Universiteit Utrecht

13th International Multilevel Conference
April 12 & 13, 2022

Conference Program
& Abstracts

Organizing committee
Emmeke Aarts
Sara van Erp
Beth Grandfield
Mirjam Moerbeek
Marianne Geelhoed (*local organization*)

Utrecht University
Department Methodology & Statistics

Conference program

Day 1 (April 12)

09:00	**Registration**
09:25	Opening
09:30	**Keynote 1**: Terrence Jorgensen *Variety is the Spice of Life: A Taxonomy of Reliability Indices When Measuring Cluster-Level Constructs*
10:15	Anson, Yonathan (Jon) *Serendipity: Hidden information in random effect coefficients*
10:35	Jak, Suzanne *Modeling cluster-level constructs with individual-level measures*
11:00	**Coffee and Tea Break**
11:20	Kesteren van, Erik-Jan *Is it the driver or the car? Modeling Formula 1 race performance using a Bayesian multilevel Beta regression approach*
11:55	Innocenti, Francesco *Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative*
12:30	**Lunch**
	Young Researchers Award Nominees (4 oral presenters)
13:15	Post, Richard *On the distribution of individual causal effects of binary exposures using flexible multilevel models*
13:45	Baurne, Yvette *Modelling the Emergence of Consensus in Groups With Nonlinear Dynamics*
14:15	**Short Break (15 min)**
14:30	Kerkhoff, Denise *Sampling Strategies to ensure estimation quality for latent and manifest predictors and ICCs in 3-Level-Models*
15:00	Mildfiner Moraga, Sebastian *Go multivariate: a Monte Carlo study of a multilevel hidden Markov models with categorical data of varying complexity*
15:35	Poster Session (Coffee and Tea)
16:20	**End of Day 1**
18:00	**Conference Dinner (for those who registered)**

Day 2 (April 13)

09:00	**Doors Open**
09:30	Mirjam Moerbeek *Optimal design of multi-period cluster crossover trials for treatments offered in groups.*
09:50	Alkema, Leontine *(Multilevel) Model-Based Estimates in Demography and Global Health: Quantifying the Contribution of Population-Period-Specific Information*
10:35	Spreitzer, Carina *Problem areas in multilevel modelling of school and classroom effects on student achievement: More questions than answers?!*
11:00	**Coffee and Tea Break**
11:20	Grilli, Leonardo *Conditioning on the pre-test versus gain score modeling: revisiting the controversy in a multilevel setting*
11:55	Qi, Hongchao *Sample size determination for clinical trials based on the meta-analytic-predictive approach*
12:30	**Lunch**
13:30	Ofili, Samantha *Partitioning Variance for Zero Inflated Count Data in Early Child Development*
13:50	**Short Break (15 min)**
14:05	**PhD-award ceremony**
14:15	**Keynote 2**: George Leckie *Multilevel modelling of school effects on student achievement and their application in school accountability systems: models, assumptions, innovations, debates, and challenges*
15:00	**Closing remarks and End of Day 2**

Abstracts Oral Presentations

Variety is the Spice of Life: A Taxonomy of Reliability Indices When Measuring Cluster-Level Constructs

Jorgensen, T.D.[1*]

[1] Methods and Statistics, University of Amsterdam, the Netherlands

Suggested talk duration: 45 minutes

Summary (max. 500 words)

Generalizability theory (GT) provides a variety of intraclass correlation coefficients (ICCs) to quantify reliability in terms of the degree to which observations generalize across different conditions of measurement—for example, scale, interrater, and test–retest reliability quantify generalization across items, raters, and occasions, respectively. Sometimes, a construct of interest is a group-level characteristic. In a multilevel modeling (MLM) context, an ICC is often used to quantify the proportion of a subject-level variable's variance that is attributable to cluster-level differences (e.g., between schools or neighborhoods). From a GT perspective, this ICC could be interpreted as the degree to which observations of a cluster generalize across measurements provided by members of that cluster, who could be considered raters of a cluster-level attribute. Structural equation modeling (SEM) becoming a popular framework for modeling GT, and the omega indices of composite reliability—which are functions of SEM parameters—are special cases of GT (i.e., how measurements of a construct generalize across indicators). The merger of MLM and SEM (i.e., MLSEM) provides a unique opportunity to explore how well measurements of cluster-level constructs generalize across multiple measurement conditions (indicators, cluster members, or both). We use GT to propose new ICCs that appropriately quantify reliability across aspects of a procedure for measuring cluster-level constructs, which add value to the limited ICCs currently considered in MLSEM applications.

Serendipity: Hidden information in random effect coefficients

Jon Anson

Dept. of Social Work

Ofra Anson

Faculty of Health Sciences

Ben-Gurion University of the Negev,
Beer Sheba, Israel

Suggested talk duration (15-60 minutes)

15 minutes

Summary (max. 500 words)

Random effects in multilevel analysis offer a way of dividing up the error component of the analysis. Error scores are broken down into variation which is due to assignment to a particular category at Level 2 (or above), and residual error for the individual case. Often, these random effect coefficients are ignored, though they may well contain important information. In the present example, we explore the evolving pattern of older age mortality over time.

The upper tail of the mortality curve can be broadly modelled as a Gompertz (log-linear) curve. If mortality at young and middle-age is declining faster than at the oldest ages, then as mortality declines, the slope of this curve will increase in order to attain the same level of mortality by age 100. The random effects of a simple, multi-level model enable us to examine whether this relationship between young and older age mortality has been constant over time.

Our cases (Level 1) are national mortality rates for 4236 life tables drawn from the Human Mortality Database (HMD), from the middle of the 19th century to the present day, classified (Level 2) by Year (270) and by Country (45). These are mainly European and European-overseas countries for which there are reliable, long-term, mortality data. The random effects for year show, as expected, that the mortality rate, by age, at older ages has declined consistently over time, reflecting the general decline in mortality over time. However, when we add a control variable, l_{50}, the proportion of people surviving to age 50, to the model, a different picture emerges. For about a century, from the mid 19th to the mid-20th century, this residual mortality increased. Since the mid-1970's, however, this residual mortality has been consistently declining, indicating that the mortality transition has entered a new phase, one in which the rate of ageing is accelerating more slowly, even as young and middle age mortality continue to decline.

Multilevel modelling separates out variation which is attributable to membership in a particular category (Level 2 and above). As can be seen from this example, this variation may hold important information, offering previously unseen insights into the structure of

the data: In the present case, the pattern of mortality decline during the Demographic Transition.

Relevance to conference theme

Substantive (but unexpected) insights gained from the use of multilevel analysis

Keywords (max. 3)

Mortality; Transition; Serendipity

Modeling cluster-level constructs with individual-level measures

Jak, S.[1*], Jorgensen, T.D. [1], Nevicka, B. [1], Ten Hove, D. [1]

[1] University of Amsterdam, Netherlands

* Presenting author

Suggested talk duration: 15 minutes

Summary (max. 500 words)

Researchers frequently use the responses of individuals in clusters to measure constructs at the cluster level. For example, student's evaluations may be used to measure the teaching quality of instructors, patient reports may be used to evaluate social skills of therapists, and residents ratings may be used to evaluate neighborhood safety.

When multiple items are used to measure such cluster-level constructs, multilevel confirmatory factor models are useful. These models allow for the evaluation of the factor structure at the cluster level (modeling the (co)variances among item means across clusters), and at the individual level (modeling the (co)variances across individuals within clusters).

If the cluster-level construct, for example teacher quality, would be perfectly measured using the responses of students, all students evaluating the same teacher would have exactly the same item scores. In that case, there will not be any systematic variance in the item scores within clusters (only sampling error), so there will be nothing to model at the individual level.

In practice, individuals do not all provide the same responses to the items, leading to systematic variance (and covariance) to be explained at the individual level. The question then arises how the variance within clusters should be modeled. In this talk, I will review some of the interpretational difficulties related to existing two-level models for cluster-level constructs and I will discuss possible alternative options.

Keywords Multilevel SEM, CFA, cluster-level constructs

Is it the driver or the car? Modeling Formula 1 race performance using a Bayesian multilevel Beta regression approach

Van Kesteren, E.-J.[1*], Bergkamp, T.L.G.[2]

[1] Utrecht University, Netherlands

[2] Rijksuniversiteit Groningen, Netherlands

* Presenting author

Suggested talk duration (15-60 minutes)

30 minutes

Summary (max. 500 words)

Contrary to many individual sports, successful performance in F1 racing is not only determined by the driver's skill, but also by the race-car constructer. This makes key performance questions in the sport difficult to answer. For example, who is the best Formula One driver? Which is the best car constructor? What is their relative contribution to race success?

In this presentation, we aim to answer these questions based on data from the hybrid era in Formula One (2014 - 2021). We take a step-by-step approach to construct a race success model using Bayesian multilevel Beta regression. Specifically, we model race success as the proportion of outperformed competitors. We show visually that our approach describes our data well, and that its parameters have an intuitive interpretation, which allows for precise inferences for the above questions.

We then use this model to make inferences about hybrid-era Formula One. In addition to producing rankings to answer the questions about driver skill and constructor advantage, we show how draws from the posterior distribution of the model's parameters can be used to perform full Bayesian inference about counterfactual situations. This allows us to answer questions such as "would Lewis Hamilton beat Kimi Räikkönen if they were driving the same car?". We argue that this approach may prove useful for sports beyond

Formula One, as it creates performance ratings for independent components contributing to success.

Relevance to conference theme

This presentation is an advanced application of Bayesian generalized multilevel modeling to sports.

Keywords (max. 3)

Bayesian analysis, generalized linear model, sports statistics

Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative

Innocenti, F.[1*], Candel, M.J.J.M.[1], Tan, F.E.S. [1], & van Breukelen, G.J.P.[1,2]

[1] Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, the Netherlands

[2] Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University, the Netherlands

* Presenting author

Suggested talk duration (15-60 minutes)

15-30 minutes

Summary (max. 500 words)

In multilevel populations, there are two types of population means of an outcome variable: the average of all individual outcomes ignoring cluster membership, and the average of cluster-specific means. To estimate the first mean, individuals can be sampled directly with simple random sampling or with two-stage sampling (TSS), that is, sampling clusters first, and then individuals within the sampled clusters. When cluster size varies in the population, three TSS schemes can be considered: (i) sampling clusters with probability proportional to cluster size and then sampling the same number of individuals per cluster; (ii) sampling clusters with equal probability and then sampling the same percentage of individuals per cluster; and (iii) sampling clusters with equal probability and then sampling the same number of individuals per cluster. Unbiased estimation of the average of all individual outcomes is discussed under each sampling scheme allowing cluster size to be related to the outcome variable of interest (i.e. informative cluster size). For each sampling scheme, optimal sample sizes are derived under a budget constraint. The three optimal TSS designs are compared, in terms of efficiency, with each other and with simple random sampling of individuals. Sampling clusters with probability proportional to size is recommended. To overcome the dependency of the optimal design on unknown nuisance parameters, maximin designs are derived. The results are illustrated with

the planning of a hypothetical survey to compare adolescent alcohol consumption between France and Italy.

References:

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S., & van Breukelen, G. J. P. (2019). Relative efficiencies of two-stage sampling schemes for mean estimation in multilevel populations when cluster size is informative. *Statistics in Medicine*, 38(*10*), 1817–1834. https://doi.org/10.1002/sim.8070

Innocenti, F., Candel, M. J. J. M., Tan, F. E. S., & van Breukelen, G. J. P. (2021). Optimal two-stage sampling for mean estimation in multilevel populations when cluster size is informative. *Statistical Methods in Medical Research*, 30(*2*), 357–375. https://doi.org/10.1177/0962280220952833

Relevance to conference theme

The aim of this talk is to present guidelines on how to design surveys to estimate the overall mean of a quantitative variable of interest (e.g. alcohol consumption, test score) in a multilevel population (e.g. adolescents nested within schools) or to compare two multilevel populations (e.g. adolescents of different countries) in terms of their population means.

Keywords (max. 3)

Informative Cluster Size; Optimal Sample Sizes, Two-Stage Sampling

On the distribution of individual causal effects of binary exposures using flexible multilevel models

Post, Richard A.J.[1*], van den Heuvel, Edwin R.[1,2]

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands

[2] Department of Preventive Medicine and Epidemiology, School of Medicine, Boston University, USA

* Presenting author – **PhD-student** (supervisor prof. E.R. van den Heuvel)

Suggested talk duration (15-60 minutes)

30 minutes *(or 45 minutes if a more elaborate introduction to causal inference is preferred)*

Summary (max. 500 words)

In recent years the field of causal inference from observational data has emerged rapidly. However, this literature has mainly focused on (conditional) average causal effect estimates. When (remaining) variability of causal effects is considerably these estimates are not informative, and possibly misleading, at an individual's level. The fundamental problem of causal inference precludes estimation of the joint distribution of potential outcomes without making additional assumptions, while the latter is necessary to describe heterogeneity of causal effects. Here, we present how multilevel structural equation models can be used to study individual effect modification and estimate the (conditional) distribution of causal effects by comparing the observed variability in the exposed and unexposed groups.

How our thinking could be applied and (partly) validated in practice is illustrated in a case study on the effect of Hepatic Steatosis on a clinical precursor to heart failure. Assuming no unmeasured confounding and independence of the individual effect modifier and residual, we conclude that the individual causal effect (ICE) distribution deviates from Gaussian. We estimate that, despite an on average harming effect, 21.2% (95% Bayesian credible interval: 5.3%; 34.1%) of the population won't be harmed. Furthermore, we illustrate how misspecification of the error distribution or ignoring confounding effect heterogeneity can affect the estimated ICE distribution.

Relevance to conference theme

The field of causal inference from observational data has emerged rapidly over the past two decades. However, the focus of the developments is on estimation of average causal effects. Multilevel structural equation models can be used to draw inference on the variability of individual causal effects by comparing the variability between groups of exposed and unexposed individuals (after adjusting for confounding). In this work we bridge the field of multilevel modelling with the field of causal inference. Such connection will be crucial to inform applied researchers how multilevel models can be used for etiological research from observational data.

Keywords (max. 3)

Causal inference; Latent modifiers; Heterogeneity of treatment effect

Modelling the Emergence of Consensus in Groups With Nonlinear Dynamics

Yvette Baurne[1*] (PhD-student, supervisor: Jonas Wallin), Frédéric Delmar[2], Jonas Wallin[1]

[1] Department of Statistics, Lund University, Sweden

[2] Department of Entrepreneurship and Innovation, emlyon business school, France

* Presenting author

Suggested talk duration

20 minutes

Summary

The study of emergent, bottom-up, processes has long been of interest within organizational research. Emergent processes refer to how dynamic interactions among lower-level units (e.g. individuals) over time form a new, shared, construct or phenomena at a higher level (e.g. work group). Recent advances suggest multilevel models to study the emergence of shared constructs, accounting for both multilevel and longitudinal aspects of importance to capture such processes (Lang et al. 2018, Lester et al. 2021).

We develop the modelling of consensus emergence by making two contributions. First, we distinguish between different patterns of consensus emergence; homogeneous and heterogeneous. In a multilevel setting, we can distinguish between these patterns in how individual random effects change over time, i.e. what the individual trajectories towards consensus within the group look like. Homogeneous consensus emergence is characterised by gradual and almost deterministic adjustments of the individual trajectories, whereas heterogeneous consensus emergence show more randomly oscillating trajectories towards consensus.

Second, we show how Gaussian Processes (GPs) can be used to further extend the consensus emergence models, allowing them to capture nonlinear dynamics in emergent processes. GPs can be incorporated on both individual and group level. When GPs are included on the individual level, the models can capture consensus emergence patterns of both homogeneous and heterogeneous character. When GPs are included on the group level, the models can capture dynamics such as nonlinear changes in group mean over time, which unaccounted for could conflate the estimation of consensus emergence.

Using an established data set, we show that conclusions on the pattern of consensus emergence can change depending on whether the nonlinear group mean change over time is adequately modelled or not. Thus it is crucial to correctly capture the group dynamics to properly understand the consensus emergence.

Relevance to conference theme

The research on emergence and similar multilevel group processes has been severely limited by the lack of methodological tools. We propose such tools to estimate different types of consensus emergence. We introduce the possibility to account for nonlinear group dynamics (using Gaussian Processes) that can arise in the context of emergent processes, which unaccounted for may conflate the estimation of consensus emergence.

Keywords

variance conflation, nonlinear dynamics, gaussian processes

References

Lang, J. W., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence in groups using longitudinal multilevel methods. *Personnel Psychology*, *71*(2), 255-281.

Lester, H. F., Cullen-Lester, K. L., & Walters, R. W. (2021). From Nuisance to Novel Research Questions: Using Multilevel Models to Predict Heterogeneous Variances. *Organizational Research Methods*, *24*(2), 342–388.

Sampling Strategies to ensure estimation quality for latent and manifest predictors and ICCs in 3-Level-Models

Kerkhoff, D.[1*], Nussbeck, F.-W.[1]

[1] Department of Psychology, University of Konstanz, Konstanz, Germany

* Presenting author, PhD-student (supervisor: Fridtjof W. Nussbeck)

Suggested talk duration: 30-40 minutes

Summary

Three-level modelling has become a popular approach to model multiply clustered data, such as patients nested within therapists nested within clinical sites. These models usually require researchers to pay particular attention to the sampling strategy, i.e., the overall number of sampled units and especially the number of units at each of the three levels. Especially in 3-level models, identifying an advantageous sampling strategy is not trivial and may not simply be adapted from findings based on research on 2-level models.

We present results of a series of extensive Monte-Carlo simulation studies to assess the estimation quality in linear 3-level models for a variety of fixed and random effect sizes and sampling strategies. We focus on two key topics that are central to most modelling approaches:

First, we present findings on the estimation quality of the intraclass correlation coefficients (ICCs) and underlying random variances. Computing the ICCs to identify how much of the variance in the criterion is located at each data level is a central initial step in any multilevel modelling strategy. However, the underlying random variances are typically prone to being biased if the overall sample size and sampling strategy is insufficient. We investigated the influence of 125 sampling strategies (overall sample size: 20 to 180,000) on estimation bias of nine ICC values (min = .059; max = .353) and statistical power of the variance components, according to the

Satorra-Bentler χ^2-test. Results show that the level-3 ICC is considerably underestimated. However, given an advantageous sampling strategy, statistical power of the variances and unbiased estimation of large ICCs can be achieved with 1,000 observations, while smaller ICCs require at least 100 clusters and several thousand observations.

Second, we present findings for the estimation quality of predictors that explain variance at multiple levels. For example, the patient's age may explain variance at level-1, the mean age of the patients within therapists may explain variance as a level-2 predictor, and the mean age of all patients within a clinical site may serve as a level-3 predictor. For such predictors, the classic approach is to use the appropriate mean values as manifest predictors on the higher levels. As an alternative approach, the predictor is modeled as a latent variable with a variance that is decomposed across levels. We investigated the parameter bias, power, and coverage for both approaches across 86 sampling strategies (overall sample size: 375 to 10,000) and found that overall estimation quality is higher if the predictor is modeled as a latent variable, and that the effects of the manifest predictors are considerably underestimated. Notably, both approaches require considerably different sampling strategies to yield better estimation quality.

We report advantageous sampling strategies for all models and discuss challenges regarding the practical meaning of parameter estimation bias and the assessment of statistical power for the random variances. We discuss further issues regarding sample and model characteristics in 3-level models, such as variable scaling and unbalanced designs.

Relevance to conference theme: statistical and/or methodological aspect of multilevel modelling

Keywords: estimation quality, 3-level modelling, sample size

Go multivariate: recommendations on multilevel hidden Markov models with categorical data of varying complexity

Mildiner Moraga, Sebastián[1*], Aarts, Emmeke[1]

[1] Utrecht University, the Netherlands

* PhD student (supervisor: Emmeke Aarts)

Suggested talk duration (15-60 minutes)

20 minutes

Summary (max. 500 words)

The multilevel hidden Markov model (MHMM) is a promising method to investigate intense longitudinal data obtained within the social and behavioral sciences. The MHMM quantifies information on the latent dynamics of behavior over time. In addition, heterogeneity between individuals is accommodated with the inclusion of individual-specific random effects, facilitating the study of individual differences in dynamics. However, the performance of the MHMM has not been sufficiently explored. Currently, there are no practical guidelines on the sample size needed to obtain reliable estimates related to categorical data. We performed an extensive simulation to assess the effect of the number of dependent variables (1-4), the number of individuals (5-90), and the number of observations per individual (100-1600) on the estimation performance of a Bayesian MHMM with categorical data of various levels of complexity. We found that using multivariate data generally alleviates the sample size needed and improves the stability of the results. Regarding the estimation of group-level parameters, the number of individuals and observations largely compensate for each other. However, only the former drives the estimation of between-individual variability. We conclude with guidelines on the sample size necessary based on the complexity of the data and study objectives of the researcher.

Relevance to conference theme

Our simulation study fills a gap on the literature: the sample size required to obtain reliable estimations with the novel multilevel hidden Markov model. We believe that our findings are relevant to both methodologists developing tools to deal with multilevel time-series, and applied researchers looking to analyze intensive longitudinal data.

Keywords (max. 3)

Intensive longitudinal data, Hidden Makov Models, Bayesian statistics

Optimal design of multi-period cluster crossover trials for treatments offered in groups

Moerbeek, M.[1*]

[1] Department of Methodology and Statistics, Utrecht University, the Netherlands

Suggested talk duration: 30 minutes

Summary (max. 500 words)

Purpose: In multi-period cluster randomized trials the clusters switch between treatment conditions for a duration of at least three periods. Within each period, a number of subjects is selected from each cluster and receives treatment within a group setting. An example is the evaluation of a group treatment to increase the changes on the labour market, with unemployed participants nested within social benefit agencies. The number of clusters and time periods for such a trial is often limited and the optimal design finds the optimal combination of number of clusters and time periods.

Method: The multilevel regression model is used to model the relation between treatment and outcome. A realistic correlation structure that allows for higher correlations within than across time periods is considered. The optimal design is the one that estimates the effect of treatment with highest efficiency. It is sought under a budgetary constraint: costs are assigned to including clusters, subjects and time periods.

Results: A Shiny App has been developed to find the optimal design. It also compares alternative designs to the optimal one on the basis of their relative efficiency. The Shiny App will be demonstrated and it will be shown how costs and correlation structures affect the optimal design.

Conclusions: The Shiny App is a user-friendly tool to find the optimal design. It avoids wasting resources in an inefficient way.

Relevance to conference theme

In cluster randomized trials subjects are nested within clusters, hence the data have a multilevel structure.

Keywords: Longitudinal data, Multilevel Designs, Optimal design

(Multilevel) Model-Based Estimates in Demography and Global Health: Quantifying the Contribution of Population-Period-Specific Information

Alkema, L.[1*], Yang, G.[1], Gile, K.[2]

[1] Department of Biostatistics, University of Massachusetts Amherst, USA

[2] Department of Statistics, University of Massachusetts Amherst, USA

* 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)

Suggested talk duration (15-60 minutes)

45 minutes; I can also make it shorter as needed.

Summary (max. 500 words)

Sophisticated statistical approaches, often using multilevel structures, are used to produce model-based estimates for demographic and health indicators even when data are limited, very uncertain or lacking. Examples in family planning and fertility-related research include the Family Planning Estimation Model (Cahill et al., 2018) that is used to estimate contraceptive use and unmet need for contraceptives worldwide, and a Bayesian accounting model (Bearak et al, 2020), used to estimate unintended pregnancies and abortions. Both models use multilevel and temporal model structures such that data-rich population-periods can help to inform estimates in settings where data are limited, very uncertain or lacking.

To facilitate interpretation and use of model-based estimates, we aim to provide a standardized approach to answer the question: To what extent is a model-based estimate of an indicator of interest informed by data for the relevant population-period as opposed to information supplied by other periods and populations and model assumptions? We propose a data weight measure to calculate the weight associated with population-period data set $\boldsymbol{y}$ relative to the model-based prior estimate obtained by fitting the model to all data excluding $\boldsymbol{y}$. In addition, we propose a data-model accordance measure which quantifies how extreme the population-period data are relative to the prior model-based prediction.

We illustrate the insights obtained from the combination of both measures in toy examples and for estimates produced by the family planning and accounting models.

Relevance to conference theme

We propose general measures to quantify shrinkage in multilevel models and accordance (or lack thereof) between population-specific data and the hierarchical distribution used. While motivated by applications in demography and global health, the measures are useful in any application of multilevel models, especially in settings where closed-form solutions to calculate shrinkage are not available.

Keywords (max. 3)

Multilevel models; shrinkage; mixture modeling

Problem areas in multilevel modelling of school and classroom effects on student achievement: More questions than answers?!

Spreitzer, C.[1*]

[1] University of Klagenfurt, Austria

* Presenting author ('PhD-student'; Supervisors: Krainer, K. and Alexandrowicz, R.)

Summary (max. 500 words)

'Children in schools' is one of the most cited textbook examples for multilevel analysis (e.g., Ditton, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). And there is, in fact, already a large body of literature modelling school effects (e.g., Gustafsson et al., 2018; Mohammadpour et al., 2015) or classroom effects on student achievement (e.g., Kruger et al., 2017; Lipowsky et al., 2009). However, studies dedicated to a three-level modelling approach (school-class-student) are scarce (e.g., Creemers & Kyriakidēs, 2008; Vanlaar et al., 2015). The effects of both school and class level factors (and possible interactions) has hardly been investigated so far (Ditton & Müller, 2015). Therefore, the present study examines a comprehensive three-level model taking into account contextual factors at the student, the class, and the school level. Predictors were selected after extensive literature reviews. From a methodological perspective, this approach revealed two major problem areas:

1. Multicollinearity
   Few studies deal with multicollinearity issues in two-level models (Shieh & Fouladi, 2003; Yu et al., 2015), but I have found no concrete study that analysis the effect of multicollinearity in a three-level context. Therefore, two-level model recommendations were adapted for the three-level model. Factors at the second and third level were highly correlated. The higher level of multilevel models can have more influence than that on the lower level (Yu et al., 2015). To deal with

this multicollinearity in the data only a model re-specification was possible. Therefore higher order factors were built, which integrated categorical and numeric information in scale construction. I combined latent dimensions into higher order factors. The person parameters obtained in a multidimensional Rasch model were used to build higher order factors through confirmatory factor analysis (CFA). Although this approach helped dealing with multicollinearity, it disregards much information pertaining to the third level 'school' and the second level 'class'.

2. Linking the three levels
   Data for student achievement was available from the educational standards survey (8th grade in mathematics). In contrast, data at school and classroom level was collected independently. Therefore, linking schools and teachers was straightforward, but there is no possibility of linking between teachers and pupils due to data protection regulations in Austria. The pupils can be assigned to the schools (level 3) but not to the teachers (level 2). Therefore, a "random match" approach was chosen to match levels 1 (pupils) to levels 2 and 3 (i.e., classes and schools). The pupils are assigned to the schools and each pupil is matched with each teacher of the school. This results in an "enlargement" of the data set because each pupil occurs several times. To correct this oversizing, a weighting factor is calculated for each observation, based on the relative frequency of teachers in the school. This approach ensures linkage, but some of the actual variance is lost through weighting.

In the lecture, the problem areas themselves and their approaches to solutions should be presented. For this purpose, they will be critically discussed taking into account the results in the multi-level analysis.

Words: 497

Relevance to conference theme

As the keynote-speaker George Leckie in his abstract for the conference states (2021): “Studies of school effects on student achievement were some of the very first applications of multilevel modelling, motivating and illustrating many early methodological developments.”

The lecture presents an application of multilevel design in the context of schools, which is an extensive field of application of multilevel models.

However, the focus refers to problem areas in this field of application, which are caused by data protection regulations or given conditions in the school system. Therefore, approaches to solutions are needed from a methodological point of view. The presented approaches are first attempts to tackle the problem areas. The Multilevel Congress brings together methodological expertise with application expertise. The problem areas and their approaches to solutions should be presented and discussed in this interdisciplinary field.

Keywords (max. 3)

School effectiveness, multicollinearity, data merging

References

Creemers, B. P. M., & Kyriakidēs, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. *Contexts of learning*. Routledge.

Ditton, H. (1998). *Mehrebenenanalyse: Grundlagen und Anwendungen des hierarchisch linearen Modells*. *Juventa Paperback*. Juventa Verlag.

Ditton, H., & Müller, A. (2015). Schulqualität. In H. Reinders, H. Ditton, C. Gräsel, & B. Gniewosz (Eds.), *Empirische Bildungsforschung: Gegenstandsbereiche* (2nd ed., pp. 121–134). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-93021-3_9

Gustafsson, J.-E., Nilsen, T., & Hansen, K. Y. (2018). School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation*, *57*, 16–30. https://doi.org/10.1016/j.stueduc.2016.09.004

Kuger, S., Klieme, E., Luedtke, O., Schiepe-Tiska, A., & Reiss, K. (2017). Student learning in secondary school mathematics classrooms: On the validity of student reports in international large-scale studies. *Zeitschrift für Erziehungswissenschaft*, *20*(2), 61–98. https://doi.org/10.1007/s11618-017-0750-6

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, *19*(6), 527–537. https://doi.org/10.1016/j.learninstruc.2008.11.001

Mohammadpour, E., Shekarchizadeh, A., & Kalantarrashidi, S. A. (2015). Multilevel Modeling of Science Achievement in the TIMSS Participating Countries.

Journal of Educational Research, *108*(6), 449–464. https://doi.org/10.1080/00220671.2014.917254

Ozberk, E. B. U., Findik, L. Y., & Ozberk, E. H. (2018). Investigation of the Variables Affecting the Math Achievement of Resilient Students at School and Student Level. *Egitim Ve Bilim-Education and Science*, *43*(194), 111–129. https://doi.org/10.15390/EB.2018.7153

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). *Advanced quantitative techniques in the social sciences series: Vol. 1*. Sage Publications.

Shieh, Y.-Y., & Fouladi, R. T. (2003). The Effect of Multicollinearity on Multilevel Modeling Parameter Estimates and Standard Errors. *Educational and Psychological Measurement*, *63*(6), 951–985. https://doi.org/10.1177/0013164403258402

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Vanlaar, G., Kyriakides, L., Panayiotou, A., Vandecandelaere, M., McMahon, L., Fraine, B. de, & van Damme, J. (2015). Do the teacher and school factors of the dynamic model affect high- and low-achieving student groups to the same extent? a cross-country study. *Research Papers in Education*, *31*(2), 183–211. https://doi.org/10.1080/02671522.2015.1027724

Yu, H., Jiang, S., & Land, K. C. (2015). Multicollinearity in hierarchical linear models. *Social Science Research*, *53*, 118–136. https://doi.org/10.1016/j.ssresearch.2015.04.008

Conditioning on the pre-test versus gain score modeling: revisiting the controversy in a multilevel setting

Arpino, B.[1], Bacci, S.[1], Grilli, L.[1*], Guetto, R.[1], Rampichini, C.[1]

[1] Department of Statistics, Computer Science, Applications, University of Florence, Italy

Suggested talk duration: 20 to 30 minutes

Summary (max. 500 words)

We consider estimating the effect of a treatment on a given outcome measured on subjects tested both before and after treatment assignment. A vast literature compares the competing approaches of modeling the post-test score conditionally on the pre-test score versus modeling the difference, namely the gain score. Our contribution resides in analyzing the merits and drawbacks of the two approaches in a multilevel setting. This is relevant in many fields, for example education with students nested into schools. The multilevel structure raises peculiar issues related to the contextual effects and the distinction between individual-level and cluster-level treatments. We derive approximate analytical results and compare the two approaches by a simulation study.

For a treatment at individual level, our results confirm the findings of the literature for a single-level setting. Specifically, the conditioning approach gives a biased estimator of the treatment effect whenever the pre-test is affected by measurement error, though the bias disappears if the pre-test and post-test scores are affected by a common source of error of the same magnitude. As a consequence, designs with the same instrument at pre-test and post-test should be preferred as they help reducing the bias. The gain score approach provides an unbiased estimator if the common trend assumption holds at the individual level, regardless of the assumption holding at cluster level. For an individual-level treatment, including the

cluster mean of the pre-test score as a regressor is not recommended, as it introduces further measurement error without reducing the bias.

On the other hand, the findings for a treatment at cluster level are different because the cluster mean of the latent ability acts as a confounder. Thus, its observable counterpart, namely the cluster mean of the pre-test score, should be inserted as a regressor. However, this is not always sufficient to completely eliminate the bias, because the cluster mean of the pre-test is affected by measurement error. The issue may be relevant with small clusters (e.g., size 4 in our simulation study). Anyway, also in this context using the cluster mean as a regressor is generally convenient because it reduces the bias. Moreover, it is worth noting that, if the cluster mean of the pre-test score is used as a regressor, then the conditioning and gain score approaches provide the same estimates of the treatment effect, regardless of the cluster size.

Relevance to conference theme

The contribution casts the classical controversy of “conditioning on the pre-test versus gain score modeling” in a multilevel setting, highlighting new issues arising from cluster-level treatments and the role of the cluster mean of the pre-test.

Keywords (max. 3)

common trend assumption, measurement error, treatment effect

Sample size determination for clinical trials based on the meta-analytic-predictive approach

Hongchao Qi[1*] (PhD student, supervisors: Dimitris Rizopoulos, Joost van Rosmalen), Dimitris Rizopoulos[1], Joost van Rosmalen[1]

[1] Department of Biostatistics, Erasmus University Medical Center, the Netherlands

Suggested talk duration (15-60 minutes)

30

Summary (max. 500 words)

The meta-analytic-predictive (MAP) approach is a Bayesian method to incorporate historical control arms in the design and analysis of new trials that aims to increase the statistical power and reduce the required sample size. Here we investigate how to determine the sample size of the new trial when the historical data is available and the MAP approach is used in the analysis.

In previous applications of the MAP approach, the prior effective sample size (ESS) acted as a metric that quantifies the number of subjects the historical information is worth and was used to determine the sample size of the new trial with the MAP approach. However, the validity of using the prior ESS in sample size calculation is not clear, because it only represents the amount of information for a single parameter and different approaches could yield inconsistent prior ESS values.

In this work, we propose an alternative and straightforward Monte Carlo sample size calculation approach to determine the sample size that achieves the desired power in the new trial given available historical controls. The control arm parameters are not taken as a point estimate but sampled from

the MAP prior, to fully incorporate the available historical information in the simulation of the new trial data. The simulated new trial data and the historical data are also analyzed with the MAP approach to derive the statistical power for the treatment effect and the resulting required sample size. The proposed sample size calculation framework was illustrated with real life data sets with different outcomes from three studies and proven to be a straightforward and generic approach to determine the required sample size for the MAP analysis.

Relevance to conference theme

The sample size calculation framework proposed in this study is based on the MAP approach, which is a hierarchical (multilevel) model incorporating historical data in the analysis of the current data. In addition to their implementation in clinical trials, the MAP approach and the proposed sample size calculation framework could also be implemented in experiments in other fields, e.g. education, psychology. Therefore, we believe the study fits the scope of the conference.

Keywords (max. 3)

Sample size calculation; clinical trials; meta-analytic-predictive (MAP) approach

Partitioning Variance for Zero Inflated Count Data in Early Child Development

Ofili, S.[1*], Barry, S J E.[2]

[1] PhD-student, University of Strathclyde, United Kingdom

[2] Supervisor & Strathclyde Chancellor's Fellow, University of Strathclyde, United Kingdom

Suggested talk duration (15-60 minutes)

15 minutes

Summary (max. 500 words)

Background

Exposure to different physical and social environments can result in neighbourhood variation in childhood development. Using multilevel modelling, the importance of the neighbourhood context can be interpreted by partitioning the variance at each contextual level (e.g. individual, preschool, neighbourhood).

Few studies have been conducted on neighbourhood variation in social, emotional and behavioural development in childhood, especially in early childhood for which there is often limited population level data. Research conducted in 2015 as part of the Child Mental Health in Education (CHiME) Study in Glasgow, U.K. was, as far as we are aware, the first attempt at estimating neighbourhood variation in early social and emotional difficulties within a city using spatial multilevel analysis.

Until recently, count distributions (used for skewed and discrete outcomes such as the number of social and emotional difficulties) were considered too complex to derive or interpret partitioned variance and this was omitted from reporting. As a result, there are still unanswered questions relating to the contextual variation of social and emotional difficulties in early childhood:

1. How important is the neighbourhood compared to the preschool and individual contexts in developmental variation?
2. How do individual level covariates affect the influence of the neighbourhood context on developmental variation?

This work aims to apply recently developed approximations of partitioned variance for count data to the CHiME Study data to answer these questions.

Methodology

The CHiME Study collected Goodman's 'Strengths and Difficulties' Questionnaire (SDQ) scores of children in Glasgow, UK, who attended local authority nurseries from 2010-2017. The data is linked to demographic characteristics and postcodes.

In a null model, 37,491 children were nested in 181 preschools and 21 electoral wards following a cross-classified structure. There was no residual spatial correlation. The adjusted model included sex, deprivation, age and a linear time trend. Scores were modelled assuming a zero-inflated negative binomial model and using R-INLA for Bayesian inference.

Partitioned variance was calculated using recently developed approximations. In this work, the equations were adapted to account for zero inflation and tested on toy data before application to the CHiME dataset.

Results

Variance approximations produced reliable results using toy data. Without adjustment for covariates, neighbourhood explained 1.03% and preschool explained 6.24% of the total variance of SDQ scores. Individual/family accounted for the remaining 92.71%. After adjustment for covariates (all were significantly associated with SDQ scores except for the time trend), the role of individual, preschool and neighbourhood contexts became 92.62%, 6.14% and 1.24%, respectively.

Conclusion

Variation between neighbourhoods and schools in Glasgow was largely unaffected by their individual composition. Given the small proportion of variance attributed to neighbourhood defined as electoral wards, an alternative geographical boundary may be more important in early child development. Preschools were more important than neighbourhoods. Most of the variance is explained by individual/family level components, highlighting the importance of the household context on child development.

Relevance to conference theme

Application of a statistical approximation of variance partition for zero inflated count data

Keywords (max. 3)

Count Distributions, Partitioning Variance, Child Development

Multilevel modelling of school effects on student achievement and their application in school accountability systems: models, assumptions, innovations, debates, and challenges

Leckie, G.[1*]

[1] Centre for Multilevel Modelling, University of Bristol, United Kingdom

Suggested talk duration: 45 minutes

Summary (max. 500 words)

Studies of school effects on student achievement were some of the very first applications of multilevel modelling, motivating and illustrating many early methodological developments. I have worked on and off in this area for the past 15 years and it continues to result in new substantive studies and methodological innovations. However, what I have found most interesting is the way various school accountability systems around the world have tried to adopt this approach and the policy and communication tensions and unintended consequences that this can lead to. In this talk, I will reflect on this area of application, reviewing some of the key multilevel models, assumptions, innovations, debates, and challenges. While my substantive focus is an education one, the more general themes of my talk are relevant to many other areas where multilevel models are used to study organisation, neighbourhood, or other group effects on individual outcomes, especially where accountability, choice, or other decisions are made off the back of such data.

Abstracts Posters

Searching for the effect of multiple uncontrolled interventions in BRMS, any tips?

Ackermans, Kevin[1*], Camp, Gino[1], Ackermans, Kevin[1*] Anne-Marieke van Loon[2], Marijke Kral[2].

[1] The Open University of the Netherlands, the Netherlands

[2] HAN University of Applied Sciences, the Netherlands

Briefly Explain Your Question (max. 100 words)

We search for the effects of 8 different (uncontrolled) interventions (1 intervention per school) on the subconcepts of learner's (4^{th} to 8^{th} grade) motivation, self-regulation, and ICT-competency data over the past three years. Data marking for intervention (yes/no), ICT competence of teachers and the presence of specially trained teachers are added to the formulas. Assessment of ICT competency in 3^{rd} grade can be used as prior. Smooths illustrate if the (motivation or self-regulation) concept grows over time grouped by intervention, school or grade. Grades are nested within schools and data is grouped by student. Are we missing anything?

Scientific field(s) of the author(s)

Educational Science

Relevance to conference theme (max. 50 words)

Bayesian multilevel analysis in BRMS

Keywords (max. 3)

BRMS, multilevel

Stepwise, bottom-up simulation for determining sample-size for longitudinal clustered RCTs: appropriately applied or not?

den Hollander, W. [1*]

[1] Trimbos Institute, The Netherlands

Briefly Explain Your Question (max. 100 words)

I have concocted a stepwise, bottom-up simulation approach in R to determine the required sample-size for longitudinal clustered RCTs. For each combination of supplied intraclass correlation coefficients, between timepoint Pearson correlation coefficients, timepoints, effect sizes, drop-out rates, number of clusters and their respective sizes 1000 datasets are generated and analyzed with the appropriate linear mixed model. To determine power, the fraction of runs that have resulted in the detection of a significant effect in the hypothesized direction is computed. The required sample-size can then be inferred from the set threshold. I'd like to get feedback on the methodology.

Scientific field(s) of the author(s)

bioinformatics, medical statistics, mental health

Relevance to conference theme (max. 50 words)

While conventional analytical approaches are occasionally applicable, applying the appropriate formulae to determine the required sample-size to detect intervention effects in longitudinal clustered RCTs might still be difficult or merely constitute approximations. In contrast, simulation is an approach that can be exactly matched to the actual study design and thereby offer study-tailored calculations.

Keywords (max. 3)

Simulation, sample-size, longitudinal clustered RCTs

How to compare two multilevel models that have dependent variables at different Scales?

Author: Marcos Roberto Machado, PhD-student

Advisor: Maria Fernanda Diamantino

Co-Advisor: Luísa Canto e Castro Loura

CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal.

Briefly Explain Your Question

This question arises within the scope of the doctoral paper entitled "Hierarchical Linear Models in the study of determining factors for the school performance of students in Brazil and Portugal", which aims to analyze the determinants that interfere in the school performance in Mathematics of Portuguese and Brazilian students in the 9th grade of Basic Education. From the provided data, it appears that the students' proficiency in Mathematics (latent traits) are measured on different scales: in Brazil the SAEB uses the IRT – Item Response Theory, while in Portugal the National Exams use the CTT - Classical Test Theory.

Scientific field of the author

- Item Response Theory
- Hierarchical Linear Models
- Generalized Linear Models
- Educational statistics

Relevance to conference theme

The present question assumes importance since the scope of the conference is entirely related to the theme of my doctoral thesis, that is, the use of multilevel models to investigate an educational phenomenon, this being a precursor of research of these methodologies.

Keywords

Mathematical proficiency; Equalization Scale; Multilevel Model

A personalized remote patient monitoring system based on daily measurements of body weight, heart rate, and blood pressure to early detect deterioration of heart failure

Mehran Moazeni[1], Emmeke Aarts[2], Folkert W. Asselbergs[3], Linda van Laake[4], Daniel Oberski[5], Lieke Numan[6]

[1,2,5] Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

[3,4,6] Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht, The Netherlands

Presenting Author: Mehran Moazeni (PhD-student)

* 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)

Briefly Explain Your Question (max. 100 words)

Non-invasive remote patient monitoring (RPM) of biometric measurements is a promising technique to detect early clinical deterioration in heart failure (HF) patients. However, conventional methods defining one overall threshold over all patients fail to show predictive value.

Here, we propose a novel personalized RPM algorithm for HF patients using a combination of multilevel linear model (MLM) and statistical process control (SPC) chart. The model is informed by both cross-sectional information on the regular longitudinal pattern, and patient-specific longitudinal information.

Comparing receiver operator curve (ROC) shows that the proposed approach outperforms the conventional approaches in terms of specificity and sensitivity.

Scientific field(s) of the author(s)

Longitudinal models, heart failure

Relevance to conference theme (max. 50 words)

Utilizing multilevel modelling to obtain personalized estimates of biometric measurements over time

Keywords (max. 3)

Remote patient monitoring, longitudinal measurements, statistical process control charts.

How to Represent Different, Yet Overlapping, Timescales in Intensive Longitudinal Data and Longitudinal Social Media Data?

*Pasquini, G.[1] (PhD-student; supervisor: Dr. Stacey B. Scott), Schwartz, H. A.[1], Clouston, S.[1], Scott, S. B.[1]

[1] Stony Brook University, United States

* 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)

Briefly Explain Your Question (max. 100 words)

How did individual and community affect align before and during the pandemic? During 2019-2020, participants from a single community reported their affect 5 times per day across 16 days each year using ecological momentary assessment (EMA) and community affect was measured using Twitter sentiment aggregations to produce one community score per week during each year. Initial attempts at analyzing these separately and correlating week-to-week loses the temporal sequencing and aggregates over within-week variation. We need guidance to model time-varying, seasonal and lead/lag effects, and year-to-year change, considering that participants' EMA periods coincided with up to four Twitter weeks each year.

Scientific field(s) of the author(s)

Developmental psychology; Natural-language processing; Public health

Relevance to conference theme (max. 50 words)

We aim to model longitudinal EMA and social media data together while preserving the timescales over which measurements were collected. The goal is to detect within-person change and variability with time-varying predictors sampled at different frequencies to accurately represent the multiple levels of nesting in the data.

Keywords (max. 3)

Time-varying; Affect; Timescales

How to Incorporate Moderators in a Multilevel Multiple Mediation Model with Repeated Measures Data Using BSEM?

Tingting Reid[1], Charee Thompson[2], Emiko Taniguchi[3]

The current study aims at investigating the pathways mediating the effect of mental illness uncertainty on support provision via fear and anxiety as well as support communication efficacy using repeated measures data consisting of 201 college students. Data collection occurred once a week for six weeks. Covariates were measured at weeks 1 and 6, respectively, and moderators were measured only at week 1. Due to the relatively small sample size and missing data issues (approximately 25% missingness at each week), we employed a full Bayesian model approach. In our hypothesized models (Figure 1), we tested two separate sets of path analyses in the multiple mediation models which allowed us to understand the relative contribution of each specific indirect effect.

Briefly Explain Your Question (max. 100 words)

1. Given that the 'a' paths are conditioned on the moderator only in Figure 1, would this be considered as a moderated mediation model? If so, how to model the conditional direct and indirect effects at both levels?
2. Given that covariates have both within-person and between-person variability, would it be best to treat them as latent using Multilevel SEM?
3. Regarding the missing data, would a fully Bayesian approach using MCMC properly handle it? Or would we need to first perform multiple imputation for missing data and then apply the Bayesian estimator? Should we use different priors?

Scientific field(s) of the author(s)

Educational Psychology, Health Communication

Relevance to conference theme (max. 50 words)

We are seeking to understand how to apply multilevel SEM to a longitudinal mediation model given the clustering nature of our data. We are also looking to apply BSEM to obtain robust results given the complexity of our model, issues relating to a small sample and missing data.

Keywords (max. 3)

BSEM, multilevel, longitudinal mediation

[1] University of Hildesheim, Germany
[2] University of Illinois at Urbana-Champaign, U.S.
[3] University of Hawaii at Manoa, U.S.

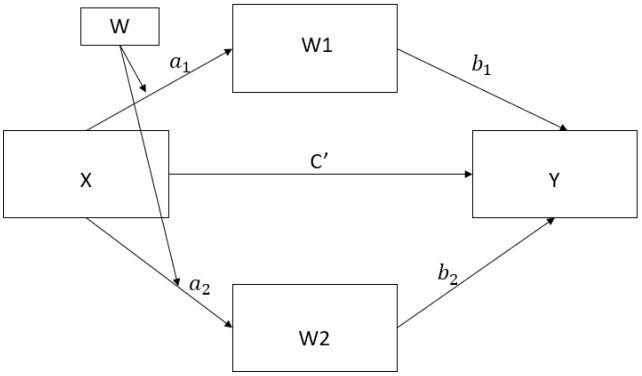

Figure 1. Hypothesized model

What is the best analytical strategy for identifying multiple school-level predictors of SES achievement gap?

Sevalneva, D.[1*], Erentaitė, R.[1], Raižienė, S.[1,2], Vosylis, R.[1,3]

[1] Kaunas University of Technology, Lithuania

[2] Vilnius University, Lithuania

[3] Mykolas Romeris University, Lithuania

* Presenting author

Briefly Explain Your Question (max. 100 words)

We aim to identify school-level factors that explain the variation of SES achievement gap in Lithuania. The population-based dataset covers a cohort of 8th-grade students for the year 2020-2021. The dataset includes a range of school characteristics and student-level data on achievement and socio-economic background. Results show that SES achievement gap varies across schools and is larger in schools with lower SES. As we seek to identify school-level factors that may help to reduce SES achievement gap, we want to know the potential pitfalls for including multiple school-level predictors to account for SES cross-level interaction effects.

Scientific field(s) of the author(s)

Developmental psychology, educational sciences

Relevance to conference theme (max. 50 words)

In our study we apply two-level random effects models with Mplus 8.5 on a large population-based dataset ($N_{students}$ = 25402, $N_{schools}$ = 670). We test compositional and interaction effects of student SES on achievement. We also have a possibility to include multiple school-level factors and distinguish school-level and classroom-level effects.

Keywords (max. 3)

SES achievement gap, SES, school-level factors

Explaining the trait-state discrepancy in teachers' well-being: What effect size measure should be used to report the cross-level interaction effect?

Schriek, Josina (PhD-student) [1*]; Soellner, Renate (supervisor)[1]; Keller, Melanie[2]; Klusmann, Uta[2]

[1] University of Hildesheim, Germany

[2] Leibniz Institute for Science and Mathematics Education, Kiel, Germany

* Presenting author

Briefly Explain Your Question (max. 100 words)

We collected trait and state data on teachers' well-being in a diary study over two weeks. Our data reflects a two-level structure with trait assessments and multiple state observations (Level 1) nested within persons (Level 2). Multilevel analyses showed that, as compared to the state assessment, teachers reported higher levels of trait well-being. Additionally, we specified slopes-as-outcomes models and included neuroticism as a moderator. The observed trait-state discrepancy was accounted for by neuroticism, with high neuroticism levels corresponding with a bigger trait-state discrepancy. We encountered the issue what effect size measure should be used to report the cross-level interaction effect?

Scientific field(s) of the author(s)

Health psychology, methodology, educational psychology

Relevance to conference theme (max. 50 words)

To evaluate a model, e.g. in OLS regression researchers report the proportion of variance explained by an interaction effect, R^2. We would like to discuss effect size measures in multilevel modeling that allow to understand the practical significance and strengths of the cross-level interaction effect.

Keywords (max. 3)

Cross-level interaction – effect size