great variations over the teacher profiles. In further analyses, the role of covariates for the classification of both student and teachers to distinct latent profiles will be investigated.

Intra- and Interindividual Variability in the Affective, Behavioural, and Perceptual Effects of Alcohol Consumption in a Social Context

Franzen, M.[1*], Sadikaj, G.[2], Moskowitz, D.S. [2], Ostafin, B.D.[1], aan het Rot, M.[1]

[1] Department of Psychology, University of Groningen, the Netherlands

[2] Department of Psychology, McGill University, Canada

* Presenting author, PhD-student. Supervisor: aan het Rot, M.; Contact Information: Minita Franzen, Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands. Phone: +31-50-363.6763 Email: m.franzen@rug.nl

Summary

Objectives: Responses to alcohol can be desirable (e.g., better mood) or undesirable (e.g., aggressive behaviour). Because such heterogeneity in responses can be observed within a person as well as between persons, the study of drinking and alcohol-related problems at the intraindividual and interindividual level is warranted. In this study, we examined the influence of interindividual differences in alcohol use on the intraindividual associations between drinking occurrence and interpersonal behaviours, perceptions, and affect during naturally occurring social interactions.

Methods: For 14 consecutive days, 219 Psychology freshmen (55% female; M_{age} = 20.7 years, *SD* = 2.18) used their personal smartphones to record how they felt, behaved, and perceived others in social interactions soon after they occurred. Interpersonal behaviours and perceptions were assessed in terms of dominance, submissiveness, agreeableness, and quarrelsomeness. Participants also reported the number of alcoholic drinks consumed within 3 hours of each interaction. In separate analyses, we considered the intraindividual associations of (1) having a drinking episode or (2) the number of drinks during an episode with interpersonal behaviours, perceptions, and affect. Further, we examined interindividual differences in drinking frequency and intensity as potential moderators of these intraindividual effects. Data were analysed in SAS 9.4 using PROC MIXED with a maximum likelihood estimation. The data had a two-level structure. Affect, behaviours, perceptions, drinking, and number of drinks were event-level variables, and frequency and intensity of drinking were person-level variables.

Results: At the intraindividual level, results suggested that when persons either had consumed alcohol or had more alcohol than usual, they reported (1) behaviour that was simultaneously more quarrelsome and more agreeable and behaviour that was simultaneously less dominant and less submissive; (2) experiencing more positive affect, and (3) perceiving others as more agreeable.
At the interindividual level, when they had drunk more than usual, more frequent drinkers perceived others as more dominant than less frequent drinkers. Furthermore, during a drinking episode in which more alcohol was consumed than usual, more intense drinkers reported behaving more dominantly and experiencing less pleasant affect than less intense drinkers.
Conclusion: The present results indicate intraindividual variability in how alcohol affects interpersonal behaviours, perceptions, and affect. Our findings are consistent with laboratory-based research indicating subjective responses depending on rising or falling blood alcohol levels.
The present results also suggest a differential susceptibility to the effects of alcohol during naturally occurring social interactions among drinkers with varying drinking frequency and intensity. It appears that people who consume alcohol more often or in greater quantities than others experience effects of drinking that are less pleasant than the effects reported by more moderate drinkers. These findings are not in line with previous research showing that frequent and intense drinkers display a greater stimulating and hedonic subjective response to alcohol intoxication and perceive less sedative and aversive effects than more moderate drinkers.

Keywords
Alcohol, social interactions, individual differences

Does the household context matter for smoking? A European overview

Goffette, C.[1]
[1] CREST, ENSAE, Université Paris-Saclay; France

Summary
One out of four European Union citizens aged 15 and over is a current smoker, but country variations of smoking prevalences are strong. According to Eurostat, the proportion of current smokers ranges from approximately one-sixth in Sweden (16.7%) and the United Kingdom (17.3%) to nearly one-third in Greece (32.6%) and Bulgaria (34.8%). Explanations of those differences have mainly been driven by socio-economic considerations and have hitherto mainly focused on micro (individual level) and macro (national level) determinants of smoking. It seems however that the interactions, norms transmission and social control within the family are a crucial dimension of tobacco consumption.
This paper aims at investigating whether the close context in which people live affects their practices, beyond their individual characteristics. Is there evidence for a household effect on smoking? In other words, do household factors (both observable and unobservable) affect individual probabilities of smoking, all other individual characteristics being equal? And if there is a household effect at play, is it sensitive to the national context? The timing of cigarette adoption and cessation has been described in terms of a diffusion process, and European countries are at different stages of this process. Does the household effect vary according to the stage of diffusion? Data from the European Community Household Panel (ECHP) are used. The dataset provides information on smoking practices for 10 countries (Austria, Belgium, Denmark, Finland, France, Greece, Ireland, Italy, Portugal and Spain). The ECHP recorded characteristics and practices of all members of the interviewed households. This hierarchical structure (individuals nested within households) allows identifying and quantifying the sources of the variation of smoking practices. Two-level random intercept logistic models with a Mundlak specification are implemented for each country in order to disentangle household-level from individual-level contributions to smoking patterns.

Results show that a substantial part of the variability in the propensity to be a daily smoker can be attributed to the household level and that there are important variations across countries (from 30% in Greece to 50% in Denmark). There is a correlation between the strength of the household effect and the stage of diffusion of smoking practices in a country (ρ=0,65): the later the stage of diffusion, the greater the household effect. I therefore hypothesize that a process of polarization between households goes with smoking diffusion: the growing social stigma smokers face leads to processes of sorting, and the household context becomes increasingly important. Further research is needed to test this hypothesis: while this paper compares countries at one time point, complementary analyses should examine data sets that cover a period of several decades for selected countries.

Keywords
Smoking, household, European comparative analysis

Predicting Individual-level effects in a click-stream

Ippel, L.[1*], Kaptein, M.C.[1], Vermunt, J.K.[1]
[1] Tilburg University, the Netherlands
* Presenting author, PhD supervised by dr M.C. Kaptein and Prof. dr. J.K. Vermunt

Summary
In the last few years, it has become increasingly easy to collect data from individuals over long periods of time. Examples include smart-phone applications used to track movements with GPS, web-log data tracking individuals' browsing behavior, and longitudinal (cohort) studies where many individuals are monitored over an extensive period of time. All these datasets cover a large number of individuals and collect data on the same individuals repeatedly, causing a nested structure in the data. Moreover, the data collection is essentially never `finished' as new data keep streaming in.
It is well known that predictions that use the data of the individual whose individual-level effect is predicted in combination with the data of all the other individuals, are better than those that just use the individual average. However when data are nested and streaming, and the outcome variable is binary, computing these individual-level predictions is computationally challenging.
In this presentation, we introduce four computationally efficient estimation methods which do not revise “old” data and do take into account the nested data structure. The four methods that we developed are based on four existing shrinkage factors: the James Stein estimator, (approximate) Maximum likelihood based shrinkage factor, Beta Binomial shrinkage factor, and a heuristic shrinkage factor determined by the number of observations per individual. A shrinkage factor predicts an individual-level effect (i.e., the probability to score a 1), by weighing the individual mean and the mean over all data points. In an extensive simulation study, we compared the performance of existing and newly developed shrinkage factors. We find that the existing methods differ in their prediction accuracy, but the differences in accuracy between our novel shrinkage factors and the existing methods are small. Our online implementation of the well-known shrinkage factors are however computationally feasible in the context of streaming data.

Keywords Data streams, Online learning, Shrinkage factors

Relating measurement invariance, cross-level invariance and multilevel reliability

Jak, S.[1*], Jorgensen, T.D.[1]
[1] University of Amsterdam, The Netherlands
* Presenting author

Summary
Data often have a nested, multilevel structure, for example when data are collected from children in classrooms. These kind of data complicate the evaluation of reliability and measurement invariance, because several properties can be evaluated at both the individual level and the cluster level, as well as across levels. For example, cross-level invariance implies equal factor loadings across levels, which is needed to give latent variables at the two levels a similar interpretation. Reliability at a specific level refers to the ratio of true score variance over total variance at that level. This paper aims to shine light on the relation between reliability, cross-level invariance and strong factorial invariance across clusters in multilevel data. Specifically, we will illustrate how strong factorial invariance across clusters implies cross-level invariance and perfect reliability at the between-level in multilevel factor models.

Keywords
Measurement invariance, cross-level invariance, multilevel reliability

Gaussian process panel modeling – A new flexible modeling approach for longitudinal data

Karch, Julian D.[1*], Voelkle, Manuel C. [2], Brandmaier, Andreas M. [1]
[1] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany
[2] Psychological Research Methods Group, Department of Psychology, Humboldt University of Berlin, Berlin, Germany
* Presenting author

Summary
In the past decades, state-space modeling has become a valuable analysis method for psychological data, in particular for the analysis of time series and longitudinal panel data. Originally developed as a time series analysis method, e.g., in engineering applications, state-space modeling allows only formulating assumptions for the within-subject variability over time. Gaussian processes time series modeling, popular in physics and computer science, is a nonparametric Bayesian time-series analysis method and is an even more flexible approach than state-space modeling. In order to be also applicable for the analysis of longitudinal panel data, a between-subject model is required to allow for multi-subject modeling: For each subject, a parametric model for the time series is formulated using Gaussian process time series modeling. On top, a between-subject model for the distribution of the subject-level model parameters is specified. This approach has already proven useful for multi-subject state-space models. Drawing upon this idea, we here present an extension of Gaussian Process time series modeling to multi-subject models. We call the result Gaussian process panel modeling (GPPM).

The main advantage of GPPM is that it is a very flexible method. Most popular modeling approaches for both time-series and longitudinal panel data can be considered a special case of GPPM. This does not only include state-space modeling both in its time-discrete and its time-continuous variant but also hierarchical linear modeling and structural equation modeling. Thus, GPPM can, for example, represent continuous-time dynamic multilevel models as well as continuous-time structural equation models, both relatively recent developments.
In addition, the generality of GPPM allows formulating novel models by either combining approaches from different traditions or by relying on the wealth of models used in Gaussian process time series modeling. As an

example, we present the squared exponential Gaussian process model, which implements the generic assumption of smooth process trajectories. We show that it is related to the continuous-time autoregressive model as well as the generalized additive model, which has recently been proposed for the analysis of psychological time series data.

In summary, GPPM is a generalization as well as an extension of existing multilevel techniques for the analysis of longitudinal panel data. We are therefore confident to enable researchers to answer new sets of challenging research questions in a unified framework that were more difficult to address before.

Keywords

Longitudinal data, continuous-time modeling, Bayesian analysis

When the mimic approach fails - Transforming explanatory variables in *g*-factor models for single- and multilevel data

Authors: Tobias Koch[1], Jana Holtmann[2], Johannes Bohn[2], & Michael Eid[2]
[1] Leuphana University Lüneburg, Germany
[2] Freie Universität Berlin, Germany

Summary

Psychologists typically relate (untransformed) explanatory variables to latent factors in single- or multilevel confirmatory factor models. This modeling strategy is often referred to as multiple indicator multiple causes (mimic) approach. Because of its intuitive appeal, the mimic approach has been recommended in many textbooks and is commonly applied in practice. However, whenever explanatory variables are simultaneously correlated with general and specific factors in *g*-factor types of models, the classical mimic approach fails. In this talk, we show which consequences can be expected when using the classical mimic approach in combination with *g*-factor models. Specifically, we demonstrate that model misspecification and parameter bias are direct consequences of relating untransformed explanatory variables to the general and specific factors in *g*-factor models (e.g., bi-factor models, latent state-trait models, multilevel CFA models, or CTCM-1 models). We present two alternative modeling strategies that can be used to circumvent these methodological problems: The multiconstruct bi-factor and the residual approach. The multiconstruct bi-factor approach is recommended for explaining general and specific factors in multilevel designs. The residual approach is most useful for explaining general and specific factors in singlelevel designs. Using real data from a multimethod and longitudinal study, the two modeling approaches are illustrated. The advantages and limitations of both modeling approaches are discussed.

Key words

G-factor models, bi-factor models, multitrait-multimethod analysis, longitudinal analysis, state-trait models, MIMIC models, CTC(M-1) model.

Studying multi-person time-lagged effects using real-life-self-report-data: It is time to go continuously (in multilevel modeling)

Kuiper, R.M.[1*]

[1] University Utrecht, the Netherlands

* Presenting author

Summary

The emergence of devices, such as smartphones, led to an exponential increase in real-time self-report data studies in the social and medical sciences. These data are collected with the *Experience Sampling Method* (ESM) in which participants register their experiences (e.g., feelings or symptoms) multiple times a day for several consecutive days. ESM-data offer the unique opportunity to model everyday processes as they unfold over time and to investigate *cross-lagged relationships*, that is, the effects variables have on each other. The latter are of interest when researchers want to *examine hypotheses* such as 'Stress *causally dominates* anxiety'. Moreover, researchers' hypotheses can be at subgroup-level and/or at person-level (e.g., 'Stress is the driving force for Anne'). Allowing for such *differentiation* is important in the *development of person-tailored treatments*: e.g., if Anne's driving force is stress, while Bill's is anxiety, then Anne is likely to benefit from stress-management, whereas Bill probably improves more when his anxiety is treated.

Unfortunately, the existing techniques for analyzing cross-lagged relationships in ESM-data fall short. On one hand, *discrete-time models* may lead to erroneous conclusions, because it is based on the assumption that all the intervals between observations are equidistant, while they are characteristically not in ESM-data. On the other hand, *multilevel continuous-time models*, which are suited for unequally spaced data, are underdeveloped: e.g., they have to assume that individually-varying cross-lagged effects equate, which forecloses causal dominance. In addition, *statistical techniques for evaluating hypotheses* regarding causal dominance, such as posed above, are lacking.

The *aim of my research in the next few years* is to enable i) modeling of individually-varying cross-lagged relationships without undesirable assumptions and ii) evaluation of hypotheses regarding those relationships (at different levels). Therewith, I provide researchers with *tools to gain better insight* in the way processes affect each other over time and to evaluate their hypotheses regarding these relationships.

Keywords Cross-lagged relationships, multilevel (multivariate) continuous-time model, model selection

Avoiding bias when estimating the consistency and stability of value-added school effects

Leckie, G.[1*]
[1] Centre for Multilevel Modelling and Graduate School of Education, United Kingdom

Summary
The traditional approach to estimating the consistency of school effects across subject areas and the stability of school effects across time is to fit separate value-added multilevel models to each subject or cohort and to correlate the resulting empirical Bayes predictions. We show that this gives biased correlations and these biases cannot be avoided by simply correlating 'unshruken' or 'reflated' versions of these predicted random effects. In contrast, we show that fitting a joint value-added multilevel multivariate response model simultaneously to all subjects or cohorts directly gives unbiased estimates of the correlations of interest. There is no need to correlate the resulting empirical Bayes predictions and indeed we show that this should again be avoided as the resulting correlations are also biased. We illustrate our arguments with separate applications to measuring the consistency and stability of school effects in primary and secondary school settings.

Keywords
multivariate response, school effects, value-added

A cautionary note on centering lower level interaction effects in multilevel models

Loeys, T.[1*], Josephy, H.[1]

[1] Department of Data Analysis, Faculty of Psychology and Educational Science, Belgium

Summary

In hierarchical data the effect of a lower level predictor on an outcome may often be confounded by an (un)measured upper level factor. When such confounding is left unaddressed, the effect of the lower level predictor is estimated with bias. Separating this effect into a within- and between-component removes such bias in a linear random intercept model under a specific set of assumptions for the confounder. When the effect of the lower level predictor is moderated by another lower level predictor, an interaction between both lower level predictors is included in the model. To address unmeasured upper level confounding, that interaction term should be separated into a within- and between-component as well. This can be achieved by first multiplying both predictors and centering that product term next, or vice versa. We show that the former centering approach is much more efficient and more robust against misspecification of the effects of cross-level and upper level terms as compared to the latter.

Keywords

Centering, Confounding, Interactions

Waiting for Godot: A multilevel framework for peer-review waiting times and research impact

Manzi, Giancarlo[1*]
[1] Department of Economics, Management and Quantitative Methods, University of Milan, Italy
* Presenting author

Summary (max. 500 words)
Multilevel research frameworks are scarcely encountered in the scientometric literature. Only during the last few years have scientometric journals started to consider and encourage the use of multilevel methods to analyze the characteristics of science and scientific research (a recent example can be found in Bornmann et al. (2011), where a multilevel meta-analysis is performed on the correlation between the h index and multiple h index variants). The structure article<-author<-journal<-publisher can be viewed as a multilevel structure and therefore exploited to reveal hidden characteristics of the scientific research production.
In the last years a dramatically increasing waiting time is in front of researchers aiming at submitting articles to top journals. Figure 1 describes the yearly average number of days between submission and last revision for published articles for the case of the Advances in Data Analysis and Classification journal, and depicts a scenario where it has more than doubled from 224 in 2013 to 474 in 2016. This is a situation that many journals are experiencing, due to lack of peer reviewers, increasing number of submissions, increasing competition, 'publish or perish' policies among universities, etc.

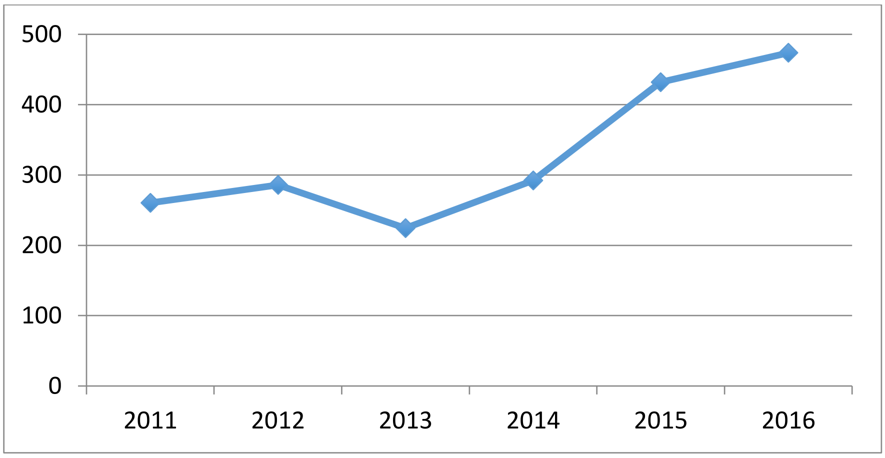

Figure 1. Average number of waiting days between submission and acceptance – "Advances in Data Analysis and Classification" – 2011/2016 (Impact Factor: 1.707; Thompson WOS Statistics & Probability rank: 23)

This paper illustrates the results of a multilevel analysis focused on finding determinants affecting peer-review waiting times until acceptance and research impact across more than 1,000 articles published in the last six years in 9 statistics & probability journals (chosen among the top 25 journals in the 2016 Thomson Web of Science Statistics & Probability ranking) from different publishers. The analysis is subdivided between a short-term analysis (focusing on the period 2014-2016) when article citations can still be considered "a matter of chance", and a medium-term analysis (focusing on the 2011-2013 period) in order to respectively account for an immediate impact and evaluate the "maturity" of the articles. Further to an overall multilevel analysis where first level units are articles, second-level units are journals and third-level units are publishers, a within-journal analysis is also conducted. Statistical multilevel tools chosen for this analysis are both frequentist and Bayesian, whereas a longitudinal analysis on the duration of the waiting times is performed through classical survival methods. Results reveal a huge variability across journals, revealing different policies not always perfectly adhering to independence and blinding of the peer review process, and discovering the importance of authors' networks to improve the impact of the research production, especially in countries where authors *need 'catch-up strategies' to aid them in publishing papers in international journals* (Lopaciuk-Gonczaryk, 2016). Software used for this analysis comprises STATA for the frequentist multilevel part, WinBUGS for the hierarchical Bayesian part and SPSS for the survival analysis.

References

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H. D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics*, 5(3), 346–359.

Lopaciuk-Gonczaryk, B. (2016). Collaboration strategies for publishing articles in international journals – A study of Polish scientists in economics. *Social Networks*, 44: 50-63.

Keywords

Peer-review process, Hierarchical methods, Survival methods

Partitioning measurement error and residual variance in longitudinal analysis with observed scores.

Matta, Tyler H.[1], Lathrop, Quinn N.[2]
[1] Centre for Educational Measurement, University of Oslo, Norway
[2] Pearson, USA

Summary

The analysis of repeated measures data is commonplace across the behavioral, health, and social sciences. Under most conventional applications of longitudinal data analysis, the implementation of a factor analysis parameterization (Meredith & Tisak, 1990) or a random coefficients (multilevel, mixed-effect) parameterization (e.g., Laird & Ware, 1982) results in identical estimates (Curran, 2003). When the outcome of interest is an observed score measured with error, the model's residual variance term confounds variance due to error in the observed score and deviation between the true score and predicted time trend (Skrondal & Rabe-Hesketh, 2004).

One way to partition these two sources of variance is the so-called *second-order growth model* (Hancock, Kuo, & Lawrence, 2001). Through the establishment of a measurement model for the true scores, and a structural model for the growth in true scores, the two sources of error are no longer conflated. While the second-order growth model appears to be an ideal solution, there are many situations when its implementation is not possible or is impractical. Access to item response data and item parameters are often unavailable to the researcher. Without this information, the second-order model is impossible. Furthermore, with the increasing complexity in measurement models (e.g., Reckase, 2009), a fully structural model would require an extremely high level of technical skill to implement, making the second-order model impractical for most applied researchers.

This paper will show how an observed score and its conditional standard error can be incorporated in a multilevel modeling framework so that the model reflects growth in the true score. Preliminary results from an applied example of English language development of 277 students shows that, when compared to a model that ignores the conditional standard error, the residual variance from a model that incorporates the conditional standard error is reduced by the average measurement error. For the 11th International Multilevel Conference, this applied example will be accompanied by a simulation study that further illustrates the workings of this true-score growth model.

References

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529–569.
Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling*, *8*(3), 470–489.
Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.
Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122.
Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall/CRC.

Keywords
Longitudinal data, Measurement error

Multilevel Mediation with Small Samples: A Cautionary Note on Multilevel Structural Equation Modeling

McNeish, D.
University of North Carolina, Chapel Hill, USA

Summary
Mediation is one of the most common statistical procedures employed in psychology and in science more broadly. Based on Google Scholar citations, Baron and Kenny (1986) is the 18th most cited paper of all-time, in any discipline. Mediation was initially restricted to independent data although it has since been extended to multilevel data and Preacher, Zyphur, and Zhang (2010) unified various models in to a single framework via structural equation models. Though multilevel mediation is growing in popularity in empirical studies, the issue of small samples has received little consideration despite being an explicit limitation discussed in Preacher et al. (2010). Recent literature has advocated addressing multilevel mediation from a multilevel structural equation (ML-SEM) perspective. Though 100 clusters have been the recommended minimum for ML-SEM (Hox & Maas, 2001), a review of 70 multilevel mediation studies finds that 89% fail to meet this threshold with the median number of clusters being 44. A drawback of ML-SEM lies in oft-used full maximum likelihood estimation which is known to yield biased estimates unless the number of clusters is large. Although more restrictive, multilevel regression (MLM) features small sample methods such as restricted maximum likelihood estimation and Kenward-Roger corrections that are unavailable in SEM. This paper performs a simulation to explore the performance of ML-SEMs and MLMs with few clusters in addition to a handful of Bayesian conditions. The simulation explores Level-2 sample sizes from 10 to 100 for a three variable mediation model where the independent variable is at Level-2 and the mediating and dependent variables are at Level-1 (i.e., a 2-1-1 mediation model). Results show that ML-SEM approach to multilevel mediation performs extremely poorly with small samples – 95% coverage intervals from the simulation were as low as 63%, indicating that the Type-I error rates are wildly inflated. In small cluster size conditions (about 10 observations per cluster), Type-I error rates were not well-behaved until 100 clusters were present. As anticipated, Bayesian estimation in M*plus* with default prior distributions did not fare much better – sampling variability estimates were highly inflated with 25 or fewer clusters which severely impacted the ability of the model to detect a non-null indirect

mediation effect. For example, in one condition of the simulation (25 clusters, about 10 observations per cluster, and a medium indirect effect size), Bayesian estimation detected non-null effects in 30% of replications compared to 90% for frequentist methods. Although seemingly outdated, the best performing method was a series of univariate multilevel regression models are fit in separate steps. Type-I error rates were well-behaved in *all* conditions of the simulation and power was the highest of any competing method. This method worked well because the models can be fit with REML and Kenward-Roger corrections rather than full maximum likelihood. The approach of using a series of multilevel models is necessarily limited to a subset of multilevel mediation models, however, so possible general strategies for multilevel mediation with small samples are discussed.

Keywords:
Multilevel Mediation; Small Sample, Multilevel Structural Equation Modeling

The consequences of treatment non-compliance in cluster randomized trials.

Moerbeek, M.
[1] Utrecht University, Faculty of Social and Behavioural Sciences, Department of Methodology and Statistics

Summary
In randomized controlled trials subject will not always adhere to the treatment they have been assigned to. This may cause the estimated effect of treatment to be biases and also affect the statistical power. In cluster randomized trials non-compliance may occur at the individual level but also at the cluster level. In the latter case all subjects within the same clusters do not comply. This presentation describes the results of a simulation study with varying degrees of non-compliance at either the cluster level or individual level. The probability of non-compliance depends on a covariate. Four methods to deal with non-compliance are taken into account: intention to treat, as treated, per protocol and compliers average causal effect. The results show non-compliance may result in biased estimates of the treatment effect and an under- or overestimate of its standard deviation. The coverage of the confidence intervals and empirical power for the test on treatment effect may be too small. Results get even worse when the covariate that affects non-compliance is not included in the model for data analysis.

Understanding and exploring the results of models with random slopes on polynomial terms

Pillinger, R.[1*], Bates, E.[1], McVie, S.[1]
[1] University of Edinburgh, United Kingdom
* Presenting author

Summary
When fitting multilevel models with random slopes on the terms of a polynomial time trend, interpreting the results is not always straightforward. It is not possible to answer directly from the parameter estimates and their standard errors questions such as whether any cases experience an increase between the beginning and end of the period of observation, whether any cases experience an increase at any time during the period, what the average difference is between each case's minimum and maximum value, what the average across cases is of the steepest negative gradient experienced during the period, or when on average the cases experience their minimum value. These questions can be answered by calculating predicted values across time for each case, but this does not take into account the uncertainty in the parameter estimates.

When using MCMC estimation, it is simple to use the chains of the parameter estimates and residuals to calculate a predicted value across time for each case for every iteration, giving a distribution of predicted trajectories for every case. These can be examined to determine, for example, whether there was a period of increase in at least 95% of them for a particular case, allowing us to identify cases for which we can reject the hypothesis that the case did not experience any increase during the period. Similarly, the timing of the minimum value can be calculated for each of the predicted trajectories for every case, and these values can be pooled across all cases to give a distribution of times for the minimum, from which we can obtain a mean and upper and lower limits. Anything else we may wish to discover about the trajectories can be established in the same way.

When not using MCMC estimation, obviously chains of parameter estimates and residuals are not available. We propose using the variance-covariance matrix of the parameter estimates to take draws of the parameters from their joint sampling distribution, and using these draws to establish the distribution of the residuals for each draw, then take a draw of the residuals from this distribution. This allows us to calculate a predicted trajectory for each case for each set of drawn parameters, and we can then proceed to answer any questions we wish, using the predicted trajectories in the same way as if they had been calculated from MCMC chains.

We illustrate this technique using recorded crime data for England, Wales, and Scotland, between 2004 and 2014, fitting a cubic time trend with all coefficients random. We are able to explore the differences in trends between violence and burglary, see whether, despite the average downward trend for both crime types, any areas experienced periods of increase, and establish whether those areas which saw a change in the direction of their trajectory all did so at a similar time.
The method can also be used with polynomials on variables other than time, and is easily extended to handle growth mixture models.

Keywords
Repeated measures; polynomial random slopes; crime trends

Stability of Peer Victimization: A Multi-Level Meta-Analysis of Longitudinal Research

Pouwels, J. Loes[1]* (PhD student), Souren, Pierre M.[2], Lansu, Tessa A. M. [3], Cillessen, Antonius H. N.[4]

[1] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands. PhD student. Supervised by Antonius H. N. Cillessen & Tessa A. M. Lansu

[2, 3, 4] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

Summary

A meta-analysis was conducted of 77 longitudinal studies that contained at least one over-time correlation (range 1 to 36) between scores for peer victimization. The overall stability of self-reported peer victimization was determined at centered value of age 10, across a one-year interval. The effects of interval length, age, and type of informant (self, peer, teacher, other/combined) on the stability of victimization also were examined.

Given the hierarchical structure of the data (correlations nested within studies) (Raudenbush, 2002), a multilevel analysis was conducted in MlWin 2.23 (Rasbash, Charlton, Browne, Healy, & Cameron, 2011). This approach also made it possible to separate within and between study variance by means of a random effects model and to add predictors to explain these variances (Borenstein, Hedges, Higgins, & Rothstein, 2010; Hox, 2002).

The conceptual model had two levels. Level 1 consisted of the correlations that were nested in Level 2, the studies. Stability of victimization (Level 1) was predicted by the variables interval length and age (Level 1) and type of informant (Level 2). Type of informant was a nominal variable consisting of self-, peer-, teacher-, and other/combined-reported victimization. It was coded by three dummy variables using the category 'self' as the reference category (Cohen & Cohen, 1983).

The conceptual model was implemented by a statistical model with some special features. First, when stability correlations were available for three time points, say 1, 2, and 3, we expected that the correlation between 1 and 3, r_{13}, would depend on r_{12} and r_{23}. This is analogous to the effect calculation in a path model, such as a mediation model (Hayes, 2013). To correct for this dependency we included r_{12} and r_{23} as predictors for r_{13} in the model by means of dummies. Second, in order to have a more suitably distributed variable for the stability of peer victimization than Pearson's r itself, we performed Fisher's r-to-Z transformations (Hayes, 1978). The Z

transformation also provided us with a measure for the standard error of the transformed stability of peer victimization. Specifically, this means that the Level 1 error variance of the dependent variable is known. We applied the method demonstrated by Maas et al. (2004) to model this known error variance. This has the advantage that heteroscedasticity could be modeled and that it was possible to distinguish within and between study variance, according to the random effects model (Hox, 2002). An additional level was added to the statistical model in order to model the error variance. Together, this resulted in a statistical model with three levels and several additional (dummy) variables.

The final model showed moderate overall stability of self-reported victimization at age 10 across a 1-year interval. Stability decreased with larger longitudinal intervals. Peer- and other/combined-reports of peer victimization yielded higher stability estimates than self-reports. Teacher-reports yielded stability estimates that were equal to those for self-reports. An interaction was found between age and informant type (peer vs. self), indicating an increase in the stability of victimization with age for peer-reports, but not for self-reports.

Keywords

Meta-analysis, Longitudinal Research, Peer Victimization

Individual differences in contrasting similar [s] sounds across languages

Quené, H.[1*], Orr, R.[1,2]
[1] Utrecht inst of Linguistics OTS, Utrecht University, the Netherlands
[2] University College Utrecht, Utrecht University, the Netherlands
* Presenting author (non-student)

Summary
According to Flege's Speech Learning Model, speech sounds that are similar across languages, such as English and Dutch [s], are merged into a single category (phoneme). Nevertheless, proficient speakers may produce the appropriate variant sound in each language: a relatively sharp [s] in English, and a less sharp [s] in Dutch. Moreover, less proficient speakers may acquire this subphonemic contrast while using English intensively as a second language (L2).
This study investigates whether Dutch speakers produce one or two phonetically distinct sounds for [s] in their L1 Dutch and L2 English, and whether and how this acoustic-phonetic categorization of [s] across languages develops over time. This issue is investigated here using data from our Longitudinal UCU English Accents corpus. Participants in this speech corpus were 282 students at University College Utrecht (UCU), who were recorded 5 times during their stay on campus. From these materials we selected 17 speakers (L1 Dutch, L2 English), and the first, second and fifth (last) interview of these speakers. During these interviews speakers had contributed a 2-minute spontaneous monologue in L1 Dutch as well as one in L2 English. (See lucea.wp.hum.uu.nl for further background on the corpus).
From the selected monologues, the relevant [s] sounds were first detected automatically by means of a customized speech sound recognizer. This yielded over 5000 [s]-like tokens, with the spectral centre of gravity (COG) of each token. Second, all tokens were manually validated and word transcriptions were added. Thirdly, the present analysis focuses on an interesting subset, viz those tokens of [s] sounds that occurred in words used both in the Dutch and English interviews (e.g. “festival, student, semester”). The phonetic environment of these [s] tokens is approximately the same in both languages.

The resulting COG data were analyzed statistically by means of a cross-classified multilevel model, with random intercepts for speakers and for carrier words, and with time as a predictor. The random slope of time (over speakers) was also included in the model, in order to explore individual differences in speakers' development over time. Most speakers were already proficient in their first recording, i.e. they produced two phonetically distinct variants of [s] in the first recording, and maintained this contrast to the last recording. A few speakers *gained* the contrast, i.e. they made no distinction in the first recording but a significant distinction in the last recording. Most puzzling are the few speakers who followed the reverse trajectory, i.e. over time they *lost* the contrast between English and Dutch [s] variants, for reasons yet unknown. We will explore possible explanations for the individual differences in speakers' learning trajectories.

Keywords
speech, foreign accent, individual differences

The impact of ignoring multiple membership structures in multilevel models with survival endpoints

Rosche, B.[1]*
[1] Utrecht University, The Netherlands
* MA student, supervisors: Mirjam Moerbeek, Jeroen Weesie

Summary
Multilevel models are commonly used to deal with clustered data structures (Snijders & Bosker, 2011). The ordinary multilevel model can be applied to pure hierarchical structures, in which lower-level units belong to a single higher-level unit. The lines connecting second and third level units in Figure 1 depict such a pure hierarchical structure. In some applications, however, it is unrealistic to require that all clustered data structures are purely hierarchical.
In political science, for instance, parties are nested in coalition governments. However, the pure hierarchical nesting is broken up because parties are member of more than one coalition government over time. This kind of data exhibits multiple membership structures. The lines connecting first and second level units in Figure 1 display such a non-pure hierarchical structure. The first-level unit A1, for example, is a member of three second-level clusters (G11 to G13) and a single third-level cluster. Sources of multiple membership structures are encountered in various sciences, including medical (Browne et al., 2001), socio-economic (Goldstein et al., 2000), and political research (Rosche, forthcoming). Most applied research ignores multiple membership structures to use the hierarchical multilevel model (or ignores hierarchical data structures altogether). However, simulation studies demonstrate that regression coefficient, variance component, and standard error estimates are biased if multiple membership structures are disregarded (Chung & Beretvas, 2012). Software limitations and computational complexity thwarted the use of models that recognize complex multilevel structures in the past. The constant growth in computational power rendered these restraints obsolete nowadays.

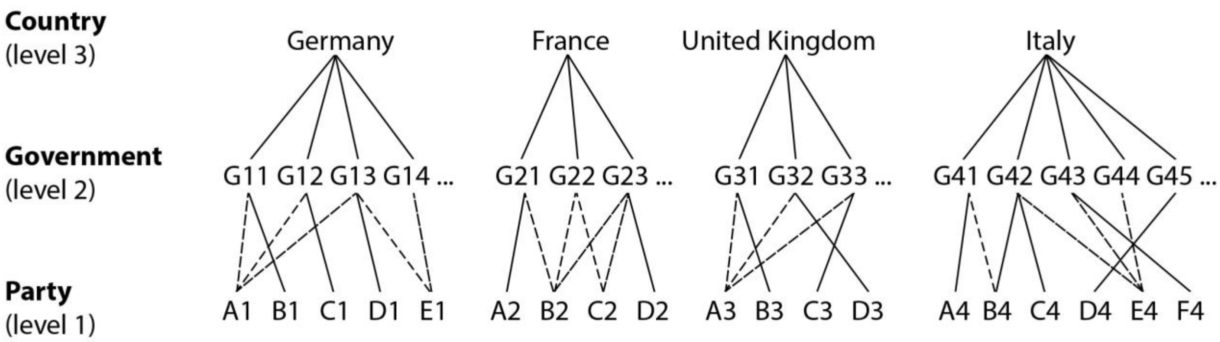

Figure 1: Pure hierarchical multilevel structure (between level 2 and 3) and multiple membership structure (between level 1 and 2)

The multiple membership multilevel (MMML) model is specifically designed to handle this complex kind of multilevel structure (Goldstein, 2011). While the MMML model is well-established in the Gaussian case, it has only recently been introduced in survival analysis to deal with time-to-event response variables. Elghafghuf, Stryhn & Waldner (2014) are the first (and only) researchers to date to describe and apply a MMML Cox model. They mention the complex estimation techniques as reason for this lacuna in the multilevel literature. The statistical properties of the MMML Cox model are not properly studied yet. Admittedly, in their paper, Elghafghuf and colleagues conducted two simulations that demonstrate good model performance in terms of low relative bias. However, they neither compare the MMML Cox to simpler variants to evaluate whether it is worthwhile to estimate a more complex model, nor do they conduct power analyses. Moreover, they study calf mortality. It is unclear whether conclusions drawn from a veterinary science application hold for parameter configurations (e.g. cluster sizes, degree of multiple 'membershipness') found in the social sciences.

To further the efforts of Elghafghuf and colleagues, a simulation study is conducted investigating bias and power of the MMML model on survival data. That is, we compare the model performance of *MMML survival model* to the model performance of *hierarchical multilevel (HML) survival model* and *simple unilevel (UL) survival model* on data that exhibits a hierarchical multilevel structure on the third level and a multiple membership structure on the second level. The simulation scenario is derived from a political science application where simple unilevel (Cox) regression is the predominant method even though the employed data features multiple membership structures and thus allows for using MMML (Cox) regression.

On the conference, we would like to present the results of the simulation study and discuss how the multiple membership model can be an effective tool in the social sciences to study micro-to-macro transformations.

Keywords
Multiple membership structure, survival analysis, simulation study

References

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. Statistical Modelling, 1(2), 103–124. doi:10.1177/1471082X0100100202

Chung, H., & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. The British journal of mathematical and statistical psychology, 65(2), 185–200. doi:10.1111/j.2044-8317.2011.02023.

Elghafghuf, A., Stryhn, H., & Waldner, C. (2014). A cross-classified and multiple membership Cox model applied to calf mortality data. Preventive veterinary medicine, 115(1-2), 29–38. doi:10.1016/j.prevetmed.2014.03.012

Goldstein, Harvey; Rasbash, Jon; Browne, William; Woodhouse, Geoffrey; Poulain, Michel (2000): Multilevel models in the study of dynamic household structures. In European Journal of Population 16 (4), pp. 373–387.

Goldstein, H. (2011). Multilevel statistical models (4th ed.). Wiley series in probability and statistics. Oxford: Wiley.

Rosche (forthcoming): Putting Parties back into the Analysis of the Survival of Coalition Governments. Unpublished master's thesis.

Snijders, T. A. B.; Bosker, R. J. (2011): Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling. 2nd ed. London: Sage Publications.

Multilevel SEM: history, computational approaches and software

Rosseel, Y.,[1*], Barendse, M. T. [2]
[1, 2]:
Department of Data Analysis
Ghent University
Belgium

Summary
In the psychometric literature, the first (technical) papers on multilevel SEM appeared in the late 1980s and early 1990s. Many of these early papers were an attempt to frame multilevel SEM (with random intercepts only) as a special form of a multiple group SEM analysis (for which standard software was available). In later work, more general algorithms were proposed in order to handle more than 2 levels, random slopes, missing data, and categorical data.

Software for multilevel SEM includes (but is not limited to): LISREL, EQS, Mplus, the Stata module gllamm, and the R packages OpenMx', xxM, and lavaan.
A brief overview will be given of the technical capabilities of these software packages, with a focus on computational aspects, together with some advantages and disadvantages. It shall be noted that technical documentation for some software packages is simply not (publicly) available.

Finally, by means of some illustrative examples, we will discuss the current multilevel capabilities of lavaan (version 0.6), and reveal some plans for future versions of lavaan.

Keywords
multilevel SEM, computational aspects, software

Bayesian vs Maximum Likelihood Estimation for Multilevel Models with Small Samples: A Systematic Review

Smid, S.C.[1*(phd-student)], McNeish, D.[2], Van de Schoot, A.G.J.[1,3]
[1] Department of Methods and Statistics, Utrecht University, The Netherlands
[2] University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, USA
[3] Optentia Research Focus Area, North-West University, South Africa
* Presenting author (also indicate if the presenting author is a PhD-student by adding the text 'PhD-student' and add the name of your supervisor)
The presenting author is a PhD student.
Supervisors:
- Rens van de Schoot, Department of Methods and Statistics, Utrecht University, The Netherlands; Optentia Research Focus Area, North-West University, South Africa
- Herbert Hoijtink, Department of Methods and Statistics, Utrecht University

Summary

In social sciences, researchers often experience difficulties collecting enough data, due to small or hard to access target groups or prohibitive costs, resulting in small data sets. With multilevel modeling this is a familiar problem too: small number of clusters and/or small sample sizes within groups can cause estimation problems because there is not enough information available. The use of Bayesian statistics has increased in the last few years, and it is often mentioned as a solution for small sample problems. In the current study, a systematic review is carried out following PRISMA guidelines, to investigate whether it is valid to use Bayes instead of Maximum Likelihood for structural equation models when the sample size is small. In this review, we included papers in which a simulation study was used to investigate and compare the performance of Bayesian parameter estimation to Maximum Likelihood estimation in structural equation models with small sample sizes. A total of *n* = 4977 records was identified in different searches. After removal of duplicates, *n* = 3548 abstracts were screened and *n* = 475 full-text articles were retrieved. We identified *n* = 29 simulation studies that met our inclusion criteria. Of these included simulation studies, *n* = 9 studies investigated multilevel models. In these *n* = 9 studies, we found contradicting results and conclusions regarding the performance of Bayes and Maximum Likelihood. We present our findings and give recommendations for applied researchers. We conclude that

Bayesian estimation can have advantages for small samples in comparison to Maximum Likelihood estimation. However, researchers should never rely on default non-informative priors when the sample size is small.

Keywords
Bayesian estimation, small samples, systematic review

Interrater-reliability in case of missings and many raters with multilevel analysis

Souren, Pierre M.[1]*
[1] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands; email:p.souren@socsci.ru.nl

Summary
When several raters rate several objects the researcher wants to know if the
ratings are independent of the raters. This independency can be expressed as interrater reliability or interrater agreement. When the rating is of interval
measurement level often a type of ICC (Shrout, 1979) is used as an index for interrater reliability or interrater agreement.
When many objects are to be rated it is not feasible to have all the raters rate all the objects. There will be missings (sometimes design-dependent); some raters have not rated some objects. Nevertheless the researcher wants
to assess whether the ratings are independent of the raters.
Several packages that use an ANOVA cannot cope with these missings. For instance: in SPSS one can circumvent the loss of records due to listwise deletion by using a correlation matrix as input for ICC calculation, but the number of raters is limited to 500 and errors (such as an overflow) can occur. When there are not much missings (about 1%) one may use SPSS but the ICC can show a substantial bias (about .2). Further: the manual for package IRR in R (M. Gamer et al., 2012) states:" Missing data are omitted in a listwise way". And Hallgren (2012) writes:" ICCs use list-wise deletion for missing data".
A multilevel model (MLM) might be able to cope with the missings and a large number of raters. (Other advantages are the possibility to test the measurement model assumptions and the (unlike in ANOVA) unbiased estimates of the variances).
This work focuses on:

> Q1. How well can a MLM estimate the several ICC types, how to calculate, how should one model the data?
> Q2. What is the effect of missings on the estimate?
> Q3. When the data are ordinal, how well do MLM's perform in estimating the true value of the ICC??

The simulation design for case 1 (Shrout, 1979) was:
Size(LARGE; SMALL) * ICClevel(0.7; 0.1) * Missings(50%; 20%; 0%)

LARGE refers to: 100 objects and 100 raters; SMALL to: 80 objects and 10 raters.
Resulting in the following 12 sets:
LARGE7.5 LARGE7.2 LARGE7.0 LARGE1.5 LARGE1.2 LARGE1.0 and
SMALL7.5 SMALL7.2 SMALL7.0 SMALL1.5 SMALL1.2 SMALL1.0

Outcomes
@Q1. An intercept only model with raters nested within objects showed that
the ICC calculated with SPSS and in a MLM are quite identical (6th meaningful digit (MeDi) may differ; for all datasets with 0% missings), given
that RIGLS or REML estimation is used. The latter confirms what Snijders and Bosker (1999) write:"REML … produces less biased estimates for the random part.." and "…literature suggests that REML method is better with respect to the estimation of the variance parameters".
@Q2. 50% missings hardly (changes in 3rd MeDi) had an influence on the ICC for the LARGE set (not really surprising because about 5000 data points were left). For the SMALL dataset the deviations where somewhat larger (2nd
MeDi differs one unit after rounding to 2 digits) but not dramatic.
@Q3. Future work including:
-setting up a MLM for ICC(2,#) and ICC(3,#),
-expanding the simulation design to more different amounts of raters, objects and missings and more different levels of ICC.
-use design dependent missings
-effect of data being ordinal.

References
Patrick E. Shrout, Joseph Fleiss,(1979). *Intraclass correlations: Uses in assessing rater reliability.* Psychological Bulletin 86(2):420-428 American Psychological Association 1979-03 0033-2909
Snijders, T. en Bosker, R. (1999).*Multi-level analysis: An introduction to basic and advanced multi-level modeling*. Sage Publications; London; 1999.
Matthias Gamer, Jim Lemon, Ian Fellows, Puspendra Singh (2012) *Various Coefficients of Interrater Reliability and Agreement*
2015: https://cran.r-project.org/web/packages/irr/irr.pdf
KA Hallgren (2012) *Computing inter-rater reliability for observational data:
an overview and tutorial.* Tutorials in quantitative methods for psychology 8
(1), 23
Keywords Interrater reliability, Multilevel-analysis

Bootstrap versions of the Hausman test for cluster-level endogeneity in random slope models

Talloen, W.[1*], Loeys, T. [2] and Moerkerke, B.[2]
[1] PhD-student, Ghent University, Belgium
[2] Supervisors, Ghent University, Belgium
* Presenting author (also indicate if the presenting author is a PhD-student by adding the text 'PhD-student' and add the name of your supervisor)

Summary

Introduction
In random intercept models, the random effect is typically assumed to be independent from the predictors. The bias that is induced by violation of this assumption due to unmeasured cluster-level confounding, can be avoided by treating cluster-specific intercepts as fixed. The Hausman test (Hausman, 1978) contrasts the fixed effect estimator (FE) with the random effect estimator (RE) to test for the presence of such cluster-level endogeneity. In literature, the focus on the Hausman test has been confined to random intercept models. However, unmeasured cluster-level heterogeneity may also interact with predictors. We therefore study two extensions of the Hausman test that rely on bootstrap and that can be used to test for cluster-level endogeneity in random slope models.

Methods
The first extension uses bootstrapping to obtain an estimator for the variance of the difference between the FE and RE estimator. The latter is then used in the calculation of the test statistic (Kaiser, 2014), which is assumed to follow a χ^2-distribution. For the bootstrapping, cases are resampled at the cluster level. We therefore refer to this method as Cluster-level Resampling Bootstrap (CRB). The second extension uses bootstrapping to obtain the distribution of the difference between the FE and RE estimator under the null hypothesis of no upper-level endogeneity. More specifically, upper and lower level residuals of the RE model are resampled to create new bootstrap samples, implying that upper and lower level residuals are indeed uncorrelated with the predictor(s). Since we rely on non-parametric resampling from the observed distribution of residuals, we refer to this method as the Non-Parametric Residual Bootstrap (NPRB). We conduct a simulation study to compare the performance of both bootstrap approaches with the original Hausman test and its robust version

under two true data-generating processes (a random intercept and random slope model, respectively).

Results
For random intercept models, all methods perform equally well in terms of achieving the nominal type I error rate and with respect to power. For random slope models, the original Hausman test can not be used, but one could rely on the robust Hausman test. We find however that when the random slope model is the true data-generating process, the CRB and NPRB Hausman test outperform the latter. Somewhat surprisingly CRB performs equally well as NPRB with respect to the type I error rate and power, even when the number of clusters is small. In addition we find that the CRB and the NPRB Hausman test perform equally well when heteroscedasticity and temporal correlation are present but unmodelled.

Keywords
Hausman test, random slope, cluster-level endogeneity

Sample size and power calculations for Stepped Wedge trials with >2 levels

Teerenstra, S.[1*], Haenen, A.[2,3], Kooi, T van der[2], Huis, A.[3], Hulscher, M.[3]
[1] Biostatistics group, Radboud Institute for Health Sciences, department for Health Evidence, Nijmegen, the Netherlands
[2] National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands
[3] IQ Healthcare, Radboud Institute for Health Sciences, department for Health Evidence, Nijmegen, the Netherlands
* Presenting author

Summary
In 2007, Hussey and Hughes published their seminal paper about power and analysis of Stepped Wedge trials focusing on 2 levels (subjects within clusters). In this talk we will discuss how to easily extend their model and resulting formulas to >2 levels. This will be illustrated using the CHANGE trial which recruits health organizations (level 1) with nursing homes (level 2) in which health care workers (level 3) are measured in several waves as to their hygiene compliance (level 4). We investigate the impact on power of the correlations (of nursing homes within health organization, of health care workers within nursing homes, of compliance evaluations within health care workers) and sample sizes at different levels.

Keywords
Stepped wedge trials, power, sample size

Bayesian Multilevel Latent Class Models for the Multiple Imputation of Nested Categorical Data

Vidotto, D.[1*], Vermunt, J. K.[1], van Deun, K.[1]
[1] Department of Methodology and Statistics, Tilburg University, The Netherlands
* Presenting author, PhD-student. Supervisor Vermunt, J. K.

Summary
In the literature, a number of imputation models have been proposed to deal with missingness in multilevel datasets, especially for continuous data. As far as multilevel categorical data are concerned, existing techniques assume underlying multivariate normality of the items with missingness. This approach has two limitations: first, it does not allow to estimate the imputation model, nor to perform imputations, in the original scale type, which can result in biased estimates of the analysis model parameter. Second, these models can correctly capture pairwise associations in the data, but fail to capture higher-order relationships, showing a lack of flexibility. With the present work, we propose using Multilevel Latent Class models to perform multiple imputation of missing multilevel categorical data. The model is flexible enough to retrieve original (complex) associations in the data at both the first and second level of the hierarchy, as well as to respect the original scale type of the data. The model is implemented under a Bayesian framework and estimated via Gibbs sampling, a natural choice for multiple imputation applications. After formally introducing the model, we carry out a simulation study with complex relationships in the data in order to assess its performance, and compare it with the commonly used listwise deletion method and an already available R-routine. Results indicate that the Bayesian Multilevel Latent Class model is able to recover unbiased and efficient parameter estimates of the analysis model considered in our study, outperforming in this way the competing methods.

Keywords
latent class models, missing data, multilevel multiple imputation.

Latent Markov Factor Analysis for Exploring Within-Subject Measurement Model Differences in Experience Sampling Studies

Vogelsmeier, Leonie V.D.E. [1*], Vermunt, Jeroen K.[2], De Roover, Kim[3]
[1] Tilburg University, The Netherlands
[2] Tilburg University, The Netherlands
[3] Tilburg University, The Netherlands
* Presenting author

Summary
Experience Sampling (ES) is increasingly popular for assessing dynamics of psychological attributes or processes within subjects. Through ES, researchers obtain systematic self-reports of participants over a certain period in everyday life via smartphone apps. The validity of the findings may be compromised by between- and within-person differences in measurement quality (e.g., response styles or measurement errors caused by distraction) and by differences in how questionnaire items are measuring the same psychological constructs (e.g., true differences in item interpretation due to cultural background differences). Both types of differences can be traced as differences in the 'measurement model' (MM) which represents a certain factor structure underlying a participant's answers.
In ES studies, it is common practice to either simply assume that the MM is the same across time points and subjects or to apply methods that only test for a priori hypotheses about MM differences or changes. However, typically we have no prior information on between-/within-subject differences in MMs. Therefore, an exploratory approach is required to investigate whether researchers can validly compare subjects/time points to draw valid conclusions. Additionally, when MM differences are found, it would be most useful to learn from differences in MMs for future research.
In this study, we present a method called latent Markov factor analysis (LMFA), which models MM differences in ES studies, without the need for prior assumptions on the MM. LMFA builds upon mixture simultaneous factor analysis (MSFA; De Roover, Vermunt, Timmerman, & Ceulemans, in press) that captures differences in latent variables between higher-level units of multilevel data. In LMFA, a multilevel structure also exists in that time points are nested within subjects. With subjects as the higher-level units, in contrast to MSFA, LMFA allows subjects to switch between different MMs over time and is thus more suitable for ES data. In LMFA, a latent Markov chain per subject clusters observations into states and, per state,

the data are factor-analysed. Within-subject MM differences are then captured by using a time-specific clustering where each latent state corresponds to a different MM. A simulation study shows very good results in recovering parameters under a wide range of conditions. The value of LMFA is illustrated with an empirical example.

Keywords
experience sampling, measurement invariance, latent markov modeling