Springer Proceedings in Mathematics & Statistics

Mark Stemmler
Alexander von Eye
Wolfgang Wiedermann *Editors*

Dependent Data in Social Sciences Research

Forms, Issues, and Methods of Analysis

Springer

Springer Proceedings in Mathematics & Statistics

Volume 145

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Mark Stemmler • Alexander von Eye
Wolfgang Wiedermann
Editors

Dependent Data in Social Sciences Research

Forms, Issues, and Methods of Analysis

Springer

Editors
Mark Stemmler
Friedrich-Alexander
University of Erlangen
Nürnberg (FAU), Erlangen, Germany

Alexander von Eye
Department of Psychology
Michigan State University
East Lansing, MI, USA

Wolfgang Wiedermann
Department of Educational
School, and Counseling Psychology
College of Education
University of Missouri
Columbia, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-20584-7 ISBN 978-3-319-20585-4 (eBook)
DOI 10.1007/978-3-319-20585-4

Library of Congress Control Number: 2015950842

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This volume presents contributions on handling data in which the postulate of independence in the data matrix is violated. When this postulate is violated and the methods assuming independence are applied nevertheless, the estimated parameters are likely to be biased, and inference statistical conclusions are very likely to be incorrect. Cook (2012) describes four contexts in which the postulate of independence is violated:

1. Repeated measures (longitudinal data)
2. Clustered data (e.g., siblings in schools, children in families, patients in hospitals)
3. Data from individuals who live closely together (e.g., people from the same neighborhood)
4. People in social networks (e.g., dyads, triads)

Cook elaborates on the significance of the problems with dependent data that "unlike some assumptions of statistical theory (e.g., normal distribution), which can sometimes be violated without very serious consequences, violation of the independence assumption typically has serious consequences" (2012, p. 522). This problem has been known for some time, which is reflected in the development of tailored methods for the analysis of dependent data (e.g., methods for the analysis of repeated measures), in corrections, taking into account the extent of dependence, adjustments of test statistics (e.g., adjustment of F values in repeated measures ANOVA), or adjustments of degrees of freedom. Examples of such developments can be found in various areas of statistics.

Solutions for handling serious violations of assumptions for dependent data are being developed and created constantly, but they are in many areas not yet completely satisfying. This volume is an effort to present the status quo of the progress in various statistical areas in managing dependence. We present modern up-to-date statistical methods for dealing appropriately with problems related to dependent data, including real data examples. These methods also reveal the power of those modern techniques. At the same time, examples are presented that illustrate problems from not dealing appropriately with assumptions of independence. All

authors of this volume are leading experts in their field of applying or developing new statistical methods for dependent data scenarios.

This book consists of five parts: (1) growth curve modeling, (2) directional dependence in regression models, (3) dyadic data modeling, (4) item response modeling, and (5) other methods for the analysis of dependent data such as multidimensional scaling techniques, methods for modeling cross-section dependence in panel data, and mixed models. In the following paragraphs, we briefly introduce the content of each part.

Part I: Growth curve modeling. Jack McArdle starts with a discussion of approaches to modeling change from the Cognition in the USA (CogUSA) survey. He tests multiple factorial invariance over time by estimating various models of latent change. *Paolo Ghisletta*, *Eva Cantori*, and *Nadège Jacot* demonstrate how to handle latent curve models including data with serious forms of nonlinearity. *Jost Reinecke*, *Maike Meyer*, and *Klaus Boers* apply a stage-sequential growth mixture model to the data of their study of Crime in the Modern City (CRIMOC), a criminological panel dataset. *Mark Stemmler* and *Friedrch Lösel* present a latent change model that includes five mixture groups in the real life example of the Erlangen-Nuremberg Development and Prevention Study (ENDPS). The first part of this volume concludes with a contribution by *Jang Schiltz* who extends Nagin's mixture models by adding a slope component.

Part II: Directional dependence in regression models. This part discusses issues related to causality. In the first chapter of this part, *Alexander von Eye*, *Wolfgang Wiedermann*, and *Ingrid Koller* present the concept of Granger causality. Granger causation is interesting from a developmental perspective. It allows researchers to test hypotheses concerning the causal relations between two series of observations which may develop simultaneously. In the second chapter, *Wolfgang Wiedermann* proposes decisions concerning the direction of effects in linear regression models based on fourth central moments.

Part III: Dyadic data modeling. Numerous techniques have been developed for the analysis of dyadic data. The most prominent of these involve regression, path, and structural equation models. *Rainer Alexandrowicz* extends these approaches by considering Item Response Theory (IRT) Models. His approach combines the advantages of metric dyadic data analysis with a model for discrete data, thus allowing for categorical items while drawing inferences based on the estimated true scores on an interval scale. In the second chapter of this part, *Heather Foran* and *Sören Kliem* apply models for latent variables in longitudinal analysis of dyads. Several competing models and their applications are demonstrated. In the final chapter of this part, *Ting Wang*, *Phillip K. Wood*, and *Andrew C. Heath* discuss the application of psychometric measurement models (with a focus on Bayesian estimation of random intercept models) to quantify environmental and genetic components in behavior genetic models.

Part IV: Item response modeling. More data examples and solutions for problems dealing with dependent data in Item Response Theory (IRT) are discussed in the fourth part. *Ingrid Koller*, *Wolfgang Wiedermann*, and *Judith Glück* exhibit quasi-exact tests for the investigation of pre-conditions for measuring change.

Steffi Pohl, *Kerstin Haberkorn*, and *Claus Carstensen* illustrate how to measure competencies across the lifespan using IRT models. *Ferdinand Keller* and *Ingrid Koller* demonstrate the use of mixed Rasch models for analyzing the stability of response styles across time. In their data example, the authors use data of the Beck Depression Inventory (BDI-II).

Part V: Other methods for the analysis of dependent data. Finally, the last part introduces various methods for the analyses of dependent data that did not belong to any of the above four topics. *Cody Ding* shows a data example from educational research using Multidimensional Scaling for the analysis of growth patterns. *Harry Haupt* and *Joachim Schnurbus* use a nonparametric approach to modeling cross-section dependence in panel data. Finally, *Christof Schuster* and *Dirk Lubbe* contrast MANOVA to Mixed Models and discuss the advantages and disadvantages of each method in terms of handling within-subject dependency.

Cook, W. L. (2012). Foundational issues in nonindependent data analysis. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 521–536). New York: The Guilford Press.

Erlangen, Germany — Mark Stemmler
East Lansing, MI, USA — Alexander von Eye
Columbia, MO, USA — Wolfgang Wiedermann
Summer 2015

Acknowledgements

In all, this volume is the result of contributions that were presented at an international meeting in Erlangen at the Friedrich-Alexander University Erlangen-Nuremberg from December 6 to 7, 2013. This meeting was financially supported by the German Research Foundation (DFG; GZ: STE 923/6-1). The topic of the meeting was the same as the title of this oeuvre. At the meeting, there were presentations and discussions of up-to-date developments and applications for the analysis of data where the postulate of independence was violated. Our thanks go to the German Research Foundation for supporting our meeting. In addition, we thank Hannah Bracken of Springer Press for her enduring effort for making this endeavor possible and for publishing this volume in the highly respected Springer Series of Proceedings in Mathematics and Statistics.

The first editor wants to express his gratefulness to Susanne and Quincy for their support, comfort, and love. The second editor of this volume wishes to acknowledge that he is dependent upon Donata, her love, and her support, and this acknowledgement is without bias. The third editor of this volume is grateful to Anna and Linus for making him a part of the most wonderful triad on earth.

Contents

About the Editors

Mark Stemmler Since 2011 Mark Stemmler is a Chair of Psychological Assessment, Quantitative Methods and Forensic Psychology at the Institute of Psychology at the Friedrich-Alexander University of Erlangen-Nuremberg (FAU), Germany, and an adjunct Professor at the College of Health and Human Development at the Pennsylvania State University, USA. He received his master's degree from the Technical University Berlin in 1989 and his PhD from the Pennsylvania State University in 1993. In 2002, he received his postdoctoral lecture qualification (Habilitation) from the FAU. From 2007 to 2011 he was a full professor for quantitative methods at the Bielefeld University, Germany. His research interests encompass developmental psychology and methodology. He has worked on longitudinal studies in the USA and Germany. His research emphasis in methodology is on person-centered methods.

Alexander von Eye received his degrees in Psychology from the University of Trier, Germany. He held positions at the University of Trier, the University of Erlangen-Nürnberg, the Max Planck Institute for Human Development in Berlin, Penn State, Michigan State, and the University of Vienna, Austria. He is accredited by the American Statistical Association as Professional Statistician, and he is Fellow of the APA and the APS. Alexander von Eye is a developmentalist whose work centers around statistical methods for longitudinal research, for categorical data analysis, and for modeling, and he conducts computer simulations to explore the performance of statistical methods under various conditions. He has published over 400 book chapters and articles and authored has edited or written over 20 scholarly books. Currently, he enjoys life in Montpellier, Southern France.

Wolfgang Wiedermann received his Ph.D. in Psychology from the Alpen-Adria University of Klagenfurt, Austria. He is Assistant Professor at the University of Missouri. His research interests include the development of methods for causal inference, methods of person-oriented research, methods for intensive longitudinal data, and methods for the psychometric analysis of preference data.

Part I
Growth Curve Modeling

The Observed Dependency of Longitudinal Data

John J. McArdle

Abstract It is well known that longitudinal data can deal with different concepts than cross-sectional data (see Baltes & Nesselroade, 1979; McArdle & Nesselroade, 2014). The key is in the observed dependency—that allows us to examine individual changes. Thus, all of the individual changes that can be examined are due to the longitudinal models (see McArdle, 2008) allowing dependencies among the observed scores at various time points. It is demonstrated here that the statistical power to detect changes is an explicit function of the positive dependencies and the timing of the observations. A lot of time is spent on the move to the *latent curve model* (LCM) from the basic regression structural model and the repeated measures model (RANOVA) because the latter seems standard in the field now. This LCM is introduced in this chapter as a principle that does have power to detect many more changes than the usual regression analysis but it comes along with several (to be discussed) assumptions.

The four articles to follow in this volume are reviewed with longitudinal dependency in mind, and the highlights of each chapter are brought out. The chapter "Nonlinear Growth Curve Models" extends the LCM to handle serious forms of nonlinearity, and this is clearly prevalent in Psychology. The chapter "Stage-Sequential Growth Mixture Modeling" extends this work to include multistage models, Poisson relations, all in the context of a multiple mixture model. This is a fairly complex example. The chapter "General Growth Mixture Modeling: The Study of Developmental Pathways of Externalizing Behavior from Preschool Age to Adolescence" is a real-life example that includes LCMs for five mixture groups. The chapter "A Generalization of Nagin's Finite Mixture Model" extends the mixture models further, mainly by adding a slope component.

But what is also important in this regard is "measurement invariance" and how this can be crucial to understanding changes. Some elaboration of the early work

A contribution for a book on "Dependent Data in Social Science Research" Edited by M. Stemmler, A. von Eye and W. Wiedermann.

J.J. McArdle (✉)
Department of Psychology, University of Southern California, Los Angeles, CA, USA
e-mail: jmcardle@usc.edu

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer
Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_1

on scales is further developed for selected items. The data to be considered here for LCM are a subset of the full set of data collected in the Cognition in the USA (CogUSA survey; McArdle & Fisher, 2015). These scales were chosen in a way that would be consistent with the principles of *multiple factorial invariance over time* (MFIT) but the result of the age-related changes over two waves was largely unknown and in need of establishment. Basically, we first try to establish MFIT over the two waves and then look for latent changes in these scales over age. Thus there are only eight scales to consider here (four cross-sectional scales by two longitudinal occasions), so there is still a lot of work to do!

It is well known that longitudinal data can deal with different concepts than cross-sectional data (see Baltes & Nesselroade 1979; McArdle & Nesselroade 2014). That is, cross-sectional data has many good opportunities for "between person differences" but it cannot deal with "within a person changes." The first dependency that is created and observed is that the same person is used at multiple occasions. This dependency has been used in multivariate modeling a great deal. Because the same person has multiple inputs and outcomes we can deal with this in different ways. All of the individual changes that can be examined are due to the longitudinal models (see McArdle 2008) allowing dependencies among the observed scores at various time points. This dependency is also responsible for the popularity of multilevel modeling (see Bryk & Raudenbush, 1987, 1992). It is demonstrated here that the statistical power to detect changes is an explicit function of the positive dependencies and the timing of the observations.

The typical lack of dependency is monitored in statistics by a careful assessment of the original scores, typically using linear regression with an outcome score (Y_n) and a predictor (X_n) score and usually written as

$$Y_n = \beta_0 + \beta_1 X_n + e_n, \tag{1}$$

where the regression terms β_0 and β_1 are thought to apply to everyone, and the residual term (e_n) is an individual characteristic that is unmeasured and supposedly follows a normal distribution. This is an effort to find the relationships between some outcome Y and the input variable X. If X is a group then this model provides a way to determine group differences on the outcome (the usual ANOVA as a between groups *t*-test). But this is not an effort to deal with observed dependency in traditional regression analysis (see Fox 1999).

But some people noticed that having an individual measured more than once created a statistical virtue. Indeed this was the stimulus for progressively repeated measures. One classical representation of longitudinal data can be found in the *repeated measures model for the analysis of variance* (*RANOVA*; see Fisher 1925). In this first model the individual score at any time point ($Y[t]_n$) is assumed to be decomposed as

$$Y[t]_n = \beta_{0n} + \beta_1 X_n + e[t]_n \tag{2}$$

where the individual (n = 1 to N) is allowed to differ at all throughout the time series (t = 1 to T) in two ways: (1) Individuals are different from one another at all times, and (2) there are random normal fluctuations at each time point ($e[t]_n$). The use of the X weighted function is an adjustment in the mean of the scores for group differences in the trends over time. This model can give correct statistics for the mean of the individuals and the effect of X (assuming it is the same over all occasions) as long as the contrast questions are "spherical" in shape (among others, see Davidson 1972; Huynh & Feldt 1976).

The repeated measures model permits the power to detect differences between treatment groups in means (or over time) as a function of the standard deviations of the scores (as usual, with the sample size included as the square root of N at the end). But in repeated measures, the variance at the second occasion is also based on the correlation of the observed scores over time:

$$\mu_d = (m[1] - m[2]) / \left(s[1]^2 + s[2]^2\right) - 2\ ((s[1] + s[2])\, r[1,2]) \tag{3}$$

where we have symbolized the estimated mean difference as μ_d, using the two observed means as m[1] and m[2], the two observed variances as $s[1]^2$ and $s[2]^2$, and the observed correlation over time as r[1,2]. This is nothing more than the mean difference over the standard deviation, but the correlation is for the same measure at two occasions. So for the same mean difference (m[1] − m[2]) as found in a cross section we can say we have found a significant different from zero if the correlation of the two measures is positive (which it typically is; see Bonate 2000; Cribbie & Jamieson 2004). For this reason, it is typically far better (depending on the sign of the correlation) to measure a person twice than to measure twice as many people just once. That is, *the longitudinal case is far more powerful than the cross-sectional case*. This is not the only issue of statistical power (see Tu et al. 2005) that could be considered, but it is relevant here. Of course, there are more than two time points over which change is to be measured, and this typically increases our power.

The Move to a Latent Curve Model

A straightforward generalization of this *RANOVA* model allows the move to a *latent curve model* (*LCM)* and makes it not very hard to understand. This LCM was first used by Tucker (1958, 1960 1966) and Rao (1958), and later Meredith and Tisak (1990) gave it a *structural equation model* (*SEM*) interpretation (also see McArdle 1986 and McArdle & Epstein 1987) to determine the best fitting curve to the observed data. Basically, the slope can vary along with any way the individual changes. Each individual is assumed to have three latent variables, defined as

$$Y[t]_n = L_n + S_n\Omega\,[t] + u[t]_n \tag{4}$$

so the three sources of variation in any response are: (1) A constant change for the individual over all times (the latent level = L), (2) a systematic change (based on a slope score = S, which is systematic with the set of basis coefficients = $\Omega[t]$), and (3) a unique change = $u[t]$, which is essentially random with respect to the other changes. We can examine that the set of basis coefficients ($\Omega[t]$ is not necessarily linear) to determine the slope of the best fitting line or trajectory of the data, but this line supposedly has the same coefficients for everyone.

All sources of individual differences are indexed by variance ($\phi_L{}^2$, $\phi_S{}^2$, and ψ^2). In addition, the constant change is allowed to have covariance (ϕ_{LS}) or be correlated (ρ_{LS}) with the systematic changes. The variance that remains (the uniquenesses, ψ^2) is assumed to be uncorrelated with the changes or the starting point and is furthermore assumed to be equal over time.

We can also have the observed group effects on these individual coefficients, and we can do what we want with them. What is usually done follows the usual regression logic with two of the latent variables as new outcomes:

$$L_n = \alpha_0 + \alpha_1 X_n + e_{Ln} \quad \text{and} \quad S_n = \beta_0 + \beta_1 X_n + e_{Sn} \tag{5}$$

in which case the e_L and e_S account for the residual variance and covariance. This kind of mixed model function, including both fixed (α_0, α_1, β_0, β_1, and $\Omega[t]$) and random ($\phi_L{}^2$, $\phi_S{}^2$, ψ^2, and ϕ_{LS}) effects, can be evaluated for goodness of fit using the standard SEM statistical logic (see Meredith & Tisak 1990; McArdle 1986). If the model fits the data of means and covariances we assume that the score model (of [4] and [5]) is reasonable.

The kind of change we will test is dependent largely on the set of basis coefficients we employ. We can force the systematic change to be linear with the time simply by fixing the coefficients $\Omega[t] = [0,1,2,3 \ldots T]$. This is often done, but it is only one option, and there are many others. We can even estimate some of the coefficients (T-2 in the one factor case) so that they form an optimal curve for the data. This is basically what the earliest pioneers (Tucker, Rao, Meredith, etc.) did. But there are many more ways to examine the curves and a lot can be done here. Using the basic logic, we can also consider more than one curve for these data (as done in later chapters).

The LCM is considered useful now because it can describe both, group (i.e., fixed) and individual (i.e., random). For this reason it is popular in psychology where we often are interested in group effects but individual differences from the same perspective. We should note that it is not widely used in other areas of science (e.g., Econometrics) where the dominant paradigm uses time as a causal hinge, so which measure came last in time is regressed on all the prior instances. The same longitudinal data can be used in this way (see McArdle 2008; McArdle & Nesselroade 2014).

We note immediately that the LCM does not try to explain how the prior time points (if measured) impact the subsequent events. This makes the procedures of LCM more descriptive than inferential. But all is not lost because there is some savings in the number of parameters used to define these differences.

Model Fit and Model Selection

A good question can be asked about "Does the model fit the data?" This question can be answered in a number of ways. But what we want is a model that has easy to understand parameters and fits as well or better than others of its kind. The approach, known by the *Bayesian Information Criteria* (BIC) is used throughout this book so it is useful to investigate it further now, according to Raftery (1996) and Nagin (2005, p.64) the formula for BIC can be written as

$$\text{BIC} = log\,(L)\text{–}1/2p\;log(n) \tag{6}$$

where the *log* is the natural logarithm, and L is the model's maximum likelihood, and this is penalized (lowered) by p, the effective number of parameters used, and n, the sample size of individuals used. "If one is comparing several models we should prefer the one the lowest BIC values." (Raftery 1996, p. 145). In this way, the BIC "counterbalances" a good fitting model by the number of parameters and the sample size used. So, although it does not seem to be the fit of the model, it can help choose one model among many others. What we hope to obtain is a model where the BIC is as negative as possible, although there are several ways to use this information. Several keen insights into how this BIC behaves are given in Nagin (2005), and these will not be repeated here, but the use of Bayes factors is illustrated. The use of the BIC is obviously Nagin's favored device for model selection with groups, but he does conclude that:

Such debate is important for advancing the theoretical foundations of model selection. However, disagreement about the technical merits of alternative criteria may obscure a fundamental point—there is no correct model. Statistical models are just approximations. The strengths and weaknesses of alternative model specifications depend upon the substantive questions being asked and the data available for addressing these questions. Thus the choice of the best model specification cannot be reduced to the application of a single test statistic. To be sure, the application of formal statistical criteria to the model selection process serves to discipline and constrain subjective judgment with objective measures and standards. However, there is no escaping the need for judgment; otherwise insight and discovery will fall victim to the mechanical application of method. In the end the objective of model selection is not the maximization of some statistic of model fit. Rather it is to summarize the distinctive features of the data in as parsimonious a fashion as possible (Nagin 2005, p.77).

I can easily say I am in complete agreement about these model-fitting issues.

Potential Biases

Thus, the collection of longitudinal data is useful because: (1) They allow the study of the natural history of the development of problem behavior, such as externalizing behavior, its onset and termination. (2) They allow the study of trajectories or pathways. A pathway is defined as "when a group of individuals experience a behavioral development that is distinct from the behavioral development of another group of individuals" (Loeber & Farrington, 1994, p. 890). Trajectories or pathways provide information of processes of continuity and discontinuity and on inter-individual differences. In addition, Loeber and Farrington (1994) postulate that the best studies now rely on multiple informants. The chapter by Stemmler and Lösel (Chapter 4) meets all of these criteria and this chapter should be considered carefully.

But we need to be clear about the difference between a repeated measures design and a multivariate design because both allow correlation over time. For both, sample members are measured on several occasions, or trials. But in the repeated measures design, each trial represents the measurement of the same characteristic, in the same way, at a different time. In contrast, for the multivariate design, each trial represents the measurement of a different characteristic. It is generally inappropriate to test for mean differences between disparate measurements, so the difference score is useful (in contrast to what is stated in Cronbach & Furby 1970).

But the longitudinal method is not without some well-reasoned detractors (see Rogosa 1988). Among many critiques of the longitudinal method: (1) It is hard to get the representative sample to come back to a second testing, and the people who do come back have done very well at the first time (see McArdle 2012); (2) if they do come back, they have seen the measures before, so it is difficult to measure exactly the same constructs at a second time, without retest or practice effects; and (3) the construct or thing that we want to measure may have changed, and we will not know it by simply looking at the variance or taking the difference between measures. These are some of the many potential confounds of the longitudinal method.

The results of these problems lead us to think that a cross-sectional study had less potential confounds than a longitudinal study. This is hardly ever true because these conditions can occur in cross sections as well, and we may not know it.

Assumption 1: In the LCM, the Latent Scores Used Are Related to Latent Change Scores

It seems that all the prior work has focused on the "change" at the individual and group levels but very few researchers are willing to say so. Instead, words like "curve" or "slope" or "trajectory" are used. But there turns out to be an easy way to represent these basic change ideas and we will usually do so here.

We can define the basic model of change to isolate the functions as

$$\mathrm{Y[t]_n} = L_\mathrm{n} + \sum \mathrm{i} = 1, \mathrm{t}\ \{\Delta \mathrm{y[i]_n}\} + u\mathrm{[t]_n} \tag{7}$$

so the changes are just accumulated up to that time (i = 1 to t). This is not intended to be a controversial statement and it leads to the same fit as the prior linear models, but it is really another way to consider have the outcome at time t (after McArdle, 2008).

The change as an outcome can be strictly defined at that latent variable level (after McArdle & Nesselroade 2014) as

$$\Delta \mathrm{y[t]_n} = \mathrm{y[t-1]_n} - \mathrm{y[t]_n} \quad \text{or} \quad \mathrm{y[t]_n} = \mathrm{y[t-1]_n} + \Delta \mathrm{y[t]_n}, \tag{8}$$

so the latent score is the source of all inquiry. This can be useful in a number of interpretations, especially for the regression of latent changes. For example, we now can fit

$$\Delta \mathrm{y[t]_n} = \beta_0 + \beta_1 \mathrm{X_n} + e_{\Delta \mathrm{n}} \tag{9}$$

so the latent change score is modeled directly, and has a residual ($e_{\Delta \mathrm{n}}$). But the LCS approach is entirely consistent with the LGM approach, as stated by McArdle (2008) and this is why the same values emerge for various estimates. The LCS model is largely a clearer change-based re-interpretation of the LCM, and the LCS model can be programmed and used efficiently (see McArdle 2008; McArdle & Nesselroade 2014).

Latent changes are apparent in this model. Much more could be said about this approach, but this is all that will be needed here.

Assumption 2: In the LCM, the Model Parameters Have the Same Shape for Everyone

This assumption is also true of all regression models (see Eq. (1)) but it is most clearly not appropriate here. That is, we can control the size and sign of some parameters of the trajectory with the means and the variances of the latent variables, but the shape of the latent change is a combination that is beyond the usual reach.

The chapters listed here do distinguish between these shapes using an unobserved difference between people. That is, this clear difference between individuals is recast at the main reason they are members of a latent grouping—a mixture of different distributions. This was evidenced in the brilliant early work of Tucker (1960 1966, also see Tucker 1992), and the subsequent maximum-likelihood formalizations of Nagin (1999 2005) and Muthén and Sheddon (1999).

This logic using multiple groups is indeed a good idea, because it is focused on different kinds of changes within the person. But Tucker (1960 1966 1992) seems to have found a way to differentiate people with standard methods of factor-cluster analysis. Perhaps the first time this procedure was used in real questions and stated

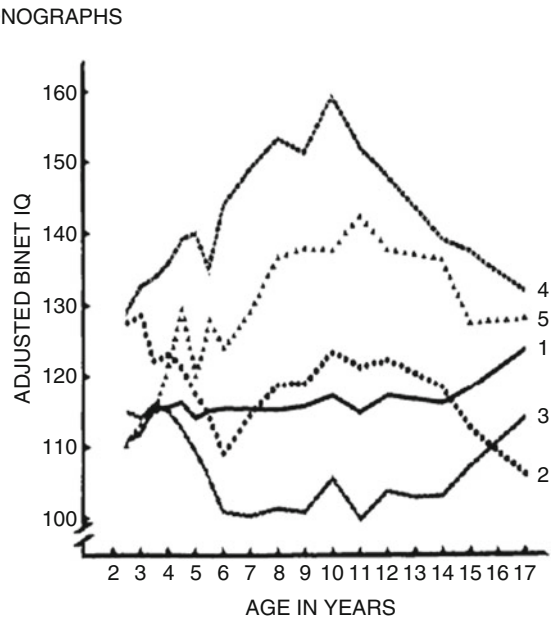


Fig. 1 From McCall, Applebaum & Hogarty (1973, p. 48)

clearly was by McCall, Applebaum, and Hogarty (1973, pp. 44–48) who suggest that there are five clusters of people based on their changes over age in IQ tests over age (see Fig. 1).

Now it is clear that Tucker (1958 1960 1966) did not have all the statistical tests (or MLE) to support these choices, nor did he have or did develop the mixture model as the possibility of a person belonging to multiple clusters (this allowing for a much better mixture), but he did distinguish large group of persons on their trajectory using multiple factors and he resolved multiple clusters, so we will generally consider Tucker's (1958 1966) work as pre-dating the more recent work of Nagin (1999 2005) and Muthén and Sheddon (1999).

Assumption 3: In the LCM, the Residuals Are Equal and Uncorrelated, and the Model Fits

There is much more that could be said about the equality of the unique variance (for details, see Grimm & Widaman 2010) but the basic idea is on must have an a priori theory about why these kind of unique but uncorrelated changes are needed. If we do have such ideas we can remove the variance terms at each time and achieve a much better fit to the data. We will not deal with these issues too much here. In this regard this is an unchallenged assumption that deserves much more scrutiny.

The simple fact that "everything else" is supposedly uncorrelated is actually never met and yet this is what is tested by the model fit. The test of goodness of fit is supposed to test whether or not the LCM can be considered viable. But the way we typically test any hypothesis is to remove all other features until all that are left are random variables. This is primarily because we do know how to test for random events (usually with the χ^2 goodness-of-fit test; but see Raftery 1996).

Assumption 4: In the LCM, the Model Has the Properties of Invariant Measurement

In all cases, it is also necessary to illustrate the loss of fit due to "multiple factorial invariance over time," (MFIT) and how this invariance can be crucial to understanding changes. That is, some things may not change while others will. Here we will only use common factor analysis in a simple example. This is a second dependency because the measures are somewhat the same within a time. Some elaboration of the early work on any scale is further developed for items. This is related to both "test bias" and "harmony." That is, if we assume that a test is a good measurement of a construct, it should behave the same way at all waves.

I do not view MFIT as a "testable hypothesis" as many others do (e.g., Meredith 1993) but I view this as a necessary feature of longitudinal data. That is, in the absence of MFIT it is not clear that we can take differences between successive occasions, and this is critical to most any accumulation model. Thus, this test would be a useful foil against a measure, and we can use it to evaluate an existing measure. But to create one, we must be accumulating something, and that something is strictly defined as the object of our MFIT. Perhaps it is best to say we can evaluate the part of the MFIT that works the way we intended. At least our intentions for MFIT are clarified in this way.

Assumption 5: In the LCM, the Model Variables All Have Normal Properties

Another kind of dependency is that due to items that are miscalculated as normal. That is, we typically assume all variables are normally distributed, even when they are highly skewed. This is also the case of a variable that can reach an upper or lower limit and should be considered censored (see Wang et al. 2008). As we do not illustrate here, but could have, this can pose a major problem for our understanding of the changes (but for an example, see Hishinuma et al. 2012; McArdle et al. 2014).

Assumption 6: In the LCM, the Individuals Have All Been Measured at Exactly the Same Developmental Time Periods

This is also probably never true in epidemiological and psychological studies. The problem comes only because the model assumes this is true. In fact, the age-at-measurement is usually not told to the analyst. This means people can be "measured on their birthdays" or at approximate yearly intervals of time, but we just never know. The word "approximate" is used here frequently, and many see this as a natural feature of longitudinal data. But it is not. The big problem that this creates is that the correlations over time, if they are not in a sequential proper timing, can yield some haphazard results. The timing is important to future studies and not enough is done about this issue yet.

The further assumption that we know the true developmental timing is quite absurd. We do not know this and we do not track it very well either. It could be age or it could be something else like puberty (see McArdle 2011), but we need to know it to state how the individuals form groups of people (see Nagin 2005). We often just use whatever longitudinal data we are given, because we are very happy to get some, and we assume we can do something with it, as is. But we cannot.

The Studies of the First Section of This Book

The studies of the first section of this book seem to criticize some of the basic assumptions of the standard LCM. This should be considered fair as a target because it is loaded with assumptions and the linear LCM was designed to be just a starting point for future work. The concepts of simultaneous estimation are also critical here to distinguish what is being done.

The first study by Paolo Ghisletta, Eva Cantoni, and Nadège Jacot as presented here is an examination of more than linear relationships in psychological research, which they term an NGCM (for nonlinear growth curve model). That is, they do not stop at the quadratic form of the prior LCM, and they do not consider the linear model to capture all the relevant variation in their outcomes (in their example, four blocks of 20 trials of time on task in a pursuit rotor task). Instead, they consider other terms (see their Eq. (6)) that are not a usual part of this basic model (our Eq. (4)).

These author(s) do fit a wide variety of nonlinear models to these data, and this is notable, and they compare each, and this is also notable. But they do drop linearity quickly as a possibility and I think this is a mistake. That is, before we deal with how nonlinear a model can be I think we ought to first see how linearity works, in terms of explained variance at each time point ($\eta^2[t]$) at least.

So I also think these claims can be made from a different perspective. That is, the LCM with a different curve may capture some of these individual changes. The curve could obviously be defined using the last 18 measurements, but an exponential

curve could be fitted with less parameters. Nevertheless, the model with the best fit for least parameters is an obvious choice. This, at least, is how I could deal with all the nonlinearity that seems to be present here. I would like to see LCM and the quadratic model as a comparison in their tables.

The second application titled "Stage-sequential growth mixture modeling with criminological panel data" is by Jost Reinecke, Maike Meyer, and Klaus Boers does exactly what this title suggests. However, it uses *General Growth Mixture Modeling* (GMM, from Muthén & Sheddon 1999) within a LCM framework to empirically distinguish between people. Expanding upon the prior work of Kim and Kim (2012) they consider three distinctive types of stage sequences: (1) stage-sequential (and linear) growth mixture models, (2) traditional piecewise GMM, and (3) discontinuous piecewise GMM and sequential process GMM. These three models are applied to a range of adolescence and young adulthood using data from the German panel study termed, *Crime in the modern City* (CrimoC, Boers et al., 2014). In the case of count variables a Poisson or negative binomial distributions (following the work of Hilbe, 2011, not Nagin 2005) can be considered which give a better model representation of the data. With the count data that criminologists seem to have, the Poisson model for measurement is used because it is more appropriate. That is, a regular regression model (but not evaluated) may still work, but the Poisson model that is used here as a measurement device because is sensitive to the use of a probability of an event. The zero-inflated Poisson (or ZIP; see Nagin 2005) model may even be a better choice because it essentially proposes that the reason for the zero counts (no criminal acts) is possibly different than the reasons for the rest of the counts (one, and so on). This can always be compared to the assumption of a continuous distribution of the LCM. And this all can be combined sequentially in a program like Mplus (Muthén & Muthén 2012).

This chapter is notable in a number of ways. First the author(s) use a three-part curve model, with knot points that are notable in terms of substance. This is a distinction that is worthwhile to make and it could be pursued further. I do not see this as quite as different as the typical LCM, so I would compare the fit of both of them. Second, they simultaneously use a measurement model based on a Poisson distribution for the scores. This is decidedly different and is most appropriate for data that comes in the form of counts. But their justification for the use in real data is not presented clearly. Third, they simultaneously use a mixture model to examine for the German Crime data. This use of multiple groups is based on the trajectory differences and they assume these cannot be accounted for otherwise. I would very much like to hear what Nagin (2005, p. 54) says about this part of the analysis. But in any case, any one of these three concerns would be a challenge to fit but they proceed as if this is all standard. This is not standard, and what they do here is quite amazing, partly because it can be done at all.

The differences between the current versions of Mplus (Muthén & Muthén 2012) and SAS PROC TRAJ (Nagin 2005) are important here. Currently, in Mplus, we can ask if any parameter is invariant over groups, and we do not need to define the group membership in advance. This can be in terms of any mean, regression, or covariance component. But in this same sense the analysis is entirely exploratory.

If we further assume that the factor loadings Ω[t], for at least t = 3, T, are different we can have different curves. This can be written with different means and variance terms so the entire placement within groups can differ. This is somewhat different than assuming different linear or polynomial coefficients for the same data. Much more could be said here (see Nagin 2005, p. 54) but Mplus 7 (now used by almost everyone here) seems much more flexible to me now. But I fully expect the debate about "groupings" will go on, and this is productive.

The third application by Mark Stemmler and Fredrich Lösel is titled, "Developmental pathways of externalizing behavior from preschool age to adolescence," and also uses general growth mixture modeling (GMM) with BIC this time to separate five categories of persons among their total sample size of $n = 541$. The goal of this study is to analyze the data of the Erlangen-Nuremberg Development and Prevention Study (ENDPS; Lösel et al., 2009) for the first time with regard to different trajectories for externalizing behavior. ENDPS is a normative sample and is a combined experimental and longitudinal study on antisocial child behavior covering a time period of nearly ten years. Social behavior was rated by multiple informants such as self, mothers, kindergarten educators, and school teachers. Using this longitudinal data, they seem to have found (1) the "*high chronics*" (2.4 %; $n = 13$), who are receiving the highest values for externalizing behavior from childhood on up to adolescence; (2) the "*low-chronics*" (58.8 %; $n = 317$) who are low on externalizing behavior throughout the years; (3) the "*high-reducers*" (7.9 %; $n = 43$) who start out high in childhood, but who reduce their externalizing behavior monotonically over time; the (4) "*late-starters-medium*" (8.7 %; $n = 47$); and the (5) "*medium-reducers*" (22.4 %; $n = 121$). The results stress the idea of a life course perspective, which enable the study of the natural history of the development of externalizing behavior, its onset, and termination.

In all, these authors give an excellent history of the GMM, and demonstrate how it has been used before in many criminological samples. They seem to show that most studies report between three and five groups (with a total range of two to seven groups), and they use the BIC. Most studies show the group of life-course persistent or chronic offenders, and one group that does not exhibit violent, aggressive, or delinquent behavior; in addition, there are existing groups of late onset or desisting. Jennings and Reingle (2012) claim that the number and shape of the groups depend on the nature of the sample (high risk versus normative sample), the life course captured, the length of the observation, and the geographical context. Among the author(s) conclusions, they postulate that further research should be based on multiple observations and across multi-informants (e.g., child/youth reports, parents and teacher report) to ensure the best results. Since this result requires expertise in criminology, we must leave it up to the reader to make sense of these trajectories.

The fourth application by Jang Schiltz is proposal for the potential extension of "the Nagin model" of multiple groups. This can be a quite useful technology because in this representation we do not have to think everyone has the same general nonlinear slope of their trajectory. The problem with Nagin's original formulation is that he only determined trajectories for the mean level and a quadratic slope, and less effort was put into the variance terms or other forms of the slope (see Nagin 2005,

p. 54). These changes are made and the basic model is extended here to include group differences in the slopes and the error terms.

Since we all believe that there will be substantial heterogeneity in real data—different change patterns for different groups—and the LCM will not be capable of dealing with these based on two means and covariances alone, it is clear that this model is more correct. This and other examples on the use of the mixture model is certainly a powerful latent variable modeling approach. But this latent variable model is not the only way to explore the groups—they can even be formed out of measured variables too (see Brandmaier et al. 2013).

The exploratory use of measured rather than latent variables is attractive on a number of counts. First, there are usually many extra ancillary variables that are measured and used as covariates for no particular reason other than they exist. As we will demonstrate, this typical usage can tell us something about their impact on mean differences or between group effects. But what we are interested in is putting them into the analysis is to see if they impact the variances and covariances also. Second, there are always extra ancillary variables that are measured and these could be selected for this exploration. That this is any mixture model is an exploration that is obvious to anyone who uses them and the selection of a group is complicated. So we do not try to handle all these assumptions at once but instead we refer to Nagin (2005) for details on this issue.

Our Cognition in the USA (CogUSA) Study

Our CogUSA study (see McArdle & Fisher 2015) was designed to do something different than those in this section—that is, the most notable feature of the design of this particular longitudinal study is the variation of age at the initial time, and the variation between time intervals for different waves of testing. As stated earlier in our last Assumption 6, this is a feature of many psychological measurements although it is hardly ever dealt with on a formal basis.

Our ability to measure similar constructs in an in-person *face-to-face* (FTF) interview and over the *telephone* (TEL) is not the key issue here, but it is important. In prior surveys (including the HRS; see Juster & Suzman, 1995; Heeringa, Berglund, & Khan 2011) the only human abilities measured over the phone (say, using the *Telephone Interview of Cognitive Status*; TICS; Fisher et al. 2013) were the very simplest ones (*Episodic Memory* and *Mental Status*; see McArdle, Fisher, & Kadlec 2007). It is not too surprising that these simple variables could be measured in the same way in either modality (FTF or TEL) and still retain MFIT (see McArdle 2010; McArdle & Nesselroade 2014).

But when we consider measuring something as important in aging research as *fluid intelligence* (*Gf*) in a survey, we remain perplexed (see Lachman & Spiro 2002). This variable needs to measure "reasoning in novel situations" and this is fairly hard to do. One of the ways this can be done in surveys is with indices that supposedly measure *numerical reasoning* (*NR*), a decided subset of all reasoning

and thinking, and the measure of *numerosity* (*NU*) from the HRS is a good indicator of this. Another way to consider *NR* this is to measure *Serial Seven's* (*S7*) from the HRS, because this takes some *NR* as well as holding specific but complex ideas in memory (see Blair 2006). Still another way to indicate *NR* is to measure something like *Number Series* (*NS*) because these are intended to be small puzzles in numerical form.

One adaptation is that we initially reasoned that people, especially older people, would not take all test items necessary for a reliable score on anything, so the items administered had to be cut down. In the case of both *Immediate Recall* (*IR*) and *Delayed Recall* (*DR*) and *Numeracy* (*NU*) and *Serial 7's* (*S7*) the work had already been done by the HRS staff. These were properly considered as short forms due to the required telephone constraints on time.

The final telephone definitions follow on Table 1. They were all administered over the telephone and this is a limitation because we do not really know what the respondent is doing. These include definitions of IR, and DR to measure a *general memory or general retrieval* (*Gr*) factor, and NU, and NS to measure a *general fluid* (*Gf*) factor at each time ([1] or [3]). We will see if the fit of this specific two factor model is different than a one *general intelligence* (*G*) factor, but we will examine the factor loadings. Clearly, McArdle et al., (2007) found the first two scales (*IR* and *DR*) to be highly correlated (r 0.80) and suggested they be added up and calculated as a single score termed *episodic memory* (*EM*) to distinguish it from another scale of cognitive measurement from the TICS, *mental status* (*MS*; $\{BC + S7 + NA + DA\}/4$), but the second factor here is much different. And we hope it is clear that several other cognitive measures obtained in CogUSA were not yet used here (see McArdle & Fisher 2015).

For common factors to retain their meaning over time, we required them to have "strict" invariance (Meredith 1993). In this case, this implies the factor loadings (Λ), unique variable intercepts (I), and unique variable variances (Ψ^2) are all assumed to be invariant over time (for each measure). We also brought all means differences to the factor score level. This is typically tested but it is clear that any differences or changes over time must go through the common factors or they are not worth using and summarizing at this level. This is basic or, indeed, fundamental to our definition of the latent variables. This does imply that the way we measure the common factors can change from time to time, but for now we assume they are identical at both occasions of measurement.

Many other researchers search for different forms of invariance (e.g., see Byrne, Shavelson, & Muthén 1989; Reise, Widaman, & Pugh, 1993; McArdle, Petway, & Hishinuma 2014), and now this is an evaluation of configural, metric, strong, or strict invariance constraints. We will not partake in this quest again here. This is primarily because we only want the number of factors (*K*) to be determined by what is comparable over time in measurement (as in McArdle & Cattell 1994; McArdle 2007) not by a lack of invariance. There is a prominent thought that the search for the type of invariance of a measure is crucial (see Byrne et al. 1989), but if this is not met then the number (or type) of common factors (can be) needs to be altered to meet this criterion. That is, the criterion of invariance should always be met before

Table 1 Selected Telephone Measures used in CogUSA (McArdle & Fisher, 2015)

All HRS/AHEAD cognitive measures were selected to satisfy the following considerations: (a) provide descriptive information on a comprehensive range of cognitive functions; (b) span all difficulty levels from competent cognitive functioning to cognitive impairment; (c) be sensitive to change over time; (d) be administrable in a survey environment with lay interviewers, over the telephone, in a short time; and (e) be valid and reliable (from the HRS documentation Report by Ofstedal, Fisher and Herzog. 2005; DR-006). As always, the IWER is asked a series of questions about the incorrect responses. In addition, several other clearly cognitive measures (BC, S7, RF, CESD) are obtained at both waves were not used in these analyses
IR = or immediate recall (IR)—One set of 20 stimulus word (from four lists) are read aloud, and the respondent (R) needs to restate these words (no credit is given for errors of any kind). The observed score is from 0 to 10. At W3 they are administered a different list of ten words (from the four lists)
DR = or delayed recall (DR)—after about 5 min (depending upon how long it took to do the eight CESD items), the R is asked if they recall any of the words from the IR. They are then asked to restate these words (no credit is given for errors of any kind). The observed score is from 0 to 10
NU = "Numeracy"—Since HS 2002, the R is asked to answer up to three numerical questions: (1) "Next I would like to ask you some questions which assess how people use numbers in everyday life. If the chance of getting a disease is 10 %, how many people out of 1,000 would be expected to get the disease?"(2) "If 5 people all have the winning numbers in the lottery and the prize is two million dollars, how much will each of them get?" (3) "Let's say you have $200 in a savings account. The account earns ten percent interest per year. How much would you have in the account at the end of two years?" The observed score is from 1 to 3
NS = Even though we wanted to, the Woodcock-Johnson "*Number Series*" items was far too long to be included in CogUSA so we cut it down from about 42 items to about 6 adaptive items. A modification of "which six" items was tried in each of the two occasions, *Wave 1* (W1) and *Wave 3* (W), but both testings supposedly yielded a W-ability estimate of NS. In the W1 testing the plan was to administer a first item of medium difficulty (for their level) and (0) if they got it incorrect an easier item about half way down the scale (based on the known difficulty of the WJ item) was presented, but (1) if the R got the item correct a harder item, about half way up the W-scale, was presented. All testing ending at six items and a WJ score was estimated from this pattern of responses. In the W3 testing s similar items were administered in a block adaptive fashion. The key idea here is to only administer six items, but the same three items are given first, spread out in difficulty, and the second set of three items are supposedly centered around the persons' ability level. In this case a W-score can be formed. Thus we assume, but do not test, MFIT

we evaluate the latent changes (as in McArdle & Cattell 1994). This is only our belief system, and we use this belief at all occasions, but we should point out that it is not one used by many others.

Methods

Available Data

The data to be analyzed are a small subset (4) of scales from recent tests of *Cognition in the USA* (CogUSA; see McArdle & Fisher 2015). These scales were chosen in

a way that would be consistent with the principles of MFIT but the result of the changes over two time points (W1 and W3 here) is unknown. Basically, we first try to establish MFIT over all ages and then look for changes in these scales over ages. We now only present eight scales in all to consider (four cross sections at two longitudinal occasions).

At each occasion, the people who took the HRS (for details, see Fisher et al. 2013) were asked to fill out the forms for all scales. Most specifically, they were asked each time to fill out a questionnaire about their own health and well-being and the full CES-D was included. We did not use sessions at Wave 2, we only use Sessions 1 and 3 primarily because this time-lag did not offer enough Age differences for Age changes to be picked up. We also do not include all items in analyses here, but we will only include eight scales from the full set (of many). We will state that 13/20 items (from the Center for Epidemiological Studies-Depression scale; CES-D) were previously analyzed by McArdle et al. (2014) who seemed to find MFIT in 13 items from the full set (of 20 items). But here there are several differences: (1) We deal with scales not items; (2) the confusion of the usual testing of MFIT was emphasized in McArdle et al. (2014); (3) in CogUSA the ages-at-testing at each occasion are substantially different.

The plots of Figs. 2a–d illustrate what we are trying to examine in the model. These are plots of the four manifest variables (*IR*, *DR*, and *NS*, *NU*) against the ages-at-testing for each person separately (i.e., joined by a line), and these illustrate lots of variability and only one kind of dependency among persons (that is, the people are largely the same ages when they are measured but the scores do change over age). They could change for a number of other reasons (such as errors of measurement or practice effects; see McArdle & Woodcock, 1997).

Models

Figure 3a is an elaboration of a latent curve model with Age differences as a double cross-sectional variable. The only variable used here is the *NS* measured at two occasions (1 and 3 for comparability) and the age-at-testing is also measured at each of these waves. The model here uses the two occasions in a double cross-sectional mode in an effort to capture the means and covariation of the *NS*-age relationship. That is

$$\begin{aligned} \mathrm{NS}[1]_n &= \beta_{01} + \beta_{11}\,(= fun\{\mathrm{Age}\,[1]\}_n) + e_{1n} \\ &\text{and} \\ \mathrm{NS}[3]_n &= \beta_{03} + \beta_{13}\,(= fun\{\mathrm{Age}\,[3]\}_n) + e_{3n} \end{aligned} \tag{10}$$

where some fixed function of age is used as a linear predictor (e.g., such as $fun\{\mathrm{AGE[t]}\} = (\mathrm{Age[t]} - 65)/10$—so the intercept is at age 65 and the difference in score is for each 10 years of Age). But using SEM we can also test whether the equations ($\beta_{01} = \beta_{03}$, $\beta_{11} = \beta_{13,}$ and the respective residual variances, $\psi 1 = \psi 3$) are

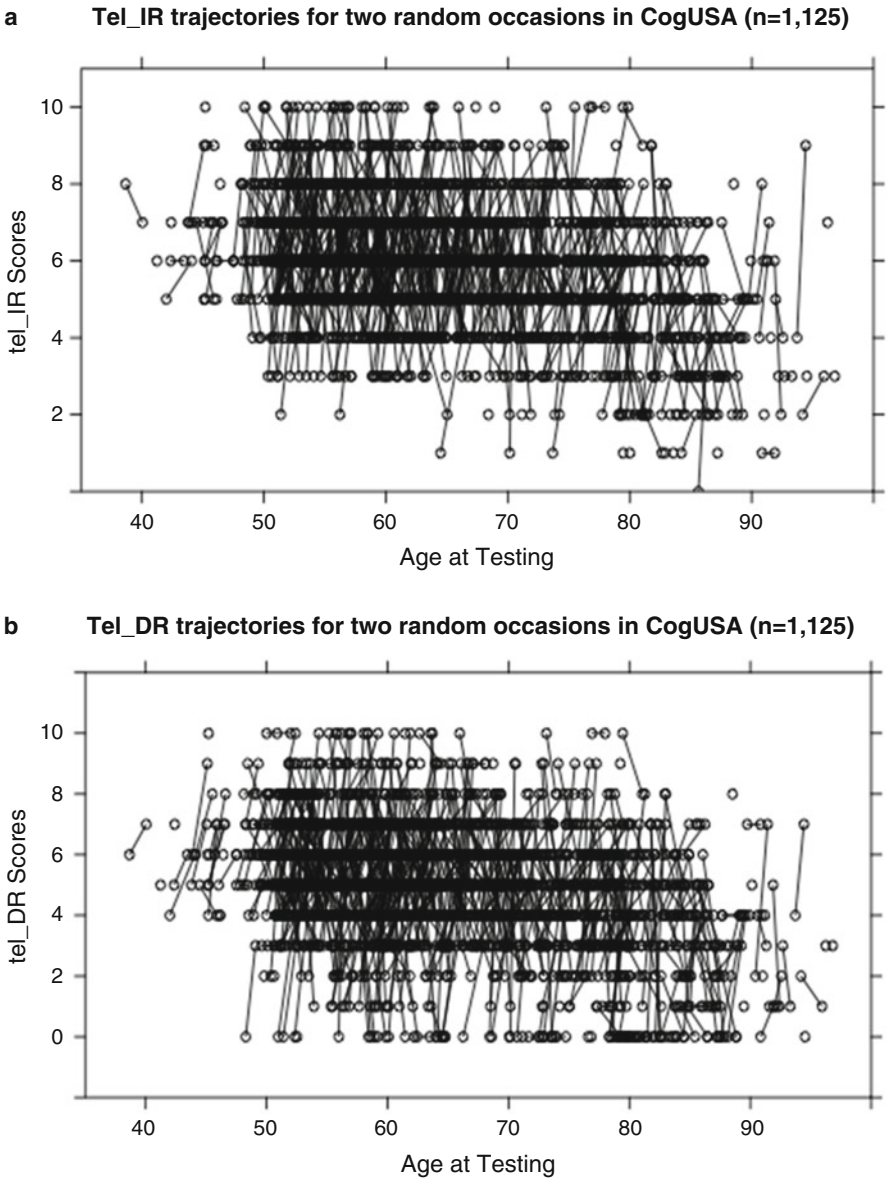


Fig. 2 (**a**) Immediate recall (IR). (**b**) Delayed recall (DR). (**c**) Number series (NS). (**d**) Numeracy (NU)

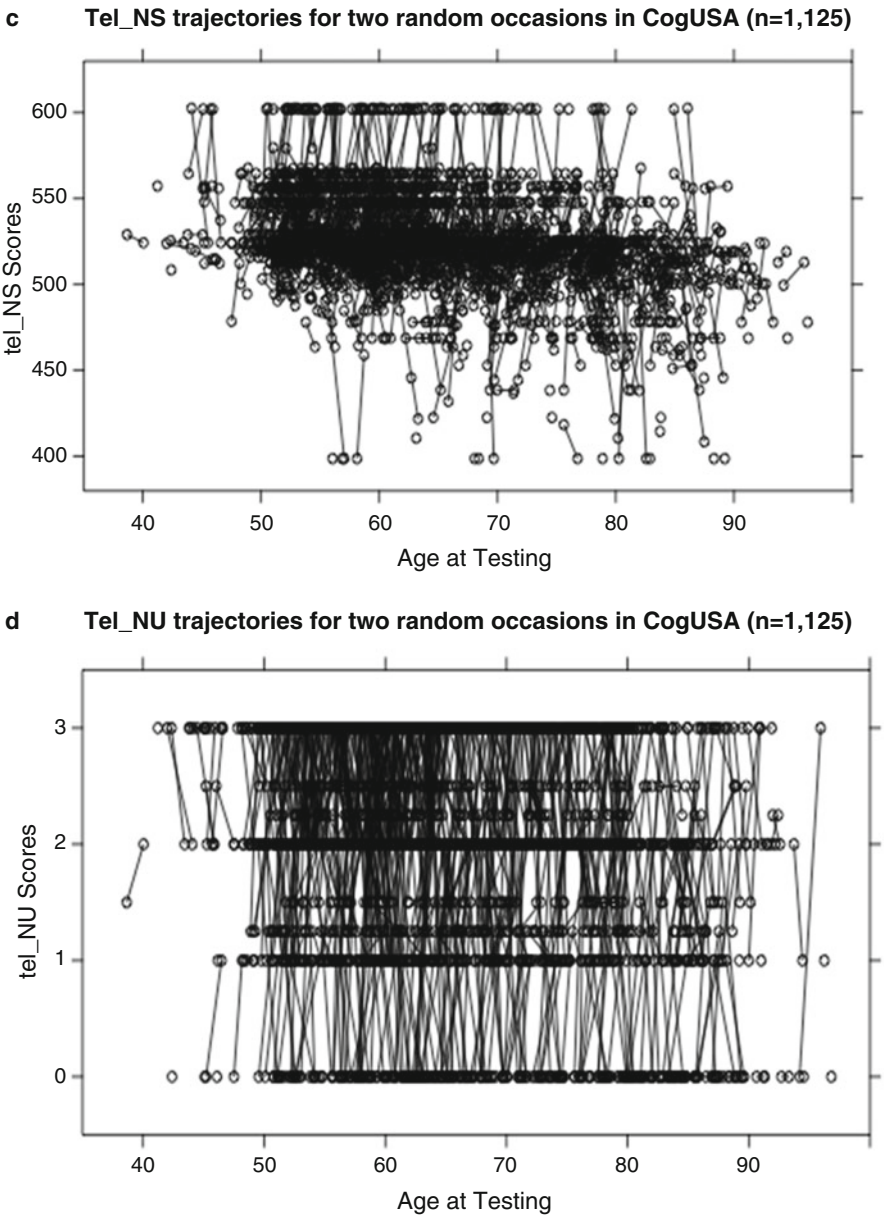


Fig. 2 (continued)

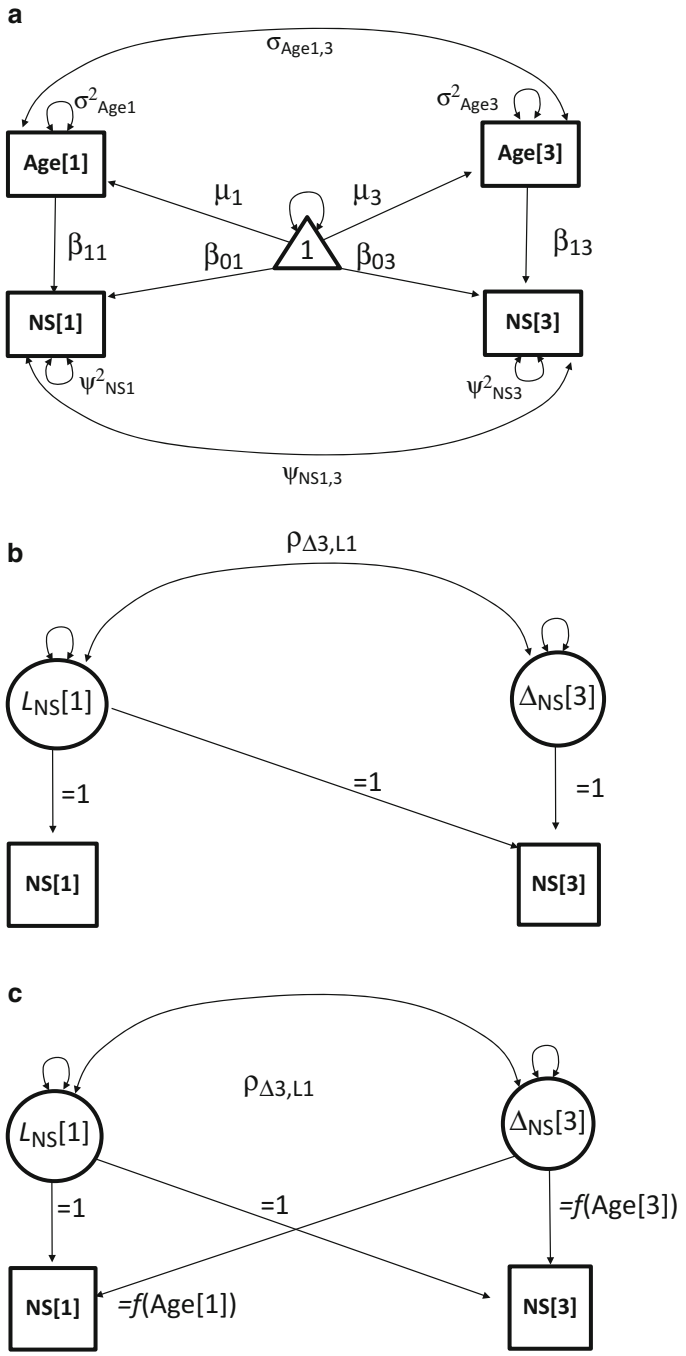


Fig. 3 (**a**) A path diagram of a one-variable model for multiple waves of measurement (W1 and W3) but as usual treated as a dual cross-section. (**b**) A path diagram of a one-variable LCM model for multiple waves of measurement (W1 and W3). (**c**) A path diagram of a one-variable LCM model for multiple waves of measurement (W1 and W3) with different ages of measurement

supposedly the same at each time of measurement. This could be useful because we may find it does not work the same way at Wave 1 and Wave 3, primarily because of the exposure at Wave 1. At very least, the prespecified *fun*{Age} becomes a testable hypothesis.

Figure 3b is a path diagram of a LCM in cases of only one change. The variable here is the *NS* measured at two occasions ([1] and [3] for comparability). But the model here is an effort to capture the mean and covariation of the *NS* test. We notice that this uses the leftover variation as the difference (or slope) and this simple representation can b credited to Joreskog (1974).

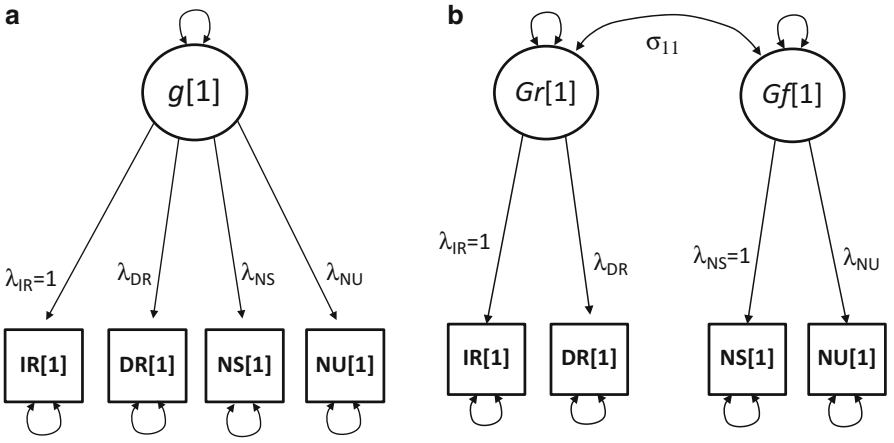
Figure 3c answers a different question about where we would add age variation to this model. Recall in CogUSA (Figs. 2, 3, 4, and 5) there is a lot of age variation at the beginning (Wave 1) and they are not measured over the same age interval over time. This variation in age was considered a random source of variation (and it was done on purpose) because we did not really know how to break up ages into groups. This is an expression of the work of the primary author of this paper (see McArdle & Woodcock, 1998). For these model to work, some predefined fixed function of age (e.g., $= f$(Age); it does not need to be linear, but it must be pre-specified) needs to be designated as a regression (or as a factor loading) that must be able to change over the individual case (because of the different ages-of-measurement). This precise feature of varying factor loadings can be used in many current computer programs (see Appendices 1 and 2 here for Mplus code). The concept of the individual loading was used by McArdle (1998, pp. 390–406) fitted together with the concept of individual likelihoods (primarily to check on individual fit). This examination of age-variation is an important concept here, but we would use this representation for any departure from the average timing that is measured (see LCM Assumption 6). This is the same concept that was subsequently used by Mehta and West (2000) and Mehta and Neale (2005) in their description of "definition" variables.

Adding a Latent Variable Measurement Model

Needless to say, these are common factor models where we assume a factor score for each person (f_n) is indicative of multiple measures at multiple occasions. This is an important addition and it can be done with SEM. Following McArdle (2007), every variable (m = 1 to M) we measure at each time (t = 1 to T) can be decomposed as

$$Y[t]_{m,n} = \lambda_m f[t]_n + u_m \tag{11}$$

into a common part (the time related common factor score f_n multiplied by a time invariant factor loading λ_m) and a unique part (the random or unique factor score u_m). We can think of the variable having an intercept of mean (ι_m) too, but this could just as well be a property of the unique factor score. This leads to a common factor model hypothesized for each time point.



Note:
Using only Time 1 latent variables *G* [1], and Time 1 observed variables IR[1], DR[1], NS[1], NU[1];

Note:
Using only Time 1 latent variables *Gm*[1], *Gf*[1] and Time 1 observed variables IR[1], DR[1], NS[1], NU[1];

Fig. 4 (**a**) A latent variable path diagram of the one-factor model at Wave 1. (**b**) A latent variable path diagram of the two-factor model at Wave 1

The specific models fitted to the one time point data are equivalent to many others, so we will not belabor the process. Needless to say, these are factor models where we assume a factor score for each person (f_n) is indicative of multiple measures at each occasions (t = 1 to T). This is presented in Fig. 4a, b for both one and two factors at one wave. We do notice that the one factor (*G*) also has several demographic influences, including scaled versions of age (and education, sex, and race). That is, in addition to the requirement that this factor account for the covariation of all the internal variables, this *G* must also account for all the covariation of all demographic influences with these measured variables. In our Fig. 4b, two common factors (of *Gf* and *Gr*) are expected, and these two factors are allowed to be correlated above and beyond the external (demographic) influences. This relaxation of the factor pattern is not the only way two common factors could be fit here, but it should fit better (see McArdle & Prescott, 1992).

In the next model we consider a single latent variable, perhaps termed a *G* factor for general intelligence. This is a very popular model for a number of good reasons (see McArdle 2012) and it can be fitted here to the four variables. In this context, the model makes the additional assumption that all four variables have a common part and a unique part. The common part is not necessarily the same for each score, and this size is as indexed by a factor loading (λ_m) or by the size of its' unique variance ($\psi_m{}^2$). The size is only relative here and this is made clear by the requirement that one of the factor loadings (or the common variance) needs to be fixed at some positive value (usually $\lambda_1 = 1$).

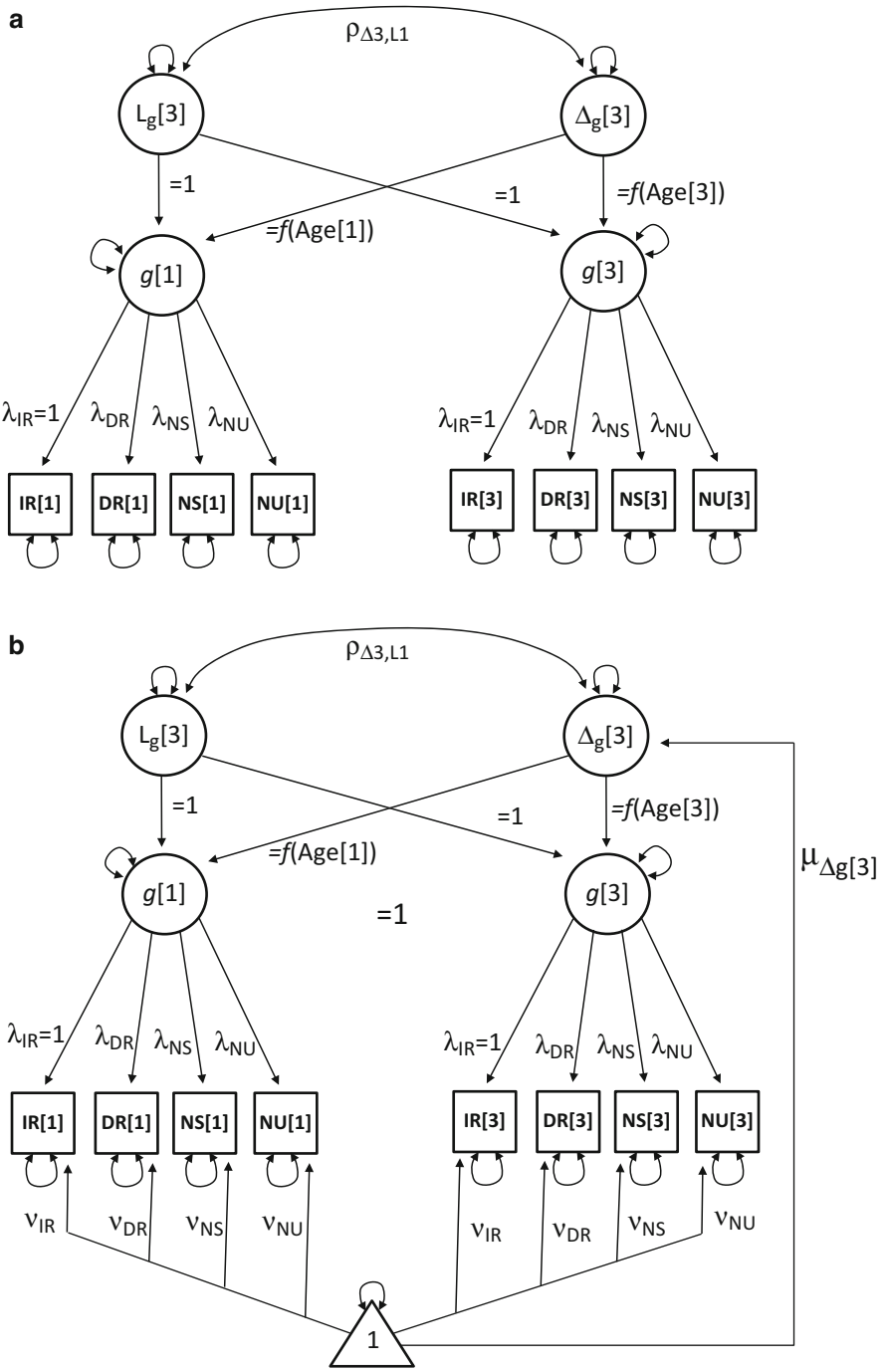


Fig. 5 (**a**) A latent variable path diagram of the invariant one-factor model for multiple waves of measurement (W1 and W3). (**b**) A full (means and covariances) latent variable path diagram of the invariant one-factor model for Wave 1 and Wave 3

As an alternative, in the next model we consider multiple latent variables termed *Gf,* for *general fluid reasoning*, and *Gr*, for a *general memory* or *general retrieval* function of memory. This multiple factor model is a very popular model for a number of good reasons (see McArdle 2012) and it can be tested (fitted) here to the four variables. This is not necessarily the best fitting model to these data (that is, other factor loadings can be estimated instead). The model requires that each factor loading have a fixed positive value (usually $\lambda_1 = 1$, and $\lambda_3 = 1$) but this is an arbitrary choice that can be altered but must be made by the investigator. Most critically, this is the same as the prior model if the correlation among the two separate factors is unity.

In this context, the model makes the additional assumption that all four variables have a common part and a unique part. The common part is not necessarily the same for each score, and this size is as indexed by a factor loading (λ_m) or by the size of its unique variance ($\psi_m{}^2$). The size is only relative here and this is made clear by the requirement that one of the factor loadings (or the common variance) needs to be fixed at some positive value (usually $\lambda_1 = 1$, and for the second factor, $\lambda_3 = 1$).

The same comparison of these two models can be examined over time and models for this type of data are drawn in Fig. 5a, b. Here the factor or factors have to do two related things: (1) Define the internal features of the covariation of the measures within a time point, and (2) account for the changes over time in the measures. Since we want the same factor at time 1 and time 2 (or W3 here), and since we define the factors by their factor loadings, we do force the factor loadings to be exactly the same over time. Although it is not necessary for this problem, to simplify our presentation here, we assign the unique variances to be the same over time as well.

Of most importance here is change over age, and the common factor part can further be decomposed as

$$f[t]_n = f_{0n} + f_{1n}\Omega[t]_n + e_n \tag{12}$$

where f_{0n} is the unobserved level or intercept of the factor score, f_{1n} is the unobserved slope of the changes due to a one-unit shift in the $\Omega[t]_n$, and e_n is the random noise or disturbance that is thought to be randomly distributed around the predicted value of the first two parts. In this way, the factor score can change and this creates change in the observed variable even with an invariant measurement model.

The differences between this and other formulations of the more standard *LCM* (Meredith & Tisak 1990; McArdle 1986) are that (1) this is a *curve of factors* model (CUFFS; after McArdle, 1988) and (2) here we explicitly assume the assignment of a factor loading that varies across the individual (McArdle & Hamagami 1996, pp. 106–112; especially p. 108). Of course, individual fitting of likelihoods is a common feature of many fitting functions now (but see McArdle 1998, pp. 390–406), so we use the program M+ here. The consistency assumption for the individual to look like the group is used to form the basic test statistics—this use of an individually measured score as a model parameter is sometimes call "adding definition variables" (from Mx manual; Neale et al., 1993). Indeed, this kind of

raw data procedure was only available in Mx in the past, and it was based on the statistical concepts of *unbalanced pedigree analysis* (from Lange et al. 1976).

If aging impacts the latent score alone we would think that it impacts both the levels and slopes in some consistent fashion. To this we add that there can be age differences at Wave 1 and Wave 3, and these are summarized for each person as

$$\begin{aligned} f_{0n} &= \beta_{01} + \beta_{11}\,(= fun\{\text{Age}\,[1]\}_n) + e_{0n} \\ &\text{and} \\ f_{1n} &= \beta_{03} + \beta_{13}\,(= fun\{\text{Age}\,[3]\}_n) + e_{1n} \end{aligned} \tag{13}$$

so the f_0 is a level over both occasions, and f_1 is a slope that is at two particular time points determined by the age of measurement. The terms are indicative of a level (or when the $fun\{\text{Age}\} = 0$; so at 65 here) and a slope (for each unit—or decade of age—of $fun\{\text{Age}\}$) of the prespecified age function. That at is, each person's unique contribution to the two ages is built up in this way. Each person has a level and a slope score and under the assumption that it is the same information about age changes is present in each variable. We can look further at this function. In other words, we have essentially taken the Age model to the latent variable level.

Figure 5a is a path diagram of this one common factor model and Fig. 5b is a more compete version (including means and unique covariances). We fit the latter one here.

Figure 6a extends this logic to having two common factors at each measurement occasion, and Fig. 6b is a full mean and covariance path diagram of this extension. Here the variable slopes and levels are all correlated, and the factor of levels does not assume a mean difference, due to lack of identification, so we do not add one.

In the final model (Fig. 6b) we use a two factor solution, but we also include: (1) the means in the diagram (as the regression from a constant triangle), and (2) the covariance of any unique features of the data ($\psi^2[1,3]$). This is simply a more complete picture of the model we will fit.

The same principles hold when we move to multiple occasions of data. The common factors are supposedly the same, but the age changes in the factor scores is examined. Here the models of 5a and 6a will be compared.

The key thing we will note about CogUSA is the staggered time lag of this longitudinal study. This is unusual for a longitudinal study (see McArdle & Woodcock, 1997), but we put in time lag as a variable because we wanted to study. That, in most cases of experimental design we vary all the things that are important to us and leave the rest as fixed quantities.

Results

The summary statistics appear in Table 2 for $n = 1{,}125$ people who supposedly took all four scales at both Wave 1 and Wave 3. These are *full information maximum likelihood* (FIML) estimates because only about 98 % participated at all times. In

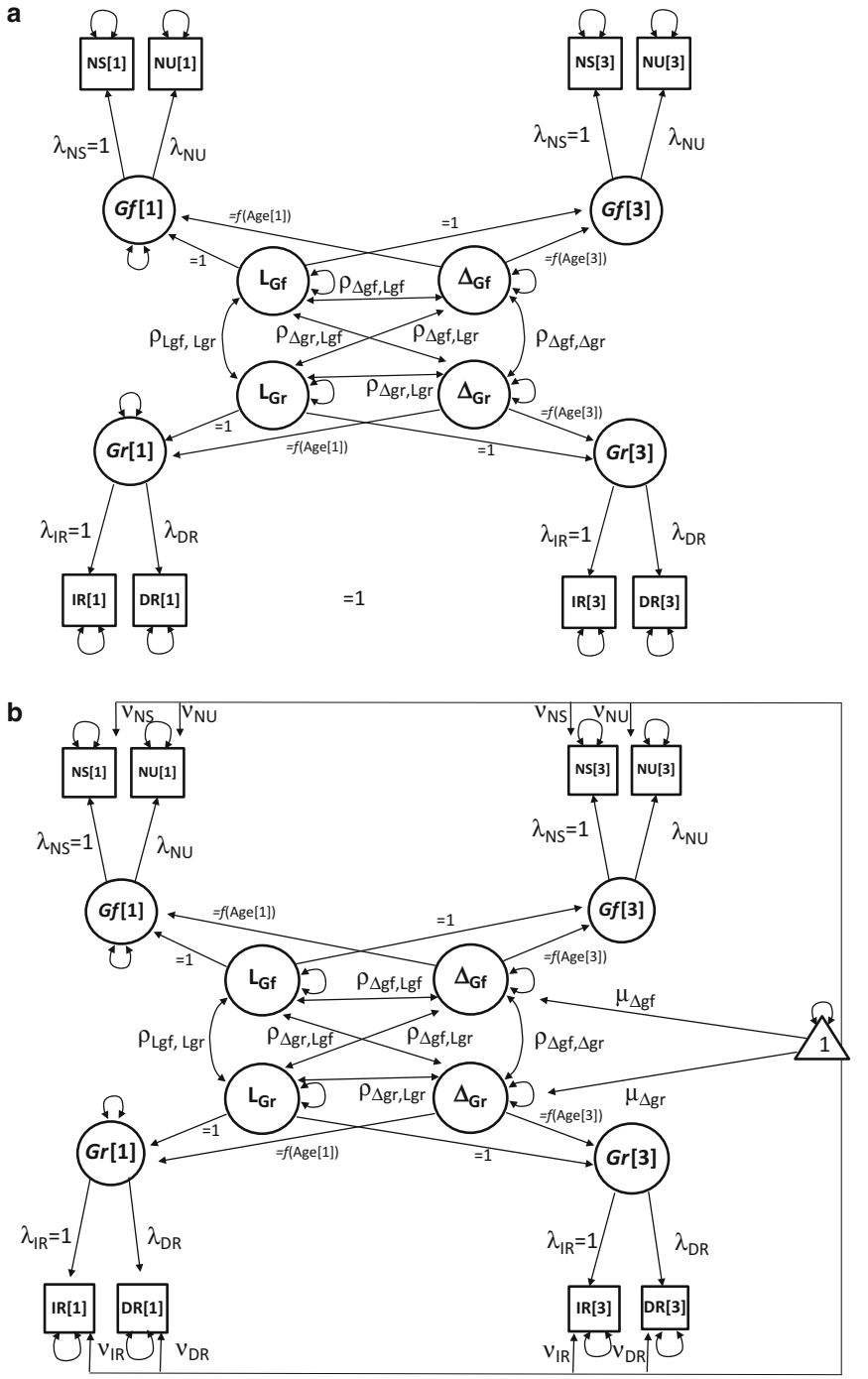


Fig. 6 (**a**) A latent variable path diagram of the two-factor invariant model for Wave 1 and Wave 3. (**b**) A full (means and covariances) latent variable path diagram of the two-factor invariant model for Wave 1 and Wave 3

Table 2 Summary statistics for $n = 1125$ participants who were all considered on four scores at both Waves 1 and 3 (and using FIML)

(2a) Means (& Variances)

	Means				
	NS[1]*	NS[3]*	NU[1]	NU[3]	IR[1]
Means	2.383	2.685	2.029	2.059	5.951
(Variances)	(7.790)	(9.444)	(0.969)	(0.912)	(2.807)

	Means				
	IR[3]	DR[1]	DR[3]	AGE[1]*	AGE[3]*
Means	6.069	4.865	5.152	-0.060	0.061
(Variances)	(2.647)	(4.099)	(3.941)	(1.103)	(1.107)

(2b) Correlations

	NS[1]*	NS[3]*	NU[1]	NU[3]	IR[1]
NS[1]	1.000				
NS[3]	0.519	1.000			
NU[1]	0.456	0.433	1.000		
NU[3]	0.383	0.502	0.504	1.000	
IR[1]	0.338	0.330	0.268	0.258	1.000
IR[3]	0.218	0.295	0.177	0.245	0.425
DR[1]	0.301	0.282	0.218	0.215	0.762
DR[3]	0.244	0.344	0.199	0.265	0.424
AGE[1]	-0.304	-0.330	-0.214	-0.192	-0.346
AGE[3]	-0.304	-0.332	-0.215	-0.193	-0.344

	IR[3]	DR[1]	DR[3]	AGE[1]	AGE[3]
IR[3]	1.000				
DR[1]	0.389	1.000			
DR[3]	0.759	0.443	1.000		
AGE[1]	-0.352	-0.310	-0.366	1.000	
AGE]3]	-0.354	-0.308	-0.366	0.999	1.000

Notes: W1_NS and W3_NS are scaled by mean = 500 and SD = 10
W1_Age and W3_Age are scaled by mean = 65 and SD = 10
To get back to the original scaling of each score we can simply multiply by the SD and add the mean

using FIML we basically assume that there is nothing special about those who did not participate again, and we use their time 1 data assuming they also follow the same general pattern as we observe in those that did come back. But, for example, NS is listed with a mean of 2.38 and a variance of 7.79 at Wave 1, and this can be recast (by the usual W Rasch-scale transformation, following McArdle & Woodcock, 1997; Here this is a transformation that basically raises the score to a power of about 9, and then adds 500) into a raw W-scale score of 524, while the same (or similar) test is listed at a mean of 2.83 with a variance of 9.46 at Wave 3, and this is a raw W-score of 528. It is thought (by many others) that the W-score will

have a linear relation with other scales while the raw score will not. The correlation (in Table 2b) is r∼0.520. Thus, the W-scores go up over time in a pattern that is related to time 1 (those that are high to begin with seem to get high scores here). The high scores can be accounted for by the ceiling of 600 on both test forms (this is the highest we would go) and it appears (as we can see in Fig. 2a) people seemed to get this score with unusually high frequency.

The resulting age-based model #1 is based on $n = 1{,}125$ who answered both scales, starting ages with a mean of 68.83 (with a variance of 106.91) and a time lag of 1.21 years (with variance 0.23). These are first fitted to *Number Series* (*NS*) at two waves (Wave 1 and Wave 3) although any one of the four scales could be used. The results are highlighted in the first column of Table 3 and all of Table 4. In addition the computer script used to assess these variables is in Appendix 1.

The numerical results of model #1 shows *NS* has a mean at age 65 of −1.36 and a slope per decade of age of 3.79. These are both significantly different from zero given their respective z-values, so we can talk about the W-scores of 484 at age 65 and with a positive increase of +3.87 (or +38.7 in W units) points per decade from that point onwards (and backwards). This should usually be contrasted with model #0 where no slope is assumed, but the equal variance assumption of the residual was assumed (in fact an equal and fixed variance assumption had to be used, so this model has only two estimated parameters; the mean and variance of the level; see McArdle, 1998). The extra parameters estimated (from the 2 in #0) are the slope mean, the slope variance, and the covariance of levels and slopes. The fact that all variables have such large variance estimates (in #1) and that increasing scores go with increasing ages (this is positive) is a surprise. The fact that model #0 has a much larger BIC is also a question that requires an answer. Perhaps the age changes are too small to count here, but the fact that they are positive is a definite difference from prior results (see McArdle et al. 2007).

The results for the one factor *G* model of behavior is listed in the next two models of Table 3 (#2 and #3). In model #3 we fit a level and slope model with one common factor (as in Fig. 4b). In model #2 we fix this changing score assumption and did not fit a slope to the *G* factor and lost substantially in fit (on $df = 4$). The single factor with factor loadings set $\lambda = 1$ for *NS*, but the factor loading is estimated as 0.77 for *NU*, 2.25 for *IR*, and 1.91 for *DR* seem to fit these data very well (with $df = 3$). The three extra parameters estimated (from the 10 in #2) are the slope mean, the slope variance, and the covariance of levels and slopes. The invariant loadings do not add anything to the misfit here. But the mean intercept of this age slope factor is 2.68 (indicating a raw W-score of 527 at age 65), and the mean of this slope factor is −0.24, and it is significant, indicating a −0.24 downhill slide in this factor for every decade of age that is increased.

This is a decidedly different result. It is not at all what the observed data seem to show (for any subscale) but this does not take the age-of-measurement into account, it uses only the common (not unique) variance, and it is a closer to what we expected. In fact, it suggests that (1) the common factor is *episodic memory* (because the loadings of 2.25 for *IR* dominates the factor), and (2) there is some decline over age. This one factor model is listed as #3 and this is given more completely in Table 5

Table 3 Numerical results for selected univariate models of NS (fitted to the data of Table 2, with the M+ computer scripts of Appendices 1 and 2)

Univariate parameters	Model 0: zero age effects at both Waves	Model 1: free age effects at both Waves	Model 2: equal age effects at both Waves	Model 3: using only NS with level only	Model 4: using only NS with level and slope
Intercept (at age 65) for NS (z)	=0	=0	=0	2.53 (33.0)	−1.36 (0.75)
Intercept (at age 65) for NS[1] (z)	2.39 (38.7)	2.35 (29.4)	2.76 (33.2)	=0	=0
Intercept (at age 65) for NS[3] (z)	2.71 (29.5)	2.76 (31.9)	2.76 (33.2)	=0	=0
Mean slope (per decade of age) for NS (z)	=0	=0	=0	=0	3.79 (3.8)
Mean slope (per decade of age) for NS[1] (z)	=0	−0.81 (10.7)	−0.89 (13.2)	=0	=0
Mean slope (per decade of age) for NS[3] (z)	=0	−0.98 (11.8)	−0.89 (13.2)	=0	=0
−ll/free p	−5369/5	−5469/7	−5469/4	−230243/2 with ψ^2=.01 fixed	−8539/5
BIC	10737	10589	10587	4602300	17113

Note: The complete baseline model of no level and no changes with all four measured variables has $\chi^2/df = 3034/28$ with $-2ll = -14327$

Table 4 Numerical results for selected multivariate models (fitted to the data of Table 2, with the M+ computer scripts of Appendices 3 and 4)

Multivariate parameters	Model 0: using only G with age level only at each Wave	Model 1: using only G with equal age level and slope at each Wave	Model 2: one common factor of G with level only	Model 3: one common factor of G with levels and slopes	Model 4: two common factors of *GF* and *Gr* with both levels and slopes
Mean level (at age 65) for G[1] (z)	2.66 (44.7)	2.67 (45.4)	=0	=0	=0
Mean level (at age 65) for G[2] (z)	2.66 (44.7)	2.67 (45.4)	=0	=0	=0
NS	=0	=0	=0	=0	=0
NU	=0	=0	=0	=0	=0
IR	=0	=0	=0	=0	=0
DR	=0	=0	=0	=0	=0
G (z)	=0	=0	2.55 (42.0)	2.68 (43.0)	=0
Gf (z)	=0	=0	=0	=0	2.75 (46.0)
Gr (z)	=0	=0	=0	=0	6.04 (155.4)
Mean slope (per decade of age) for G[1] (z)	=0 (=0)	−0.26 (15.0)	=0	=0	=0
Mean slope (per decade of age) for G[3] (z)	=0 (=0)	−0.26 (15.0)	=0	=0	=0
NS	=0	=0	=0	=0	=0
NU	=0	=0	=0	=0	=0
IR	=0	=0	=0	=0	=0
DR	=0	=0	=0	=0	=0
G (z)	=0	=0	=0	−0.25 (13.0)	=0
Gf (z)	=0	=0	=0	=0	−0.28 (9.3)
Gr (z)	=0	=0	=0	=0	−0.50 (12.3)
Factor loading for NS	=1.00	=1.00	=1.00	=1.00	=1.00
NU	0.75	0.75	0.77	0.75	0.74
IR	2.26	2.25	2.33	2.25	=1.00
DR	1.91	1.91	1.97	1.91	0.84
−ll/free p	−15288//10	−15288/11	−18287/10	−15973/13	−15560/22
BIC	32263	32046	38060	32038	31275

Table 5 Parameter estimates for the single-variable (number series) model with age changes

TESTS OF MODEL FIT

Loglikelihood

H0 Value	-8539.069

Information Criteria

Number of Free Parameters	5
Akaike (AIC)	17088.137
Bayesian (BIC)	17113.265
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	17097.384

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
L	BY				
	NS[1]	1.000	0.000	999.000	999.000
	NS[3]	1.000	0.000	999.000	999.000
S	BY				
	NS[1]	999.000	0.000	999.000	999.000
	NS[3]	999.000	0.000	999.000	999.000
L	WITH				
	S	-1618.371	76.563	-21.138	0.000
Means					
	L	-1.362	1.827	-0.746	0.456
	S	3.789	0.977	3.879	0.000
Intercepts					
	NS[1]	0.000	0.000	999.000	999.000
	NS[3]	0.000	0.000	999.000	999.000
Variances					
	L	3694.429	156.884	23.549	0.000
	S	1055.135	44.750	23.579	0.000
Residual Variances					
	NS[1]	0.000	0.000	999.000	999.000
	NS[3]	0.000	0.000	999.000	999.000

Note: See text for details. The use of 999.000 is used for empty locations and when the individual varies. This is Mplus 7.11 output from the Appendix 2 program

and the Mplus computer script that was used is in Appendix 2. We can see the BIC is a lot larger for the no-changes model (#2), but we do not assume that Age changes modeled in this way are important. This is only one latent variable after all.

A specially selected (and highly restricted) two-factor model alternative is listed as #4. This is a more complex two-factor model that allows for more covariation among the measures but also requires "strict" invariance of the loadings and unique variances over time. Similarly, all variable intercepts are set to zero here, so the mean changes have to go through the common factors (as in McArdle & Nesselroade 2014). In this approach the model has two common factors with $\lambda = 1$ the required fixed loading for *NS* and *IR* and estimated loadings of 0.74 (for *NU*) and 0.84 (for *DR*). This model has 22 parameters, largely due to the extra common factor covariances, and these extra parameters are penalized heavily by the BIC, including the level intercepts of 2.82 (or W = 528) and 6.04 (for *IR*). This still seems to have the smallest BIC value, so it could be chosen on this basis.

It appears that *Gr* as measured by *IR* and *DR* declines the most over age, with −0.49 per age decade. The function termed *Gf*, indicated by *NS* and *NU*, decline significantly but only at −0.27 per decade. In contrast to the one-factor version, the BIC for this model is smaller, and needless to say, this is far less decline than we initially expected from a normal aging population, so maybe we do not have the right factor yet. This has the smallest BIC of all used, so it could be chosen as the best model for the data. But this BIC (about 31,275) is not much smaller than the prior BIC (about 32,038), and the two-factor model has a lot of extra parameters so this substantive model is not yet considered a large improvement.

Discussion

The final model chosen did not have the smallest BIC but it seemed to fit the data the best. So there is much more to be done here. We fully realize the two-factor model did not fit as well as we would have liked. We would have liked to separate apart the aspects of G_F and G_R but this proved difficult with only 4 measures. But this is a complicated choice (made here only by BIC) and because the number of measurements is small and we want to say a lot about aging. That is, we had a difficult distinction because we were only working with four variables. But this SEM is a useful starting point even if it only deals with LCM Assumption 6.

We tried to use the standard HRS procedures (e.g., Genesys surveys) to contact households with some persons in the HRS age range (over 50). We succeeded in reaching over 3,000 people, but not everyone agreed to test further. We can only suggest the reader look carefully at Heeringa et al. (2011) and McArdle and Fisher (2015) for details. At Wave 1 we actually had measured over 1,500 people we have measured age and other demographics (like respondent education, sex, minority status, health, dyad, nursing home, currently employed), as well as administration of a telephone versions of standard cognitive tests (the TICS; defined here as *BC*, *S7*, *IR*, *DR*), as well as some additional tests HRS (*NU*) and some new telephone

Table 6 Parameter estimates for the current invariant measurement model with two common factor of changes

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Loglikelihood

H0 Value	-15973.231

Information Criteria

Number of Free Parameters	13
Akaike (AIC)	31972.461
Bayesian (BIC)	32037.793
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	31996.502

MODEL RESULTS

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
G[1]	BY				
NS[1]		1.000	0.000	999.000	999.000
NU[1]		0.748	0.017	43.113	0.000
IR[1]		2.250	0.048	46.838	0.000
DR[1]		1.907	0.042	45.848	0.000
G[3]	BY				
NS[3]		1.000	0.000	999.000	999.000
NU[3]		0.748	0.017	43.113	0.000
IR[3]		2.250	0.048	46.838	0.000
DR[3]		1.907	0.042	45.848	0.000
L_G	BY				
G[1]		1.000	0.000	999.000	999.000
G[3]		1.000	0.000	999.000	999.000
G[1]		1.000	0.000	999.000	999.000
G[3]		1.000	0.000	999.000	999.000
S_G	BY				
G[1]		999.000	0.000	999.000	999.000
G[3]		999.000	0.000	999.000	999.000
G[1]		999.000	0.000	999.000	999.000
G[3]		999.000	0.000	999.000	999.000
L_G	WITH				
S_G		0.026	0.008	3.283	0.001
Means					
L_G		2.673	0.059	45.338	0.000
S_G		-0.245	0.019	-13.001	0.000

(continued)

(continued)

Intercepts				
NS[1]	0.000	0.000	999.000	999.000
NS[3]	0.000	0.000	999.000	999.000
NU[1]	0.000	0.000	999.000	999.000
NU[3]	0.000	0.000	999.000	999.000
IR[1]	0.000	0.000	999.000	999.000
IR[3]	0.000	0.000	999.000	999.000
DR[1]	0.000	0.000	999.000	999.000
DR[3]	0.000	0.000	999.000	999.000
Variances				
L_G	0.151	0.022	6.929	0.000
S_G	0.030	0.015	1.960	0.050
Residual Variances				
NS[1]	7.509	0.228	32.890	0.000
NS[3]	7.509	0.228	32.890	0.000
NU[1]	0.872	0.028	31.166	0.000
NU[3]	0.872	0.028	31.166	0.000
IR[1]	0.378	0.045	8.469	0.000
IR[3]	0.378	0.045	8.469	0.000
DR[1]	1.556	0.055	28.213	0.000
DR[3]	1.556	0.055	28.213	0.000
G[1]	0.228	0.017	13.260	0.000
G[3]	0.228	0.017	13.260	0.000

Note: This is Mplus 7.11 program output for the input in Appendix 4

administered WJ-based adaptive tests (*NS*, *RF*) and the depression scale (*CES-D*). Our main goal here was to see if we could measure the same constructs as before, but the time over the telephone, and we basically found we could. But we are also involved in a number of selection issues (see McArdle 2013; Heeringa et al. 2011), and a small sample of 200 HRS respondents who were at the top and the bottom of the HRS Cognition scores (in 2008) were also re-measured on our instruments.

Perhaps with more measured variables we can also make finer distinctions among multiple factors. Or perhaps we can take into account the non-normality of the data (see Fig. 2a–d). In this same sense, we were evaluating only part of the measured scale and not all of it. For another example, take our previous analytic work in McArdle et al. (2015). In this research we evaluated 13 items from the CES-D. If we had been evaluating the CES-D for use in these grades (9th and 10th) we would use 20 items as listed in the typical CES-D. But we were most interested in evaluation the concept of depression, the latent factor that it represented, and we thought we could test this idea with only four items. But we looked at 13 items in this chapter only, mainly because we were trying to indicate one factor only.

Indeed an expansion or delineation of the purported common factors can be achieved with measured variables such as *S7* (a telephone administration of HRS Serial Sevens) and *RF* (a telephone administration of WJ-Retrieval Fluency). Each of these could expand the factor space in important ways and may lead to some stability. We can also add several occasions of measurement, before and after Wave 3. This use of Wave 2 was a few weeks later than Wave 1 but we gave

them a full battery of tests, about 1,200 people for 3 h in the home, including the WJ-R, and the WASI, plus some *personality scales* (e.g., *BFI*, *NCS*) and some *dispositional measures* (*RISK*). This was mainly included so we could verify the telephone measures against telephone adapted tests and several were administered here. There are many measures here and they can be useful too.

At Wave 3 we went back to the original telephone forms after an average of about 1.21 years (with 0.21 as a standard deviation). This was supposedly different for people of different ages in some time lag but other conditions need to be stated up-front as well. For example, Rodgers et al. (2003) and McArdle and Fisher (2015) make it very clear that there we naturally had eight groups of respondents by time-lag (this was not designed) and it was hard to do a second telephone test in a short time if the people had not completed the *FTF* of Wave 2. This confound does not apply to any people selected but it is there. We started this testing by verifying the Birth Date of the person being tested (the interviewee, or IV). We repeatedly tested the original people again on the same battery used in Wave 1 with, as stated, about a year and a half delay (e.g., Wave 3).

During the last 5 years we have measured the same people again as part of a second 5-year study. During these times (2009–2014) we were mainly interested in the difference (if there are any) between Telephone testing and Internet testing. This issue will not be raised or resolved here. There are currently (as of 2015) no plans by us to re-measure these same people again but we have let the CogECON and HRS teams at the University of Michigan contact them. In any case, our experiment is complex, as are our multiple assumptions, but this is the basic model of aging and invariance we will use in further analyses of these data, so comments are welcome.

Appendix 1: An example of an M+ 7.11 Computer Program for Two Time Point Data with Age Changes

```
TITLE: EX_MOD1 -- Dynamic Impact of constraints NS at
Two Times
Using Age at Testing as a cross-section
Run of NS with No Slope ALL CogUSA scores (McArdle,
2014)
DATA: FILE = CogUSA_Repeated3.dat;
!LISTWISE=ON;
VARIABLE: NAMES =
id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
```

```
W1_IR W2_IR W3_IR
W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC
W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU
W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF
;
USEVAR = W1_Age W3_Age W1_NS W3_NS;
MISSING=.;
DEFINE: W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
W1_age = (W1_age - 50)/10;
W3_age = (W3_age - 50)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
W1_NS on W1_age (B1);
W3_NS On W3_age (B3);
!equating error variances
! W1_NS W3_NS (V_U);
OUTPUT: SAMPSTAT STANDARDIZED;
```

Appendix 2: An example of an M+ 7.11 Computer Program for Two Time Point Data with Age Changes as a Loading Constraint

```
TITLE: EX_Mod: Table 3.2 -- Dynamic Impact of
constraints NS by Two Times
Run of All TEL Time CogUSA scores (McArdle, 2014)
DATA: FILE = CogUSA_Repeated3.dat;
!LISTWISE=ON;
VARIABLE: NAMES =
id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
W1_IR W2_IR W3_IR
```

```
W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC
W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU
W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF;
CONSTRAINT = W1_Age W3_Age;
USEVAR = W1_NS W3_NS;
MISSING=.;
DEFINE: W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
!frst get a level and a slope for NS;
l BY W1_NS@1 W3_NS@1;
s BY W1_NS * (LNS1)
W3_NS * (LNS3);
[l s];
l s;
l WITH s;
!eliminating original variables
[W1_NS@0 W3_NS@0];
W1_NS@0 W3_NS@0;
!equating error variances
! W1_NS W3_NS (V_U);
MODEL CONSTRAINT: ! To get at individual loadings;
LNS1 = (W1_age - 50)/10;
LNS3 = (W3_age - 50)/10;
OUTPUT: SAMPSTAT STANDARDIZED;
```

Appendix 3: M+ 7.11 Computer Program for Four Variables at Two Time Points of Data based on a Model of One Common Factor with Age Changes

```
EX_MOD: Table 4.2 -- Multivariate Dynamic Impact of
Equal Age
```

```
For four variables NS, NU, IR,DR, with and MFIT for G
(McArdle, 2015)
DATA: FILE = CogUSA_Repeated3.dat; !LISTWISE=ON;
VARIABLE: NAMES = id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
W1_IR W2_IR W3_IR W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF
;
USEVAR = W1_Age W3_Age
W1_NS W3_NS W1_NU W3_NU
W1_IR W3_IR W1_DR W3_DR
;
MISSING=.;
DEFINE:
W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
W1_age = (W1_age - 65)/10;
W3_age = (W3_age - 65)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
W1_G BY W1_NS (L_NS);
W1_G BY W1_NU (L_NU);
W1_G BY W1_IR (L_IR);
w1_G BY W1_DR (L_DR);
W3_G BY W3_NS (L_NS);
W3_G BY W3_NU (L_NU);
W3_G BY W3_IR (L_IR);
W3_G BY W3_DR (L_DR);
W1_G On W1_Age (b11); [W1_G] (b01);
W3_G ON W3_Age (b13); [W3_G] (b03);
!Equal uniquenesses at the factor level
W1_G W3_G (U2_G);
!Equal Uniqueness at the Variable Level
W1_NS W3_NS (U_NS); W1_NU W3_NU (U_NU):
```

```
W1_IR W3_IR (U_IR); W1_DR W3_DR (U_DR);
!original means of variables not used
[W1_NS@0 W3_NS@0 W1_NU@0 W3_NU@0 W1_IR@0 W3_IR@0
W1_DR@0 W3_DR@0];
OUTPUT: SAMPSTAT STANDARDIZED;
```

Appendix 4: M+ 7.11 Computer Program for Four Variables at Two Time Points of Data Based on a Model of Two Common Factors with Age Changes as a Loading Constraint

```
TITLE: EX_MOD: TABLE 4.4 -- Dynamic Impact of
constraints on
four variables NS,NU, IR,DR, with am MFIT-G
Run of 4 TEL Time CogUSA scores (McArdle, 2015)
DATA: FILE = CogUSA_Repeated3.dat;
!LISTWISE=ON;
VARIABLE: NAMES =
id
w1_age w2_age w3_age
educ female black
ldays12 ldays23 ldays13
W1_IR W2_IR W3_IR
W1_DR W2_DR W3_DR
W1_BC W2_BC W3_BC
W1_S7 W2_S7 W3_S7
W1CESDp W2CESDp W3CESDp
W1_NU W2_NU W3_NU
W1_NS W2_NS W3_NS
W1_RF W2_RF W3_RF
;
CONSTRAINT = W1_Age W3_Age;
USEVAR = W1_NS W3_NS W1_NU W3_NU W1_IR W3_IR W1_DR
W3_DR;
MISSING=.;
DEFINE:
!Imputation just to keep the time 1 people in at all
occasions...
IF (W3_age LT 0) THEN W3_age=999;
```

```
W1_NS = (W1_NS - 500)/10;
W3_NS = (W3_NS - 500)/10;
ANALYSIS: TYPE=MEANSTRUCTURE;
MODEL:
W1_G BY W1_NS (L_NS);
W1_G BY W1_NU (L_NU);
W1_G BY W1_IR (L_IR);
W1_G BY W1_DR (L_DR);
W3_G BY W3_NS (L_NS);
W3_G BY W3_NU (L_NU);
W3_G BY W3_IR (L_IR);
W3_G BY W3_DR (L_DR);
!get a level and a slope for G;
l_G BY W1_G@1 W3_G@1;
s_G BY W1_G * (L1);
s_G BY W3_G * (L3);
[l_G s_G]; l_G s_G; l_G WITH s_G;
!But with equal error variances
W1_G W3_G (U_G);
!original variables not used
[W1_NS@0 W3_NS@0];
W1_NS@0 W3_NS@0;
!Except equal error variances
W1_NS W3_NS (U_NS);
!original variables not used
[W1_NU@0 W3_NU@0];
W1_NU@0 W3_NU@0;
!Except equal error variances
W1_NU W3_NU (U_NU);
!original variables not used
[W1_IR@0 W3_IR@0];
W1_IR@0 W3_IR@0;
!Except Equal error variances
W1_IR W3_IR (U_IR):
!original variables not used
[W1_DR@0 W3_DR@0];
W1_DR@0 W3_DR@0;
!Except equal error variances
W1_DR W3_DR (U_DR);
```

```
MODEL CONSTRAINT: ! To get at individual loadings;
L1 = (W1_age - 65)/10;
L3 = (W3_age - 65)/10;
OUTPUT: SAMPSTAT STANDARDIZED;
```

References

Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesseloade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York: Academic.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29*(02), 109–125.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2012). Exploratory data mining with structural equation model trees (Chapter 4). In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining*. New York: Taylor & Frances.

Brandmaier, A.M., von Oertzen, T., McArdle, J.J. & Lindenberger, U. (2013). Structural Equation Model Trees. *Psychological Methods, 18*, 71–86 (doi: 10.1037/a0030001).

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Pub. Inc.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman & Hall.

Cribbie, R. A., & Jamieson, J. (2004). Decreases in posttest variance and the measurement of change. *Methods of Psychological Research Online, 9*(1), 37–55.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Cromwell, J. B., Labys, W. C., & Terraza, M. (1994). *Univariate tests for time series models*. Thousand Oaks, CA: Sage.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated-measures experiments. *Psychological Bulletin, 77*, 446–452.

Fisher, G. G., Hassan, H., Rodgers, W. L., & Weir, D. R. (2013). *Health and retirement study imputation of cognitive functioning measures: 1992-2010 (Final Release Version) data description*. Ann Arbor: University of Michigan, Survey Research Center.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fox, J. (1999). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.

Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent growth curve analysis. *Structural Equation Modeling, 17*, 424–442.

Heeringa, S. G., Berglund, P. A., & Khan, A. (2011). *Sampling error estimation in design-based analysis of the PSID data*. Institute for Social Research: University of Michigan Survey Research Center. Retrieved from http://psidonline.isr.umich.edu/Publications/Papers/tsp/2011-05_Heeringa_Berglung_Khan.pdf

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge, MA: Cambridge University Press.

Horn, J. L., & McArdle, J. J. (1992). A practical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117–144.

Hishinuma, E., Chang, J., McArdle, J. J., & Hamagami, A. (2012). Potential causal relationship between depressive symptoms and academic achievement in the Hawaiian High Schools Health Survey using contemporary longitudinal latent variable change models. *Developmental Psychology, 48*(5), 1327–1342.

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction from degrees of freedom from sample data in the randomized block and split plot design. *Journal of Educational Statistics, 1*, 69–82.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. Krantz et al. (Eds.), *Contemporary developments in mathematical psychology—Volume II*. San Francisco, CA: W.H. Freeman Co.

Juster, F.T., & Suzman, R. (1995). The health and retirement study: An overview. HRS Working Papers Series #94-1001. Journal of Human Resources, 1995 Supplement (JHR 30-S).

Lachman, M. E., & Spiro, A. (2002). Critique of cognitive measures in the Health Retirement Study (HRS) and the Asset and Health Dynamics among the Oldest Old (AHEAD) study. *HRS Data Monitoring Board Report,* November 2002.

Lange, K., Westlake, J., & Spence, M. A. (1976). Extensions to pedigree analysis: III. Variance component by the scoring method. *Annals of Human Genetics, 39*, 485–491.

Loeber, R., & Farrington, D. P. (1994). Problems and solutions in longitudinal and experimental treatment studies of child psychopathology and delinquency. *Journal of consulting and clinical psychology, 62*(5), 887.

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics, 16*(1), 163–200.

McArdle, J.J. (1988). Dynamic but structural equation modeling of repeated measures data. In J.R. Nesselroade & R.B. Cattell (Eds.), *The Handbook of Multivariate Experimental Psychology,* New York, Plenum Press, *2*, 561–614.

McArdle, J. J. (1998). Modeling longitudinal data by latent growth curve methods. In G. Marcoulides (Ed.), *Modern methods for business research* (pp. 359–406). Mahwah, NJ: Lawrence Erlbaum Associates.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100 years* (pp. 99–130). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McArdle, J. J. (2008). Latent variable modeling of longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J. J. (2010). Contemporary challenges of longitudinal measurement using HRS data. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 509–536). London: SAGE Press.

McArdle, J. J. (2011). Latent curve modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling*. NY: Oxford.

McArdle, J. J. (2012). Testing the idea of general intelligence. *F&M Scientist, 1*, 27–66.

McArdle, J. J. (2013). Dealing with longitudinal attrition using logistic regression and decision tree analyses. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 282–311). New York: Taylor & Frances.

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique contactor problems. *Multivariate Behavioral Research, 29*(1), 63–113.

McArdle, J. J., & Epstein, D. B. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*(1), 110–133.

McArdle, J. J., & Fisher, G.G. (2015). *New analyses of longitudinal cognitive data from The CogUSA survey.*

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analysis of age trends in tests of cognitive ability in the health and retirement survey, 1992-2004. *Psychology and Aging, 22*(3), 525–545.

McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 89–124). Hillsdale, NJ: Erlbaum.

McArdle, J. J., Hishinuma, E., Chang, J., & Hamagami, A. (2014). Longitudinal analyses of the dynamic relationships among depression and academic achievement from the Hawaiian High Schools Health Survey. *Structural Equation Modeling, 21*(4), 608–629.

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: APA Books.

McArdle, J. J., Petway, K. T., & Hishiuma, E. S. (2015). IRT for growth and change (chap. 20, p. 435–456. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling*. New York: Routledge.

McArdle, J.J. & Woodcock, J.R. (1997). Expanding test-rest designs to include developmental time-lag components. *Psychological Methods, 2* (4), 403–435.

McArdle, J.J. & Prescott, C.A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18* (3), 87–115.

McCall, R. B., Appelbaum, M. I., & Hogarty, P. S. (1973). Developmental changes in mental performance. *Monographs of the Society for Research in Child Development, 38*, 1–84.

McGaw, B., & Joreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology, 24*, 154–168.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*(2), 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrica, 55*, 107–122.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychological Methods, 10*(3), 259.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*(1), 23.

Muthén, B., & Christoffeson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407–419.

Muthén, B., & Sheddon, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463–469.

Muthén, B., & Muthén, L. (2012). *The Mplus 7.0 computer program for structural equation modeling*. Santa Monica, CA: Muthén & Muthén Publishing.

Nagin, D. S. (1999). Analyzing developmental trajectories: Semi-parametric, group-based approach. *Psychological Methods, 4*, 139–177.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Neale, M.C. (1993). *Mx Statistical Modeling.* Unpublished Manuscript, Virginia Commonwealth University, Richmond, VA

Neale, M.C., Boker, S. M., Xie, G., & Maes, H. H. (1999). Mx statistical modeling (5th ed.). Unpublished program manual, Virginia Institute of Psychiatric and Behavioral Genetics, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA.

Rao, C. R. (1958). Some statistical methods for the comparison of growth curves. *Biometrics, 15*, 1–17.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika, 83*(2), 251–266.

Reise, S. P., Widaman, K. F., & Pugh, P. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(2), 552–566.

Rodgers, W., Ofstedal, M. B., & Herzog, A. R. (2003). Trends in scores on tests of cognitive ability in the elderly U.S. population: 1993-2000. *Journals of Gerontology: Social Sciences, 52B*(6), S338–S346.

Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York: Springer.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Tu, Y. K., Blance, A., Clerehugh, V., & Gilthorpe, M. S. (2005). Statistical power for analyses of changes in randomized controlled trials. *Journal of Dental Research, 84*(3), 283–287.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19–23.

Tucker, L. R. (1960). *Determination of generalized learning curves by factor analysis*. New Jersey: Princeton University.

Tucker, L. R. (1966). Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 476–501). Chicago: Rand McNally.

Tucker, L. R. (1992). Remarks on the studies of the variety of individuals. *Multivariate Behavioral Research, 27*(4), 635–647.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research, 43*(3), 476–496.

Nonlinear Growth Curve Models

Paolo Ghisletta, Eva Cantoni, and Nadège Jacot

Abstract In the past three decades, the growth curve model (also known as latent curve model) has become a popular statistical methodology for the analysis of longitudinal or, more generally, repeated-measures data. Developed primarily within the latent variable modeling framework, the equivalent model emerged from other fields under the names of linear mixed-effects model, random-effects model, hierarchical linear model, and linear multilevel model. This methodology estimates the so-called growth parameters that describe individuals' change trajectories across time and are related via linear combinations to the dependent variable. While satisfying in many research settings, oftentimes a linear relation between dependent variable and growth parameters cannot allow for meaningful interpretation of the growth parameters, parsimonious descriptions of the change phenomenon, good adjustment to the data across all values of the time predictor, and realistic extrapolations outside the empirical range of the time predictor. Consequently, nonlinear alternatives have been proposed, for which the growth parameters can be related to the dependent variable via any mathematical function (not just linear combinations). We discuss the theoretical foundations as well as practical implications of estimating nonlinear growth curve models. We also illustrate the methodology with an example from the psychological literature.

P. Ghisletta (✉)
Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

Distance Learning University Switzerland, Brig, Switzerland
e-mail: paolo.ghisletta@unige.ch

E. Cantoni
Research Center for Statistics, University of Geneva, Geneva, Switzerland

Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland
e-mail: eva.cantoni@unige.ch

N. Jacot
Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland

Distance Learning University Switzerland, Brig, Switzerland

Research Center for Statistics, University of Geneva, Geneva, Switzerland

Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland
e-mail: nadege.jacot@unige.ch

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_2

The (Linear) Growth Curve Model

The growth curve model (GCM), also called latent curve model, has become a familiar model for repeated-measures data in many scientific disciplines. This model formally addresses ideas that were formulated a long time ago by several scientists, in several disciplines, all concerned with describing and understanding change in entities assessed repeatedly on the same outcome. In particular, in social sciences and humanities, entities were observed on outcomes as varied as children's stature, adolescents' learning abilities, and adults' memory capacities (McArdle 2001).

Relying on previous work on individual curve fitting via principal component analysis (Rao 1958; Tucker 1958), Meredith and Tisak (1985) showed how GCMs could be represented via confirmatory factor analysis and structural equation modeling (SEM). The GCM is related to other models commonly used with repeated-measures data, such as Wiener and Markov simplex models (Jöreskog 1970) or the repeated-measures analysis of variance, but can also be expanded to more complicated growth specifications (Meredith & Tisak 1985, 1990). McArdle and colleagues (McArdle 1986; McArdle & Epstein 1987) later showed how the GCM can be implemented and tested with commonly available SEM software, and how the model can further be expanded to specify known and unknown growth functions, which can be empirically compared to each other and tested on the same data.

Somewhat in parallel, stemming from different disciplines such as biostatistics and education, the linear mixed-effects models (LMMs) emerged. This model can be represented as the extension of the linear regression model to cases where multiple sources of variance are present. In particular, with respect to repeated-measures data, variations in the outcome are due to variations within an individual, but across the repeated measures (hence across time), and variations across individuals (Bryk & Raudenbush 1987; Laird & Ware 1982). This model is also known under the name of random-effects model, linear multilevel model, and hierarchical linear model, and is fully equivalent, under certain conditions, to the GCM (Ghisletta & Lindenberger 2004; Rovine & Molenaar 2000). We will use the LMM nomenclature only for simplicity.

The GCM can be represented as follows (Laird & Ware 1982):

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{1}$$

where for a total of $N = \sum_{i=1}^{m} n_i$ observations, $\boldsymbol{y}_i$ is the $(n_i \times 1)$ data vector for the ith individual (meaning that the number of repeated measurements may vary across individuals), $\boldsymbol{\beta}$ represents a $(p \times 1)$ vector of fixed effects, which are constant across all individuals, $\boldsymbol{X}_i$ is an $(n_i \times p)$ design matrix that can be specific to the ith individual, $\boldsymbol{b}_i$ is a $(k \times 1)$ vector of random effects, which varies across individuals, $\boldsymbol{Z}_i$ is a $(n_i \times k)$ design matrix linking $\boldsymbol{y}_i$ to the random effects, and $\boldsymbol{e}_i$ is a $(n_i \times 1)$ vector of within-individual errors. We assume that $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_i)$, where $\boldsymbol{R}_i$ is the $(n_i \times n_i)$ within-individual covariance matrix (typically $\boldsymbol{R}_i$ is specified by a small number of

parameters, the simplest case being $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}$, where $\boldsymbol{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix). This implies that $\mathrm{E}(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$ and $\mathrm{Cov}(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{R}_i$. However, this specification of the errors $\boldsymbol{e_i}$ can and should be tested (e.g., Grimm & Widaman 2010). Indeed, if the investigated change process is partially driven by a stochastic trend (such as a simple random walk process), the errors are no longer random, and failure to account for this dependency may result in spurious results and increase of Type-I errors (Braun, Kuljanin, & DeShon 2013; Kuljanin, Braun, & DeShon 2011).

To specify the interindividual variations we assume $\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$ (where $\boldsymbol{D}$ is a $(k \times k)$ dispersion matrix; $\boldsymbol{b}_i$ are independent of each other and of $\boldsymbol{e}_i$). If $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are normally distributed and independent, then $\boldsymbol{y}_i$ are normally distributed with mean $\boldsymbol{X}_i\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ (Davidian & Giltinan 1995). The fact that it is possible to specify in closed form the marginal distribution is due to both the Gaussianity of all random terms and the linear, additive dependence of the dependent variable $\boldsymbol{y}_i$ on the fixed effects $\boldsymbol{\beta}$ through the design matrix $\boldsymbol{X}_i$ and on the random effects $\boldsymbol{b}_i$ through the design matrix $\boldsymbol{Z}_i$. The linear dependence of the response on the covariates and the regression parameters (i.e., the fact the model is linear in its parameters) and the assumption of Gaussianity greatly facilitates both estimation and statistical inference (likelihood based inference can easily be carried out).

The GCM is a linear model because the response variable is a linear function of the parameters of the model. For instance, in the simple linear regression model we assume that the dependent variable y of individual i is expressed as $y_i = \beta_0 + \beta_1 x_i + e_i$. Accordingly, the model assumes that for each individual i the value of y is obtained by a linear combination of β_0 and β_1, where the parameters β_0 and β_1 are at most multiplied by a fixed value (e.g., 1 and x_i) and then added. Note that if x_i is a nonlinear function of time t_i, such as $x_i = t_i^2$ the model is still linear in its parameters. By contrast, a model where $y_i = \beta_0 x_i^{\beta_1} + e_i$ is said nonlinear in its parameters because β_1 is the exponent of x_i.

Probably the most common GCM is the one implementing a growth function with a linear shape:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}, \tag{2}$$

where the outcome y_{ij} for individual i at time j is specified as the sum of the individual-specific intercept β_{0i} and the individual-specific linear slope β_{1i} multiplied by the time of assessment, and a time- and individual-specific residual e_{ij}. The intercept is the predicted score for subject i at time $t_{ij} = 0$ and the linear slope is the constant amount of linear change for subject i across each unit of time.

While this model is easy to implement in existing SEM or LMM software and can easily be estimated, it is rarely satisfying from a theoretical perspective (Davidian & Giltinan 1995; Grimm, Ram, & Hamagami 2011; Pinheiro & Bates 2000). It is indeed hard to imagine that phenomena of scientific interest are truly linear, because this time relation implies constant rate of change, hence null acceleration or deceleration in change, and lack of asymptotes. In other words, any outcome

is predicted, at least in the long run, to attain levels outside of any empirically imaginable range, thereby ranging from minus infinity to plus infinity (unless $\beta_{1i} = 0\ \forall i$).

Another common specification of the GCM consists in expanding the polynomial relating time t_{ij} to the outcome. Most often the quadratic expansion is applied:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij}, \tag{3}$$

but examples include applications to cubic or even higher-order polynomials. The polynomial approach however presents several limitations (Davidian & Giltinan 1995; Pinheiro & Bates 2000). First, to increase the accuracy of the model usually several terms are added (e.g., quadratic, cubic). In general, for each curvature in the observed trajectory an additional polynomial degree is needed for a reasonable adjustment. This inevitably reduces the statistical parsimony of the model in terms of number of necessary parameters. Second, adding terms of higher degrees to the polynomial will likely increase multicollinearity among these terms and existing terms of lower degrees. For instance, if the trajectory is to be described as a function of the participants' age with a quadratic polynomial, both age and age squared will be predictors in the model. However, given that only positive values are admissible for age, the two predictors will inevitably correlate very highly. This multicollinearity presents problems that are well known in ordinary least squares regression and that are transposed to the multilevel extension (estimation issues, high standard errors of the parameter estimates, and problematic substantive interpretation). Analytical solutions are hence required. Very often the linear predictor is centered around its mean before it is squared to obtain the quadratic term. This transformation inevitably requires additional care during parameter interpretation, especially in the presence of orders greater than 2 or in the presence of interaction terms. Residual centering is also a possibility that reduces multicollinearity, but this option is not very popular in GCM applications, given it is more complicated than predictor centering (Lance 1988). Also, orthogonal polynomials (e.g., Chebyshev's, Gegenbauer's, Hermite's, Jacobi's, Laguerre's) may be specified, which, by definition, are not composed of multicollinear terms and hence avoid this issue altogether (Chihara 1978). Third, and perhaps most important from a statistical viewpoint, a polynomial function (including of order one, the linear) will usually only approximate the observed range of the data, while distancing itself noticeably beyond this range. This implies that predictions outside the observed range of the data are usually not valid with the polynomial approach. Indeed, a polynomial function is not necessarily bounded. Fourth, to maximize the overall adjustment to the data a polynomial will usually overfit the observations in center of the data and attribute large deviations to observations at the extremes. Finally, interpreting polynomial coefficients in substantive terms is usually not simple. One may estimate characteristics of polynomials, such as inflection points, but rarely are these discussed in terms of theoretical interpretation.

Rather than transforming a predictor to then apply the polynomial strategy an analyst may choose to transform the outcome variable. Here again voices from

the statistical community were raised against this practice for several reasons (Davidian & Giltinan 1995; Pinheiro & Bates 2000). First, this approach may once more render difficult any parameter interpretation. Second, oftentimes data on the same outcome collected in independent samples must be subjected to different transformations to approximate linearity (e.g., different values of λ in the Box-Cox transformation). Third, in specific applications the outcome may not be a continuous variable approximately normally distributed. A dichotomous (two-value) outcome, for instance, is best described by a Bernoulli distribution of the value of interest. No transformation will change such a distribution to become approximately normal. Finally, and most importantly, nonlinearity may arise for meaningful empirical or theoretical reasons. For instance, growth in various organisms is often well described by logistic functions; survival functions with time-dependent hazard rates may follow a Weibull distribution; compound interests follow exponential functions. In such cases the appropriate nonlinear modeling procedure (rather than a transformation to fit a linear model) holds the potential to estimate parameters that will further the understanding of the phenomenon under investigation, whereas a transformation of the outcome may obscure such understanding.

The Nonlinear Growth Curve Model

Conscious of the limitations of LMM in various applied settings, several statisticians have first explored, then established what is now a well-recognized statistical model, the nonlinear mixed-effects models (NLMMs). Dedicated books include, but are not limited to, Davidian and Giltinan (1995), Vonesh and Chinchilli (1996), and Pinheiro and Bates (2000).

Basically, LMMs have been extended to NLMMs, to include functions that are nonlinear in their parameters. Analogously to LMMs, NLMMs are represented to specify both intraindividual and interindividual variations. Intraindividual variation is characterized by a nonlinear regression of the outcome on the predictors, while the interindividual variability is represented through individual-specific regression parameters that often incorporate fixed and random effects. We suppose again that individuals $i = 1, 2, \ldots, m$ have been observed on n_i responses at time $j = 1, 2, \ldots, n_i$, so the outcome is indicated y_{ij}. We denote $\boldsymbol{x}_{ij}$ the vector of predictors and suppose a nonlinear function $f(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_{ij})$ that models the relationship between y_{ij} and $\boldsymbol{x}_{ij}$, where $\boldsymbol{\beta}_{ij}$ is a vector of parameters for individual i at time j. We suppose that f is common to all individuals but that elements of $\boldsymbol{\beta}_{ij}$ may vary across individuals (and across time).

Following these specifications we can assume the following intraindividual model:

$$y_{ij} = f(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_{ij}) + e_{ij}s \tag{4}$$

where e_{ij} is again a random error, such that $\mathrm{E}(e_{ij}|\boldsymbol{\beta}_{ij}) = 0 (e_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}_i))$.

The interindividual model can be specified as follows:

$$\boldsymbol{\beta}_{ij} = \boldsymbol{d}(\boldsymbol{a}_{ij}, \boldsymbol{\beta}, \boldsymbol{b}_i), \tag{5}$$

where $\boldsymbol{d}$ is a vector-valued function, $\boldsymbol{a}_{ij}$ is a covariate vector corresponding to individual attributes for individual i at time j (which may be other individual characteristics than $\boldsymbol{x}_{ij}$), $\boldsymbol{\beta}$ is a vector of fixed parameters (or fixed effects), and $\boldsymbol{b}_i$ is a vector of random effects associated with individual i. Normality is the most common assumption for the distribution of the random effects ($\boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$; Davidian & Giltinan 1995).

Equation (4) can be specified to obtain the LMM of Eq. (1). The more general specification of the NLMM allows both fixed and random effects to be related to potentially different sets of individual characteristics in a nonlinear fashion. This gain in flexibility is somewhat similar to that in the non-hierarchical regression models when the General Linear Model is expanded to the Generalized Linear Model (McCullagh & Nelder 1989; Nelder & Wedderburn 1972).

The analogy between the SEM implementation via GCM and the LMM does not generalize in a straightforward fashion to the case of the NLMM. Hence, technically speaking, we cannot necessarily say that an NLMM is equivalent to a nonlinear GCM (NGCM). While recent advances in GCM allow specifying models with nonlinear fixed effects, the functional expression of random effects is required to be linear. Models with nonlinear fixed and linear random effects are sometimes referred to as models with additive random coefficients, or partially nonlinear; in such a case, the NLMM is equivalent to the NGCM. These models can be specified with SEM software as long as nonlinear constraints can be specified on the estimation of the fixed parameters (for instance, where parameters λ and γ are estimated under the nonlinear constraint that $\lambda = \exp(\gamma)$; for a full example, see Ghisletta & McArdle 2001). In opposition, models where both fixed and random effects are nonlinear are said to be with multiplicative random coefficients and are also referred to as fully nonlinear. These retain the advantages of the additive random coefficients models and moreover allow specifying, hence testing for, individual difference in potentially important (nonlinear) change components (Grimm, McArdle, & Hamagami 2007). Because SEM assumes random effects to be linear (i.e., additive), fully nonlinear models cannot be estimated with SEM software if directly specified as such. Here, the NLMM is not truly equivalent to the NGCM. Nonetheless, Browne and colleagues (Browne 1993; Browne & Du Toit 1991), in their seminal work, delineate a methodology that allows approximating nonlinear, monotonic, differentiable functions (e.g., Gompertz, exponential, logistic) in the SEM framework. The function's parameters are linearized via first-order Taylor expansions, assuming the random effects are normally distributed and in linear relations with possible covariates (for applications see for instance Blozis 2004; Cudeck & Harring 2007; Ghisletta, Kennedy, Rodrigue, Lindenberger, & Raz 2010; Grimm et al. 2007, 2011). This method, however, is limited to functions that can be linearized, that is rewritten as the sum of each random effect multiplied by the partial derivative with respect to each random effect of the underlying (target)

function (Browne 1993; Browne & Du Toit 1991; Grimm et al. 2007). Not every function $f(\boldsymbol{x}_{ij}, \boldsymbol{\beta}_{ij})$ abides to these characteristics. Moreover, the approximation may fail under some circumstances (Vandergrift, Curran, & Bauer 2002; von Oertzen 2010).

Available Software and Estimation Issues

The greater modeling flexibility of the NLMM comes with a cost in terms of increased computational difficulties. As for the LMM, the marginal likelihood function is obtained by integrating the joint density of the outcome and the random effects with respect to the random effects. However, because in the (fully) NLMM the random effects are allowed to enter the model nonlinearly, a closed-form expression of the marginal likelihood, contrary to the LMM, does no longer exist. Therefore, an approximate likelihood function needs to be employed (common choices include a Laplacian approximation or adaptive Gaussian approximation with multiple quadrature points). This results in computationally more intensive algorithms and approximate inference results. As a consequence, acceptable solutions (producing realistic predictions) may not always be obtained.

Dedicated software has been developed to estimate NLMM, e.g. `PROC NLMIXED` in SAS (Wolfinger 1999) and the `nlme` and `lme4` packages in the R environment (Pinheiro et al. 2014). `PROC NLMIXED` in SAS is particularly simple to specify and offers four estimation methods for the integral of the likelihood over the random effects: adaptive Gauss-Hermite quadrature (default method; Pinheiro & Bates 1995), the first-order method (Beal & Sheiner 1982), Hardy quadrature (available only for a single random effect), and adaptive importance sampling (Pinheiro & Bates 1995). In our experience, the default adaptive Gauss-Hermite quadrature method works well but may be highly dependent on initial starting values provided by the analyst. As an alternative, we find the first-order method of Beal and Sheiner (1982) very useful. Functions that can be implemented as NGCM and estimated via the Taylor expansion method of Browne (1993) can be estimated with common SEM software that allows for nonlinear constraints. Some authors find this strategy to produce more robust results and more easily attain convergence (Grimm et al. 2011).

Modeling Strategy

Whichever estimation method is used, it is of utmost importance to first plot the data. Individual longitudinal trajectories and the sample average curve ought to be plotted, in the hope that they may guide the choice of a functional specification. For instance, data showing a clear nonlinear increase and graphical evidence for an upper asymptote may be well described by an exponential function. Second,

the analyst needs to map the parameters of the chosen function to the plotted data. For instance, the values of asymptotes can easily be guessed by eye-balling the longitudinal plot. However, guessing the values of rates of growth usually requires a deeper examination and some experience. To this end, a family of curves from the same function, obtained by altering the function's parameter values, can be simulated and plotted, to then be confronted to the empirical plot of the data, in the hope that a match emerges (e.g., Sit & Poulin-Costello 1994).

Another modeling step we find at times useful consists in fitting a series of nonlinear regressions, separately for each participant's time series. This can easily be achieved with the `nls` function of the default `stats` package in R or with `PROC NLIN` in SAS, both of which implement least-squares estimation of nonlinear regression models. The nonlinear function parameters are thus estimated for each participant and their distributions can be plotted. Although such distributions are wider than expected based on the random effects of the NLMM (because of the inferior statistical efficiency of the regression model; e.g., Pinheiro & Bates 2000), they are centered on the unbiased estimates of the fixed effects of the NLMM. Hence, good initial starting values for the fixed effects of the NLMM can be obtained. Alternatively, it is possible to consider a set of different starting values, each of which is used for a separate model estimation. The examination of the various parameter estimates (also in terms of overall statistical adjustment, such as the maximal likelihood) can be used to evaluate the dependency of the solution on the starting values and eventually lead to select a given set. In sum, one should by all means not underestimate the importance of the initial values in the parameter estimation methods of NLMM or NGCM.

Another important aspect of fitting NLMMs and NGCMs concerns the modeling of random effects. Analysts should first specify the fixed effects, as these are typically more easily estimable with any estimation method (Pinheiro & Bates 2000; Snijders & Bosker 2012). Then, one should think hard about which aspects of the change function may vary across individuals, and for those additionally estimate random effects. That is, one should resist the temptation to let every parameter of the specified change function vary across individuals by estimating all possible random effects. In fact, it is not uncommon for some random effects of the same change function to correlate very highly. For instance, in a classical memory study, where a list of nouns is repeatedly presented and asked to recall, one may find that participants differ with respect to their initial level (initial asymptote), rate of learning, and final performance (final asymptote). In this situation the variance of the rate of learning parameter and that of the final asymptote may overlap greatly. This redundant information would be operationalized as a high correlation between the two random effects, a clear sign of model overparameterization. This could be due to computation difficulties, such that more data would be needed to separately estimate the two random effects, or to a substantive feature of the change process, owing to the fact that individuals who learn many words could end up recalling many words by the end of the experiment. In general, it is best to first test the fixed effects only, then add random effects one at a time, to increasingly test their empirical relevance.

Finally, we find it very useful to plot the expected trajectories for each individual. Once a function is tested on the data, the parameter estimates need to be controlled carefully, to see whether their value appears plausible. This can easily be done for the fixed effects. It suffices to replace their estimated values in the original equation of the function to obtain the average sample trajectory. However, the random effects should also be checked. Contemporary software easily estimate individual values of the random effects (e.g., empirical Bayes estimates are available in SAS `NLMIXED`, conditional modes are available in `nlme` in R). These can be replaced in the equation of the original function to obtain individual expected trajectories, which should then be plotted for diagnostics (cf. Figs. 2 and 3 below).

Illustration

We illustrate some specifications of NGCM with an example from the psychological literature. We reanalyze the data reported in Kennedy, Partridge, and Raz (2008), which were themselves based on a previous study by Raz, Williamson, Gunning-Dixon, Head, and Acker (2000). Both studies aimed at understanding how the acquisition of new skills, in particular a perceptual-motor skill, relates to individual age-related differences in cognitive functioning and in structures of the central nervous system in healthy human adults. Participants performed the Pursuit Rotor task (Durkin, Prescott, Furchtgott, Cantor, & Powell 1995), which consists in keeping a wand with a 7-mm tip on a 2-cm light spot that rotates circularly at 40 rpm. The outcome is the number of seconds (time-on-target) each participant manages in achieving the task. To capture a large range of age-related differences, the $m = 98$ participants of the sample varied from 19 to 80 years of age.

A number of covariates of interest were also assessed. To simplify the illustration, we consider only two covariates here. The first covariate is chronological age (measured in years, and centered around the group average of 47.63 years). The second is the score on a task of spatial abilities, called Spatial Relations, from the Woodcock-Johnson Psycho-Educational Battery-Revised (Woodcock & Johnson 1989). Participants are presented with a whole shape as well as with a series of six disjointed shapes, from which they are asked to choose a correct combination. The complexity and degree of abstraction of the shapes increase according to participants' capacities. This task is quite difficult and relies not only on spatial abilities, but also on working memory, the capacity to hold in memory pieces of information while manipulating them. In fact, to successfully manage the task participants must rotate and translate the shapes to then assemble them to form the correct shape that corresponds to the whole shape presented. We considered the usual score, that is the total number of correct responses.

In the original data set, four blocks of 20 trials, each trial lasting 15 s, were administered. The trials were separated by 10-s pauses, and the blocks were

distributed across 3 days (with blocks 1 and 2 on day 1, blocks 3 and 4 on day 2 and 3, respectively). In the original publications the data were averaged either by block (over 20 trials), or over all $n_i = 80$ trials, thereby impeding a trial-by-trial analysis. Ghisletta et al. (2010) reanalyzed the data with an NGCM specifying a three-parameter exponential function (more detail in the subsection below), relating performance at each trial to the covariates of interest. By doing so, Ghisletta et al. (2010) were able to study the learning curve across all trials. The authors implemented a multivariate analysis, in which a specific growth model was estimated for each block and all growth parameters were allowed to correlate across the blocks. They concluded that within each block participants quickly learned the task and improved their time-on-target, until they reached an upper asymptote. The initial performance level, learning rate, and final asymptotic level of performance correlated strongly across blocks.

For simplicity, we only consider the 20 trials of the first block here. The observed individual and average sample trajectories are plotted in Fig. 1. Clearly, we can see that individuals start at relatively low values of time-on-target and then increase rather quickly, to become stable relatively soon. This is particularly visible in the sample average trajectory. Moreover, we see that there is high variability with respect to both performance on trial 1 and maximal performance, and that this variability appears to increase throughout the study; finally, we can guess that the rate of learning also discriminates individuals.

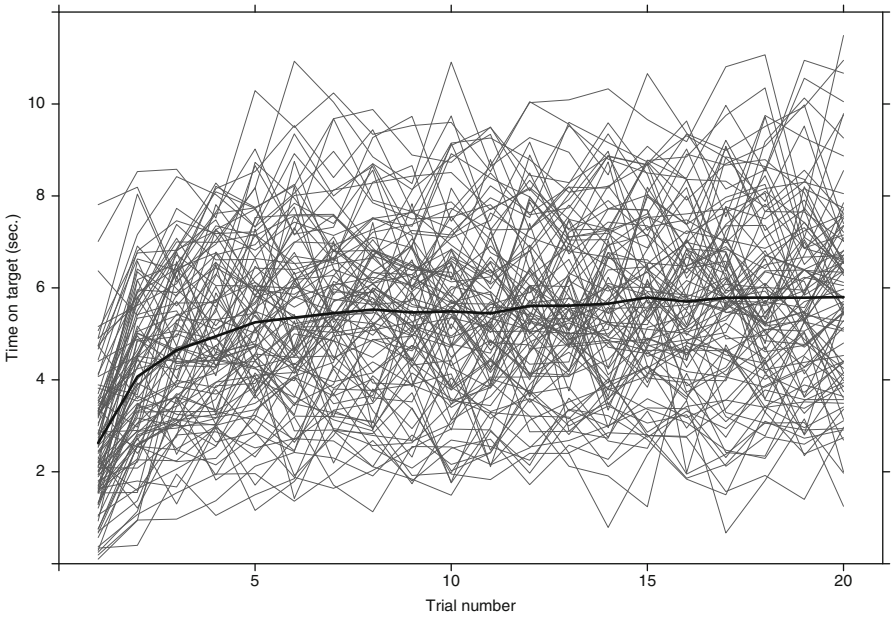


Fig. 1 Observed individual trajectories of time-on-target (in seconds—*thin lines*) and sample average trajectory (in seconds—*thick line*) by trial number (from 1 to 20)

Nonlinear Functions

We implement a number of nonlinear functions, each of which specifies parameters that allow interpreting meaningfully the learning patterns observed in the data. Clearly, learning rate is not well represented by a straight line. Thus, the linear GCM of Eq. (2) likely does not represent well the data. Moreover, the data are naturally bounded above and below by the duration of the task. Participants' admissible time-on-target scores range from 0 to 15 s. Naturally, then, functions used to describe such data should contain both a lower and an upper asymptote. That is why the quadratic GCM of Eq. (3), or other, higher-order polynomials, may not be well suited for these data either. In the end, we choose a number of nonlinear functions with parameters representing a lower and upper asymptote, or similarly initial and final performance, and rate of growth that is not constant in time. To represent the heterogeneity observed in the learning curves, the functions specify both fixed and random effects in their characteristics. All functions have the form specified in Eq. (4), where the only covariate $\boldsymbol{x}_{ij}$ is the trial number t_{ij}, going from 0 to 20. Further, we always assume $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$.

Exponential

This function contains three parameters and is presented in Meredith and Tisak (1990). McArdle, Ferrer-Caja, Hamagami, and Woodcock (2002) estimated it as an NLMM while Ghisletta et al. (2010), based on the work of Browne (1993), estimated it as an NGCM. The function is:

$$y_{ij} = \beta_i - (\beta_i - \alpha_i)\exp(-(t_{ij} - 1)\gamma_i) + e_{ij}. \tag{6}$$

For individual i, α_i represents the initial performance at time $t_{ij} = 1$ (first trial), β_i the final performance at $t_{ij} = 20$ (final asymptote, representing potential performance), and γ_i the rate of change (representing learning speed). We estimate the random effects associated with all three growth components ($\alpha_i = \alpha + U_{1i}$, $U_{1i} \sim \mathcal{N}(0, \sigma_1^2)$; $\beta_i = \beta + U_{2i}$, $U_{2i} \sim \mathcal{N}(0, \sigma_2^2)$; and $\gamma_i = \gamma + U_{3i}$, $U_{3i} \sim \mathcal{N}(0, \sigma_3^2)$).

Logistic

This function also contains three parameters and is presented in Meredith and Tisak (1990):

$$y_{ij} = \frac{\alpha_i \beta_i}{\alpha_i + (\beta_i - \alpha_i)\exp(-(t_{ij} - 1)\gamma_i)} + e_{ij}. \tag{7}$$

The interpretation of the parameters is the same as in Eq. (6) and as above we estimate the random effects of the growth components ($\alpha_i = \alpha + U_{1i}$, $U_{1i} \sim \mathcal{N}(0, \sigma_1^2)$; $\beta_i = \beta + U_{2i}$, $U_{2i} \sim \mathcal{N}(0, \sigma_2^2)$; and $\gamma_i = \gamma + U_{3i}$, $U_{3i} \sim \mathcal{N}(0, \sigma_3^2)$).

Gompertz

This function contains three parameters as well and is presented in Meredith and Tisak (1990):

$$y_{ij} = \beta_i \exp(\ln\left(\frac{\alpha_i}{\beta_i}\right) \exp(-(t_{ij} - 1)\gamma_i)) + e_{ij}. \tag{8}$$

The interpretation of the parameters is the same as in Eq. (6) and as above we estimate the random effects of the growth components ($\alpha_i = \alpha + U_{1i}$, $U_{1i} \sim \mathcal{N}(0, \sigma_1^2)$; $\beta_i = \beta + U_{2i}$, $U_{2i} \sim \mathcal{N}(0, \sigma_2^2)$; and $\gamma_i = \gamma + U_{3i}$, $U_{3i} \sim \mathcal{N}(0, \sigma_3^2)$).

Chapman-Richard

This function contains three parameters, is often used in forestry, and is presented in Sit and Poulin-Costello (1994):

$$y_{ij} = \beta_i(1 - \exp(-\gamma_i t_{ij}))^{\delta_i} + e_{ij}. \tag{9}$$

It can also be found in the literature with an alternative parameterization where $\delta_i = \frac{1}{1-m_i}$ and with an additional parameter that pre-multiplies the exponential term (e.g., Fekedulegn, Mac Siurtain, & Colbert 1999). For individual i, β_i represents the final asymptotic performance, while γ_i and δ_i determine the shape. We estimate the random effects associated with all three growth components ($\beta_i = \beta + U_{2i}$, $U_{2i} \sim \mathcal{N}(0, \sigma_2^2)$; $\gamma_i = \gamma + U_{3i}$, $U_{3i} \sim \mathcal{N}(0, \sigma_3^2)$; and $\delta_i = \delta + U_{4i}$, $U_{4i} \sim \mathcal{N}(0, \sigma_4^2)$).

von Bertalanffy

This function also contains three parameters, is presented in Draper and Smith (1998) and Fekedulegn et al. (1999), and was first used to model forest growth (Yuancai, Marques, & Macedo 1997):

$$y_{ij} = (\beta_i^{\frac{1}{\delta_i}} - \exp(-\gamma_i t_{ij}))^{\delta_i} + e_{ij}. \tag{10}$$

This function is also presented with an alternative parameterization where $\delta_i = \frac{1}{1-m_i}$ and with an additional parameter that pre-multiplies the exponential term (e.g., Fekedulegn et al. 1999). For individual i, β_i represents the final asymptotic

performance and γ_i and δ_i determine the shape of the trajectory. As for the Chapman-Richard function, we estimate the random effects of all growth components ($\beta_i = \beta + U_{2i}$, $U_{2i} \sim \mathcal{N}(0, \sigma_2^2)$; $\gamma_i = \gamma + U_{3i}$, $U_{3i} \sim \mathcal{N}(0, \sigma_3^2)$; and $\delta_i = \delta + U_{4i}$, $U_{4i} \sim \mathcal{N}(0, \sigma_4^2)$).

Schnute

This function contains four parameters and was developed by Schnute (1981) to study the growth of fish populations.

$$y_{ij} = \left((\alpha_i^{\gamma_i} + (\beta_i^{\gamma_i} - \alpha_i^{\gamma_i})) \frac{1 - \exp(-\delta_i(t_{ij} - t_1))}{1 - \exp(-\delta_i(t_2 - t_1))} \right)^{\frac{1}{\gamma_i}} + e_{ij}. \tag{11}$$

t_1 and t_2 are the first and last values, respectively, of t_{ij} in the data (hence $t_1 = 1$ and $t_2 = 20$ here). α_i and β_i correspond to $\hat{y}$ values at $t_{ij} = t_1$ and $t_{ij} = t_2$, respectively (hence $\hat{y}_{i1}$ and $\hat{y}_{i20}$, respectively, here). γ_i and δ_i define the shape of the curve. We estimate the random effects of all growth components ($\alpha_i = \alpha + U_{1i}$, $U_{1i} \sim \mathcal{N}(0, \sigma_1^2)$; $\beta_i = \beta + U_{2i}$, $U_{2i} \sim \mathcal{N}(0, \sigma_2^2)$; $\gamma_i = \gamma + U_{3i}$, $U_{3i} \sim \mathcal{N}(0, \sigma_3^2)$; and $\delta_i = \delta + U_{4i}$, $U_{4i} \sim \mathcal{N}(0, \sigma_4^2)$).

Results

Overall Adjustment and Parameter Estimates

Table 1 contains the total number of parameters (i.e., associated with the fixed effects and with the distributions of the random effects, as presented in each function's description), the -2 Log-Likelihood (-2LL), and the Bayesian Information Criterion (BIC; Schwarz 1978) of each fitted nonlinear functions. We did not

Table 1 Total number of parameters (p; fixed and random effects), -2 Log-Likelihood (-2LL) and Bayesian Information Criterion (BIC) values, and parameter estimates for each nonlinear function (only fixed effects and variances of random effects are shown for each parameter of the functions in Eqs. (6)–(11))

Function	p	-2LL	BIC	α	β	γ	δ	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_e^2
Exponential	10	5752	5798	2.82	5.75	0.39	–	1.57	3.11	0.08	–	0.82
Logistic	10	5777	5823	2.99	5.73	0.53	–	1.57	3.07	0.10	–	0.83
Gompertz	10	5765	5810	2.91	5.74	0.46	–	1.57	3.09	0.09	–	0.83
Chapman-Richard	10	5738	5784	–	5.76	0.33	0.56	–	3.19	0.11	0.14	0.81
von Bertalanffy	10	5758	5804	–	5.74	0.42	2.08	–	3.10	0.09	0.16[a]	0.82
Schnute	15	5703	5772	2.62	5.75	4.24	0.23	1.40	3.27	35.44	0.26	0.79

[a] A non-significant parameter at the 5 % level

encounter numerical/optimization difficulties for any of the functions fitted. From a strict statistical data-adjustment perspective clearly the Schnute's function is the model that best describes the data. Its -2LL value is lowest, which might be expected given the higher number of parameters, but so is also the BIC value. Thus, we may be tempted to choose the Schnute function as the most adequate among those tested here for these data. Nevertheless, it is well known that considering only information criteria such as the BIC when selecting models can be a flawed strategy. In fact, a model that fits well the data according to some statistical criterion may not offer for clear interpretation of its parameter estimates. In some extreme cases, a perfectly well-fitting model may indeed represent a highly inappropriate structure of the data (Hayduk 2014).

Thus, we proceed to carefully examine all parameter estimates. In Table 1 we also present for each function the estimated fixed effects, the variances of the random effects (for simplicity we do not show the covariances), and the residual variance. As can be seen, the Schnute function obtains a very large variance estimate (σ_3^2) for a rate of change parameter (γ). This warns us that, despite the superior statistical adjustment of this model, its solution may not be interpretable. To check, we plotted the individual estimated trajectories of this function in Fig. 2. Clearly, such predictions are not representative of the data plotted in Fig. 1.

In contrast, Fig. 3 represents the expected individual trajectories of the Chapman-Richard's equation, the second best fitting function (based on -2LL and BIC) tested

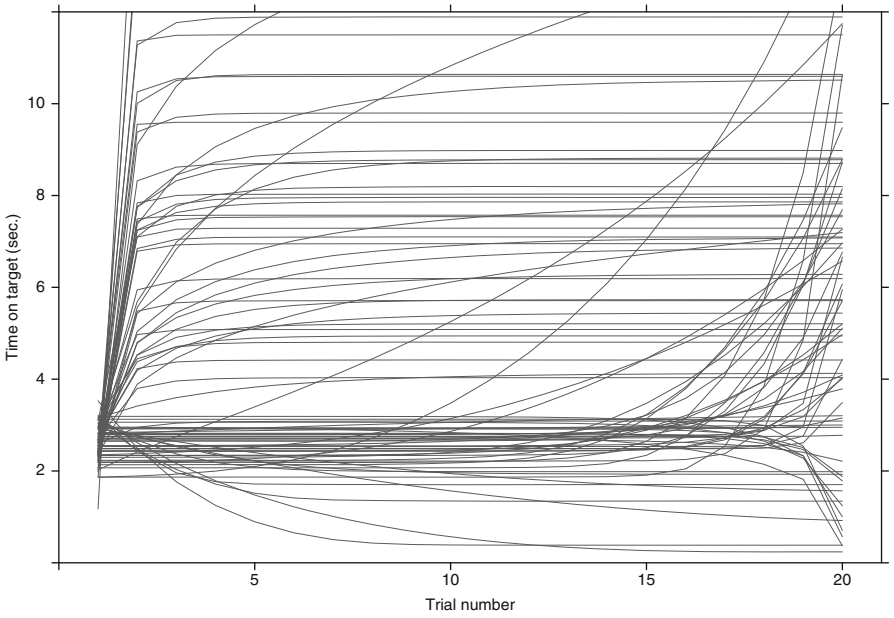


Fig. 2 Estimated individual trajectories of time-on-target (in seconds) by trial number (from 1 to 20) according to the Schnute's function

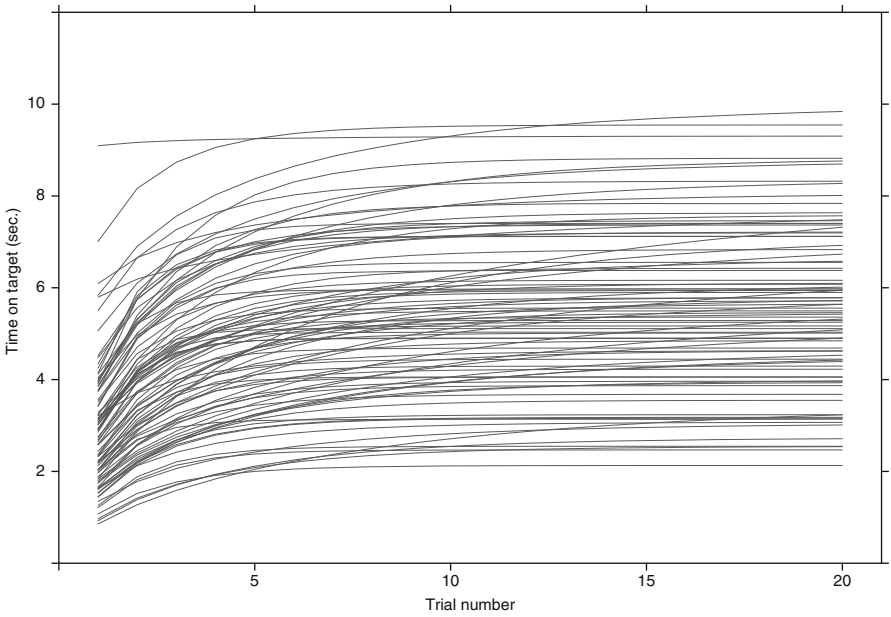


Fig. 3 Estimated individual trajectories of time-on-target (in seconds) by trial number (from 1 to 20) according to the Chapman-Richard's function

here. These trajectories mirror well the previous observations about the empirical curves. No apparent outlying trajectory emerges, and these trajectories appear feasible descriptions of the empirical data (note that the remaining functions all produced similar expected individual trajectories).

We also notice that, while the functions appear different, they share some common similarities. For instance, all but the Chapman-Richard and the von Bertalanffy explicitly define a lower asymptote of initial performance value, which we consistently called α. The estimates of this parameter are quite similar across the functions (ranging from 2.62 to 2.99; cf. Table 1) and there appears to be some amount of heterogeneity therein (from 1.40 to 1.57). All functions explicitly define an upper asymptote, or final performance level, which we consistently labeled β. This is also estimated with very high agreement across the functions (from 5.73 to 5.76). The sample appears to be more highly differentiated with respect to β (from 3.07 to 3.27) than to α. Rates of change (γ and δ) are more difficult to compare, because they intervene differently in the equations. Nevertheless, all functions estimated some degree of heterogeneity also with respect to the rate of learning.

Table 2 Percentage of variance explained of each random effect by age and spatial abilities

Function	U_{1i}	U_{2i}	U_{3i}	U_{4i}
Exponential	18.53	20.71	4.26	–
Logistic	16.99	21.16	1.72	–
Gompertz	17.44	20.99	2.69	–
Chapman-Richard	–	19.71	0.88	−11.65
von Bertalanffy	–	20.74	−0.30	−23.17

Covariates' Effects

Finally, we explore the effect of age and spatial abilities on learning characteristics as implemented in the different functions. To do so we added two time-invariant covariates, age and the spatial abilities score, to the equations considered before, which so far only contained the trial number as predictor. We did not consider the Schnute function, given the unsuccessful previous results. Table 2 shows the effect size of these predictions, in terms of percent of total variance, explained by age and spatial abilities, of the random effects associated with each growth component (U_{1i}, U_{2i}, U_{3i}, and U_{4i}). For simplicity, we do not show the parameters associated with each predictor, but instead summarize below the main findings.

With respect to final performance, across all functions about one fifth of interindividual differences were explained by age and spatial abilities. In all cases greater age affected negatively, and greater spatial abilities affected positively, final performance. With respect to initial performance, it was likewise predicted to some extent (about a sixth of the variance). However, spatial abilities represented the only significant predictor of initial performance. Finally, with respect to rates of change (γ and δ), the statistical tests associated with age and spatial abilities were shakier. For the exponential, logistic, and Gompertz functions, older people increased more than younger people, but the effect was quite weak (less than 5 % of predicted variance). Spatial relations did not affect the rate of learning. For the Chapman-Richard and the von Bertalanffy functions, the covariates' effects on the rate of change parameters yielded negative effect sizes, a likely indicator of an unstable solution (Snijders & Bosker 2012).

Conclusions

In the illustration from the psychological literature, data from a learning experiment on perceptual-motor skills were analyzed with six nonlinear functions that displayed some similarities. Among them, the first three (exponential, logistic, and Gompertz) were very similar and produced similarly stable results: an initial low level of performance, followed by a rather steep learning rate, to lead to a quite high final performance level. Moreover, the three phases were all characterized by strong variability in the sample. Finally, age and spatial relations proved to be significant predictors of final performance, while spatial relations also influenced initial performance, and age affected the rate of learning. In the end, we are not

preoccupied with choosing "the best model." Rather, we realize that any statistical model allows for alternative formulations that are statistically equivalent (hence obtain same statistical adjustment), but may possibly lead to different substantive conclusions. Here, we are comforted by the fact that three different models all agree in their final conclusions. The remaining three functions (Chapman-Richard's, von Bertalanffy's, and Schnute's) stem from forestry, fishery, biology, and related disciplines concerned with population growth. Their interpretation appears more difficult in our application, given that their results were not always interpretable. Nevertheless, they also indicated general nonlinear growth, with much heterogeneity therein, and at times revealed effect of age and spatial abilities.

NLMM and NGCM allow testing complex functions that present advantages over simpler functions. First, the parameters are often interpretable with respect to the underlying change process of the outcome. Second, such functions may provide a finer picture of the change process in terms of its components. For instance, rather than describing an overall linear change process, which presents only the concept of rate of change, more complex functions may distinguish lower and upper asymptote, rate of change, amount of change, and timing of change (Grimm et al. 2011). In many applications these are fundamental aspects of the change process that should not be merged into a sole and unique slope parameter, which would inevitably confound different aspects of change. A precise account of the components of a change process is required to test the effects of external covariates. Third, besides providing superior description of change processes, nonlinear functions may also allow for realistic extrapolations beyond the data, whereas simpler linear functions usually lead to unrealistic predictions outside of the range of the data.

Clearly most existing applications of GCM and LMM in the social sciences and humanities are linear, certainly also due to the greater software availability for linear compared to nonlinear models. Nevertheless, interest in nonlinear models for repeated-measures data is quickly increasing in many disciplines (e.g., Blozis, Conger, & Harring 2007; Boker, Schreiber, Pompe, & Bertenthal 1998; Grimm et al. 2011; Hall & Clutter 2004; Jordan, Daniels, Clarke, & He 2005; McArdle et al. 2002; Peek, Russek-Cohen, Wait, & Forseth 2002). At the same time, progress is being made in software development. We are confident that nonlinear models of change, such as the NLMM and NGCM, will become more easily estimable in the near future and will thereby increase in popularity in many disciplines, among which also the social sciences.

References

Beal, S. L ., & Sheiner, L. B. (1982). Estimating population kinetics. *Critical Review of Biomedical Engineering, 8*, 195–222.

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological Methods, 9*, 334–353.

Blozis, S. A., Conger, K., & Harring, J. (2007). Nonlinear latent curve models for longitudinal data. *International Journal of Behavioral Development, 31*, 340–346.

Boker, S. M., Schreiber, T., Pompe, B., & Bertenthal, B. I. (1998). Nonlinear analysis of perceptual-motor coupling in the development of postural control. In H. Kantz, J. Kurths, & G. Mayer-Kress (Eds.), *Nonlinear analysis of physiological data* (pp. 251–269). Berlin, Germany: Springer.

Braun, M. T., Kuljanin, G., & DeShon, R. P. (2013). Spurious relationships in the analysis of longitudinal data in organizational research. *Organizational Research Methods, 16*, 302–330.

Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions 2* (pp. 171–197). Amsterdam, The Netherlands: Elsevier Science Publisher.

Browne, M. W., & Du Toit, S. H. C. (1991). Models for learning data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 47–68). Washington, DC: American Psychological Association.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147–158.

Chihara, T. S. (1978). *An introduction to orthogonal polynomials*. New York: Routledge.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology, 58*, 615–637.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. London, UK: Chapman & Hall.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley-Interscience.

Durkin, M., Prescott, L., Furchtgott, E., Cantor, J., & Powell, D. A. (1995). Performance but not acquisition of skill learning is severely impaired in the elderly. *Archives of Gerontology and Geriatrics, 20*, 167–183.

Fekedulegn, D., Mac Siurtain, M., & Colbert, J. (1999). Parameter estimation of nonlinear growth models in forestry. *Silva Fennica, 33*, 327–336.

Ghisletta, P., Kennedy, K. M., Rodrigue, K. M., Lindenberger, U., & Raz, N. (2010). Adult age differences and the role of cognitive resources in perceptual-motor skill acquisition: Application of a multilevel negative exponential model. *Journal of Gerontology: Psychological Sciences, 65*, 163–173.

Ghisletta, P., & Lindenberger, U. (2004). Static and dynamic longitudinal structural analyses of cognitive changes in old age. *Gerontology, 50*, 12–16.

Ghisletta, P., & McArdle, J. J. (2001). Latent growth curve analyses of the development of height. *Structural Equation Modeling, 8*, 531–555.

Grimm, K. J., McArdle, J. J., & Hamagami, F. (2007). Nonlinear growth mixture models in research on cognitive aging. In K. van Monfort, H. Oud, & A. Satorra (Eds.), *Longitudinal models in the behavioural and related sciences* (pp. 267–294). Mahwah, NJ: Lawrence Erlbaum Associates.

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Developmental, 82*, 1357–1371.

Grimm, K. J., & Widaman, K. F. (2010). Residual structure in latent growth curve modeling. *Structural Equation Modeling, 17*, 424–442.

Hall, D. B., & Clutter, M. (2004). Multivariate multilevel nonlinear mixed effects models for timber yield predictions. *Biometrics, 60*, 16–24.

Hayduk, L. (2014). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement, 74*, 905–926.

Jordan, L., Daniels, R. F., Clarke, A. I., & He, R. (2005). Multilevel nonlinear mixed-effects models for the modeling of earlywood and latewood microfibril angle. *Forest Science, 51*, 357–371.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23*, 121–145.

Kennedy, K. M., Partridge, T., & Raz, N. (2008). Age-related differences in acquisition of perceptual-motor skills: Working memory as a mediator. *Aging, Neuropsychology, and Cognition, 15*, 165–183.

Kuljanin, G., Braun, M. T., & DeShon, R. P. (2011). A cautionary note on modeling growth trends in longitudinal data. *Psychological Methods, 16*, 249–264.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963–974.

Lance, C. E. (1988). Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Applied Psychological Measurement, 12*, 163–175.

McArdle, J. J. (1986). Latent growth within behavior genetic models. *Behavior Genetics, 16*, 163–200.

McArdle, J. J. (2001). Growth curve analysis. In N. J. Smelser & P. B. Baltes (Eds.), *The international encyclopedia of the behavioral and social sciences* (pp. 6439–6445). New York, NY: Pergamon Press.

McArdle, J. J., & Epstein, D. B. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*, 110–133.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology, 38*, 115–142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman and Hall.

Meredith, W., & Tisak, J. (1985, July). *"Tuckerizing" curves.* Paper presented at the Annual Meetings of Psychometric Society. Santa Barbara, CA.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122.

Nelder, J. A., & Wedderburn, R. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, 135*, 370–384.

Peek, M. S., Russek-Cohen, E., Wait, D. A., & Forseth, I. N. (2002). Physiological response curve analysis using nonlinear mixed models. *Oecologia, 132*, 175–180.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*, 12–35.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effect models in S and S-PLUS*. New York, NY: Springer.

Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D., EISPACK authors, & R-core. (2014, March). *nlme: Linear and nonlinear mixed effects models.* Available from http://cran.r-project.org/web/packages/nlme/index.html

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics, 14*, 1–17.

Raz, N., Williamson, A., Gunning-Dixon, F., Head, D., & Acker, J. D. (2000). Neuroanatomical and cognitive correlates of adult age differences in acquisition of a perceptual-motor skill. *Microscopy Research and Technique, 51*, 85–93.

Rovine, M. J., & Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research, 35*, 51–88.

Schnute, J. (1981). A versatile growth model with statistically stable parameters. *Canadian Journal of Fisheries and Aquatic Sciences, 38*, 1128–1140.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Sit, V., & Poulin-Costello, M. (1994). *Catalog of curves for curve fitting*. Victoria, BC: Ministry of Forestry.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19–23.

Vandergrift, N. A., Curran, P. J., & Bauer, D. J. (2002, August). *Methods for modeling nonlinearity in latent curve models.* Paper presented at the American Psychological Association. Chicago, IL.

von Oertzen, T. (2010). *Using the structured latent curve approximation for developmental simulations* (Tech. Rep.). Berlin, Germany: Max Planck Institute for Human Development.

Vonesh, E. F., & Chinchilli, V. M. (1996). *Linear and nonlinear models for the analysis of repeated measurements*. New York, NY: Dekker.

Wolfinger, R. D. (1999). *Fitting nonlinear mixed models with the new NLMIXED procedure*. Cary, NC: SAS Institute Inc.

Woodcock, R., & Johnson, M. B. (1989). *Woodcock-johnson psychoeducational battery - revised*. Chicago, IL: Riverside.

Yuancai, L., Marques, C. P., & Macedo, F. W. (1997). Comparison of schnute's and bertalanffy-richards' growth functions. *Forest Ecology and Management, 96*, 283–288.

Stage-Sequential Growth Mixture Modeling of Criminological Panel Data

Jost Reinecke, Maike Meyer, and Klaus Boers

Abstract The detection of distinctive developmental trajectories is of great importance in criminological research. The methodology of growth curve and finite mixture modeling provides the opportunity to examine different developments of offending. With latent growth curve models (LGM) (Meredith and Tisak, Psychometrika 55:107–122, 1990) the structural equation methodology offers a strategy to examine intra- and interindividual developmental processes of delinquent behavior. There might, however, not be a single but a mixture of populations underlying the growth curves which refers to unobserved heterogeneity in the longitudinal data. Growth mixture models (GMM) introduced by Muthén and Shedden (Biometrics 55:463–469, 1999) can consider unobserved heterogeneity when estimating growth curves. GMM distinguish between continuous variables which represent the growth curve model and categorical variables which refer to subgroups that have a common development in the growth process. The models are usually based on single-phase data which associate any event with a specific period. Panel data, however, often contain several relevant phases. In this context, stage-sequential growth mixture models with multiphase longitudinal data become increasingly important. Kim and Kim (Structural Equation Modeling: A Multidisciplinary Journal 19:293–319, 2012) investigated and discussed three distinctive types of stage-sequential growth mixture models: traditional piecewise GMM, discontinuous piecewise GMM, and sequential process GMM. These models will be applied here to examine different stages of delinquent trajectories within the time range of adolescence and young adulthood using data from the German panel study *Crime in the Modern City* (CrimoC, Boers et al., Monatsschrift für Kriminologie und Strafrechtsreform 3:183–202, 2014). Methodological and sub-

J. Reinecke (✉)
Faculty of Sociology, Bielefeld University, Postbox 100131, D-33501 Bielefeld, Germany
e-mail: jost.reinecke@uni-bielefeld.de

M. Meyer
Institute of Political Science, University of Münster, Scharnhorststraße 100, 48151 Münster, Germany

K. Boers
Faculty of Law, Universitätsstr. 14-16, 48143 Münster, Germany

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_3

stantive differences between single-phase and multi-phase models are discussed as well as recommendations for future applications.

Introduction

Investigations regarding the development of delinquency during the life course are currently of great importance in longitudinal criminological research. Over the past 20 years a variety of criminologists have argued that there are distinctive groups of offenders which can be described by different delinquent trajectories (Loeber & LeBlanc 1990; Moffitt 1993; Sampson & Laub 2003; Thornberry 2005).

> A trajectory is a pathway or line of development over the life span such as worklife, parenthood, and criminal behavior. Trajectories refer to long-term patterns of behavior and are marked by a sequence of transitions. Transitions are marked by life events (e.g. first job or first marriage) that are embedded in trajectories and evolve over shorter time spans. (Sampson & Laub 1997, p. 142)

Major methodological developments in criminological longitudinal research are influenced by the debate whether distinctive groups of criminal behavior can be identified and in which way the development of a "criminal career" can be incorporated in a statistical model. The debate is mainly enforced by Moffitt's dual taxonomy of offending behavior: The adolescent limited offenders exhibit antisocial behavior only during adolescence while life-course-persistent offenders begin to behave antisocially early in childhood and continue this behavior into adulthood (Moffitt 1993). In further analyses of data from the "Dunedin Multidisciplinary Health and Development Study" (Moffitt, Caspi, Rutter, & Silva 2011) four antisocial behavior trajectory groups were identified among females and males: life-course-persistent, adolescent-onset, childhood-limited, and low trajectory groups (Odgers et al. 2008). Furthermore, Nagin and collaborates explored population heterogeneity in behavioral trajectories using other longitudinal studies, like the "Cambridge Study" (Farrington & West 1990), the "Philadelphia Study" (Tracy, Wolfgang, & Figlio 1990), and the "Montreal Study" (Tremblay, Desmarais-Gervais, Gagnon, & Charlebois 1987). Depending on the type of the dependent variable, nature of the sample, and characteristics of the community, three to five trajectories were detected which reflect different intensity and growth of delinquency. These trajectories distinguish between non-offenders, a time-limited delinquent behavior through adolescence and a more or less chronic group of offenders (D'Unger, Land, McCall, & Nagin 1998; Nagin 1999; Nagin & Land 1993).

The reported findings suggest that there is a variety of heterogeneous trajectories which differ in the age of entry and exit in delinquency, its intensity, its duration, and its continuity (for an overview, see Piquero 2008). Furthermore the research on delinquent trajectories has shown that most people commit delinquent acts rarely or do not become delinquent at all. In most cases early intensive offenders desist from crime. There are, however, trajectories which are marked by high delinquency rates

or by increases in delinquency (Mariotti & Reinecke 2010; Moffitt 1993; Sampson & Laub 2003; Thornberry 2005). Moreover, research has shown that transition points are relevant for the analysis of delinquent trajectories (Sampson & Laub 1993).

Techniques of longitudinal statistical modeling are highly relevant and gained considerable attention in the examination of delinquent trajectories. With latent growth curve models (LGM) (Meredith & Tisak 1990), the structural equation methodology offers a strategy to examine intra- and interindividual developmental processes of delinquent behavior. It can, however, not be assumed that there is always a single population underlying the growth curves. Therefore, observed as well as unobserved heterogeneity has to be taken into account. Observed heterogeneity can be considered by relevant exogenous variables (e.g. gender) which are related to the growth curve variables explaining parts of their variances. To capture unobserved heterogeneity, the latent growth curve model has to be enlarged by a growth mixture model (Muthén & Shedden 1999) which contains the continuous manifest and latent variables as well but in addition categorical variables. The latter ones refer to particular subgroups reflecting different developmental processes. Analyses with a growth mixture model usually assume single-phase data which associate any event with a specific time period. However, longitudinal data often contain transition points which separate different phases of the development under study. An appropriate framework for multi-phase longitudinal data regarding unobserved heterogeneity is the extension of the growth mixture models (GMM) to stage-sequential growth mixture models (Kim & Kim 2012). These models can lead to a better understanding of the developmental process over several phases.

The following section "Method and Models" discusses basic conceptions of growth curve, growth mixture, and the multi-phase growth mixture models as well as the respective methods of model estimation and model evaluation. Additional attention is given to the distributional assumptions of the manifest time-variant variable under study. In the case of count variables Poisson or negative binomial distributions (Hilbe 2011) can be considered which give a better model representation compared to the assumption of a continuous distribution (Reinecke & Seddig 2011).

All models are applied to panel data from the German panel study *Crime in the Modern City* (CrimoC).[1] The data set contains 3938 adolescents and young adults who participated at least twice in a row in the eight panel waves. Data, variables and descriptive statistics are discussed in section "Data, Variables, and Descriptive Statistics".

Results are presented in section "Modeling Results". The analysis starts with single-phase growth mixture models considering up to eight classes and considers various specifications of stage-sequential growth mixture models. Finally, in section "Conclusion" models are compared and discussed with recommendations for further analyses.

[1]Principal investigators of the panel study are Klaus Boers (University of Münster) and Jost Reinecke (University of Bielefeld). Since 2002 the study is continuously funded by the German Science Foundation (DFG). Further information can be found at www.crimoc.org.

Method and Models

Latent Growth Curve Models

LGM specified with structural equations have already been discussed in several papers (McArdle 2009; McArdle & Epstein 1987; Meredith & Tisak 1990) and books (Bollen & Curran 2006; Duncan, Duncan, Strycker, Li, & Alpert 2006; Reinecke 2012). As is typical for all structural equation models, growth curve models distinguish between a measurement and a structural model. Structural model refers to the intraindividual development whereas the measurement model refers to interindividual differences of those trends. For a growth curve model with two factors, the measurement model can be formulated as follows:

$$y_t = \lambda_{1t}\eta_1 + \lambda_{2t}\eta_2 + \epsilon_t \tag{1}$$

y_t are the manifest variables at time t, which are related to the latent variables η_1 and η_2. η_1 is the initial level factor or intercept factor while η_2 is the linear growth factor or slope factor. λ_{1t} and λ_{2t} are the factor loadings on η_1 and η_2. ϵ_t is the measurement error of y_t. For each latent variable η, a structural equation has to be formulated as follows:

$$\eta_1 = \alpha_1 + \zeta_1 \tag{2}$$

$$\eta_2 = \alpha_2 + \zeta_2 \tag{3}$$

The latent variables η_1 and η_2 are described by their means (α_1 and α_2) as well as by their residuals (ζ_1 and ζ_2). ζ_1 and ζ_2 can be defined as deviations of the latent variables from their mean values.[2] Variances and covariances of the latent variables are specified in the matrix Ψ:

$$\Psi = \begin{pmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{4}$$

Assuming linear growth, the factor loadings for η_1 have to be fixed to one and the factor loadings for η_2 have to be restricted according to a linear development:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} * \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix} \tag{5}$$

[2]The structural equations can be extended by time-invariant latent variables which serve as predictors of the intercept and slope (e.g. gender). Then, ζ_1 and ζ_2 are no longer deviations from the mean values of the latent variables η_1 and η_2 (Reinecke 2012, p. 6).

To model non-linear growth curves it is possible to extend the two-factor model by additional latent variables, for instance a quadratic term. The measurement and structural equations (1)–(3) of the two-factor model described above can be extended as follows:

$$y_t = \lambda_{1t}\eta_1 + \lambda_{2t}\eta_2 + \lambda_{3t}^2\eta_3 + \epsilon_t \tag{6}$$

$$\begin{aligned} \eta_1 &= \alpha_1 + \zeta_1 \\ \eta_2 &= \alpha_2 + \zeta_2 \\ \eta_3 &= \alpha_3 + \zeta_3 \end{aligned} \tag{7}$$

Another possibility to cope with nonlinearity is the so-called piecewise growth curve model which is useful when transition points are assumed across the time range (Bollen & Curran 2006). Such a model contains two or more latent variables. Contrary to the linear growth model, those models can be used to analyze multiphase data. Piecewise growth curve models are meaningful when transition points can be found in the course of development (see, for instance, Raudenbush & Bryk 2002). Assuming one transition point, the first trajectory describes the development between the intercept and the transition point. The second trajectory describes development after the transition point. If six panel waves and a transition point for the third panel wave are assumed, the following measurement model can be formulated:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} * \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix} \tag{8}$$

η_1 is the intercept, η_2 is the linear slope for the first phase, and η_3 is the linear slope for the second phase. Because of the transition point at the third panel wave, the restricted values of factor loadings of η_1 are not changing for the subsequent waves.

Growth Mixture Models for Single-Phase Data

With GMM it is possible to control for unobserved heterogeneity in the data. If the variances of the growth factors in a linear or piecewise growth curve model are not different from zero, growth mixture modeling is not necessary. The GMM extends Eqs. (1)–(3) by a categorical variable c with $k = 1, 2, \ldots, K$ classes. Assuming a two-factor growth mixture model, the following measurement and structural equations can be formulated:

$$y_{tk} = \lambda_{1tk}\eta_{1k} + \lambda_{2tk}\eta_{2k} + \epsilon_{tk} \tag{9}$$

$$\eta_{1k} = \alpha_{1k} + \zeta_{1k} \tag{10}$$

$$\eta_{2k} = \alpha_{2k} + \zeta_{2k} \tag{11}$$

Means and variances of the latent variables are estimated for each class k ($\alpha_{1k}, \alpha_{2k}, \psi_{11k}, \psi_{22k}$). The matrix Ψ_k contains the class-specific variances and covariances:

$$\Psi_k = \begin{pmatrix} \psi_{11k} & \\ \psi_{21k} & \psi_{22k} \end{pmatrix} \tag{12}$$

The so-called latent class growth analysis (LCGA; Muthén 2004) is a submodel of the GMM which gained great importance in criminological research under the name *group-based trajectory modeling* (Nagin 2005). LCGA assumes that variances and covariances of the growth factors are restricted to zero ($\Psi_k = 0$). Consequently, there are no residual terms in the structural equations of the latent variables η_1 and η_2 and therefore all class members are treated as homogenous regarding their individual developments:

$$\eta_{1k} = \alpha_{1k} \tag{13}$$

$$\eta_{2k} = \alpha_{2k} \tag{14}$$

Previous analyses of delinquent trajectories with longitudinal data show that specifications of the LCGA lead to quite reasonable substantive results (see, for instance, Kreuter & Muthén 2008). From a methodological point of view Muthén (2004, p. 350) suggests using LCGA as starting point for the analysis of trajectories, because it can be explored how many different classes might be necessary to estimate distinct developmental trends appropriately.

In most criminological studies the longitudinal response variable is a count measure (e.g., the number of convictions). Therefore, the Poisson regression model as a special case of the generalized linear model has to be used. Let y_i be the number of observed count occurrences, x_i the vector of covariates, and ν_i the expected number of counts. The number of events in an interval of a given length is Poisson distributed and the Poisson regression model can be formulated via a log link function (Hilbe 2011, p. 31):

$$Pr(y_i|x_i) = exp(-\nu_i)\nu_i^{y_i}/y_i! \tag{15}$$

with $\nu_i = exp(\alpha + x_i'\beta)$. β is the vector of regression coefficients. The conditional mean function of the Poisson distribution is $E(y_i|x_i) = \nu_i$ with its equidispersion $Var(y_i|x_i) = \nu_i$. Small values of ν_i indicate the rarity of the event and the skewness of the distribution.

If the assumption of equidispersed data does not hold, the negative binomial regression model can be employed by introduction of latent heterogeneity in the conditional mean of the Poisson model (Hilbe 2011, p. 185):

$$Pr(y_i|x_i, \epsilon_i) = exp(\alpha + x_i'\beta + \epsilon_i) = h_i v_i \tag{16}$$

where $h_i = exp(\epsilon_i)$ is assumed to have a one parameter gamma distribution, $G(\theta, \theta)$ with a mean equal to 1 and variance $\kappa = 1/\theta$. The negative binomial distribution can be obtained by integrating h_i out of the joint distribution. The conditional mean function is still $E(y_i|x_i) = v_i$ while overdispersion can be obtained from the latent heterogeneity with the variance function $Var(y_i|x_i) = v_i^2[1 + (1/\theta)]$.

Within the context of the CrimoC study previous analyses of GMM using the assumption of a negative binomial distributed variable have shown that those models have always better model fits compared to models with the assumption of a continuous or Poisson distributed variables (Reinecke & Seddig 2011, p. 432). Therefore the negative binomial distribution assumption will be used for the current analyses.

Growth Mixture Models for Multi-Phase Data

The discussed mixture models always assume that every estimated trajectory relies on longitudinal data covering a single phase of development. In case of long repeated panel designs this assumption might not be appropriate. The larger the time span of the longitudinal data, the higher is the chance that modeling of different phases is necessary to estimate the trajectories of the particular latent classes. The difference between single-phase and multi-phase data does not depend on specific features of a panel design but on whether transitions points are likely between the particular measurement occasions.

For homogenous populations piecewise growth curve models, as discussed above, are able to consider transition points. In case of unobserved heterogeneity the piecewise growth curve model can be extended to a so-called *Traditional Piecewise Growth Mixture Model* (TPGMM, Kim & Kim 2012, p. 300). TPGMM has multiple growth components and additionally one mixture component. The growth components are the same as for piecewise growth models whereas the finite mixture component is the same as for the GMM. The growth trajectories before and after a transition point are connected at the transition point. Figure 1 illustrates the model assumption: y_1–y_8 are the measures for eight panel waves, I is the intercept and $S1$ as well as $S2$ are the particular slopes. The first and second growth trajectory are connected at the transition point (e.g. t_6). c represents the mixture component, X is a time-invariant exogenous variable, U represents outcome variables. Both X and U will not be considered in the applications (cf. section "Modeling Results").

If a larger change or a discrepancy (e.g. intervention) is expected at the transition point, the TPGMM might not be sufficient to model this effect. One possible

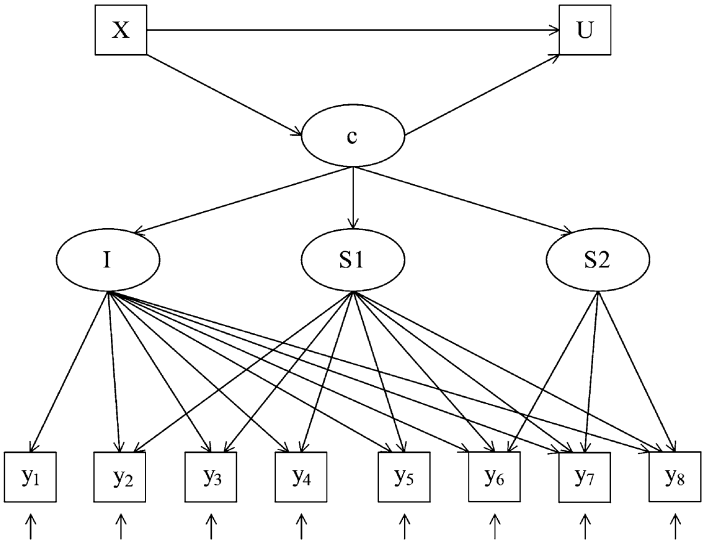


Fig. 1 Traditional Piecewise Growth Mixture Model. *Source*: Kim and Kim (2012, p. 297)

extension of the TPGMM is the so-called *Discontinuous Piecewise Growth Mixture Model* (DPGMM, Kim & Kim 2012, p. 301) in which an intercept is specified for each phase. Figure 2 shows an example with eight panel waves and a transition point between the fourth and fifth measurement: y_1–y_4 are the first-phase measures, y_5–y_8 are the second-phase measures, $I1$ is the first intercept and $S1$ is the first slope, $I2$ is the second intercept and $S2$ is the second slope. All other variables are the same as in Fig. 1. In difference to the TPGMM the trajectories of the first and the second phases are not directly connected at the transition point.

Introducing a second intercept changes the measurement part of the DPGMM compared to the TPGMM while the structural part remains the same. The measurement part of the model is given as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1\,0\,0\,0 \\ 1\,1\,0\,0 \\ 1\,2\,0\,0 \\ 1\,3\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,1\,1 \\ 0\,0\,1\,2 \\ 0\,0\,1\,3 \end{bmatrix} * \begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \end{bmatrix} \tag{17}$$

η_{11} and η_{12} are the intercepts for the first and the second phase whereas η_{21} and η_{22} are the particular slopes. Both the TPGMM and the DPGMM assume

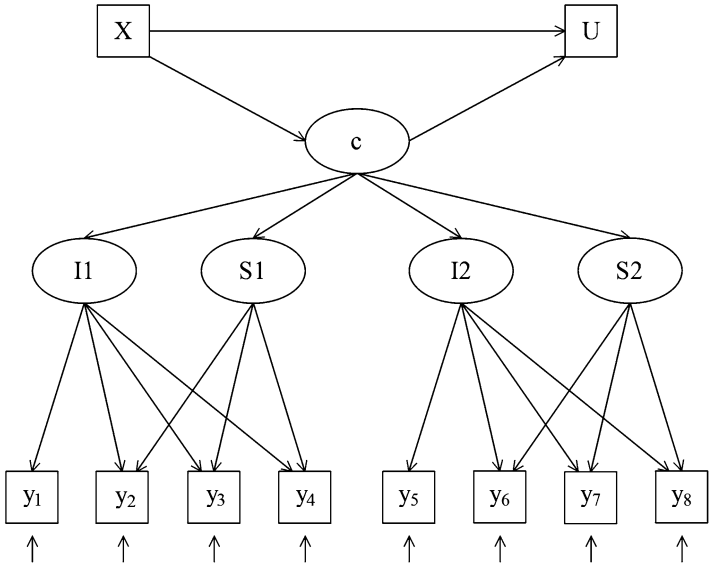


Fig. 2 Discontinuous Piecewise Growth Mixture Model

a common mixture component c for all phases. If in addition changes between the classes due to the transition between the phases have to be considered, the DPGMM can be extended to a so-called *Sequential-Process Growth Mixture Model* (SPGMM, Kim & Kim 2012, p. 303). Transition points as well as changes between latent class membership can be applied with the SPGMM. Figure 3 shows an example with the two mixture components $c1$ and $c2$. The relationship between both mixture components is specified via a transition probability matrix which contains the estimates of the probability of latent class membership of the second phase, conditional on latent class membership at the first phase. The number of intercepts and slopes and the specifications of the measurement part of the model are equal to the DPGMM.

Model Estimation and Model Evaluation

Mixture models are estimated by maximizing the log-likelihood function within the admissible range of parameter values given classes and data. The program M*plus* employs the EM-algorithm for maximization (Dempster, Laird, & Rubin 1977; Muthén & Shedden 1999). Thereby, different sets of starting values are tested for the calculation of the optimal function value and the best set is used for the estimation of the parameters. For a given solution, each individual's probability of membership in each class is estimated. Individuals can be assigned to the classes by

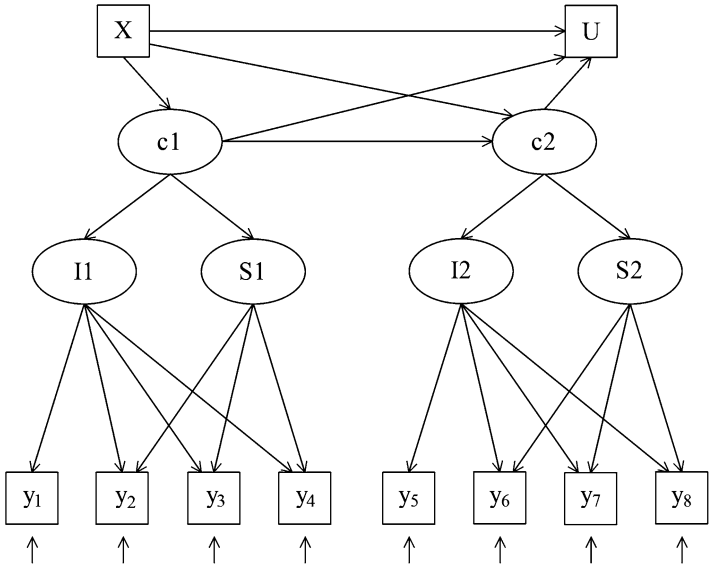


Fig. 3 Sequential-Process Growth Mixture Model

calculating the posterior probability that an individual i belongs to a given class k. Each individual's posterior probability estimate for each class is computed as a function of the parameter estimates and the values of the observed data. The number of classes has to be specified in each model variant.

Standard errors of estimates are asymptotically correct if the underlying mixture model is the true model. χ^2-differences between the particular mixture model variants, however, cannot be calculated because a k-class model is not nested within a $k + 1$-class model. Therefore, the Bayesian Information Criterion (BIC, Schwarz 1978) is used for model comparisons. Furthermore, M*plus* calculates a sample size adjusted BIC which was found to give superior performance for model selection (adj. BIC, Yang 1998). Models with the lowest BIC or adjusted BIC can be selected for further substantial interpretations. But accepting or rejecting a model on the basis of the BIC is more or less descriptive and does not imply any statistical test.

However, Lo, Mendell, and Rubin (2001) have developed a statistical test for mixture models. The so-called Lo-Mendell-Rubin likelihood test (LMR-LRT) tests a k-class model against a $k - 1$-class model. Thereby the relation of the likelihoods of a $k - 1$-class model to the ones of a k-class model is calculated. If the p-value of the test is small, the k-class model should be accepted (Reinecke 2012, p. 38). The LMR-LRT can only be calculated for GMM and TPGMM.

In addition, the entropy of a particular mixture model can be used to decide about the adequate number of classes. Entropy is a summary measure of classification quality based on the estimated posterior probabilities that ranges from zero to one:

$$E_k = 1 - \frac{\Sigma_i \Sigma_k (-\hat{p}_{ik} ln \hat{p}_{ik})}{n \; lnK} \tag{18}$$

$\hat{p}_{ik}$ is the estimated probability for each individual i to be in class k. The closer its values are to one, the better the classification.

BIC, adjusted BIC, LMR-LRT, Entropy, and the substantive interpretability of the classes should be considered for the decision process (Muthén & Muthén 2000). In the context of multi-phase mixture models three additional aspects have to be taken into account (Kim & Kim 2012, 305f): At first, the number of latent classes in each phase should be kept as small as possible, second,for multiple latent classes nearly empty patterns can be accepted (e.g., as outliers) and finally, redundant classes should be avoided as well as classes which are misleadingly omitted. All in all, it is advisable to make the decision about the number of latent classes not only on the basis of one information source, but include various statistical and substantive arguments (see also Kim 2014).

Data, Variables, and Descriptive Statistics

The data used for the current analyses are taken from the panel survey of Duisburg which is part of the ongoing German panel study CrimoC. Duisburg is an industrial city of about 500,000 inhabitants. It is located in the western part of the Ruhr area in Germany. The sample was drawn from secondary schools in Duisburg. Eight annual panel waves have been collected between 2002 and 2009, which covers the period from early to late adolescence. The self-administered questionnaires were completed in school classes as long as the students attended the particular schools. After leaving school participants were usually contacted by mail. If repeated contacts were unsuccessful personal contacts were realized to conduct the interviews. Retention rates are between 84 and 92 % (Boers et al. 2014, p. 184).

The panel data contain individuals who participated twice in a row between 2002 and 2009 (n=3938). Table 1 gives descriptive information about each panel wave (age, sex). In the first panel wave (2002) the sample's average age is 13, in 2009 it is about 20 years. The sex ratio in each panel wave is relatively balanced although there are always more female than male participants. In 2002, for instance, 48.6 % of the respondents are male and 51.4 % female. In the subsequent panel waves, however, there are larger differences. In 2009 only 42.2 % of the respondents are male. Therefore, females are slightly overrepresented in the data.

To measure deviant and delinquent behavior, about 15 different offenses are obtained in the questionnaires of each panel wave. These offenses can be classified into property offences (burglary, theft of cars, theft out of cars; fencing, theft out of vending machines, theft of bicycles, shoplifting), violent offences (robbery, purse snatching, assault with a weapon; assault), and criminal damage offences (graffiti, scratching, other criminal damage). Concerning each of those offences, the respondents were asked whether they ever committed it (lifetime prevalence)

and whether they have committed them in the past year (annual prevalence). If they committed the particular offence in the past year, they were also asked about the frequency of their offending (annual incidence). The time-variant dependent variable of the mixture models considers the annual incidence rates which is given as the sum of the particular rates of the 15 offences.

Table 2 gives a descriptive overview of the distributions of prevalence and incidence rates. The prevalence of the self-reported criminal behavior increases in early adolescence between 2002 and 2003 and decreases later on. The peak is reached in 2003 in which the adolescents were about 14 years old. In the year 2002 nearly 31 % of the respondents reported an offence. This rate increased to 40 % in 2003 and decreased continuously down to 7 % in 2009.

Incidence mean rates are based on the number of persons who reported at least one offence in the prevalence measure. Some of the respondents gave an answer to the annual prevalences but not to the annual incidences. Therefore the numbers of persons are slightly different for each of the eight panel waves. The lower half of Table 2 shows the means of the annual incidences of the offenders. The first row of means are based on the number of valid answers in each panel wave. The second row are the means estimated via the Full Information Maximum Likelihood procedure (FIML, Enders 2010, p. 88) considering unit nonresponses in each panel

Table 1 Descriptive information about the sample

Year		2002	2003	2004	2005	2006	2007	2008	2009
n		2683	3094	3105	3140	2989	2577	2410	2299
Age		13	14	15	16	17	18	19	20
Sex	Male	48.6	48.9	47.8	48.4	45.5	43.2	43.4	42.2
	Female	51.4	51.1	52.2	51.6	54.5	56.9	56.6	57.8

Table 2 Annual prevalence and incidence rates

Year	2002	2003	2004	2005	2006	2007	2008	2009
	Prevalence rates (percentages)							
n	2683	3094	3105	3140	2989	2577	2410	2299
Property	18.5	24.3	21.7	17.4	12.9	8.3	5.0	4.6
Serious	2.9	4.3	4.2	3.7	2.1	0.9	0.7	0.3
Minor	17.9	23.3	21.0	16.6	12.3	7.8	4.7	4.5
Violence	13.9	19.2	15.0	13.2	9.6	5.4	4.1	2.6
Serious	3.8	6.2	4.8	4.3	2.5	1.1	0.8	0.5
Minor	12.8	16.7	13.3	11.8	8.8	5.1	3.6	2.2
Damage	16.9	23.3	18.9	13.8	9.5	5.4	3.0	1.6
All offences	30.6	39.1	35.1	28.6	22.4	14.3	9.6	7.3
	Incidence rates (means)							
n	2483	2692	2783	2838	2751	2464	2282	2205
Mean	2.78	4.83	5.28	4.26	3.32	1.74	1.70	0.77
n	3938	3938	3938	3938	3938	3938	3938	3938
Mean (FIML)	2.92	5.24	5.97	4.78	3.73	2.13	1.92	0.81

FIML Full Information Maximum Likelihood

wave. The FIML estimated means are slightly higher compared to those based on the complete cases in each panel wave which reflects a certain underreporting of the incidence rates (see also Reinecke & Weins 2013). Nevertheless, both rows of means reflect the typical development of adolescents' delinquent behavior with the peaks at age 15 (year 2004) and a continuous decline thereafter. FIML estimated means, variances, and covariances are used for the GMM in section "Modeling Results".

Modeling Results

With different slope specifications variants of the TPGMM (see Fig. 1) are firstly evaluated. One specification assumes three phases with one turning point at the second panel wave and another turning point at the sixth panel wave. The factor loadings of the intercept and the three linear slopes are restricted as follows:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0.5 & 0 \\ 1 & 1 & 1.5 & 0 \\ 1 & 1 & 2.5 & 0 \\ 1 & 1 & 3.5 & 0 \\ 1 & 1 & 3.5 & 20.25 \\ 1 & 1 & 3.5 & 30.25 \end{pmatrix} \tag{19}$$

The first linear slope (second column in the matrix) specifies the first turning point. Therefore subsequent factor loadings are restricted to the value of one. The second linear slope (third column) specifies the continuous development of delinquency up to the sixth panel wave with a difference value of one. Therefore subsequent factor loadings are fixed to the value of 3.5. The third slope (fourth column) reflects a faster development by doubling the value of 3.5 with additional constants ($3.5^2 + 8 = 20.25$ and $3.5^2 + 10 = 30.25$). These fixed values were previously explored by different model specifications of the piecewise growth curve model.

Alternatively, a more parsimonious specification assumes only two slopes and a faster development of delinquency after the second wave. The factor loadings of the intercept and the two linear slopes are restricted as follows:

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0.25 \\ 1 & 1 & 2.25 \\ 1 & 1 & 6.25 \\ 1 & 1 & 12.25 \\ 1 & 1 & 20.25 \\ 1 & 1 & 30.25 \end{pmatrix} \tag{20}$$

The restrictions of the factor loadings of the first slope (second column in the matrix) are equal to the previous specification in Matrix (19). The second slope (third column) specifies the continuous development of delinquency by adding the constants 2, 4, 6, 8, and 10 to the value of 0.25. So, the factor loadings of the second slope for the last two panel waves do not differ to the factor loadings from the third slope in Matrix (19). In general, restrictions of the factor loadings influence the form of the trajectory pieces, the direction of the development (increase or decrease of delinquency) can only be observed from the sign of the particular slope mean estimators (see vector α in Eq. (7)).

TPGMM are calculated from two up to eight classes. Incidence rates are treated as a negative binomial distributed count variable (cf. Eq. (16)). Intercept and slopes are specified according to LCGA, i.e., all variances and covariances of the growth curve variables are fixed to zero (cf. section "Growth Mixture Models for Single-Phase Data"). Table 3 shows the particular fit information for the TPGMM with three and two linear phases according to the specifications in Matrices (19) and (20).

All the BIC and adjusted BIC values of the models with two phases are lower than the particular models with three phases. It clearly shows that the development of delinquency can be modelled sufficiently well with two phases: one for the increase and one for the decrease. Regarding the TPGMM with two phases the *p*-value of the LMR-LRT shows no redundancy up to six classes.

Table 4 and Fig. 4 give an overview of the model. The largest class in this model represents a group of adolescents who were nearly not involved in delinquent behavior during the observed period (non-offenders, 49.9 %). The second largest class is characterized by a slight increase in the early adolescence and a likewise

Table 3 Model fit information of the TPGMM with three and two phases

Class	Parameter	E_k	BIC	adj. BIC	LMR-LRT	*p*-Value
Three phases						
2	17	0.699	46,116	46,062	2182	0.00
3	22	0.624	45,656	45,586	490	0.00
4	27	0.562	45,538	45,452	168	0.00
5	32	0.555	45,462	45,360	116	0.01
6	37	0.574	45,445	45,328	56	0.26
7	42	0.569	45,437	45,304	45	0.91
8	47	0.551	45,441	45,291	37	0.00
Two phases						
2	15	0.699	46,098	46,050	2159	0.00
3	19	0.621	45,631	45,570	485	0.00
4	23	0.612	45,470	45,397	187	0.01
5	27	0.552	45,402	45,316	99	0.01
6	31	0.546	45,374	45,276	59	0.05
7	35	0.538	45,366	45,255	39	0.42
8	39	0.543	45,368	45,244	24	0.13

Table 4 Means of the growth variables (TPGMM)

Class	Variable	Mean	Standard error	*z*-Value
Class 1	I	−3.180	0.413	−7.698
Non−offenders	S1	−0.833	0.542	−1.537
(n=1966, 49.9 %)	S2	−0.055	0.027	−2.052
Class 2	I	1.253	0.435	2.878
Adolescence−limited	S1	0.729	0.333	2.191
(n=662, 16.8 %)	S2	−0.092	0.017	−5.292
Class 3	I	0.741	0.417	1.779
Low−level−decliners	S1	0.067	0.343	0.196
(n=530, 13.5 %)	S2	−1.545	0.531	−2.909
Class 4	I	−1.331	0.454	−2.931
Low−rate−offenders	S1	0.838	0.400	2.096
(n=300, 7.6 %)	S2	−0.008	0.033	−0.245
Class 5	I	2.705	0.178	15.167
Persistent offenders	S1	0.827	0.216	3.825
(n=270, 6.9 %)	S2	−0.033	0.013	−2.555
Class 6	I	2.231	0.315	7.076
High−level−decliners	S1	0.797	0.290	2.745
(n=210, 5.3 %)	S2	−0.841	0.076	−11.090

I intercept, *S1* first slope, *S2* second slope

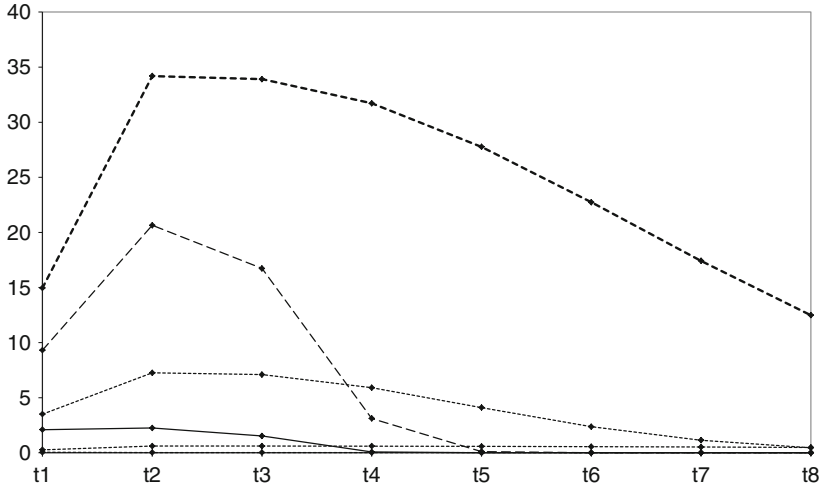


Fig. 4 Traditional Piecewise Growth Mixture Model with six classes. Labels of the classes (from the *bottom* to the *top*): non-offenders, low-rate-offenders, low-level-decliners, adolescence-limited-offenders, high-level-decliners, persistent offenders

slight decrease later on. Here, delinquency was limited to the period of adolescence (adolescence-limited offenders, 16.8 %). The third largest group comprises adolescents who committed crimes just in early adolescence (low-level-decliners, 13.5 %). The following class is a group of adolescents who reported only a few crimes during the observed period (low-rate-offenders, 7.6 %). Only a small proportion of the adolescents can be classified as persistent offenders with a large incidence rate (6.9 %). A likewise small proportion is characterized by a high crime rate in early adolescence and a low crime rate later on (high-level-decliners, 5.3 %).

This type of growth trajectory, however, can distort real growth patterns in data, when a more dynamic change or a discrepancy is expected at the transition point.

The TPGMM, however, "can distort real growth patterns in data, when a more dynamic change or a discrepancy is expected at the transition point" (Kim & Kim 2012, p. 300). The DPGMM (see Fig. 2) contains intercepts for each phase. For the substantive application a DPGMM would be assumed to have one intercept and slope for the increase of delinquency as well as one intercept and slope for the decrease of delinquency. With this model a larger discrepancy at the transition is expected which means a sufficient discontinuity between the phases. However, previous analyses with the CrimoC panel data did not support a discontinuity of the developmental process and therefore the specification of a DPGMM was rejected.

As described in section "Growth Mixture Models for Multi-Phase Data" the SPGMM extends the DPGMM by additional latent class variables assuming that the number of classes can change between the phases. According to the results of the DPGMM and in difference to Fig. 3 we do not assume two but only one intercept for the phases. In addition, it is proposed that the number of classes will decrease over time. Substantively this means that a larger unobserved heterogeneity of the trajectories is expected in early adolescence compared to late adolescence. With increasing age and increasing desistance from crime a smaller unobserved heterogeneity is expected. Similar to the TPGMM, the models are tested with three and two phases. According to the assumption of decreasing heterogeneity the number of classes is always higher in the first phase compared to the subsequent phases. Table 5 shows the model fit information for the calculated SPGMMs. Model selection is limited to the BIC and adjusted BIC (LMR-LRT is not calculated in M*plus* when different class patterns are specified). Similar to the TPGMM, results show that two phases are sufficient. The model with three classes in the first phase and two classes in the second phase can be selected for further interpretations.

The combination of the first and the second phase leads to a six-class pattern with different combinations of classes (see Table 6):

1. Class pattern 1 1: 12.3 % of the adolescents change from low-rate-offenders in the early adolescence to non-offenders later on.
2. Class pattern 1 2: 4.1 % of the adolescents are characterized by a high and increasing delinquency rate in early adolescence and a declining delinquency rate later on.
3. Class pattern 2 1: Nearly half of the adolescents are characterized as non-offenders in both phases.

Table 5 Model fit information of the SPGMM with three and two phases

Class			Parameter	E_k	BIC	adj. BIC
Three phases						
1. Phase	2. Phase	3. Phase				
2	1	1	17	0.614	46,116	46,062
2	2	1	26	0.563	45,532	45,450
3	2	1	35	0.552	45,428	45,317
Two phases						
2	1	–	15	0.699	46,098	46,050
2	2	–	19	0.621	45,631	45,570
3	1	–	22	0.614	45,477	45,407
3	2	–	29	0.538	45,365	45,272

Table 6 Number and proportion of persons in the class patterns (SPGMM)

Class	Class pattern	n	%
1	1 1: low-rate-offenders → non-offenders	484	12.3
2	1 2: high starters → decliners	162	4.1
3	2 1: non-offenders	1888	47.9
4	2 2: early increasers → decreasers	759	19.3
5	3 1: low-rate-offenders	415	10.5
6	3 2: high starters → persisters	230	5.8
Latent variable	Composition	n	%
C1	1: early starters/high starters (484 + 162)	646	16.4
(Phase 1)	2: non-offenders/early increasers (1888 + 759)	2647	67.2
	3: low rate offenders/high starters (415 + 230)	645	16.4
C2	1: non-offenders/low-rate-offenders (1888 + 415)	2787	70.8
(Phase 2)	2: decliners/decreasers/persisters (162 + 759 + 230)	1151	29.2

4. Class pattern 2 2: 19.3 % of the adolescents show a slight increase in early adolescence and a slight decrease later on.
5. Class pattern 3 1: 10.5 % of the adolescents are characterized as low-rate-offenders in both phases.
6. Class pattern 3 2: 5.8 % of the adolescents show persistent delinquency on a high level with a decreasing tendency in late adolescence.

The first phase is characterized by three classes. The first one comprises adolescents who started to behave delinquently early in the adolescence and partly on a high level (16.4 %). The second class encompasses non-offenders and adolescents whose crime rate increases slightly on a low level (67.2 %). Finally, low rate offenders and high starters can be found in the third class. The second phase comprises two classes. The first of them encompasses non- and low-rate offenders, the second one adolescents with decreasing delinquency (see Fig. 5).

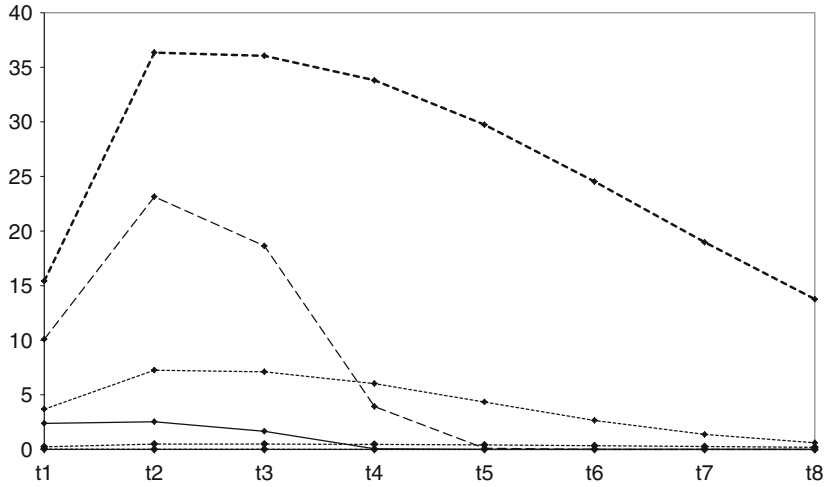


Fig. 5 Sequential-Process Growth Mixture Model. Labels of the classes (from the *bottom* to the *top*): non-offenders, low-rate-offenders, low-rate-offenders → non-offenders, early increasers → decreasers, high-starters → decliners, high-starters → persisters

With the two latent class variables C1 and C2 the SPGMM is able to estimate transition probabilities. With a probability of 66 % it is more likely for adolescents to stay as or to become non- or low-rate-offenders during the life course than to still act delinquent in late adolescence. Quite a few adolescents, however, commit crimes in late adolescence as well.

One possibility to compare and validate the results of different mixture model specifications is to look at the bivariate table with the particular proportions for the latent classes based on their most likely latent class membership. The estimated TPGMM contains six classes and one latent class variable, the estimated SPGMM also contains six classes which can be differentiated into different class patterns. Table 7 gives the result of the crosstabulation between the class distribution of the TPGMM and the SPGMM. Most of the individuals in pattern 1 1 of the SPGMM belong to the third class of the TPGMM (low-level decliners, 96.07 %), nearly all individuals in pattern 1 2 of the SPGMM belong to the sixth class of the TPGMM (high-level decliners, 99.38 %). Differences between these two patterns (SPGMM) or classes (TPGMM) refer only to the level of delinquency, both patterns or classes are characterized by processes of desistance.

Non-offenders in pattern 2 1 of the SPGMM are 100 % part of class 1 of the TPGMM. Pattern 2 2 of the SPGMM is characterized by processes of early increasing and later declining delinquency. Eighty-six percent of these individuals belong to class 2 of the TPGMM (adolescent-limited offenders). The rest of pattern 2 2 is distributed across the other classes of the TPGMM. The lowest congruence to class 4 of the TPGMM has pattern 3 1 of the SPGMM (low-rate offenders). Only 68 % of the individuals are in the particular cell of the cross-table. Nearly 17 %

Table 7 Individual latent class membership of the TPGMM (row) and SPGMM (column)

Class	1 1 low-rate-offenders → non-offenders	1 2 high starters → decliners	2 1 non-offenders	2 2 early increasers → decreasers	3 1 low-rate-offenders	3 2 high starters → persisters	∑
1	9	0	1888	0	69	0	1966
Non-offenders	0.46	0.00	96.03	0.00	3.51	0.00	100.00
	1.86	0.00	100.00	0.00	16.63	0.00	49.92
2	2	0	0	653	7	0	662
Adolescence-limited	0.30	0.00	0.00	98.64	1.06	0.00	100.00
	0.41	0.00	0.00	86.03	1.69	0.00	16.81
3	465	0	0	23	42	0	530
Low-level-decliners	87.74	0.00	0.00	4.34	7.92	0.00	100.00
	96.07	0.00	0.00	3.03	10.12	0.00	13.46
4	0	0	0	16	284	0	300
Low-rate-offenders	0.00	0.00	0.00	5.33	94.67	0.00	100.00
	0.00	0.00	0.00	2.11	68.43	0.00	7.62
5	0	1	0	39	0	230	270
Persistent offenders	0.00	0.37	0.00	14.44	0.00	85.19	100.00
	0.00	0.62	0.00	5.14	0.00	100.00	6.86
6	8	161	0	28	13	0	210
High-level-decliners	3.81	76.67	0.00	13.33	6.19	0.00	100.00
	1.65	99.38	0.00	3.69	3.13	0.00	5.33
∑	484	162	1888	759	415	230	3.938
	12.29	4.11	47.94	19.27	10.54	5.84	100.00
	100.00	100.00	100.00	100.00	100.00	100.00	100.00

of the pattern belongs to the non-offender class 1 and about 10 % to the low-level declining class 3 of the TPGMM. Similar to the class of non-offenders pattern 3 2 (persistent offenders) of the SPGMM are 100 % part of class 5 of the TPGMM. In total, the crosstabulation of both class memberships confirms the stability of the latent class distributions although the specifications of TPGMM and SPGMM are different. Non-offenders and low-rate offenders have overlaps in their particular developments and therefore their assignments can differ between the models. This has been observed in previous applications of GMM with criminological panel data (see, for example, Mariotti & Reinecke 2010; Piquero 2008; Reinecke & Seddig 2011).

Conclusion

This study has shown that with an increasing number of panel waves unobserved heterogeneity of developmental processes results not only from a mix of these developments but also from multiple phases. In difference to the TPGMM, the SPGMM has separate mixture parts with a latent class variable in each phase. Whereas the TPGMM has only one intercept over multiple phases, the DPGMM and SPGMM specify separate intercepts as well as separate slopes. Kim and Kim (2012) showed how growth and mixture models can be extended to more complex and flexible stage-sequential growth mixture models within the structural equation modeling framework. Their model applications contain continuous data related to smoking behavior. In the present study the observed variable was treated as a count variable with overdispersion. Therefore piecewise and stage-sequential growth mixture models have been applied with the specification of a negative binomial distributed variable. But all the analyses are limited to the LCGA specification meaning that no variances and covariances were estimated for the growth curve variables within classes.

With eight panel waves of self-reported delinquency obtained from the CrimoC study (Boers et al., 2014) separate intercepts could not be detected and identified while separate growth components reflect increase and decrease of delinquency through the period of adolescence and young adulthood. If only one intercept is required, the specification of the DPGMM collapses to the TPGMM. One possible explanation is that the CrimoC study contains no experimental intervention and therefore the different trajectory pieces do not reflect phases of discontinuity.

Starting with the single-phase TPGMM six distinct classes of delinquent developments could be identified: non-offenders who were nearly not involved in delinquent behavior at all, adolescent-limited offenders with the typical development of the age-crime curve, low-level-decliners who limited their delinquency in early adolescence, low-rate-offenders who reported only a few crimes during the panel study, persistent offenders with the largest incidence rate compared to the other classes and high-level decliners with a high crime rate in early adolescence and a declining tendency later on. A specification with six classes could also be verified

with the multiple-phase SPGMM. Different models with two and three phases were tested and compared. The first phase can be characterized by the development of delinquency in early adolescence, the second phase by the development in late adolescence. A possible third phase belongs to the period of young adulthood which might be detected with further panel waves. The specification of the different SPGMM variants assume always a higher number of classes in the first phase compared to the second or third phase. Heterogeneity of the development of delinquency is expected to be higher in the first panel waves and decreases thereafter. On the average this assumption was confirmed. The number of offenses decreases over time and the development of delinquency tends to be homogenized. Two class patterns of the final SPGMM are expected to be stable across the phases: the non-offenders and the low-rate offenders. One pattern shows the transition from low-rate to non-offending, two patterns show the transition from high starters to decliners or persisters and another pattern is characterized by the transition of early increasing to later decreasing delinquency. Transition parameters between the phases show that the probability to stay as or to become a non- or low-rate-offender is much higher than to persist as a delinquent persons during the life course. The crosstable of the most likely latent class memberships of the TPGMM and the SPGMM reflects the stability of the classification and serve as a proof of quality for the substantive interpretations.

Although the applications of the single and multi-phase mixture models is very useful for the longitudinal criminological research technique in various fields, some unresolved issues have to be mentioned. The complexity of the models requires not only large sample sizes but also a large number of starting values. In the initial stage, 500 random sets of starting values were generated and optimized for each of the sets. The ending values of 20 optimizations with the highest log-likelihoods were used as starting values in the final stage. With the assumption of a negative binomial distribution stable results could only be obtained with the LCGA specification. Evaluation of model fit is not the same for single-phase and multiple-phase mixture models. The LMR-LRT is only available for models with one latent class variable while the statistical evaluation of multiple-phase models is limited to descriptive information criteria with preference to the adjusted BIC (Kim 2014). In addition, the large number of zeros in the incidence rates can be accounted by an inflation part of the particular mixture model (Reinecke & Seddig 2011). This extension has to be studied in future applications of stage-sequential growth mixture models.

References

Boers, K., Reinecke, J., Daniel, A., Kanz, K., Schulte, P., Seddig, D., et al. (2014). Vom Jugend-zum frühen Erwach-senenalter. Delinquenzverläufe und Erklärungszusammenhänge in der Ver-laufsstudie Kriminalität in der modernen Stadt. *Monatsschrift für Kriminologie und Strafrechtsreform, 3*, 183–202.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach. Wiley series on probability and mathematical statistics.* New Jersey: Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplète data via the EM algorithm. *Journal of the Royal Statistical Society, 39*, 1–38.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications.* Mahwah: Lawrence Erlbaum.

D'Unger, A., Land, K. C., McCall, P. L., & Nagin, D. S. (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses of the London, Philadelphia and Racine cohort studies. *American Journal of Sociology, 103*, 1593–1630.

Enders, C. K. (2010). *Applied missing data analysis.* New York: Guilford Press.

Farrington, D. P., & West, D. J. (1990). The Cambridge study in delinquent development: A longterm follow-up of 411 London maies. In: H. J. Kerner & G. Kaiser (Eds.), *Kriminalität: Persönlichkeit, Lebensgeschichte und Verhalten* (pp. 115–138). Berlin: Springer.

Hilbe, J. M. (2011). *Negative binominal regression* (Vol. 2). Cambridge: Cambridge University Press.

Kim, S.-Y. (2014). Determining the number of latent classes in single- and multiphase growth mixture models. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 263–279.

Kim, S.-Y., & Kim, J.-S. (2012). Investigating stage-sequential growth mixture models with multiphase longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 19*, 293–319.

Kreuter, F., & Muthén, B. O. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and groupbased approaches using growth mixture modelling. *Journal of Quantitative Criminology, 24*, 1–31.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767–778.

Loeber, R., & LeBlanc, M. (1990). Toward a developmental criminology. In T. Michael & M. Norval. (Eds.), *Crime and justice* (Vol. 12, pp. 275–473). Chicago: University of Chicago Press.

Mariotti, L., & Reinecke, J. (2010). *Wachstums- und Mischverteilungsmodelle unter Berücksichtigung unbeobachteter Heterogenität: Empirische Analysen zum delinquenten Verhalten Jugendlicher in Duisburg.* Technical Report Sozialwis- senschaftliche Forschungsdokumentation 21. Munster: Institut für sozialwis- senschaftliche Forschung e.V.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*(1), 110–133.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674–701.

Moffitt, T. E., Caspi, A., Rutter, M., & Silva, P. A. (2001). *Sex differences in antisocial behavior.* Cambridge: Cambridge University Press.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 345–368). Thousand Oaks: Sage.

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable- centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism, Clinical and Experimental Research, 24*, 882–891.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463–469.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods, 4*, 139–157.

Nagin, D. S. (2005). *Group-based modeling of development.* Cambridge, MA: Harvard University Press.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology, 31*, 327–362.

Odgers, C. L., Moffitt, T. E., Broadbent, J. M., Dickson, N., Hancox, R. J., Harrington, H., et al. (2008). Female and male antisocial trajectories: From childhood origins to adult outcomes. *Development and Psychopathology, 20*, 673–716.

Piquero, A. R. (2008). Taking stock of developmental trajectories of criminal activity over the life course. In A. M. Liberman. (Ed.), *The long view of crime: A synthesis of longitudinal research* (pp. 23–78). New York: Springer.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks: Sage.

Reinecke, J. (2012). *Wachstumsmodelle.* Sozialwissenschaftliche Forschungsmeth- oden Band 3. München und Mering: Rainer Hampp Verlag.

Reinecke, J., & Seddig, D. (2011). Growth mixture models in longitudinal research. *Advances in Statistical Analysis, 95*, 415–434.

Reinecke, J., & Weins, C. (2013). The development of delinquency during adolescence: A comparison of missing data techniques. *Quality & Quantity, 47*(6), 3319–3334.

Sampson, R. J., & Laub, J. H. (1993). *Crime in the making: Pathways and turningpoints through life.* Cambridge: Harvard University Press.

Sampson, R. J., & Laub, J. H. (1997). A life-course theory of cumulative disadvantage and the stability of delinquency. In T. P. Thornberry. (Ed.), *Developmental theories of crime and delinquency. Advances in criminological theory* (Vol. 7, pp. 133–161). New Brundwick, London: Transaction Publishers.

Sampson, R. J., & Laub, J. H. (2003). Life-course desisters? Trajectories of crime among delinquent boys followed age 70. *Criminology, 41*, 555–592.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

Thornberry, T. P. (2005). Explaining multiple patterns of offending across the life course and across generations". In R. J. Sampson & J. H. Laub. (Eds.), *Developmental criminology and its discontents. The annals of the American Academy of Political and Social Science* (Vol. 602, pp. 156–195). Thousand Oaks: Sage.

Tracy, P. E., Wolfgang, M. E., & Figlio, R. M. (1990). *Delinquency in two birth cohorts.* New York: Plenum.

Tremblay, R. E., Desmarais-Gervais, L., Gagnon, C., & Charlebois, P. (1987). The preschool behavior questionnaire: Stability of its factor structure between cultures, sexes, ages and socioeconomic classes. *International Journal of Behavioral Development, 10*, 467–484.

Yang, C. C. (1998). Finite mixture model selection with psychometric applications. Unpublished doctoral dissertation. University of Groningen.

Developmental Pathways of Externalizing Behavior from Preschool Age to Adolescence: An Application of General Growth Mixture Modeling

Mark Stemmler and Friedrich Lösel

Abstract This study applies a developmental and life-course perspective on the data of the Erlangen-Nuremberg Development and Prevention Study (ENDPS; Lösel, Stemmler, Jaursch, and Beelmann, Monatsschrift für Kriminologie und Strafrechtsreform 92:289–308, 2009) to find interindividual differences in intraindividual change of externalizing problem behavior. Based on a sample of $N = 541$ boys and girls, general growth mixture modeling (GGMM; Nagin, Psychological Methods 4:139–177, 1999; McArdle, The handbook of research methods in developmental psychology. New York: Blackwell Publishers, 2005) was applied. In a prospective longitudinal design measurements with multiple informants were analyzed from preschool to adolescence. The results of the GGMM showed five groups representing different developmental trajectories: (1) "*high-chronics*" (2.4 %; $n = 13$), who had the highest scores of externalizing behavior at all times; (2) "*low-chronics*" (58.8 %; $n = 317$) who were low on externalizing behavior throughout the years; (3) "*high-reducers*" (7.9 %; $n = 43$) who started out high, but reduced their externalizing behavior monotonically over time; (4) "*late-starters-medium*" who increased externalizing problems at later age (8.7 %; $n = 47$); and (5) "*medium-reducers*" whose problems decreased from an originally medium level (22.4 %; $n = 121$). The results are in accordance with international studies on developmental trajectories of offending and suggest that a perspective on a broad

This research was supported by a grant from the German Federal Ministry of Family Affairs, Seniors, Women and Youth.

M. Stemmler (✉)
Friedrich-Alexander, University of Erlangen, Nürnberg (FAU), Erlangen, Germany
e-mail: mark.stemmler@fau.de

F. Lösel
University of Erlangen-Nuremberg, Erlangen, Germany

University of Cambridge, Cambridge, UK

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_4

range of behavioral problems can be fruitful. The findings are discussed with regard to other studies on latent group-based modeling, non-statistical taxonomies, and practical applications.

Introduction

Prospective longitudinal studies enable the analysis of interindividual differences in intraindividual change and are therefore the preferred research design in developmental psychology (McCartney, Burchinal, & Bub 2006; Nesselroade & Baltes 1979). This approach, also called developmental and life-course perspective, acknowledges the basic assumption that human behavior and its connected social context are changing over time. Due to progress in longitudinal studies and statistical methodology (e.g., growth curve modeling) life-course research became particularly important in the study of antisocial behavior and led to the field of "developmental and life-course criminology" (e.g., Boers, Lösel, & Remschmidt 2009a; Farrington 2002).

Since the 1990s, statistical tools such as latent group-based modeling or general growth mixture modeling (GGMM) have been successfully applied to longitudinal datasets to describe the number and shape of violence, aggression and delinquency trajectories (see Piquero, Farrington, & Blumstein 2007; Jennings & Reingle 2012). By using GGMM or related tools it is possible to find different groups with individual change curves leading to different developmental outcome in terms of antisocial behavior or delinquency. In an early study Nagin and Tremblay (1999) used the data of the Montréal Study to analyze trajectories of boys' physical aggression, oppositional behavior, and hyperactivity from ages 6 to 15. Four developmental trajectories were identified for the three problem behaviors under study. The group sizes varied depending on the particular behavior: a *chronic problem trajectory* (4–6 %), a *high-level near-desister trajectory* (25–30 %), a *moderate-level desister trajectory* (45–52 %), and a *no problem trajectory* (17–25 %). D'Unger et al. (1998) analyzed the data of three renowned longitudinal studies: the Cambridge Study in Delinquent Development (Farrington et al. 2009), the Philadelphia Birth Cohort Study (Tracy et al. 1990) and the Racine Birth Cohorts Study (Shannon 1988). The data were used to detect different trajectories with regards to official police records. The British data suggested four different trajectories: *nonoffenders* (64 %) with almost zero police contacts, one *adolescence-peaked trajectory* (12.7 %), and two c*hronic trajectories*, one on a low (9.9 %) and the other on a high level (13.4 %). The data from Philadelphia came up with five different groups: *nonoffenders* (60.8 %), *adolescence-peaked trajectories* (*low rate*) (8.6 %), *adolescence-peaked trajectories* (*high rate*) (1.0 %), *chronic offenders* (*low rate*) (21.3 %), and *chronic offenders* (*high rate*) (8.3 %). And the Racine data came up with four or five classes depending on the birth cohort: *nonoffenders* (1942: 34.6 %; 1945: 35.4 %, 1955: 44.5 %), *adolescence-peaked trajectories* (1942: 20.1 %; 1945: 39.8 % (low-rate), 19.4 % (high rate); 1955: 2.2 % (early onset),

15.4 % (late onset)), and *chronic offenders* (1942: 31.4 % (low rate), 8.8 % (high rate); 5.1 % (late onset); 1945: 5.4 %; 1955: 30.1 % (low rate), 7.8 % (high rate)).

Bushway et al. (2003) used self-reported data of the Rochester Youth Development Study (RYDS; Thornberry 1997). Seven groups were identified: *very low-level offenders* (38.6 %), *low-level offenders* (22.5 %), *late starters* (9.8 %), *intermittent offenders* (8.6 %), *bell-shaped desisters* (8.5 %), *slow uptake chronic offenders* (7.8 %), and *high-level chronic offenders* (4.2 %). Hoeve et al. (2008) analyzed self-reported delinquency and conviction rates of youth who participated in the youngest cohort of the Pittsburgh Youth Study (PYS; Loeber & Hay 1997). Development was followed through age 20 and five different groups were found: *non-delinquents* (27.2 %), *minor persisting* (27.6 %), *moderate desisting* (6.8 %), *serious persisting* (24.2 %), and *serious desisting* (14.3 %). Bongers et al. (2004) studied problem behavior in children and adolescents aged 4–18 years in the Netherlands and found three types of parent-reported development of aggressive behavior: a *near-zero trajectory* (71.0 %), a *low decreaser trajectory* (21.4 %), and a *high decreaser trajectory* (7.7 %). The high-level trajectory showed the highest probability for predicting adult DSM-IV disorders (Reef et al. 2011).

Although the vast majority of studies on developmental trajectories of antisocial behavior has been carried out in the Anglo-American context, there is also research on this topic in Germany: Reinecke (2006) analyzed the data from the panel study Crime in the Modern City (CRIMOC; Boers, Seddig, & Reinecke 2009b) to identify different classes of deviant and delinquent behavior (self-report). From nine data waves starting at age 13, three classes evolved: *Adolescents with almost no deviant or delinquent activities* (58.2 %), a *medium proportion of adolescents with a low increase of delinquency* (33.3 %), and *a small number with a larger growth starting on a higher level* (8.5 %).

Overall, these and other studies suggest that there are no consistent numbers and types of developmental trajectories of delinquency, violence and crime. The most common results support Moffitt's (1993) theory-driven typology of an early starting and relatively persistent development of antisocial behavior versus an adolescence-limited pathway. In addition, nearly all studies show a large group of youngsters who are low in antisocial behavior across all measurement points. A recent systematic review of studies on developmental trajectories points in the same direction (Jennings & Reingle, 2012). Depending on age, type of sample (e.g., high risk vs. normative), kind of problem behavior, mode of measurement, method of analysis, geographical context and other issues the results varied between two and seven trajectories, but three to five were most common. Jennings and Reingle (2012) made a number of suggestions for further progress in this field research. In addition to more research on the explanation of different pathways, the authors suggest more studies on broader topics of developmental psychopathology, different cultural contexts, and data from multiple informants.

The present study follows the latter proposals. We analyzed the data of the Erlangen-Nuremberg Development and Prevention Study (ENDPS; Lösel et al. 2009, 2013) with regard to different trajectories for the broad category of externalizing problems. ENDPS is based on a normative sample and is a combined

experimental and longitudinal study on child behavior covering a time period of approximately ten years. Social behavior was rated in standardized reports from multiple informants such as mothers, kindergarten educators, school teachers, and the youngsters themselves. Therefore, the ENDPS can provide information on prototypical developments of a broad range of problem behaviors in a European context that may be relatively less biased by specific outcome measurements. As this publication is embedded in a method-oriented volume, the following section contains details of our statistical model and analysis.

Overview of Statistical Models

From a statistical point of view, one can treat latent growth curve modeling as multi-level models with the repeatedly measured observed variables on the first level and the latent variable on the second level (cf. McArdle 1988, 2005; Stemmler & Petersen 2012). If the assumption does not hold, that the underlying modeling of the growth over time is valid for a homogeneous population under investigation, growth curve models with latent classes come into play, to explain the "unobserved heterogeneity" (Nagin 1999; Muthén & Shedden 1999). The mathematical generalizations were described in a book on "finite mixture models" by McLachlan and Peel (2000). Nagin (1999) was the first scholar to apply growth curve modeling for different classes in the field of criminology. Nagin called his approach *semi-parametric, group-based modeling approach*, whereas Muthén (2004) used the term *latent class growth modeling* to underline the fact that in this model the random coefficient of the growth curve was fixed to zero, indicating no within class variation. However, this model is a special case of the general growth mixture models (GGMM) which can be analyzed with MPLUS (Muthén & Muthén 2010) or the LAVAAN package (Rosseel 2012) of the R statistical programing environment (R Core Team 2015).

The traditional growth curve model is based on the following equation (cf. Reinecke 2006, 2012):

$$y_t = \lambda_{t1}\eta_1 + \lambda_{t2}\eta_2 + \epsilon_t \tag{1}$$

In this formula y_t are the observed variables measured at time t, which are determined by the two latent variables η_i representing the *level* and *slope* of the growth curve, and ε_t the residuals (see Fig. 1).

The coefficients of the level are usually fixed to the value 1.0, whereas the coefficient of the slope may represent either linear growth (i.e., $\lambda_{12} = 1,\ \lambda_{22} = 2 \cdots \lambda_{t1} = t$) or any other combination, as long as the necessary coefficients are fixed. The equations for the latent variables are

$$\eta_{\text{level}} = \alpha_1 + \zeta_1 \tag{2a}$$

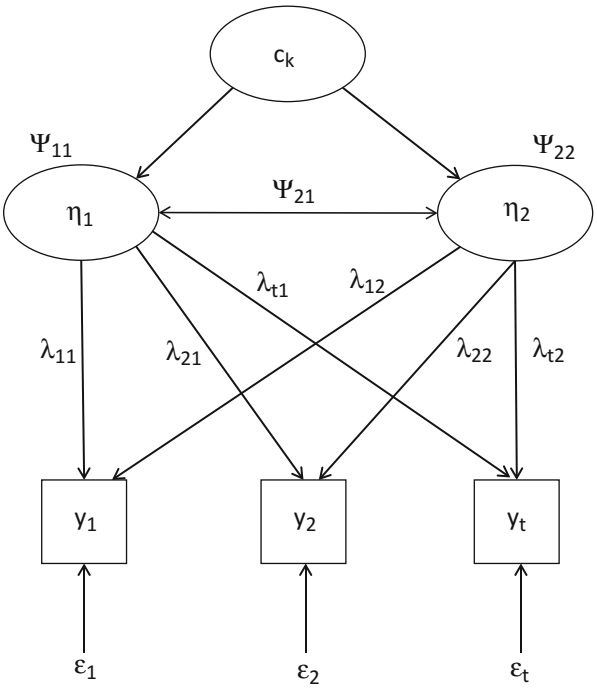


Fig. 1 A general growth mixture model (GGMM), which is basically a growth curve model with latent classes (c_k). *Note*: c_k are the different latent classes; η_i represents the latent variables for *level* and *slope*; the y_i are the observed variables for each measurement point; the ϵ_i are the residuals or error terms; the λ_i are the coefficients for the latent variables and ψ_i are the variance and covariances of *level* and *slope*

$$\eta_{\text{slope}} = \alpha_2 + \zeta_2 \tag{2b}$$

The above traditional growth curve can easily be extended to a conditional growth curve model if an exogenous variable, functioning as a predictor is included in the model. The extension is as follows:

$$\eta_m = \alpha_m + \Gamma\xi_n + \zeta_m \tag{3}$$

where the matrix Γ $(m \times n)$ contains the regression coefficients of the exogenous variable ξ on the endogenous variables η_i. The variances and the covariance of the latent variables can be found in the psi-matrix (see Reinecke 2006, 2012):

$$\Psi = \begin{pmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{4}$$

with ψ_{11} representing the variance of the level variable, ψ_{22} represents the variance of the slope variable, and ψ_{21} indicates the covariance between the two latent variables. If the growth model represents the different trajectories of different subpopulations, the statistical parameters vary across classes (see Fig. 1). According to Muthén (2004) such a general growth mixture model can be written as

$$y_{1k} = \lambda_{1tk}\eta_{1k} + \lambda_{2tk}\eta_{2k} + \epsilon_{tk} \tag{5a}$$

$$\eta_{1k} = \alpha_{1k} + \zeta_{1k} \tag{5b}$$

$$\eta_{2k} = \alpha_{2k} + \zeta_{2k} \tag{5c}$$

The variances of the η variables are estimated separately for each class, as well as their covariances. The parameters of GGMM can be estimated in MPLUS using the EM algorithm to obtain maximum-likelihood (ML) estimators (Dempster et al. 1977; Muthén & Shedden 1999). At the end, individuals are assigned a particular class based on their established posterior probabilities. This class membership may be used for further statistical analysis validating the obtained results of the GGMM; however there is controversy about this issue because class membership is based on probabilities and a pretended fixed class membership may overlook possible misclassifications due to error variance (Clarke & Muthen 2009).

There is no statistical test for the evaluation of the required number of necessary classes (Reinecke 2006, 2012), but there are useful statistical parameters such as the entropy measure E_k which varies between 0 and 1, with values close to 1 indicating a reasonable classification. And there are the Bayesian Information Criterion (BIC) or the adjusted BIC which are based on the maximum likelihood of the model. Of two comparing models the one with the lowest BIC or adjusted BIC is preferred. Finally, the Lo-Mendell-Rubin likelihood ratio test (LMR-LRT) compares the ratio of the likelihoods of two competing models, that is the (k-1)-classes model with k-classes model. The null hypothesis (H_0) states that the (k-1) model should be preferred. Therefore, significant or small p-values of the LMR-LRT are in support of the k-classes model. Another statistical parameter is the BRT (i.e., bootstrapped likelihood ratio test) which also compares the (k-1)-classes model with the k-classes model. The larger the likelihood the better the BRT. However, all statistical parameters are proxies that are used to select the best model, the final decision should also take theoretical issues into account.

In case of missing data MPLUS uses the full information maximum likelihood (FIML) estimator (Reinecke 2005). This estimator, which does not require Missing Completely at Random (MCAR) but Missing at Random (MAR), is well established in all currently available SEM programs. With a reasonably large sample size FIML produces unbiased parameter estimates.

Based on the abovementioned review of the life-course criminological literature we expected two groups with relatively stable levels of externalizing symptoms: those who are chronically high and those who are chronically low, with the latter group being larger. In addition, we envisioned groups with time-limited externalizing behavior and/or a later start of problems.

Method

Sample

The data were taken from the Erlangen-Nuremberg Prevention and Development Study (ENDPS; Lösel, Stemmler, Beelmann, & Jaursch 2005; Lösel, Stemmler, Jaursch, & Beelmann 2009; Lösel, Stemmler, & Bender 2013). The ENDPS is a combined prospective longitudinal and experimental prevention study with a multi-informant and multi-method approach. The original sample of the core study consisted of 675 kindergarten children (336 boys, 339 girls) from 609 families. The project is a longitudinal study that started at preschool age and is now containing seven waves of data collection. The sample was nearly representative of young families living in Erlangen and Nuremberg (Franconia). According to an index of the socioeconomic status (SES; Geißler 1994) which included income, education, profession, and housing conditions, 13.3 % of the families were lower class, 32.3 % were lower middle class, 30.6 % middle class, 15.4 % upper middle class, and 3.0 % upper class. Approximately 86 % of the parents were married at *Time 1*. The retention rates varied over time; in the most recent wave (nearly 10 years after the first one) approximately 90 % of the original sample participated (Lösel & Stemmler 2012; Stemmler & Lösel 2012).

For the analyses below, the data was structured according to age so that homogeneous age groups were assessed at the various measurement points. Data were collected when the study child was at the ages of 4 or 5, 6 or 7, 8 or 9, 10–12, and 13 or 14. Children were included if they had at least data on 3 out of the 5 measurement points. The data of the other two missing data points were imputed. Overall, the longitudinal sample contained $N = 541$ children. The cross-sectional sample sizes were as follows: $n = 525$ (4–5 years), $n = 424$ (6–7 years), $n = 422$ (8–9 years), $n = 486$ (10–12 years), and $n = 377$ (13–14 years).

Measures

The children's social behavior in kindergarten and at school was assessed by our German adaptations of the Social Behavior Questionnaire (SBQ; Tremblay et al. 1987; Tremblay et al. 1992). The SBQ is available in multiple versions. Here, kindergarten educators', school teachers', and mothers' ratings were used (Lösel, Beelmann, & Stemmler 2002). The content and format of the teacher's SBQ versions are identical and consist of 46 items. The mother's version has two additional items. The teacher's version item "stealing things" is divided for the mothers' version into "stealing things at home" and "stealing things outside home." Each item is rated on a 3-point scale ranging from "*0*" = *never/not true* to "2" = *almost always/true most of the time*. In the present study we only used items on externalizing behavior problems. Our *Externalizing Problems* scale was formed

of four primary scales: *Physical Aggression*, *Destroying Things/Delinquency*, *Indirect Aggression*, and *Hyperactivity/Attention Problems.* The reliabilities for the different informants were $\alpha = .89$ (preschool teachers/kindergarten educators), $\alpha = .91$ (school teachers), and $\alpha = .74$ (mothers).

To enhance the validity of measurement, at each wave the data from two informants were combined (mean of *z*-scores), that is weighing the teachers' and mothers' ratings equally. At preschool age we used the information from the mothers and kindergarten educators and at elementary school age the mothers' and school teachers' ratings. In secondary school we added the children's self-reports to the mothers' SBQ data, again using the mean ratings of the two informants. To assess externalizing behavior through the child's self-report we used the German version of the Strength and Difficulties Questionnaire (SDQ; Goodman 1999; German adaptation: Hölling, Erhart, Ravens-Sieber, & Schlack 2007). The items are answered on a 3-point scale ranging from "*1*" = *does not apply* to "*3*" = *does clearly apply*. The *Externalizing Scale* consists of five items. The reliability in our sample was rather low ($\alpha = .50$), but similar to the results from a nationwide German sample of the Robert-Koch Institute (Hölling et al. 2007). The mothers' SBQ ratings and the children's SDQ ratings were combined by averaging *z*-scores.

Results

Linear and quadratic latent class growth analyses with an increasing number of classes were tested. MPLUS, version 6, was used (Muthén & Muthén 2010). Models with within-class variation as well as with no-within-class variations were analyzed. Hundred random sets of starting values were generated in the initial stage and ten optimizations were carried out. The OPTSEED option was applied to specify the random seed that has been found to result in the highest log-likelihood in the previous analyses (Muthén & Muthén 2010). The fit of different latent classes ranging between one and six can be taken from Table 1. The statistical results suggest a linear GGMM of five classes according to Nagin (1999) with no-within-class variation. Here, a $BIC = 4051.39$ and an *adj. BIC* $= 3991.02$ were obtained. The *LMT-LRT* suggested that compared to a k-1 = 4-class solution the five-class solution should be preferred (*LMR-LRT* $= 79.75$, $p = .08$). The corresponding *BRT* generated the smallest value (*BRT* $= 83.97$) of all solutions with a likelihood ratio *LRT* $= -2007.89$. The smallest *BIC* and *adj. BIC* were found for the six-classes model, but there were very small classes ($n < 10$) and the *LMT-LRT* and the *BRT* revealed a lesser fit (Table 1).

The five classes represent different developmental trajectories from childhood to adolescence. Figure 1 depicts the different developmental trends. Squares indicate the "observed" means and triangles the estimated means. The upper dashed-dotted lines are the "*high-chronics*" (2.4 %; $n = 13$), who are receiving the highest values in externalizing behavior from childhood on up to adolescence. The opposite class are the "*low-chronics*" (dashed lines; 58.8 %; $n = 317$) who are low on externalizing

Table 1 Results of the general growth mixture models (GGMM) with different classes

Test	1 class	2 classes	3 classes	4 classes	5 classes	6 classes
BIC	5107.10	4353.58	4192.29	4116.48	4051.39	4021.67
Adjusted BIC	5084.88	4321.84	4151.02	4065.69	3991.02	3951.83
LMR-LRT	–	733.54	171.11	89.93	79.75	129.16
p-value	–	0.00	0.15	0.27	0.08	0.27
Likelihood	–	−2531.52	−2145.33	−2055.24	−2007.89	−2007.89
BRT	–	772.39	180.18	94.69	83.97	132.58
p-value	–	0.00	1.00	1.00	1.00	1.00
Class Sizes	541 (100)	435 (80.4)	375 (69.3)	346 (64.0)	317 (58.6)	315 (58.2)
N (%)		106 (19.6)	127 (23.5)	116 (21.4)	121 (22.4)	115 (21.2)
			39 (7.2)	48 (8.9)	47 (8.7)	44 (8.1)
				31 (5.7)	43 (7.9)	30 (5.5)
					13 (2.4)	29 (5.3)
						8 (1.5)

BIC Bayesian Information Criterion, *LMR-LRT* Lo-Mendell-Rubin likelihood ratio test, $N = 541$, *BRT* Bootstrapped likelihood ratio test

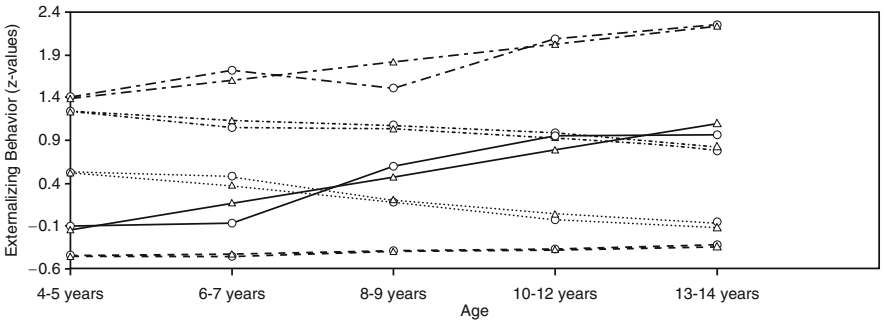


Fig. 2 Results of the general growth mixture model (GGMM) resulting in five developmental trajectories. *Note*: The *y-axis* displays the values for "Externalizing Behavior," and all values were z-transformed. The *x-axis* shows the age of the juveniles at each measurement point. The *square*s represent "observed" means and the *triangles* "estimated" means. The *upper dashed-dotted lines* are the "high-chronics" (2.4 % of the sample), the *dashed-dot-dotted lines* are the "high-reducers" (7.9 %), the *dotted lines* are the "medium-reducers" (22.4 %), the *ascending black lines* are the "late-starters-medium" (8.7 %), and the *dashed lines* represent the "low-chronic" (58.6 %)

behavior throughout the years; including the majority of the sample. The dashed-dot-dotted lines are the "*high-reducers*" (7.9 %; $n = 43$) who start out high in childhood, but who reduce their externalizing behavior monotonically over time. By adolescence they are passed by the "*late-starters-medium*" (ascending black lines; 8.7 %; $n = 47$). Finally, the dotted lines show the trends of the "*medium-reducers*" (22.4 %; $n = 121$) who include about one-quarter of the sample. Their externalizing is medium high in kindergarten but decreases linearly up to adolescence (Fig. 2).

Discussion

Prospective longitudinal studies on problem behavior have a number of advantages (Loeber & Farrington 1994): They allow the study of the natural history of the development of problems such as onset, increase, decrease and termination. Based on individual data they enable the study of trajectories or pathways. A pathway is defined as "when a group of individuals experience a behavioral development that is distinct from the behavioral development of another group of individuals" (p. 890; Loeber & Farrington 1994). The identification of distinctive groups of trajectories enables one to estimate the proportion of the population following each trajectory group and to relate group membership probability to personal and social characteristics. Valid distinctions of developmental pathways can guide policy, e.g., with regard to risk-based early prevention programs (Farrington & Welsh 2007; Lösel et al. 2013). Loeber and Farrington (1994) also postulate that the best studies should rely on multiple informants. This is in accordance with numerous findings that showed rather low agreement between different informants from different social contexts (e.g., Achenbach 2006; Lösel 2002).

This research meets the abovementioned criteria. We adopted a developmental and life-course perspective by using the data of the Erlangen-Nuremberg Development and Prevention Study (ENDPS). We applied general growth mixture modeling (GGMM) to data from early childhood to adolescence, covering a 10-year period, on externalizing behavior problems rated at each measurement point by two different informants (kindergarten educators, mothers, school teachers, and self-report). The results suggested a five-class solution representing five different developmental trajectories.

Although our study contained data on a broad range of externalizing symptoms and a community sample of boys and girls from Germany the results were relatively similar to Anglo-American studies that used Nagin's (1999) approach on *semiparametric group-based modeling*. As mentioned in the introduction, most studies showed between three and five classes depending on the type of outcome measures and samples used (Jennings & Reingle 2012). The small group of "high-chronics" and the largest group of "low-chronics" (no problems at all times) are in accordance with the well-replicated trajectories of delinquency, aggression, and violence (Jennings & Reingle 2012). The group of "high-reducers" confirms that not all children who exhibit early antisocial behavior enter on a persistent pathway. In contrast, various international studies have shown that a half or more recover within a short period of time (e.g., Moffitt et al. 1996; Nagin & Tremblay 1999; Werner & Smith 1992). Even in the presence of various risk factors abstaining or early desistance from problem behavior seems to be more the rule than an exception (Lösel & Bender 2003; Lösel & Farrington 2012). Our fourth trajectory of "late-starters-medium" may indicate an early phase of the adolescent-limited pathway that has been found in studies that covered the whole range of youth and young adulthood (e.g., Moffitt et al. 2002). Further waves of the ENDPS may show whether the increase of externalizing problems continues until late adolescence and then be

followed by a decrease. The fifth trajectory we found in our study is insofar plausible as it shows a moderate level of behavioral problems that decreased from early childhood to youth. These "medium-reducers" show a similar trend as the "high-reducers," but are a larger group that decreases from a more normative lower level of externalizing problems. Both pathways may indicate positive influences of cognitive competences, self-control, and social skills that reduce physical aggression and other antisocial behavior from early childhood onwards (e.g., Tremblay et al. 2004).

Overall, our findings fit well into the international criminological literature. However, there seems to be a difference with regard to the size of the group with intensive and persistent problem behavior. Whereas in criminological trajectory studies often approximately 5 % of a cohort belonged to this category, in our study only 2.4 % belonged to this group. This lower prevalence may have been partially due to the comparatively young age when our sample was first assessed. In addition, less serious problems of externalizing behavior in a "normal" community sample may be more temporary and thus not lead to a larger group with high problem stability. Taking together the "high-chronics" and the "high-reducers" the respective proportion was about 10 %. This is within the range of point prevalence rates for externalizing child behavior in Germany (e.g., Hölling et al. 2007).

One should also mention that our study contained both boys and girls. As boys show more externalizing problems than girls the relatively small size of the "high-chronics" group is plausible. Because we investigated a nearly representative sample of the local area we included both sexes in the trajectory analysis. As boys show more externalizing problems than girls (see Lösel & Stemmler 2012; Moffitt, Caspi, Rutter, & Silva 2001; Moretti & Odgers 2002), mixed-gender studies on this issue may contain problems. However, different prevalence rates do not necessarily imply that there are different risk variables and developmental processes. Although gender is a sound predictor of delinquency and offending (Ryder, Gordon, & Bulger 2009), most risk variables for boys and girls seem to be similar (see Moffitt et al. 2001; Silverthorne & Frick 1999). Boys simply show more risks for externalizing problems and girls may also benefit from more protective factors and mechanisms (e.g., Lösel & Bender 2003; Lösel, Stemmler, & Bender 2013; Werner & Smith 2001).

In sum, the results of our study are consistent with international research that concentrated on more specific forms of antisocial behavior. Addressing a broad range of externalizing problems bears the advantage of a relatively sensitive detection of early needs for intervention and prevention. In the ENDPS we found encouraging effect sizes in predictive validity with Odds Ratios of up to 10 (Wallner, Lösel, Stemmler & Corrado, submitted). More detailed analyses on the prediction of trajectories are in progress.

However, the present study underlines the methodological progress due to the invention of GGMM. It allows the empirical and statistical driven search and identification of different developmental pathways that overcomes more or less arbitrary definitions of groups. For example, Moffitt (1993) defined boys as "life-course persistent antisocial" if they had above average scores (by at least one standard deviation) on a scale of antisocial behavior. Elevated scores by three raters

(parents, teachers, and self) were required at each of seven biennial assessments from age 3 to 15. However, for various reasons, the algorithm had to be changed later to at least three elevated scores out of the 5 assessments from ages 5 to 11 years (Moffitt et al. 2002). In another high-quality study Elliott and Huizinga (1980) defined youngsters as *high delinquents* if they had more than 12 crimes per year, as *exploratory delinquents* if they had equal or less than five crimes per year. Such a priori definitions always involve some kind of arbitrariness. In our view, such group definitions are well justified as long as they are to some degree theory driven. It is encouraging that such original groupings were supported by advanced statistical analyses (Nagin, Farrington, & Moffitt 1995). Insofar, GGMM has provided a tremendous progress in finding the most adequate number of groups or pathways leaving behind scientific capriciousness.

However, in spite of the convergent validity of our results with studies from North America one must acknowledge various limits. First, although the algorithm for the selection of different trajectories is fully objective, the final solution still required some subjective decisions (i.e., the exclusion of a pathway with very small group size). Second, GGMM leads to pathways of *relative* and not *absolute* homogeneity in development; that is, one must assume individual cases in each trajectory that are rather similar to some cases in another pathway. Third, GGMM provides a descriptive developmental grouping of a specific data set that requires cross validation. Fourth, it needs to be emphasized that the labeling of the different groups is data-driven and not based on theoretically or clinically relevant distinctions. For example, the children on the "high-chronic" pathway in our community sample may still differ in many characteristics from a persistent group of offenders in a high-risk sample. This points to a general problem with GGMM. The question is whether the identified latent classes are real existing subpopulations or just different statistically generated groups with rather general labels made up by researchers. Therefore, further investigation of differential predictors of various developmental pathways is an important task for our own and other future research. If one is not interested in finding discrete latent classes or if one does not assume the existence of subpopulations one could use the so-called heterogeneous growth curve modeling (HGM; Brandt & Klein in press). HGM models growth curves while using covariates like gender or school type to explain the unobserved heterogeneity in the slope variance. Further limits of the abovementioned GGMM are the use of categorical variables or extremely non-normal data (for solutions see Bauer & Curran 2004).

References

Achenbach T. (2006) As others see us: Clinical and research implications of cross-informant correlations for psychopathology. Current Directions in Psychological Science 15:94–98

Bauer D. J., Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. Psychological Methods 9:3–29

Boers, K., Lösel, F., & Remschmidt, H. (Eds.) (2009). Developmental and life-course criminology (Special Issue). *Monatsschrift für Kriminologie und Strafrechtsreform* [Monthly Journal of the Criminal and Penal Law Reform], 2–3.

Boers, K., Seddig, D., & Reinecke, J. (2009). Sozialstrukturelle Bedingungen und Delinquenz im Verlauf des Jugendalters. Analysen mit einem kombinierten Markov-und Wachstumsmodell [Social structural circumstances and delinquency during the course of adolescence. Analyses with a combined Markov- and growth curve model]. *Monatsschrift für Kriminologie und Strafrechtsreform* [Monthly Journal of the Criminal and Penal Law Reform], (2–3), 267–288.

Bongers I. L., Koot H. M., van der Ende J, Verhulst, C. F. (2004). Developmental trajectories of externalizing behaviors in childhood and adolescence. Child Development 75(5):1523–1537

Brandt, H., & Klein, A. (in press). A heterogeneous growth curve model for non-normal data. *Multivariate Behavioral Research*.

Bushway S. D., Thornberry T. P., Krohn, M. D. (2003). Desistance as a developmental process: A comparison of static and dynamic approaches. Journal of Quantitative Criminology 19(2): 129–153

Clarke, S. L., & Muthen. B. (2009). Relating latent class analysis results to variables not included in the analysis. Unpublished manuscript. Retrieved from http://www.statmodel.com/download/relatinglca.pdf

Dempster A. P., Laird NM, Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39(1):1–38

D'Unger A. V., Land K. C., McCall P. L., Nagin, D. S. (1998). How many latent classes of delinquent/criminal careers? Results from mixed Poisson regression analyses. The American Journal of Sociology 103(6):1593–1630

Elliott, D. S., & Huizinga, D. (1980). *Defining pattern delinquency: A conceptual typology of delinquent offenses*. Paper presented at the meeting of the American Society of Criminology, San Francisco, CA.

Farrington, D. P. (2002). Developmental criminology and risk-focused prevention. In: Maguire M., Morgan R., Reiner R. (eds) The Oxford handbook of criminology, 3rd edn. Oxford University Press, Oxford, pp 657–701

Farrington D. P., Coid J. W., West, D.J. (2009). The development of offending from age 8 to age 50: Recent results from The Cambridge Study in Delinquent Development. Monatsschrift für Kriminologie und Strafrechtsreform [Monthly Journal of the Criminal and Penal Law Reform] 59:160–173

Farrington D. P., Welsh, B.C. (2007). Saving children from a life of crime. Oxford University Press, Oxford, UK

Geißler R. (ed) (1994) Soziale Schichtung und Lebenschancen in Deutschland [Social strata and life opportunities in Germany]. Enke, Stuttgart

Goodman, R. (1999). The extended version of the strengths and difficulties questionnaire as a guide to child psychiatric caseness and consequent burden. Journal of Child Psychology and Psychiatry 40:791–801

Hölling, H., Erhart, M., Ravens-Sieber, U., & Schlack, R. (2007). Verhaltensauffälligkeiten bei Kindern und Jugendlichen: Erste Ergebnisses aus dem Kinder- und Jugend gesundheitssurvey (KIGGS) [Problem behavior in children and adolescents: First results of a children and youth health survey (KIGGS)]. *Bundesgesundheitsblatt-Gesundheitsforschung- Gesundheitsschutz [Federal health journal- health research- and health protection], 50,* 784–793.

Hoeve M., Blokland A., Semon Dubas J., Loeber R., Gerris J.R.M., van der Laan, P.H. (2008). Trajectories of delinquency and parenting styles. Journal of Abnormal Child Psychology 36:223–235

Jennings W. G., Reingle, J.M. (2012). On the number and shape of developmental/life-course violence, aggression, and delinquency trajectories: A state-of-the-art review. Journal of Criminal Justice 40:472–489

Loeber R, Farrington, D.P. (1994). Problems and solutions in longitudinal and experimental treatment studies of child psychopathology and delinquency. Journal of Consulting and Clinical Psychology 62(5):887–900

Loeber R, Hay D. (1997). Key issues in the development of aggression and violence from childhood to adolescence. Annual Review of Psychology 48:371–410

Lösel F (2002) Risk/need assessment and prevention of antisocial development in young people: Basic issues from a perspective of cautionary optimism. In: Corrado R, Roesch R, Hart SD, Gierowski J (eds) Multiproblem violent youth, NATO SPS series. Ios Press, Amsterdam, pp 35–57

Lösel, F., Beelmann, A., & Stemmler, M. (2002). *Skalen zur Messung sozialen Problemverhaltens bei Vorschul- und Grundschulkindern: Die deutschen Versionen des Eyberg Child Behavior Inventory (ECBI) und des Social Behavior Questionnaire (SBQ)*. [Scales for measuring problem behavior in preschool- and elementary school children: The German versions of the Eyberg Child Behavior Inventory (ECBI) und des Social Behavior Questionnaire (SBQ)]. Universität Erlangen-Nürnberg: Institut für Psychologie.

Lösel F., Bender, D. (2003). Protective factors and resilience. In: Farrington DP, Coid JW (eds) Early prevention of adult antisocial behaviour. Cambridge University Press, Cambridge, pp 130–204

Lösel F, Farrington, D.P. (2012). Direct protective and buffering protective factors in the development of youth violence. American Journal of Preventive Medicine 43(2, supplement):8–23

Lösel F., Stemmler, M. (2012). Preventing child behavior problems at pre-school age: The Erlangen-Nuremberg development and prevention study. International Journal of Violence and Conflict 6(2):214–224

Lösel F., Stemmler M, Beelmann A, Jaursch, S. (2005). Aggressives Verhalten im Vorschulalter: Eine Untersuchung zum Problem verschiedener Informanten [Aggressive behavior at preschool age: A study on the issue of different informants]. In: Seiffge-Krenke I (ed) Aggressionsentwicklung zwischen Normalität und Pathologie [Development of aggression between normality and pathology]. Vandenhoeck & Ruprecht, Göttingen, pp 141–167

Lösel F., Stemmler M, Bender, D. (2013). Long-term evaluation of a bimodal universal prevention program: Effects on antisocial development from kindergarten to adolescence. Journal of Experimental Criminology 9(4):429–449

Lösel F., Stemmler M, Jaursch S, Beelmann, A. (2009). Universal prevention of antisocial development: Short- and long-term effects of a child and parent-oriented program. Monatsschrift für Kriminologie und Strafrechtsreform [Monthly Journal of the Criminal and Penal Law Reform], 92:289–308

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In: Nesselroade JR, Cattell RB (eds) The handbook of multivariate experimental psychology, vol 2. Plenum Press, New York, pp 561–614

McArdle J. J. (2005). Latent growth curve analysis using structural equation modeling techniques. In: Teti DM (ed) The handbook of research methods in developmental psychology. Blackwell Publishers, New York, pp 340–466

McCartney, K., Burchinal, M. R., & Bub, K. (2006). Best practices in quantitative methods for developmentalists. *Monographs of the Society for Research in Child Development, Serial No 285, 71*(3).

McLachlan G. J., Peel D. (2000). Finite mixture models. Wiley, New York

Moffitt, T. E. (1993). "Life-course persistent" and "adolescence limited" antisocial behavior: A developmental taxonomy. Psychological Review 100:674–701

Moffitt T. E., Caspi A, Dickson N, Silva P, Stanton, W. (1996). Childhood-onset versus adolescent-onset antisocial conduct problems in males: Natural history from ages 3 to 18 years. Development and Psychopathology 8(02):399–424

Moffitt, T. E., Caspi A, Harrington H, Milne, B. J. (2002). Males on the life-course-persistent and adolescence-limited antisocial pathways: Follow-up at age 26 years. Development and Psychopathology 14:179–207

Moffitt, T. E., Caspi A, Rutter M, Silva, P. A. (eds) (2001). Sex differences in antisocial behavior: Conduct disorder, delinquency, and violence in the Dunedin longitudinal study. Academic, New York, pp 53–70

Moretti M, Odgers C (2002) Aggressive and violent girls: Prevalence, profiles and contributing factors. NATO Science Series Sub Series I Life and Behavioural Sciences 324:116–129

Muthén, B.O. (2004). Latent variable analysis: Growth mixture modeling and related technique for longitudinal data. In: Kaplan D (ed) The Sage handbook of quantitative methodology for the social science. Sage, Thousand Oaks, CA, pp 345–368

Muthén, L. K., Shedden K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 55(2):463–469

Muthén, L. K., Muthén, B. O. (2010). Mplus user's guide, 6th edn. Muthén & Muthén, Los Angeles, CA

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi parametric, group based approach. Psychological Methods 4:139–177

Nagin, D. S., Farrington, D. P., Moffitt, T. E. (1995). Life-course trajectories of different types of offenders. Criminology 33:111–139

Nagin, D. S., Tremblay, R. E. (1999). Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. Child Development 70:1181–1196

Nesselroade J, Baltes, P. (1979). Longitudinal research in the study of behavior and development. Academic, New York

Piquero A. R., Farrington D. P., Blumstein, A. (2007). Key issues in criminal career research. Cambridge University Press, Cambridge

Reef J, Diamantopoulou S, van Meurs I, Verhulst, F. C., van der Ende, J. (2011). Developmental trajectories of child to adolescent externalizing behavior and adult DSM-IV disorder: results of a 24-year longitudinal study. Social Psychiatry and Psychiatric Epidemiology 46(12): 1233–1241

Reinecke, J. (2005). Strukturgleichungsmodelle in den Sozialwissenschaften [Structural equation modeling in the social sciences]. Oldenbourg Verlag, München, Wien

Reinecke, J. (2006). *Delinquenzverläufe im Jugendalter: Empirische Überprüfung von Wachstums- und Mischverteilungsmodellen [Trajectories of delinquency in adolescence: Empirical analyses of growth curve models and mixture models]*. Sozialwissenschaftliche Forschungsdokumentationen 20. Münster: Institut für sozialwissenschaftliche Forschung e.V. *[Social science research documentation 20]*.

Reinecke, J. (2012). Wachstumsmodelle [Growth curve modeling]. Rainer Hampp Verlag, Mering

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software 48:1–36

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Ryder, J. A., Gordon C, Bulger, J. (2009). Contextualizing girls' violence: Assessment and treatment decisions. In: Andrade JT (ed) Handbook of violence risk assessment and treatment: New approaches for mental health professionals. Springer Publishing Company, New York, pp 449–493

Shannon, L. W. (1988). Criminal career continuity: Its social context. Human Science Press, New York

Silverthorne P, Frick, P. J. (1999). Developmental pathways to antisocial behaviour: The delayed-onset pathway in girls. Development and Psychopathology 11:101–126

Stemmler M, Lösel, F. (2012). The stability of externalizing behavior in boys from preschool age to adolescence: A person-oriented analysis. Psychological Test and Assessment Modeling 54(2):195–207

Stemmler M, Petersen, A. C. (2012). Latent growth curve modeling and the study of problem behavior in girls. In: Bliesener T, Beelmann A, Stemmler M (eds) Antisocial behavior and crime: Contributions of developmental and evaluation research to prevention and intervention. Hogrefe Publishing, Cambridge, MA, pp 315–332

Thornberry, T. P. (ed) (1997). Developmental theories of crime and delinquency. Transaction Publishers, New Brunswick, NJ

Tracy, P. E., Wolfgang M. E., Figlio, R. M. (1990). Delinquency careers in two birth cohorts. The Plenum series in crime and justice. Plenum Press, New York

Tremblay, R. E., Desmarais-Gervais L, Ganon C, Charlebois, P. (1987). The Preschool Behavior Questionnaire. Stability of its factor structure between cultures, sexes, ages and socioeconomic classes. International Journal of Behavioral Development 10:467–484

Tremblay, R. E., Nagin, D. S., Séguin, J. R., Zoccolillo M, Zelazo P. D., Boivin M, Pérusse D, Japel C. (2004). Physical aggression during early childhood: Trajectories and predictors. Pediatrics 114:43–50

Tremblay, R. E., Vitaro F, Gagnon C, Piché C, Royer, N. (1992). A prosocial scale for the Preschool Social Behavior Questionnaire: Concurrent and predictive correlates. International Journal of Behavioral Development 15:227–245

Wallner, S., Lösel, F., Stemmler, M., & Corrado, R. Prediction of antisocial development in preschool children using the Cracow instrument: A follow-up of five years in a community sample. Manuscript submitted for publication.

Werner, E. E., Smith, R. S. (1992). Overcoming the odds: High risk children from birth to adulthood. Cornell University Press, Ithaca, NY

Werner E. E., Smith R. S. (2001). Journeys from childhood to midlife: Risk, resilience, and recovery. Cornell University Press, Ithaca, NY

A Generalization of Nagin's Finite Mixture Model

Jang Schiltz

Abstract We present a generalization of Nagin's finite mixture model that allows non-parallel trajectories for different values of covariates. We investigate some mathematical properties of this model and illustrate its use by giving typical salary curves for the employees in the private sector in Luxembourg between 1981 and 2006, as a function of their gender, as well as of Luxembourg's gross domestic product (GDP).

Introduction

Longitudinal data are the empirical basis of research on various subjects in sociology, psychology, economics, criminology, and medicine and a host of statistical techniques are available for analyzing them (see Singer & Willet 2003). The common statistical aim of these various application fields is the modelization of the evolution of an age or time based phenomenon (Nagin 2002). Hence, the study of developmental trajectories is a central theme (Ferguson, Lynskey, & Horwood 1996; Jones, Nagin, & Roeder 2001; Moffitt 1993; Patterson, DeBaryshe, & Ramsey 1989; Sampson & Laub 2005). The objective of these approaches is to capture information about interindividual differences in intraindividual change over time (Nesselroade 1991). In the 1990s, the generalized mixed model assuming a normal distribution of unobserved heterogeneity (Bryk & Raudenbush 1992), multilevel modeling (Goldstein 1995), latent growth curves modeling (Muthén 1989; Willett & Sayer 1994), and the nonparametric mixture model, based on a discrete distribution of heterogeneity (Jones et al. 2001) have emerged. There has been a growing interest in this approach to answer questions about atypical subpopulations (see Eggleston, Laub, & Sampson 2004).

Growth mixture modeling, introduced by Muthén and Shedden (1999), is a very suitable framework to handle the issue of unobserved heterogeneity. They can be seen as an extension of the structural modeling approach with techniques of latent

J. Schiltz (✉)
University of Luxembourg, LSF, 4, rue Albert Borschette, L-1246 Luxembourg, Luxembourg
e-mail: jang.schiltz@uni.lu

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer
Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_5

class analysis (Muthén 2001). The inferred membership of each individual to a certain class is produced with the information of the estimated class probabilities (Reinecke & Mariotti 2009). Extensive applications of different growth curve models with structural equations are discussed by Duncan, Stryker, Li, and Alpert (2006) and a general nonlinear multilevel structural equation mixture model, that combines recent semiparametric nonlinear structural equation models with multilevel structural equation mixture models for clustered and non-normally distributed data is presented in Keleva and Brandt (2014). An overview of the different concepts in mixture modeling, on the other hand, can be found in Young (2008).

Latent class growth analysis, also called nonparametric mixed model or semiparametric mixture model, is the simplest specification of a growth mixture model. It allows no variation across individuals within classes. It was originally discussed by Nagin and Land (1993), Nagin (1999), and Roeder, Lynch, and Nagin (1999) and is actually specifically designed to detect the presence of distinct subgroups among a set of trajectories and represents an interesting compromise between analysis around a single mean trajectory and case studies (von Eye & Bergman 2003). Compared to subjective classification methods, the nonparametric mixed model has the advantage of providing a formal framework for testing the existence of distinct groups of trajectories. This method does not assume a priori that there is necessarily more than one group in the population. Rather, an adjustment index is used to determine the number of sub-optimal groups. This is a significant advance over other categorical methods which determine the number of groups only subjectively (von Eye & Bergman 2003). Andruff, Carraro, Thompson, Gaudreau, and Louvet (2009) conclude that latent class growth analysis serves as a steppingstone to growth mixture modeling analyses in which the precise number and shape of each trajectory must be known a priori in order for the researcher to impute the requisite start values for the model to converge in software packages such as Mplus (Jung & Wickrama 2008).

While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the model's estimated parameters are not the result of a cluster analysis but of maximum likelihood estimation (Nagin 2005). Moreover, this method allows to evaluate the accuracy of the assignment of the individuals to the different sub-groups and to consider the variation of this accuracy in subsequent analyses (Dupéré, Lacourse, Vitaro, & Tremblay 2007). Nagin and Odgers (2010) document numerous applications of group-based trajectory modeling in criminology and clinical research. They state that the appeal of group-based trajectory modeling for the future lies in the potential for the innovative application of trajectory models on their own, in conjunction with other statistical methods or embedded within creative study designs while carefully considering the perils and pitfalls inherent in the use of any methodology.

The remainder of this article is structured as follows. In section "Nagin's Finite Mixture Model," we present the basic version of Nagin's finite mixture model, as well as one of his generalizations, allowing to add covariates to the trajectories and we show two drawbacks of the model. In section "Our Model," we present a generalization of the model that overcomes these drawbacks and we discuss model

selection and group member probabilities for the new model. Section "Statistical Properties" presents some basic statistical properties of the model. In section "A Data Example," finally, we highlight typical features of the new model by means of a data example from economics.

Nagin's Finite Mixture Model

Starting from a collection of individual trajectories, the aim of Nagin's finite mixture model is to divide the population into a number of homogenous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population (Nagin 2005).

More, precisely, consider a population of size N and a variable of interest Y. Let $Y_i = y_{i_1}, y_{i_2}, \ldots, y_{i_T}$ be T measures of the variable Y, taken at times $t_1, \ldots, t_T$ for subject number i.

To estimate the parameters defining the shape of the trajectories, we need to fix the number r of desired subgroups. Denote the probability of a given subject to belong to group number j by π_j. Then π_j is also the size of group j and

$$P(Y_i) = \sum_{j=1}^{r} \pi_j P^j(Y_i), \tag{1}$$

where $P^j(Y_i)$ is the probability of Y_i if subject i belongs to group j.

This model is called a finite mixture model, because we suppose that the population is composed of a mixture of unobserved groups and Eq. (1) sums across this finite number of discrete groups that compose the population.

If we suppose conditional independence for the sequential realizations of the elements y_{it} over the T periods of measurement for each group, we obtain

$$P^j(Y_i) = \prod_{i=1}^{N} p^j(y_{it}), \tag{2}$$

where $p^j(y_{it})$ is the probability distribution function of y_{it} given membership in group j.

Nagin specified his model for three different kinds of distributions (Nagin 2005). For count data, $P(Y_i)$ is specified as the Poisson distribution, for binary data it is specified as the binary logit distribution, and for censored data it is specified as the censored normal distribution.

In any case, the objective is to estimate a set of parameters $\Omega = \{\pi_j, \beta_0^j, \beta_1^j, \ldots; j = 1, \ldots, r\}$ which allow to maximize the probability of the measured data. The particular form of Ω is distribution specific, but the β parameters always perform the basic function of defining the shapes of the trajectories. In both

standard growth curve modeling and Nagin's finite mixture model, the shapes of the trajectories are described by a polynomial function of age or time (Nagin 2005).

In this paper, we suppose that the data follow a normal distribution (not necessarily censored). Assume that for a subject in group j

$$y_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k + \varepsilon_{it}, \tag{3}$$

where s denotes the order of the polynomial describing the trajectories in group j and ε_{it} is the disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ. If we denote the density of the standard centered normal law by ϕ and $\beta^j t_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k$, the likelihood of the data is given by

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi\left(\frac{y_{it} - \beta^j t_{it}}{\sigma}\right). \tag{4}$$

The disadvantage of the basic model is that the trajectories are static and do not evolve in time. Thus, Nagin introduced several generalizations of his model in his book (Nagin 2005). Among others, he introduced a model allowing to add covariates to the trajectories. Let $z_1, \ldots, z_M$ be M covariates potentially influencing Y.

We are then looking for trajectories

$$y_{it} = \sum_{k=0}^{s} \beta_k^j t_{it}^k + \alpha_1^j z_1 + \ldots + \alpha_M^j z_M + \varepsilon_{it}, \tag{5}$$

where ε_{it} is normally distributed with zero mean and a constant standard deviation σ. The covariates z_m may depend or not upon time t.

But even this generalized model still has two major drawbacks.

First, the influence of the covariates in this model is unfortunately limited to the intercept of the trajectory. This implies that for different values of the covariates, the corresponding trajectories will always remain parallel by design, which does not necessarily correspond to reality.

Secondly, in Nagin's model, the standard deviation of the disturbance is the same for all the groups. That too is quite restrictive. One can easily imagine situations in which in some of the groups all individual are quite close to the mean trajectory of their group, whereas in other groups there is a much larger dispersion.

Our Model

Definition

To address and overcome these two drawbacks, we propose the following generalization of Nagin's model.

Let $x_1 \ldots x_M$ and $z_{i_1}, \ldots, z_{i_T}$ be covariates potentially influencing Y. Here the x variables are covariates not depending on time like gender or cohort membership in a multicohort longitudinal study and the z variable is a covariate depending on time like being employed or unemployed. They can of course also designate time-dependent covariates not depending on the subjects of the data set which still influence the group trajectories, like gross domestic product (GDP) of a country in case of an analysis of salary trajectories.

The trajectories in group j will then be written as

$$y_{it} = \sum_{k=0}^{s} \left(\beta_k^j + \sum_{m=1}^{M} \alpha_{km}^j x_m + \gamma_k^j z_{i_t} \right) t_{it}^k + \varepsilon_{it}, \tag{6}$$

where the disturbance ε_{it} is normally distributed with mean zero and a standard deviation σ_j constant inside group j but different from one group to another. Since, for each group, this model is just a classical fixed effects model for panel data regression (see Woolridge 2002), it is well defined and we can get consistent estimates for the model parameters.

Our model allows obviously to overcome the drawbacks of Nagin's model. The standard deviation of the uncertainty can vary across groups and the trajectories depend in a nonlinear way on the covariates. In practice this dependence of all the power coefficients of the polynomials may considerably extend the computation time for the parameters, so it can be useful just to work with a first or second order dependence instead of using the full model.

On the other hand, it is even possible to further generalize the model and consider trajectories that are not polynomial. In economics, for instance, there is often the need to consider exponential trajectories.

Let f^j be a function describing the trajectory in group j and depending on parameters $\beta_0, \ldots, \beta_s$. Then the trajectories in group j can be written as

$$y_{it} = f\left(t_{it}, z_{i_t}; \beta_0^j, \ldots, \beta_s^j, \gamma_0^j, .., \gamma_s^j\right) + \varepsilon_{it}, \tag{7}$$

where the disturbance ε_{it} is normally distributed with mean zero and a standard deviation σ_j constant inside group j but different from one group to another. To avoid too complicated notations, we have left aside here covariates not depending on time, but these could also be included in the trajectories. The important fact is that as long as the description of the trajectories remains deterministic, we can choose all kinds of shapes we like. The econometrical properties just depend on the disturbance term which remains unchanged.

Since our model is just a generalization of Nagin's finite mixture model, a lot of its main features and properties remain the same as in Nagin's model.

Model Selection

The problem of how many components to include in a finite mixture model is among the most challenging in statistics (Nagin 2005). Bauer and Curran (2003) even cautioned that the existence of multiple classes may simply be due to skewed or otherwise non-normally distributed data. There have been a host of propositions for a criterion for deciding the correct number of groups, but there is not really a common acceptance of the best criteria. This is seen as a critical issue in the application of mixture modeling, because classes are used for interpreting results and making inferences (Nylund, Asparouhov, & Muthén 2007). One widely recommended option is the Bayesian Information Criterion (Kass & Raftery 1995; Nagin 2005; Raftery 1995; Schwarz 1978). If k denotes the number of parameters in the model, BIC is calculated as

$$\text{BIC} = \log(L) - 0.5k\log(N). \tag{8}$$

The bigger the BIC, the better the model.

Closely related to the BIC is Akaike information criteria (AIC) (Akaike 1974), defined by

$$AIC = -2\log(L) + 2k. \tag{9}$$

Here the optimal number of groups is the one minimizing AIC.

Recently, Nielsen et al. (2014) proposed the methodology of leave-one-out cross-validation error (CVE). CVE is calculated as

$$CVE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\left|y_{it} - \hat{y}_{it}^{[-i]}\right|, \tag{10}$$

where $\hat{y}_{it}^{[-i]}$ is the forecast for individual i if the model is estimated based on data for all the other individuals except for himself. The decision rule is to take the number of groups corresponding to the smallest CVE. This looks like a nice idea, but it is not always computationally feasible for large data sets. Since instead of estimating the model once, it necessitates to estimate the model N times!

Besides, it should not be forgotten that there does not always exist a "best" number of groups. If the groups are seen as a statistical device for approximating an unknown but continuous population distribution of trajectories, the question of what constitutes an optimum number of groups is ill-posed (Nagin 2005).

There has been an extensive debate about the optimal number of groups in the literature (Nagin & Tremblay 2005; Sampson & Laub 2005) with the conclusion that analogous to determining the number of factors using explanatory factor analysis, the researcher should ultimately use a combination of factors in addition to fit indices, including his research question, parsimony, theoretical justification, and

interpretability (Jung & Wickrama 2008) and that often subject-specific judgment is more important than statistical considerations to choose the number of groups (Bloklad, Nagin, & Nieuwbeerta 2005; Eggleston et al. 2004).

Group Membership Probability

The posterior probability of individual i's membership in group j $P(j/Y_i)$ can easily be computed with Bayes's theorem.

$$P(j/Y_i) = \frac{P(Y_i/j)\,\hat{\pi}_j}{\sum_{j=1}^{r} P(Y_i/j)\,\hat{\pi}_j}. \tag{11}$$

Hence, bigger groups have on average larger probability estimates. Besides, to be classified into a small group, an individual really needs to be strongly consistent with it (Nagin 2005).

These probabilities can then be used to create balance on lagged outcomes and other covariates established prior to t for the purpose of inferring the impact of first-time treatment on the outcome of interest (see Haviland & Nagin 2005) and deciding thus whether a therapeutic intervention (or a turning-point event) alters the trajectories under study.

Statistical Properties

The model's estimated parameters are the result of maximum likelihood estimation. As such, they are consistent and asymptotically normally distributed (Cramér 1946; Greene 1995; Theil 1971).

In our model, for a given group, the trajectories follow in fact a nonlinear regression model. As such, exact confidence interval procedures or exact hypothesis tests for the parameters are generally not available (Graybill & Iyer 1994). There exist, however, approximative solutions. The standard error can be approximated, for instance, by a first-order Taylor series expansion (Greene 1995). This approximate standard error (ASE) is usually quite precise if the sample size is sufficiently large.

Consider model (6), for which $(2 + M)s$ regression parameters have to be estimated. Then confidence intervals of level α for the parameters β_k^j are just

$$CI_\alpha(\beta_k^j) = \left[\hat{\beta}_k^j - t_{1-\alpha/2;N-(2+M)s} ASE(\hat{\beta}_k^j); \hat{\beta}_k^j + t_{1-\alpha/2;N-(2+M)s} ASE(\hat{\beta}_k^j)\right], \tag{12}$$

where $t_{1-\alpha;n}$ denotes as usual the $1-\alpha$ quantile of the Student distribution with n degrees of freedom.

The confidence intervals for the α_{kl}^{j} and γ_{k}^{j} are obtained in the same way.
The confidence intervals of level α for the disturbance factor σ_j is given by

$$CI_{\alpha}(\sigma_j) = \left[\sqrt{\frac{(N-(2+M)s-1)\hat{\sigma_j}^2}{\chi^2_{1-\alpha/2;N-(2+M)s-1}}}; \sqrt{\frac{(N-(2+M)s-1)\hat{\sigma_j}^2}{\chi^2_{\alpha/2;N-(2+M)s-1}}} \right], \qquad (13)$$

where $\chi^2_{1-\alpha;n}$ denotes the $1-\alpha$ quantile of the Chi-Square distribution with n degrees of freedom.

A Data Example

For the following example, we use Luxembourg administrative data originating from the General Inspectorate of Social Security, IGSS (Inspection générale de la sécurité sociale). The data have previously been described and exploited with Nagin's basic model by Guigou, Lovat, and Schiltz (2010, 2012). The file contains the salaries of all employees of the Luxembourg private sector who started their work in Luxembourg between 1980 and 1990 at an age of less than 30 years. This choice was made to eliminate people with a long carrier in another country before moving to Luxembourg. The main variables are the net annual taxable salary, measured in constant (2006 equivalent) euros, gender, age at first employment, residentship and nationality, sector of activity, marital status, and the years of birth of the children. The file consists of 1,303,010 salary lines corresponding to 85,049 employees. In Luxembourg, the maximum contribution ceiling on pension insurance is five times the minimum wage, currently 7577 EUR (2006 equivalent euros) per month. Wages in our data are thus also capped at that number.

We will not present here an exhaustive analysis of the whole dataset, but just two illustrations of the possibilities of our generalized mixture model and its differences from Nagin's model. We concentrate on the first 20 years of the careers of the employees who started working in Luxembourg in 1987. That gives us a sample of 1716 employees. We will first compute typical salary trajectories for them, taking into account the gender of the employees and then typical salary trajectories as a function of the GDP of the country.

Since we are in a somewhat special situation where we work with the complete population and not just a sample, it may seem a bit strange to speak about parameter significance and confidence intervals for this example. But first, this is just an illustration of the possibilities and main features of our model, so it makes sense to show what results we would get in a classical situation. And more importantly, in case of a use of the results to predict the future salary evolution, we are dealing in fact with just a subsample of the whole population. If we argue that for a reasonable time horizon, the typical salary trajectories just depend on the covariates that we included in our equations, then the complete set of people starting to work in

2006 is just a part of the whole population of people starting to work in 2006 and the subsequent years. Confidence intervals for the salary trajectories then indicate prediction bounds.

First Illustration

Figure 1 shows a three group solution modeled by Nagin's generalized model representing the salary of employees in Luxembourg during the first 20 years of their professional career. We see that for the low salary group women and men are gaining exactly the same salary (with the consequence that there appears just one salary trajectory for the two lower salary groups on the graph instead of two) whereas in the middle and high salary groups, men earn more than women. Due to the limitations of the model, the evolution of the salaries seems to be exactly the same for men and women; their salary trajectories are strictly parallel.

Figure 2 shows the three group solution for the 20 first year of Luxembourg employees calibrated with our model. We see a somewhat different and more realistic pattern emerging. For the high salary group the income of men and women remain more or less parallel, except for a short time interval around year five. This is however no longer the case for the middle and low salary groups. Here, we observe that the women in these groups have higher salaries than the men at the beginning

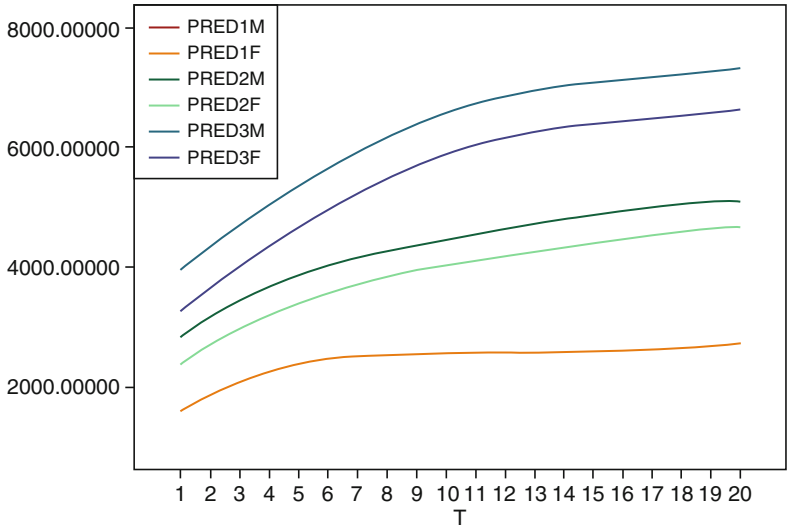


Fig. 1 Salary evolution by gender, modelized by Nagin's model

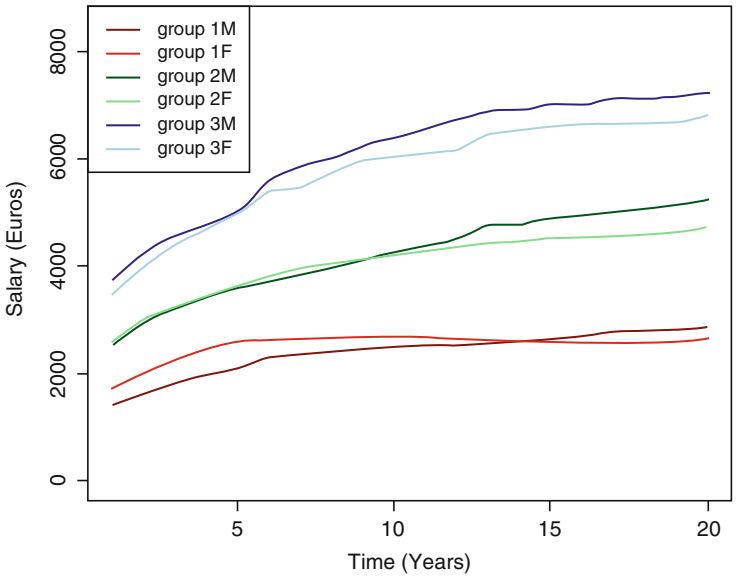


Fig. 2 Salary evolution by gender, modelized by our model

of their career, but this is reversed somewhere in the middle and after 10 years for the middle salary group and 15 years for the low salary group the income of the men becomes higher than the one of the women.

We obtained this results by calibrating the model

$$S_{it} = (\beta_0^j + \alpha_0^j x_i) + (\beta_1^j + \alpha_1^j x_i)t + (\beta_2^j + \alpha_2^j x_i)t^2 + (\beta_3^j + \alpha_3^j x_i)t^3, \qquad (14)$$

where S denotes the salary and x the gender. Table 1 shows the values of the parameters for a 3-group solution.

We observe that all parameters are significant, with the exception of β_3 for the middle and higher salary group. Hence there really seems to be a nonlinear relation between the salaries and the gender and a simple parallel shift is not enough to explain what is going on.

The disturbance terms for the three groups are $\sigma_1 = 33.11$, $\sigma_2 = 54.18$ and $\sigma_3 = 78.85$, respectively. The dispersion is thus higher in the groups with higher salaries than in those with lower salaries. This makes sense, since in the low salary group a lot of employees just earn the minimal wage. Hence, a lot of them have the same salary.

Table 1 Parameter estimates for model (14)

Parameter	Estimate	Standard error	95 % confidence interval	
			Lower	Upper
Results for group 1				
β_0	1166.353	37.910	1085.987	1246.719
α_0	256.208	3.490	248.809	263.607
β_1	275.505	15.254	243.169	307.841
α_1	99.868	1.405	84.845	114.891
β_2	−19.076	1.666	−22.608	−15.543
α_2	11.826	0.153	11.501	12.151
β_3	0.484	0.052	0.372	0.594
α_3	0.325	0.005	0.315	0.335
Results for group 2				
β_0	2397.209	76.051	2235.987	2558.430
α_0	−79.595	33.731	−151.103	−8.087
β_1	275.972	30.600	211.103	340.842
α_1	82.874	13.572	50.293	115.455
β_2	−10.238	3.343	−17.325	−3.151
α_2	−11.024	1.483	−15.047	−7.001
β_3	0.178	0.104	−0.044	0.400
α_3	0.287	0.047	0.150	0.424
Results for group 3				
β_0	3289.495	90.003	3098.698	3480.119
α_0	−258.446	7.752	−292.977	−223.915
β_1	464.349	36.214	387.580	541.119
α_1	48.97	3.119	34.954	62.986
β_2	−17.111	3.956	−25.498	−8.724
α_2	−8.343	0.341	−14.398	−2.288
β_3	0.181	0.124	−0.082	0.444
α_3	0.279	0.011	0.273	0.285

Second Illustration

The second example illustrates the dependence of the trajectories on time-varying covariates. We use the same data as before and analyze the influence of Luxembourg's GDP on the salary trajectories. GDP denotes here in fact Luxembourg's GDP of the previous year, since standard economical theory tells us that there is a time lag of nearly a year for the influence of GDP on the salaries. We use the model

$$S_{it} = (\beta_0^j + \gamma_0^j z_{it}) + (\beta_1^j + \gamma_1^j z_{it})t + (\beta_2^j + \gamma_2^j z_{it})t^2 + (\beta_3^j + \gamma_3^j z_{it})t^3, \qquad (15)$$

where S denotes the salary and z_t is Luxembourg's GDP in year t of the study. The first question to settle is how many groups we want to use in our solution.

Table 2 BIC of solutions for various number of groups

Number of groups	BIC	Number of empty groups
3	−285193.1	0
4	−282444.2	0
5	−282197.7	0
6	−279710.3	0
7	−279415.1	1
8	−279238.3	2
9	−278162.3	2
10	−277312.7	2
11	−277335.1	3
12	−276637.7	3

We calibrate the model for solutions between 3 and 12 groups and compute the BIC for each of them (see Table 2). Besides the BIC, we also indicate the number of "empty groups" in the solution. The term "empty group" is used here for groups of size smaller than 0.1 %. Since we are interested in typical salary trajectories, we consider those empty groups as outliers and prefer solutions containing just noticable groups, meaning groups with larger sizes.

We see that in this example the BIC is in fact an increasing function of the number of groups, which is not astonishing since the salary trajectories form a continuum. But we also observe, that up to 6 groups, the solutions contain no empty groups, whereas from a 7 group solution onwards, there are empty groups in the solution. In the 7 group solution, there is one such group, in the 8, 9, and 10 group solutions, there are two and in the 11 and 12 group solutions, there are three.

We finally decide on a 6 group solution. Group sizes are quite balanced, the different group sizes are indeed $\pi_1 = 15.9\,\%$, $\pi_2 = 16.6\,\%$, $\pi_3 = 21.2\,\%$, $\pi_4 = 14.4\,\%$, $\pi_5 = 14.9\,\%$, and $\pi_6 = 16.9\,\%$

Figure 3 shows the salary trajectories of the 6 groups (scale at the left side of the y axis), as well as the GDP of Luxembourg (in black, scale on the right side of the y axis) during the same time. Group one contains mainly the employees that gain the legal minimum wage. Groups two and six represent employees with rather flat careers. Their salary is more or less constant from year five on. They are just distinguished by their starting salary. Groups three, four, and five, on the other hand, represent more dynamical careers, again characterized mainly by the differences in their starting salaries.

Tables 3 and 4 show the values of the parameters for a 3 group solution. Significant parameters are given in bold. We recall that the beta parameters modelize the evolution in time, independently of the GDP, whereas the gamma parameters modelize the part of the salary varying with the GDP.

We observe that in most groups, there is no significant influence of the GDP. In group three, we have an influence from GDP, as well as a combined influence from GDP and time. In group six, we observe a combined influence from GDP and time. In the four other groups, the salary trajectories are just a polynomial function of time.

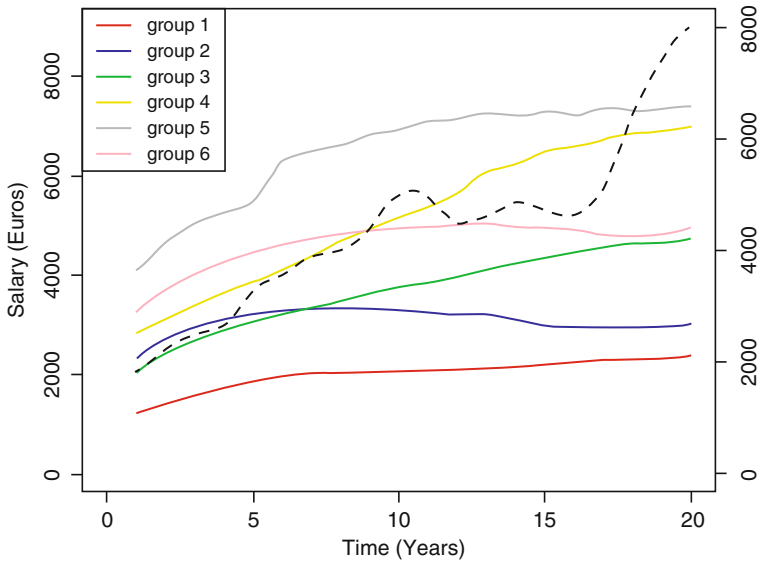


Fig. 3 Typical salary trajectories for the 6 group solution and evolution of Luxembourg's GDP (*dashed line*)

The disturbance terms for the six groups are $\sigma_1 = 21.29, \sigma_2 = 41.9, \sigma_3 = 40.02$, $\sigma_4 = 58.81, \sigma_5 = 124$, and $\sigma_6 = 31.22$, respectively. Again, we observe that the minimal wage group exhibits the smallest variability, whereas the high salary groups four and five also have the highest disturbance term.

Discussion

In this article, we presented Nagin's finite mixture model and some of its generalizations and showed some inherent shortcomings for some possible application. We addressed these by proposing a new generalized finite mixture model. A key characteristic is its ability to modelize nearly all kinds of trajectories and to add covariates to the trajectories themselves in a nonlinear way.

We illustrated these possibilities through a data example about salary trajectories. In the first part, we showed how to add a classical group membership predictor variable to the trajectories and in the second part, we added a time series that does not depend on the subjects of the analysis but influences the shape of the trajectories in some of the groups.

When adding covariates to the trajectories in growth mixture modeling, an important question is whether these covariates are predictors of group membership or not. Nagin (2005) and Jones and Nagin (2007) present some statistical tests to

Table 3 Parameter estimates for model (15)

Parameter	Estimate	Standard error	95 % confidence interval	
			Lower	Upper
Results for group 1				
$\boldsymbol{\beta_0}$	1038.164	225.623	546.958	1529.815
γ_0	−0.086	0.017	−0.462	0.291
$\boldsymbol{\beta_1}$	265.204	35.930	186.927	343.423
γ_1	0.028	0.029	−0.035	0.091
$\boldsymbol{\beta_2}$	−25.520	6.035	−38.670	−12.370
γ_2	−0.002	0.002	−0.006	0.002
$\boldsymbol{\beta_3}$	0.914	0.024	0.395	1.432
γ_3	0.000025	0.000039	−0.000060	0.000197
Results for group 2				
$\boldsymbol{\beta_0}$	1558.955	44.380	590.638	2525.115
γ_0	0.244	0.340	−0.497	0.985
$\boldsymbol{\beta_1}$	516.538	70.704	362.521	670.647
γ_1	−0.076	0.057	−0.070	0.047
$\boldsymbol{\beta_2}$	−43.103	11.872	−68.974	−17.230
γ_2	0.006	0.003	−0.001	0.001
$\boldsymbol{\beta_3}$	0.949	0.468	−0.071	1.969
γ_3	−0.000149	0.000076	−0.000315	0.000018
Results for group 3				
β_0	731.828	423.905	−191.737	1655.324
$\boldsymbol{\gamma_0}$	0.708	0.329	0.001	1.416
$\boldsymbol{\beta_1}$	496.510	67.526	349.482	643.674
$\boldsymbol{\gamma_1}$	−0.169	0.054	−0.286	−0.056
β_2	−20.551	11.342	−45.254	4.160
$\boldsymbol{\gamma_2}$	0.012	0.003	0.005	0.019
β_3	0.253	0.447	−0.721	1.227
$\boldsymbol{\gamma_3}$	−0.002542	0.000073	−0.000414	−0.000095

check this. In case a covariate is a predictor of group membership, it not only influences the shape of the trajectories but also group membership itself, as well as the composition of the different groups. If it is not, there is an alternative way to see our model. It is then in fact equivalent to perform the clustering and compute the number and composition of the groups with Nagin's basic finite mixture model and use standard regression models for each groups to get the trajectories as a function of the covariate.

Table 4 Parameter estimates for model (15)

Parameter	Estimate	Standard error	95 % confidence interval	
			Lower	Upper
Results for group 4				
$\boldsymbol{\beta_0}$	1933.257	622.902	575.626	3289.048
γ_0	0.503	0.477	−0.537	1.543
$\boldsymbol{\beta_1}$	341.610	99.233	125.387	557.892
γ_1	−0.139	0.079	−0.213	0.003
β_2	13.051	16.673	−23.272	49.368
γ_2	0.009	0.005	−0.001	0.019
β_3	−0.993	0.657	−2.424	0.438
γ_3	−0.000152	0.000107	−0.000385	0.000082
Results for group 5				
$\boldsymbol{\beta_0}$	3662.004	1313.374	800.711	6523.683
γ_0	0.004	1.006	−2.188	2.196
β_1	357.604	209.216	−98.294	813.335
γ_1	0.065	0.168	−0.299	0.430
β_2	−2.738	35.134	−79.281	73.808
γ_2	−0.008	0.010	−0.030	0.014
β_3	−0.296	1.384	−3.312	2.721
γ_3	0.000256	0.000226	−0.000237	0.000749
Results for group 6				
$\boldsymbol{\beta_0}$	2278.347	330.711	1557.590	2998.241
γ_0	0.442	0.253	−0.110	0.994
$\boldsymbol{\beta_1}$	495.228	52.683	380.412	610.014
$\boldsymbol{\gamma_1}$	−0.099	0.042	−0.191	−0.007
β_2	−16.037	8.851	−35.314	3.274
γ_2	0.004	0.003	−0.001	0.010
β_3	−0.266	0.349	−1.026	0.494
γ_3	−0.000005	0.000057	−0.000129	0.000119

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Andruff, H., Carraro, N., Thompson, A., Gaudreau, P., & Louvet, B. (2009). Latent class growth modelling: A tutorial. *Tutorials in Quantitative Methods for Psychology, 5*(1), 11–24.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*, 338–363.

Bloklad, A. A., Nagin, D. S., & Nieuwbeerta, P. (2005). Life span offending trajectories of a Dutch conviciton cohort. *Criminology, 43*, 919–954.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Duncan, T. E., Stryker, L. A., Li, F., & Alpert, A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues and applications*. Mahwah, NJ: Lawrence Erlbaum.

Dupéré, V., Lacourse, E., Vitaro, F., & Tremblay, R. E. (2007). Méthodes d'analyse du changement fondées sur les trajectoires de développement individuel: modèles de régression mixtes paramétriques et non paramétriques. *Bulletin de MÉthodologie Sociologique, 95*, 26–57.

Eggleston, E. P., Laub, J. H., & Sampson, R. J. (2004). On the Robustness and Validity of Groups. *Journal of Quantitative Criminology, 20*(1), 37–42.

Ferguson, D. M., Lynskey, M. T., & Horwood, L. J. (1996). Factors associated with continuity and change in disruptive behavior patterns during childhood and adolescence. *Journal of Abnormal Child Psychology, 24*, 533–553.

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Graybill, F. A., & Iyer, H. K. (1994). *Regression analysis: Concepts and applications*. Belmont,CA: Duxbury Press.

Greene, W. H. (1995). *Econometric analysis*. New York: Macmillan.

Guigou, J.-D., Lovat, B., & Schiltz, J. (2010). The impact of ageing population on pay-as-you-go pension systems: The case of Luxembourg. *Journal of International Finance and Economics, 10*(1), 110–122.

Guigou, J.-D., Lovat, B., & Schiltz, J. (2012). Optimal mix of funded and unfunded pension systems: The case of Luxembourg. *Pensions, 17*(4), 208–222.

Haviland, A. M., & Nagin, D. S. (2005). Causual inferences with group based trajectory models. *Psychometrika, 70*(3), 557–578.

Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociological Methods & Research, 35*(4), 542–571.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29*(3), 374–393.

Jung, T., & Wickrama, K. A. S (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317.

Kass, R. E., & Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association, 90*, 773–795.

Keleva, A., & Brandt, H. (2014). A general non-linear multilevel structural equation mixture model. *Frontiers in Psychology, 5*, article 748.

Moffitt, T. E. (1993) Adolescence-limited and life-course persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674–701.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557–585.

Muthén, B. O. (2001). Latent variable mixture modeling. In G.A. Marcoulides & R.E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Lawrence Erlbaum.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biomectrics, 55*, 463–469.

Nagin, D. S. (1999). Analyzing developmental trajectories: Semi-parametric. groupe-based approach. *Psychological Method, 4*, 139–157.

Nagin, D. S. (2002). Analyse des trajectoires de développement: vue d'ensemble d'une méthode semiparamétrique fondée sur le groupement. In *Recueuil du Symposium de Statistique Canada Modélisation des données d'enquête pour la recherche sociale et économique*.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers and population heterogeneity: Specifiction and estimation of a nonparametric, mixed Poisson model. *Criminology, 31*, 327–362.

Nagin, D. S., & Odgers, C. L. (2010). Group-based trajectory modeling (nearly) two decades later. *Journal of Quantitative Criminology, 26*, 445–453.

Nagin, D. S., & Tremblay, R. E. (2005). Response to methodological sensitivities to latent class analysis of long-term criminal trajectories. *Journal of Quantitative Criminology, 20*, 27–35.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L.A. Collins & J.L. Horn (Eds.), *Best methods for the analysis of change* (pp. 92–106). Washington, DC: American Psychological Association.

Nielsen, J. D., Rosenthal, J. S., Sun, Y., Day, D. M., Bevc, I., Duchesne, T. (2014). Group-based criminal trajectory analysis using cross-validation criteria. *Communications in Statistics - Theory and Methods, 43*(20), 4337–4356.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535–569.

Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989) A developmental perspective on antisocial behavior. *American Psychologist, 44*, 329–335.

Raftery, A. E. (1995). Baysian model selection in social research. *Sociological Methodology, 25*, 111–164.

Reinecke, J., & Mariotti, L. (2009). Detection of unobserved heterogeinity with growth mixture models. *Revista de Matemática: Teoria y Applicaciones, 16*(1), 16–29.

Roeder, K., Lynch, K. G., & Nagin, D. S. (1999). Modeling uncertainty in latent class membership: A case study in criminology. *Journal of the American Statistical Association, 94*, 766–776.

Sampson, R. J., & Laub, J. H. (2005). Seductions of method: Rejoinder to Nagin and Tremblay's "developmental trajectory groups. fact or fiction?". *Criminology, 43*(4), 905–913.

Schwarz, G. (1978). Estimating dimensions of a model. *Annals of Statistics, 6*, 461–464.

Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Theil, H. (1971). *Principles of econometrics*. New York, NY: Wiley.

von Eye, A., & Bergman, L. R. (2003). Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology, 15*, 553–580.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*(2), 363–381.

Woolridge, J. (2002). *Econometric analysis of cross-section and panel data*. Cambridge, MA: MIT press.

Young, D. (2008). An overview of mixture models. *Statistics Surveys, 0*, 1–24.

Part II
Directional Dependence in Regression Models

Granger Causality: Linear Regression and Logit Models

Alexander von Eye, Wolfgang Wiedermann, and Ingrid Koller

Abstract Granger causality models are very popular when it comes to making decisions on which of a number of series of scores is on the dependent versus the independent side. With this chapter, we pursue two goals. First, we specify Granger causality models in terms of logit models and compare these with the routinely applied linear regression models. The comparison shows that, in order to make the models parallel, either model assumptions must be changed or model terms must be removed from (or inserted into) the model specification. The second goal involves extending Granger causality modeling. We propose conditioning terms on measures within the observed series. By implication, these models require higher-order interactions. In addition, model terms can be conditioned on covariates. Issues concerning parameter interpretation are discussed. Data examples are given from the fields of aggression development in adolescents and intimate partner violence.

Causality assumptions are central for all attempts to perform intervention or to change behavior. Without such assumptions, these efforts would be pointless. Causality is also a concept that has been discussed by scholars at least since Aristotelian metaphysics. Most prominent are Hume's tenets of *regularity* and *temporal priority*. Regularity implies that antecedents exist that are necessary, sufficient, or both for subsequent events. Temporal priority implies that antecedents occur temporally prior to subsequent events. The classical, essentialist perspective of causality posits that the antecedents be both necessary and sufficient to be considered causes of subsequent effects. In contrast, mechanistic concepts of causality also consider contemporary causes, even causes that are located in the future (see, e.g.,

A. von Eye (✉)
Psychology Department, Michigan State University, East Lansing, MI 48824-1116, USA
e-mail: voneye@msu.edu

W. Wiedermann
Department of Educational, School, and Counseling Psychology College of Education, University of Missouri, Columbia, MO 65211, USA
e-mail: wiedermannw@missouri.edu

I. Koller
University of Vienna, Wien, Austria

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_6

Williamson, 2011). Similarly, in the discussion of causality in history, Foucault (1966) considers concepts of contemporary effects. Cook and Campbell (1979) also cast doubt on the inevitability element involved in the definition of causality by temporal order, noting that this element may be inappropriate for the social sciences. The authors marshal a probabilistic concept of causality that links antecedents and consequences in a probabilistic fashion. In contrast, Sobel (1994, 1996) considers probabilistic concepts of causality, in particular Suppes' (1970) theory, not tenable. In this chapter, we focus on statistical methods for the analysis of hypotheses that are compatible with causality assumptions. For a discussion of causality from a philosophical perspective, see Stegmüller (1983; see also Beebee, Hitchcock, & Menzies, 2012; Lynd-Stevenson, 2007).

Statistical methods to estimate the probability of observed data under hypotheses that are compatible with causal assumptions or theories have been developed for experimental and nonexperimental research. These methods are graphical, consider counterfactuals, use manifest or latent variables, are frequentist or Bayesian, and require various sets of assumptions, some parametric, some concerning the nature of data (see, e.g., Foster, 2012; Matsuada, 2012; Pearl, 2000, 2012). Among the most frequently discussed and employed approaches to the empirical analysis of causation are path models (e.g., mediation models; Baron & Kenny, 1986) and *Granger causation* (Granger, 1969; cf. von Eye, Wiedermann, & Mun, 2013). Granger causation is interesting from a developmental perspective. It allows researchers to test hypotheses concerning the causal relations between two series of observations which can develop simultaneously. By the same token, the concept is also interesting because it uses, in its original form (Granger, 1969), only past observations to predict the later observations (note that attempts have been made to incorporate tests of hypotheses concerning observations that are located within a series; see Sims, 1980; and models have been discussed that include putative contemporary causes; see, e.g., von Eye et al., 2013).

Concepts of Granger causality were adopted first in econometric research (see, e.g., the textbooks by Bourbonnais, 2011; Lütkepohl & Krätzig, 2004; Mignon, 2008). Recently, however, there have been developments of the methodology and applications in behavioral sciences (see, e.g., Gates, Molenaar, Hillary, Ram, & Rovine, 2010; Kalimeri et al., 2012; von Eye et al., 2013). Granger causality has been discussed from statistical and philosophical perspectives. Points of critique relate to the fact that the statistical approach to Granger causality is based on the assumption that the process under study is stationary (for an overview, see Liu & Badahori, 2012). This, as is well known, is rather unlikely, in particular in developmental processes of growth and decline (see Molenaar & Campbell, 2009). Another point of discussion concerns the symmetry that is inherent in regression models (McArdle, 2012; von Eye & DeShon, 2012). Third, the temporal order of causal events continues to be an issue. Standard Granger causality does not enable researchers to test hypotheses on instant causes, because it is based on Humean concepts of causality, according to which the cause precedes the effect in time. Human anticipation of events, however, makes is very likely that contemporaneous or future events can be causes of current action. Examples include the saving of

resources for retirement, or the building up of defense against (possible) moves of the other player in chess. In addition, the effects of unobserved confounders need to be discussed in more detail, in the context of Granger causality.

Another topic that needs discussion concerns methods of analysis of categorical variables. This is the topic we begin to address in the current chapter. Specifically, we ask in this chapter whether logit models that can be used to test causal hypotheses correspond with the linear regression models that are typically used.

The remainder of this chapter is structured as follows. We first describe the regression models used in the original approach to Granger causality. We then discuss logit models for the analysis of Granger causality-compatible hypotheses. We then introduce the notion of higher-order interactions into the context of methods for the analysis of Granger causality and propose new hypotheses in the context of Granger causality. Empirical data examples are presented from research on development of aggression in adolescence and on effects of intimate partner violence.

Elements of Granger Causality

Granger causality methodology is used to test hypotheses about the causal relations between two series of scores. The question asked is whether one of the two series causes the other. If this is the case, this series is said to *Granger-cause* the other. The methods used to test such hypotheses are mostly linear regression methods. The regression models used are *v*ector *a*uto*r*egressive models (VAR models; for an overview, see Lütkepohl & Krätzig, 2004). In the present context, consider a variable, Y, observed T times, with lag p. A VARp process for this series of observations can be defined by

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \cdots + \Phi_p Y_{t-p} + \epsilon_t,$$

where t indicates the last observation, p indexes the sequence of observations, and Φ contains the model parameters. Under standard GLM conditions, the $T+$ first observation can be estimated as

$$E\left[Y_{T+1} \mid \underline{Y}_T\right] = \sum_{i=0}^{T-p+1} \Phi_i,$$

where $\underline{Y}_T$ contains all observations of Y, including the one at Time T. Now, consider two series of scores, Y_1 and Y_2. Heuristically, let Y_2 be the putatively dependent series, and Y_1 the putatively independent series. It can be said that the series Y_1 *Granger-causes* the series Y_2 when Y_1 makes a contribution to the prediction of the last score in the series Y_2 above and beyond the contribution that is made by past observations of Y_2 alone. Let Y_{1t} and Y_{2t} be the observations of Y_1 and Y_2, at Time t. Then, the VARp process for the two variables Y_1 and Y_2 can be described by the well-known equation

$$\begin{bmatrix} Y_{1,t} \\ Y_{2,t} \end{bmatrix} = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} + \begin{bmatrix} a_1^1 & b_1^1 \\ a_1^2 & b_1^2 \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a_p^1 & b_p^1 \\ a_p^2 & b_p^2 \end{bmatrix} \begin{bmatrix} Y_{1,t-p} \\ Y_{2,t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix},$$

where the a and the b are regression parameter estimates (the superscripts denote the corresponding time series), and the ε are random residuals, uncorrelated with the predicted scores and the predictors. In the null hypothesis, it is posited that there exists no causal relation between Y_1 and Y_2. There are four possible outcomes for tests of this null hypothesis:

1. Y_{1t} can be considered Granger-causing Y_{2t} if the following null hypothesis is rejected: H_0: $b_1^1 = \cdots = b_p^1 = 0$.
2. $Y_{2\mathrm{t}}$ can be considered Granger-causing $Y_{1\mathrm{t}}$ if the following null hypothesis is rejected: H_0: $a_1^2 = \cdots = a_p^2 = 0$.
3. When both null hypotheses are rejected, one can consider processes of reciprocal causation.
4. When none of these null hypotheses is rejected, one can consider independence of the two series.

All this applies accordingly when more than two series are included in the causal hypotheses. There are many options for testing these hypotheses. For example, when standard regression models are estimated, F-tests are suitable, and when manifest variable structural models are estimated, chi-square difference tests can be used. In either case, two comparison models are estimated. In the first, the Y_{2t} observation is regressed onto its p past values. In the second, the past values of Y_{1t} are also included in the regression equation. These two models are nested and can be statistically compared. In a manifest variable regression context, the first model is

$$Y_{2\mathrm{t}} = b_{01} + \sum_{i=1}^{T-p+1} b_{2\mathrm{i}} Y_{2\mathrm{i}} + \epsilon_1,$$

and the second model is

$$Y_{2\mathrm{t}} = b_{02} + \sum_{i=1}^{T-p+1} b_{2\mathrm{i}} Y_{2\mathrm{i}} + \sum_{i=1}^{T-p+1} b_{1\mathrm{i}} Y_{1\mathrm{i}} + \epsilon_2.$$

If the additional portion of variance of Y_{2t}, which is accounted for by including Y_{1t} in the second model, amounts to a significant increase over the amount explained by the first model, one calls the series Y_2 *Granger-caused* by the series Y_1.

Linear Regression and Logit Models for Granger Causality

In this section, we discuss logit models for Granger causality. Logit models have been used before to test Granger causality-compatible hypotheses (see, e.g.,

Christopoulos & Leon-Ledesma, 2008; Wang, 2011). Therefore, there is no need to ask whether logit models can be used for this purpose. Instead, we pursue two goals. First, we compare linear regression models with logit models in their application to testing Granger causality-compatible hypotheses. The question we ask is whether linear regression models and logit models can be specified such that they can be used to test comparable hypotheses. Second, we present new models.

To simplify interpretation of the comparison models, we use, without loss of generality, a particular selection of models; these are models in which the last measure in a series is predicted instead of the entire series. The main reason for doing this is that the number of parameters to be interpreted in this selection of models is much smaller than in standard Granger causality models. To illustrate, consider two series of p measures each. To predict one series from itself in a standard Granger causality model, one needs $2(p-1)$ parameters for the regression of each measure on its predecessor (one intercept + one slope parameter each), $2(p-2)$ parameters for the regression of each measure on the measures two occasions before, etc., or, in sum $2\sum_{i=1}^{p}(p-i)$ parameters. The same number is needed to predict the measures of the second series from themselves. The same numbers of parameters again are needed to predict one series from the other and vice versa. In sum, for two series of p measures each, one estimates, in a standard Granger causality model, $4\left(2\sum_{i=1}^{p}(p-i)\right)$ parameters.

In the present chapter, we only estimate the following parameters:

- $2\,(2\,(p-1))$ parameters for regressing the measures of each series on their immediate predecessors, and
- $2(p-1)$ parameters for predicting the last measure in each series from all of their predecessors in the respective other series.

In sum, we only estimate $6(p-1)$ parameters instead of $4\left(2\sum_{i=1}^{p}(p-i)\right)$, a savings of $8\sum_{i=2}^{p}(p-i)+2\,(p-1)$ parameters. This savings has the potential of obtaining a more parsimonious model to explain observed variability and simplifying interpretation considerably.

It is important to note that the models that we consider are still models of Granger causality. von Eye and Wiedermann (2014) proposed a taxonomy of Granger causality models. This taxonomy results from completely crossing the four binary variables: (1) order of lag considered (coded as 1 = lag 1, 2 = lag >1) (O), (2) type of contemporaneous effect considered (1 = correlation, 2 = regression) (T), (3) direction of effect hypothesized (1 = yes, 2 = no) (D), and (4) segment of dependent variable targeted (1 = entire series, 2 = segment of series) (S). The models that we use for illustration are indexed as 1 2 1 2, where the indexes are in the order of variables listed. These models include regressions of both series of measures onto the respective other, contemporaneous relations are either fixed or estimated as correlations, lags cover time-adjacent relations, and, most important for the present examples, predict only a segment of the series of measures. In the present examples,

only the last measure is predicted. We also consider models in which the lag is greater than 1, that is, models of the type 2 1 2 2.

To address the above question concerning the Granger-caused relation between two series of scores, we consider two series of scores of the variables X and Y. Let both variables be dichotomous so that logistic regression models can meaningfully be applied, and let both series consist of three observations, $X1$, $X2$, and $X3$, and $Y1$, $Y2$, and $Y3$. In the first model that is estimated in the analysis of hypotheses that are compatible with Granger causation, one regresses $Y3$ onto the past observations of Y, $Y1$ and $Y2$. The corresponding linear regression model is

$$Y3 = \beta_0 + \beta_1 Y1 + \beta_2 Y2.$$

To evaluate the contribution made by the past observations of X, one also includes $X2$ and $X3$ in the second regression model,

$$Y3 = \beta_0 + \beta_1 Y1 + \beta_2 Y2 + \beta_3 X1 + \beta_4 X2.$$

Now, one could be tempted to consider the following two logit models as parallel to the linear regression models. Regressing $Y3$ onto its past observations yields

$$\log\left(\frac{p_{Y3=1}}{1-p_{Y3=1}}\right) = \beta_0 + \beta_1 Y1 + \beta_2 Y2,$$

and regressing $Y3$ also onto $X1$ and $X2$ yields

$$\log\left(\frac{p_{Y3=1}}{1-p_{Y3=1}}\right) = \beta_0 + \beta_1 Y1 + \beta_2 Y2 + \beta_3 X1 + \beta_4 X2.$$

Based on this formulation, the linear regression models and the logit models do look parallel. In the following paragraphs, we examine these equations in more detail and discuss whether the impression of parallel equations, that is, the impression of equations that allow one to answer equivalent questions, can be defended.

To answer the question whether the equations of linear regression and logit models given in the last paragraphs are equivalent, we look at the regression and the logit models in more detail. These were the models for the six measures $X1$, $X2$, and $X3$, and $Y1$, $Y2$, and $Y3$. The first model discussed for the analysis of hypotheses that are compatible with Granger causation regresses $Y3$ onto the past observations of Y, $Y1$ and $Y2$. The second model includes $X1$, $X2$, and $X3$, in addition. Using the log-linear notation (see Agresti, 2013; von Eye & Bogat, 2005; von Eye, Mair, & Bogat, 2005; von Eye & Mun, 2013), the first logit model can equivalently be recast as follows. We use, for this model specification, the cross-classification of all six variables. The model is

$$\begin{aligned}\log\,\widehat{m} = \lambda + \lambda^{X1} + \lambda^{X2} + \lambda^{X3} + \lambda^{Y1} + \lambda^{Y2} + \lambda^{Y3} \\ +\lambda^{Y3,Y1} + \lambda^{Y3,Y2} + \lambda^{Y1,Y2} \\ +\lambda^{X1,X2} + \lambda^{X1,X3} + \lambda^{X2,X3} + \lambda^{X1,X2,X3}.\end{aligned}$$

In its first row, this equation contains the main effects of all variables in the model. In its second row, the equation contains the effects needed for the regression of $Y3$ onto $Y1$ and $Y2$. The interaction between the two predictors, $Y1$ and $Y2$, is needed because the model makes no assumption concerning the relations among predictors. Therefore, these relations cannot simply be set to zero; they are estimated. In its third row, this equation contains all possible interactions among the X observations. Relations among X and Y observations are not part of this model, because we first regress $Y3$ solely onto $Y1$ and $Y2$. These relations are needed in the second model, in which $Y3$ is not regressed solely onto $Y1$ and $Y2$ but also onto $X1$ and $X2$. The second model, therefore, is

$$\begin{aligned}log\,\widehat{m} = \lambda + \lambda^{X1} + \lambda^{X2} + \lambda^{X3} + \lambda^{Y1} + \lambda^{Y2} + \lambda^{Y3} \\ +\lambda^{Y3,Y1} + \lambda^{Y3,Y2} + \lambda^{Y1,Y2} \\ +\lambda^{X1,X2} + \lambda^{X1,X3} + \lambda^{X2,X3} + \lambda^{X1,X2,X3} \\ \lambda^{Y3,X1} + \lambda^{Y3,X2}.\end{aligned}$$

The first and the second log-linear models are nested. Therefore, they can be compared by using, for example, the chi-square difference test.

The comparison of the present log-linear models with the present linear regression models reveals an important difference. In the linear regression models, the interactions among the predictors are not included. In log-linear models, they are. To illustrate, compare the first of the log-linear models in the present section with the linear regression model for $Y3$ in the last section. The log-linear model does contain the interaction between $Y1$ and $Y2$, but the linear regression model does not. The logit model discussed in the last section is equivalent to the log-linear model discussed here. Therefore, the difference to the linear regression model applies to the logit model as well.

When models are specified to test hypotheses that are compatible with Granger causation, researchers, therefore, make different assumptions for linear regression and log-linear models. For the former, the assumption is made that terms not included in the model do not exist, for example interactions among past observations of the predicted variable. If these interactions exist—which is very likely for repeated observations—and there are no corrections for these effects, parameter estimates can be imprecise. Indeed, methods have been proposed to correct the residual terms in linear regression Granger causality models (e.g., Engle & Granger, 1987; Granger, 1981; Granger & Newbold, 1974). For the log-linear (or logit) model, this assumption is not required. The offending terms are part of the model, and one reason for biased parameter estimates is thus eliminated.

However, there may be additional reasons for bias in parameter estimation for Granger causation. These reasons are discussed in the next section, in which we propose extensions of the log-linear models for Granger causality.

Higher-Order Interactions in Models for Granger Causality

In this section, we propose extending the methodology for testing hypotheses that are compatible with Granger causality for categorical variables by also considering interactions higher than first order. Higher-order interactions can be considered both for the Y observations and the X observations. Consider the series of four observations $Y1$, $Y2$, $Y3$, and $Y4$, with $Y4$ being the predicted observation. In standard logit modeling of Granger causality hypotheses, the first of the two models to be estimated contains the interactions [$Y1$, $Y4$], [$Y2$, $Y4$], [$Y3$, $Y4$], or, in words, the last Y observation is predicted from the past Y observations. In logit models, all interactions among the past observations are part of the model as well. These are the interactions [$Y1$, $Y2$], [$Y1$, $Y3$], [$Y2$, $Y3$], and [$Y1$, $Y2$, $Y3$]. Now, in the extended approach, the prediction of $Y4$ from its past observations can be conditioned on Y observations that are not part of a particular interaction. For example, the association [$Y1$, $Y4$] can be conditioned on $Y2$, $Y3$, or both. This can result in the interactions [$Y1$, $Y2$, $Y4$], and [$Y1$, $Y3$, $Y4$]. Similarly, the three-way interaction [$Y2$, $Y3$, $Y4$] and the four-way interaction [$Y1$, $Y2$, $Y3$, $Y4$] can be considered for the first model. Including the interaction of all four Y observations in the first model will not result in a saturated model if the table under study is spanned by all X and Y observations. However, the first model will be saturated in the X observations.

In a second example, let there be two X observations, $X1$ and $X2$, observed at the same points in time as $Y1$ and $Y2$. Then, the first model will include the main effects [$X1$] and [$X2$] as well as the interaction [$X1$, $X2$]. The second model will routinely include the effect [$Y2$, $X1$]. That is, the last Y observation is predicted from the past X observations. Here, we propose considering additional terms, specifically terms that condition the prediction of the last Y observation from one of the X observations on one or more of the other X observations. In the present example, there is only one possible additional term, [$Y2$, $X1$, $X2$]. For a series of three or more X observations, additional terms are conceivable, and terms with higher than three-way interactions. Appendix 2 explains parameter interpretation in an example in which interactions higher than first order are part of the models.

The interpretation of the two- and higher-order interactions can be based on the magnitude of the estimated parameter, significance tests of the null hypothesis that a parameter is zero, and the equation $\lambda = (X'X)^{-1}X' \log \widehat{m}$ (von Eye & Mun, 2013; for an example and parameter interpretation, see Appendix 1, below). An additional and often used option involves transforming the log-linear parameters into odds ratios and then interpreting these (cf. Rudas, 1997; von Eye & Schuster, 2000).

Additional extensions can be considered. For example, the prediction of the last Y observation from past Y observations can be conditioned on past X observations.

Similarly, the prediction of the last Y observation from past X observations can be conditioned on past Y observations. Covariates can be included in the models, predictions can be conditioned on covariates, and additional series of scores can be included in the set of causal hypotheses.

Data Examples

In this section, we present two data examples. In the first example, we illustrate how a three-way interaction between observations from the putative causal series of scores can be included in the model. In the second example, we again illustrate the use of higher-order interaction terms that allow researchers to test specific hypotheses that are compatible with Granger causality, and we discuss related log-linear models.

Data example 1. For the first example, we use data collected by Finkelstein, von Eye, and Preece (1994), on the development of aggressive behavior. Sixty-seven adolescent girls and 47 boys responded to a questionnaire concerning aggressive behavior, at three points in time, spaced in 2-year intervals. The questionnaire addressed the four dimensions: Aggressive Impulse, Aggression-Inhibitory Response, Verbal Aggression against Adults, and Physical Aggression against Peers. In addition, physical pubertal development was assessed using Tanner scores. In the following analyses, we use the variables physical aggression against peers (P) and aggressive impulses (A). We use the repeated observations P83, and P87, and A83, and A87, where the numbers indicate the years in which the data were collected. When the data were collected, the respondents were, on average, 11, and 15 years of age.

Substantively, we ask whether the developmental change in self-rated physical aggression against peers is Granger-caused by developmental change in self-rated aggressive impulses (cf. von Eye et al., 2013). To answer this question, we estimate three logistic regression models. The first regresses P87 onto P83. The model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 P83 + \epsilon,$$

where π is the probability that $P87 = 1$, that is, that the P87 rating is below the median. The second model is estimated to determine whether including the series of two scores of aggressive impulses Granger-causes the series of scores of physical aggression against peers. This model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 P83 + \beta_2 A83 + \beta_3 A87 + \epsilon,$$

where π is defined as in the first model. Note that, in this model, the contemporary measure A87 is part of the model. This option has been considered in recent discussion of Granger causality (von Eye, & Wiedermann, 2014; von Eye et al., 2013,

and is compatible with mechanistic concepts of causality; see, e.g., Williamson, 2011). If this model is significantly better than the first, development of aggressive impulses can be considered Granger-causing development of physical aggression against peers. In the third model, we include the [A83, A87] interaction. If this interaction is significant, the effect of contemporary aggressive impulses depends on the strength of aggressive impulses 4 years prior. The model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 P83 + \beta_2 A83 + \beta_3 A87 + \beta\,(A83, A87) + \epsilon.$$

Table 1 displays the P83 × P87 × A83 × A87 cross-classification.

The overall model fit X^2 for the first model is 11.76. For $df = 1$, this value suggests that significant effects exist ($p < 0.01$). Specifically, we estimate $\lambda^{P83} = -1.55$, se = 0.48, $z = -2.50$, and $p < 0.01$. We conclude that P83 is predictive of P87. To Granger-cause this development, the model that includes the putative causal series of scores of aggressive impulses must significantly improve this first model. Table 2 displays the parameter estimates for the second logit model.

The overall model fit X^2 for the second model is 24.63. For $df = 3$, this value suggests that at least one significant effect exist ($p < 0.01$). P83 is still a significant predictor of P87, and so is the contemporaneous measure of A87. Aggressive impulses that were self-rated 4 years prior are not significant predictors. The Nagelkerke R^2 for the first model is 0.142; the Nagelkerke R^2 for the second model is 0.282, an increase by almost 100 %. The ΔLR-X^2 for the comparison of the two corresponding log-linear models is 16.29. For $\Delta df = 2$, this difference is significant ($p < 0.01$), and we conclude that the development of aggressive

Table 1 Cross-classification of P83, P87, A83, and A87

P83	P87	A83	A87	Frequency	Cumulative frequency	Percent	Cumulative percent
1	1	1	1	24	24	21.053	21.053
1	1	1	2	5	29	4.386	25.439
1	1	2	1	10	39	8.772	34.211
1	1	2	2	9	48	7.895	42.105
1	2	1	1	1	49	0.877	42.982
1	2	1	2	4	53	3.509	46.491
1	2	2	1	1	54	0.877	47.368
1	2	2	2	1	55	0.877	48.246
2	1	1	1	9	64	7.895	56.140
2	1	1	2	3	67	2.632	58.772
2	1	2	1	12	79	10.526	69.298
2	1	2	2	11	90	9.649	78.947
2	2	1	1	4	94	3.509	82.456
2	2	1	2	3	97	2.632	85.088
2	2	2	1	2	99	1.754	86.842
2	2	2	2	15	114	13.158	100.000

Table 2 Parameter estimates for logistic regression of P87 onto P83, A83, and A87

Parameter	Estimate	Standard error	z	p	95 % Confidence interval	
					Lower	Upper
Constant	5.342					
A83	0.560	0.540	1.038	0.299	−0.497	1.618
A87	−1.769	0.525	−3.368	0.001	−2.799	−0.739
P83	−1.551	0.539	−2.880	0.004	−2.606	−0.495

Table 3 Parameter estimates for logistic regression of P87 onto P83, A83, and A87 and the A83 × A87 interaction

Parameter	Estimate	Standard error	z	p	95 % Confidence interval	
					Lower	Upper
Constant	5.803	2.809	2.066	0.039	0.298	11.308
P83	−1.561	0.544	−2.871	0.004	−2.626	−0.495
A83	0.257	1.734	0.148	0.882	−3.142	3.655
A87	−2.060	1.673	−1.231	0.218	−5.338	1.219
A83 × A87	0.192	1.043	0.184	0.854	−1.853	2.237

impulses Granger-causes the development of physical aggression against peers in adolescence. It should be noted, however, that this interpretation does not conform with the classic notion of Granger causality. In this notion—it is based on the Humean tradition of temporal order in causality—contemporaneous effects cannot be considered causes. Therefore, this interpretation requires a different concept of causality than the one propagated in the Humean tradition.

With the third model, we ask whether the contemporary element in the causal relation between the two series of scores depends on the strength of aggressive impulses 4 years prior. The third logit model, therefore, contains the A83 × A87 interaction. Table 3 displays the parameter estimates for this model.

The overall model fit LR-X^2 for this model is 24.67. For $df = 4$, this value suggests that significant effects exist ($p < 0.01$). P83 is still a significant predictor of P87, but none of the parameters for aggressive impulses is significant. In addition, the improvement of this model over the second is zero. The Nagelkerke R^2 for the third model is 0.282 as well. The improvement in model fit is nonsignificant. We, therefore, retain the more parsimonious second model.

Data example 2. For the second example, we use data from a longitudinal project on intimate partner violence (Bogat, Levendosky, DeJonghe, Davidson, & von Eye, 2004). Two hundred and four women responded, in yearly intervals, to questions concerning the frequency with which they suffered violence perpetrated by intimate partners. In addition, they filled a questionnaire that was administered to assess the degree to which they showed symptoms of post-traumatic stress disorder (PTSD scale for battered women; Saunders, 1994). 62 % of the respondents were Caucasian, 25 % African American, and 13 % other or mixed racial backgrounds. At the beginning of the study, the women were, on average, 27 years old.

For the following analyses, we use the information regarding violence and PTSD assessed at the first and the second observation points. We ask whether initial violence allows one to predict PTSD 1 and 2 years later. Violence was scored as 1 = did not experience violence and 2 = did experience violence. PTSD was scored as 1 = symptoms below the clinical cut off and 2 = symptoms above the clinical cut off. The questions concerning both violence and PTSD were formulated such that they covered the period since the last interview, that is, the year before the interview. In the following sections, we abbreviate violence at Time 1 with V1, violence at Time 2 with V2, PTSD at Time 1 with P1, and PTSD at Time 2 with P2. Table 4 displays the V1 × V2 × P1 × P2 cross-classification.

For the analysis of the V1 × V2 × P1 × P2 cross-classification, we specify three models. The first model regresses P2 onto P1. The model is, in log-linear notation,

$$\log \widehat{m} = \lambda + \lambda^{V1} + \lambda^{V2} + \lambda^{P1} + \lambda^{P2} + \lambda^{V2,V1} + \lambda^{P1,P2}.$$

The interaction between P1 and P2 is needed in this model because it represents the auto-regression of P2 onto P1. The interaction between V1 and V2 is needed because the model must be saturated in the variables that are not part of the regression of P2 onto P1. The second model also regresses P2 onto V1, the past observation of the second series of scores. The model is

$$\log \widehat{m} = \lambda + \lambda^{V1} + \lambda^{V2} + \lambda^{P1} + \lambda^{P2} + \lambda^{V2,V1} + \lambda^{P1,P2} + \lambda^{V1,P2}.$$

Table 4 Cross-classification of V1, V2, P1, and P2

V1	V2	P1	P2	Frequency	Cumulative frequency	Percent	Cumulative percent
1	1	1	1	82	82	40.196	40.196
1	1	1	2	7	89	3.431	43.627
1	1	2	1	47	136	23.039	66.667
1	1	2	2	13	149	6.373	73.039
1	2	1	1	3	152	1.471	74.510
1	2	1	2	3	155	1.471	75.980
1	2	2	1	0	155	0.000	75.980
1	2	2	2	11	166	5.392	81.373
2	1	1	1	0	166	0.000	81.373
2	1	1	2	0	166	0.000	81.373
2	1	2	1	12	178	5.882	87.255
2	1	2	2	7	185	3.431	90.686
2	2	1	1	0	185	0.000	90.686
2	2	1	2	0	185	0.000	90.686
2	2	2	1	2	187	0.980	91.667
2	2	2	2	17	204	8.333	100.000

If the second model is significantly better than the first, one can say that the first series of scores, P1 and P2, is Granger-caused by the second series, V1, and V2. In the third model, we ask, in addition, whether changes in intimate partner violence affect PTSD at Time 2. This question can be answered by including the three-way interaction V1 × V2 × P2 in the model, or

$$\log \widehat{m} = \lambda + \lambda^{V1} + \lambda^{V2} + \lambda^{P1} + \lambda^{P2} + \lambda^{V2,V1} + \lambda^{P1,P2} + \lambda^{V1,P2} + \lambda^{V1,V2,P2}.$$

If this model is significantly better than the second, we will talk about *higher-order Granger causality*. The second and the third model are hierarchically related to the first model and to each other. Table 5 displays goodness-of-fit information and the results of the chi-square difference tests.[1]

The first model serves as reference model. Hypotheses that are compatible with the concept of Granger causality can be retained only if at least the second model represents a significant improvement over the first model. The parameter of interest in the first model is the association between P1 and P2. This parameter is significant ($z = 5.257$; $p < 0.01$), thus supporting the hypothesis that P2 can be predicted from P1. The second model is used to answer the question whether, over time, intimate partner violence causes PTSD. This model is significantly better than the first model (see Table 5). The parameter of interest in this model is the interaction between V1 and P2. It is significant ($z = 4.890$; $p < 0.01$). We conclude that the series of PTSD observations is Granger-caused by the series of intimate partner violence observations.

With the third model, we ask whether the effect of V1 on P2 is conditional on V2. In other words, we ask whether the effect of V1 on P2 depends on whether, at the second observation point, the respondent was a victim of intimate partner violence or not. The significant V1 × V2 × P2 interaction ($z = 2.505$; $p = 0.017$) allows us to answer this question in the affirmative. We conclude that the effect of V1 on P2 is moderated by V2 (von Eye & Schuster, 2000).

To interpret this three-way interaction, we first note that the odds of exhibiting above clinical-level PTSD symptoms at the second observation point are higher for respondents who were victims of intimate partner violence in the last trimester of pregnancy (the first observation point). This relation remains unchanged for respondents who were victims in the year before the second observation point as well (1 year later). If, however, a respondent was not victim at the first observation point but she was victim at the second observation point, she will exhibit such symptoms as well. The odds ratio for this moderator effect is $\theta = 0.110$ (se = 0.050; $z = 2.2$; $p = 0.035$), and is, thus, significant.

[1] All of the models reported in this chapter were estimated with the log-linear and the logit modules of SYSTAT, lem, and R. SYSTAT reported estimation issues with some of the models. Therefore, we invoked the Delta option with $\Delta = 0.1$. The results without Delta differ only minimally from those presented here. For example, the overall goodness-of-fit LR-X^2 for the first model in Table 5 is 109.39 without Delta and 110.51 with Delta.

Table 5 Goodness-of-fit of three models for the estimation of parameters for Granger causation of PTSD by intimate partner violence

Model	χ^2	*df*	*p*	$\Delta\chi^2$—Model 1	Δdf—Model 1	*p*	$\Delta\chi^2$—Model 2	Δdf—Model 2	*p*
1	110.51	9	<0.01						
2	85.36	8	<0.01	25.15	1	<0.01			
3	79.26	7	<0.01	31.25	2	<0.01	6.10	1	<0.01

Interesting tidbit. In most software packages for logit models, the overall goodness-of-fit test indicates whether there are significant predictor effects at all. These tests do not indicate whether the model can be used to describe the data in a cross-classification. Therefore, it can be the case that ill-fitting models are retained. In the present example, none of the models fits (Table 5). With increasing complexity, the models explain increasing portions of the variability. They are, however, far from explaining the data well. To answer the question how the data can be explained, we need a more complex model. Including the two contemporary effects [V1, P1] and [V2, P2] in addition to the effects already in the third model results in a well-fitting model. Its overall goodness-of-fit LR-X^2 is 3.062 ($df = 5$; $p = 0.690$). This model contains the elements needed for the interpretation that the series of intimate partner violence Granger-causes the series of PTSD symptoms. It is, however, enriched by effects not considered before. For an alternative parameterization, see Appendix 3.

Discussion

In this chapter, we discuss logit modeling for the analysis of hypotheses that are compatible with the notion of Granger causality. As is well known, logit models for categorical variables and, in particular, models of logistic regression can be specified that are parallel to the linear regression models used for metric variables. Interestingly, when the models are recast in terms of equivalent log-linear models, three issues become apparent. The first issue is that, although the parameters of interest may be strong and significant, a model may not fit. In the application of logistic regression, this issue is often ignored. In log-linear modeling, however, parameter interpretation requires that a model describe the data well (Agresti, 2013; von Eye & Mun, 2013). We, therefore, recommend using the more general log-linear approach to logistic regression and interpreting causality parameters only when a model can be retained.

The second issue is that the log-linear model approach allows one to enrich the models that are specified to analyze hypotheses that are compatible with Granger causality with additional terms. We propose considering higher-order interactions even if the resulting models become nonhierarchical (see Mair & von Eye, 2007). With these terms, interactions can be conditioned on observations from within the series of comparison. In other words, differential paths of development can be taken into account.

The third issue concerns the incorporation of once-observed variables as covariates. Using covariates, causality hypotheses can be conditioned on strata of the population. Most interestingly, causality hypotheses can be conditioned on treatments. This way, hypotheses can be tested that posit that specific treatments are effective but others are not. Hypotheses can also be investigated, according to which treatments are effective in particular when they are administered before certain junctures or to clients with particular profiles.

A central point of the discussion in this chapter is the comparison of logit models with linear regression models for Granger causality. This discussion has shown that standard application is based on different sets of assumptions. In linear regression models, terms not explicitly included in a model are considered non-existing. In other words, the parameters that correspond to these terms are set to zero. In contrast, logit models do contain terms that are not explicitly included in the model specification. For example, interactions of every order among the measures of the dependent series and among the measures of the independent series are automatically part of the model. This can be made explicit by translating logit models into corresponding log-linear models.

From this observation, we deduce two implications. First, to make linear regression models and logit models equivalent in specification, the same terms must be part of a model. This can be achieved either by including the interactions in linear regression models that are automatically part of logit models or by removing these terms from logit models. This can be achieved by estimating log-linear models without these terms instead of routine logit models.

In this chapter, we do not intend to make a contribution to the general discussion causality theory. However, we propose new models that can be interesting when Granger causality hypotheses are considered, that is, hypotheses that relate series of scores to one another. One important deviation from the original concepts of Granger causality should be made explicit. In addition to past observations, we also considered contemporary observations as possible causal agents (this was discussed before by von Eye et al., 2013). This approach is compatible with mechanistic concepts of causality, which also discuss contemporary causes, even causes that are located in the future (Williamson, 2011).

Appendix 1: Design Matrix and Parameter Interpretation for Granger Causality Models for the $Y1 \times Y2 \times Y3 \times X1 \times X2$ Cross-classification with Y3 as the Putative Dependent Variable

In this appendix, we illustrate models that can be used to test hypotheses that are compatible with the notion of Granger causality. We use the two series of observations $Y1$, $Y2$, and $Y3$, and $X1$ and $X2$. $Y3$ is the last observation, and the numbers indicate the temporal order of observations. For the sake of simplicity, let all five variables be binary. The cross-classification of these five variables contains 32 cells and can be analyzed under the following two models. The first model represents the hypothesis that $Y3$ can be predicted from $Y1$ and $Y2$. This is the model

$$\begin{aligned} log\,\widehat{m} = \lambda + \lambda^{X1} + \lambda^{X2} + \lambda^{Y1} + \lambda^{Y2} + \lambda^{Y3} \\ +\lambda^{Y3,Y1} + \lambda^{Y3,Y2} + \lambda^{Y1,Y2} \\ +\lambda^{X1,X2}. \end{aligned}$$

In its first row, this model equation displays the intercept and the main effects of all five variables. In the design matrix given in Table 6, the intercept is implied, and the effect coding vectors for the main effects are given in the first five columns. In its second row, the equation displays all pairwise interactions among the Y observations. This row represents the prediction of $Y3$ from $Y1$ and $Y2$ that is the key element of the first model that is estimated for the analysis of Granger causation. A logistic regression model would include these interactions as well. In the design matrix given in Table 6, the coding vectors for these effects appear in columns 6, 7, and 8. In its last row, the equation displays the $X1 \times X2$ interaction. This interaction and the main effects of $X1$ and $X2$ are needed because the first model must not fail just because these effects might exist. In the design matrix given in Table 6, the vector for this effect appears in column 9.

These nine vectors are orthogonal. Therefore, the interpretation of each of them via

$$\lambda = (X'X)^{-1}X' \log \widehat{m}$$

is straightforward (cf. von Eye & Mun, 2013). To give an example, the interpretation of the parameter for the $Y1 \times Y3$ interaction is

$$\begin{aligned} \lambda^{Y1,Y3} = \frac{1}{32} (\log m_{11111} + \log m_{11112} + \ldots \\ - \log m_{11211} - \cdots + \log m_{22221} + \log m_{22222}) . \end{aligned}$$

Alternative models are illustrated in Appendix 2.

The second model to be estimated includes $X1$ and $X2$ as additional predictors of $Y3$. The model is

$$\begin{aligned} log\,\widehat{m} = \lambda + \lambda^{X1} + \lambda^{X2} + \lambda^{Y1} + \lambda^{Y2} + \lambda^{Y3} \\ +\lambda^{Y3,Y1} + \lambda^{Y3,Y2} + \lambda^{Y1,Y2} \\ +\lambda^{X1,X2} + \lambda^{X1,Y3} + \lambda^{X2,Y3}. \end{aligned}$$

The new terms can be seen in the third row of this equation. The corresponding coding vectors appear in columns 10 and 11 of the design matrix in Table 6. Evidently, the two models are nested. The second model contains all terms of the first, and two additional terms (the last two in the equation). Therefore, these models can be compared using, for example, the chi-square difference test.

Table 6 Design matrix for Granger causality modeling of the $Y1 \times Y2 \times Y3 \times X1 \times X2$ cross-classification with $Y3$ as the putative dependent variable

Y1	Y2	Y3	X1	X2	Y1Y2	Y1Y3	Y2Y3	X1X2	X1Y3	X2Y3	Y1Y2Y3	X1X2Y3
1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	−1	1	1	1	−1	1	−1	1	−1
1	1	1	−1	1	1	1	1	−1	−1	1	1	−1
1	1	1	−1	−1	1	1	1	1	−1	−1	1	1
1	1	−1	1	1	1	−1	−1	1	−1	−1	−1	−1
1	1	−1	1	−1	1	−1	−1	−1	−1	1	−1	1
1	1	−1	−1	1	1	−1	−1	−1	1	−1	−1	1
1	1	−1	−1	−1	1	−1	−1	1	1	1	−1	−1
1	−1	1	1	1	−1	1	−1	1	1	1	−1	1
1	−1	1	1	−1	−1	1	−1	−1	1	−1	−1	−1
1	−1	1	−1	1	−1	1	−1	−1	−1	1	−1	−1
1	−1	1	−1	−1	−1	1	−1	1	−1	−1	−1	1
1	−1	−1	1	1	−1	−1	1	1	−1	−1	1	−1
1	−1	−1	1	−1	−1	−1	1	−1	−1	1	1	1
1	−1	−1	−1	1	−1	−1	1	−1	1	−1	1	1
1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	−1
−1	1	1	1	1	−1	−1	1	1	1	1	−1	1
−1	1	1	1	−1	−1	−1	1	−1	1	−1	−1	−1
−1	1	1	−1	1	−1	−1	1	−1	−1	1	−1	−1
−1	1	1	−1	−1	−1	−1	1	1	−1	−1	−1	1
−1	1	−1	1	1	−1	1	−1	1	−1	−1	1	−1
−1	1	−1	1	−1	−1	1	−1	−1	−1	1	1	1
−1	1	−1	−1	1	−1	1	−1	−1	1	−1	1	1
−1	1	−1	−1	−1	−1	1	−1	1	1	1	1	−1
−1	−1	1	1	1	1	−1	−1	1	1	1	1	1
−1	−1	1	1	−1	1	−1	−1	−1	1	−1	1	−1
−1	−1	1	−1	1	1	−1	−1	−1	−1	1	1	−1
−1	−1	1	−1	−1	1	−1	−1	1	−1	−1	1	1
−1	−1	−1	1	1	1	1	1	1	−1	−1	−1	−1
−1	−1	−1	1	−1	1	1	1	−1	−1	1	−1	1
−1	−1	−1	−1	1	1	1	1	−1	1	−1	−1	1
−1	−1	−1	−1	−1	1	1	1	1	1	1	−1	−1

Appendix 2: Granger Causality Models and Parameter Interpretation for the $Y1 \times Y2 \times Y3 \times X1 \times X2$ Cross-classification with $Y3$ as the Putative Dependent Variable; Including Higher-Order Interactions

In this appendix, we resume the example from Appendix 1, extend the models for Granger causality, and include interactions higher than first order. As in the example

in Appendix 1, we consider the two series of observations $Y1$, $Y2$, and $Y3$, and $X1$ and $X2$, with $Y3$ being the last observation, and the numbers indicating the temporal order of observations. For both models, we assume that the cross-classification under study is spanned by all five variables. In the first model, we predict $Y3$ from past observations of Y. However, in contrast to the model in Appendix 1, we hypothesize that, for the prediction of $Y3$, we need the three-way interaction [$Y1$, $Y2$, $Y3$] in addition to the two-way interactions considered before. The first model thus becomes

$$\begin{aligned} log\,\widehat{m} = \lambda + \lambda^{X1} + \lambda^{X2} + \lambda^{Y1} + \lambda^{Y2} + \lambda^{Y3} \\ +\lambda^{Y3,Y1} + \lambda^{Y3,Y2} + \lambda^{Y1,Y2} + \lambda^{Y1,Y2,Y3} \\ +\lambda^{X1,X2}. \end{aligned}$$

In this model, we propose that the interaction between $Y1$ and $Y3$ varies across the categories of $Y2$. The design matrix given in Table 6 contains the vector that is needed for this additional interaction in column 12. This vector is orthogonal to all other vectors in the design matrix. Therefore, the corresponding parameter can be interpreted as specified in the coding vector.

For the interpretation of this effect, the odds ratio approach can also be used. We first consider the odds ratio for the [$Y1$, $Y3$] interaction. The odds ratio for this effect is

$$\theta^{Y1,Y3} = \frac{m_{1.1..}m_{2.2..}}{m_{1.2..}m_{2.1..}},$$

with

$m_{1.1..} = m_{11111} + m_{11112} + m_{11121} + m_{11122} + m_{12111} + m_{12112} + m_{12121} + m_{12122}$ etc. Conditioning this odds ratio on the categories of $Y2$ results in

$$\theta^{Y1,Y3|Y2} = \frac{\dfrac{m_{111..}m_{212..}}{m_{112..}m_{212..}}}{\dfrac{m_{121..}m_{222..}}{m_{122..}m_{222..}}},$$

with

$m_{1.1..} = m_{11111} + m_{11112} + m_{11121} + m_{11122}$ etc.

For the second model, we now condition the [$Y3$, $X2$] interaction on $X1$. The resulting model equation is

$$\begin{aligned} log\,\widehat{m} = \lambda + \lambda^{X1} + \lambda^{X2} + \lambda^{Y1} + \lambda^{Y2} + \lambda^{Y3} \\ +\lambda^{Y3,Y1} + \lambda^{Y3,Y2} + \lambda^{Y1,Y2} \\ +\lambda^{X1,X2} + \lambda^{X1,Y3} + \lambda^{X2,Y3} + \lambda^{X1,X2,Y3}. \end{aligned}$$

The coding vector for this effect appears in column 12 of the design matrix in Table 6. This vector is also orthogonal to all other vectors in the design matrix. An interpretation along the lines used for the interpretation of the effect that conditions the [$Y1$, $Y3$] interaction on $Y2$ is, therefore, possible.

Appendix 3: Conditional Probability Parameterization of the Granger Causality Model for the Cross-classification of V1, V2, P1, and P2 (Table 4)

The models discussed in the context of Table 4 can also be interpreted as path models. Path models can be parameterized as conditional probability models (Goodman, 1973; Vermunt, 1997; von Eye & Mun, 2013). In the present example, the model would be |

$$\pi_{V1,V2,P1,P2} = \pi_{V1}\pi_{P1V1}\pi_{V2,V1}\pi_{V1,V2,P1},$$

where the effects considered in the model are listed in the subscripts. Note that these effects are considered hierarchical. Therefore, the models expressed as log-linear models can be more flexible. In the present example, this path model comes with LR-$X^2 = 3.65$, which, for $df = 2$, suggests a well-fitting model ($p = 0.16$). The Delta options were invoked again with $\Delta = 0.1$. As the log-linear model, this model describes the data well, but the parameters needed for an interpretation in the sense of the series V1–V2 predicting the series P1–P2 fail to be significant. For example, the z-score for the parameter that represents the connection between P1 and V1 is 0.027 ($p = 0.98$). The lem code used for this model appears below.

```
* Example 2: conditional probability parameterization
man 4
dim 2 2 2 2
lab A B C D
mod A
C|A
D|ACB
B|A
add 0.1
dat [82 7 47 13 3 3 0 11
0 0 12 7 0 0 3 17]
```

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York, NY: Wiley.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1256–1258.

Beebee, H., Hitchcock, C., & Menzies, P. (Eds.). (2012). *The Oxford handbook of causation*. Oxford, UK: Oxford University Press.

Bogat, G. A., Levendosky, A. A., DeJonghe, E., Davidson, W. S., & von Eye, A. (2004). Pathways of suffering: The temporal effects of domestic violence on women's mental health. *Maltrattamento e abuso all'infanzia, 6*, 97–112.

Bourbonnais, R. (2011). *Économétrie* (6th ed.). Paris: Dunod.

Christopoulos, D. K., & Leon-Ledesma, M. (2008). Testing for Granger (non)-causality in a time varying coefficient VAR model. Retrieved April 20, 2013, from ftp://ftp.ukc.ac.uk/pub/ejr/RePEc/ukc/ukcedp/0802.pdf

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Engle, R. E., & Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica, 55*, 251–276.

Finkelstein, J. W., von Eye, A., & Preece, M. A. (1994). The relationship between aggressive behavior and puberty in normal adolescents: A longitudinal study. *Journal of Adolescent Health, 15*, 319–326.

Foster, E. M. (2012). Causal inference, identification, and plausibility. In B. Laursen, T. D. Little, & N. E. Card (Eds.), *Handbook of developmental research methods* (pp. 17–30). New York, NY: The Guilford Press.

Foucault, M. (1966). *Les mots et les choses—Une archéologie des sciences humaine*. Paris: Gallimard.

Gates, K. M., Molenaar, P. C. M., Hillary, F. G., Ram, N., & Rovine, M. J. (2010). Automatic search for fMRI connectivity mapping: An alternative to Granger causality testing using formal equivalences among SEM path modeling, VAR, and unified SEM. *NeuroImage, 50*, 1118–1125.

Goodman, L. A. (1973). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology, 78*, 1135–1191.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*, 424–438.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics, 33*, 121–130.

Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in Econometrics. *Journal of Econometrics, 26*, 1045–1066.

Kalimeri, K., Lepri, B., Aran, O., Jayagopi, D. P., Gatica-Perez, D., & Pianesi, F. (2012). Modeling dominance effects on nonverbal behaviors using granger causality. Retrieved April 20, 2013, from http://www.idiap.ch/~gatica/publications/KalimeriLepriAranJayagopiGaticaPianesi-icmi12.pdf

Liu, Y., & Badahori, M. T. (2012). A survey on granger causality: A computational view. http://www-bcf.usc.edu/~liu32/granger.pdf

Lütkepohl, H., & Krätzig, M. (Eds.). (2004). *Applied times series econometrics*. Cambridge, UK: Cambridge University Press.

Lynd-Stevenson, R. M. (2007). Concerns regarding the traditional paradigm for causal research: The unified paradigm and causal research in scientific Psychology. *Review of General Psychology, 11*, 286–304.

Mair, P., & von Eye, A. (2007). Application scenarios for nonstandard log-linear models. *Psychological Methods, 12*, 139–156.

Matsuada, R. L. (2012). Key advances in the history of structural equation models using path diagrams. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 43–55). New York, NY: The Guilford Press.

McArdle, J. J. (2012). Foundational issues in contemporary modeling. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 385–410). New York, NY: Guilford.

Mignon, V. (2008). *Économétrie. Théorie et applications*. Paris: Economica.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychology, 18*, 112–117.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York, NY: The Guilford Press.

Rudas, T. (1997). *Odds ratios in the analysis of contingency tables*. Newbury Park, CA: Sage.

Saunders, P. J. (1994). Empirical evidence on causal relationships between the money supply, prices and wages in the UK. *British Review of Economic Issues, 16*, 45–63.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica, 48*, 1–48.

Sobel, M. E. (1994). Causal inference in latent variable models. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 3–35). Newbury Park, CA: Sage.

Sobel, M. E. (1996). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York, NY: Plenum.

Stegmüller, W. (1983). *Erklärung, Begründung, Kausalität*. Berlin: Springer.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland.

Vermunt, J. K. (1997). *Log-linear models for event histories. Advanced quantitative techniques in the social sciences*. Thousand Oakes, CA: Sage Publications.

von Eye, A., & Bogat, G. A. (2005). Logistic regression and prediction configural frequency analysis—A comparison. *Psychology Science, 47*, 407–414.

von Eye, A., Mair, P., & Bogat, G. A. (2005). Prediction models for configural frequency analysis. *Psychology Science, 47*, 342–355.

von Eye, A., & DeShon, R. P. (2012). Directional dependency in developmental research. *International Journal of Behavior Development, 36*, 303–312.

von Eye, A., & Mun, E.-Y. (2013). *Log-linear modeling—Concepts, interpretation and applications*. New York, NY: Wiley.

von Eye, A., & Schuster, C. (2000). The odds of resilience. *Child Development, 71*, 563–566.

von Eye, A., & Wiedermann, W. (in press). Manifest variable Granger causality models for developmental research—A taxonomy. DOI: 10.1080/10888691.2014.1001512

von Eye, A., Wiedermann, W., & Mun, E.-Y. (2013). Granger causality - Statistical analysis under a configural perspective. *IPBS: Integrative Psychological & Behavioral Science, 48*, 79–99. doi:10.1007/s12124-013-9243-1.

Wang, A. (2011). Empirical study on the interaction relationship of regional logistics and regional economic growth. *Business Management and Electronic Information, 2*, 603–606.

Williamson, J. (2011). Mechanistic theories of causality. *Philosophy Compass, 6*, 421–447.

Decisions Concerning the Direction of Effects in Linear Regression Models Using Fourth Central Moments

Wolfgang Wiedermann

Abstract Direction dependence analysis is attracting growing attention in the social sciences for its potential to help decide concerning the direction of effects of linear regression models. Direction dependence analysis assumes that observed data deviate from normality. Various tests have been proposed that can be applied when observed variables are skewed. However, these tests cannot be used when data are nonnormal and symmetric. The present chapter discusses direction dependence approaches for symmetric nonnormal data based on the fourth central moment. A new direction dependence approach based on regression residuals obtained from competing linear regression models is proposed. Three significance tests are described which can be used to test hypotheses compatible with direction dependence when data are nonnormal and symmetric. Results of a Monte Carlo simulation are reported which suggest that the significance tests perform well under various data scenarios. An empirical example from research on intimate partner violence is given to illustrate the application of the direction dependence tests.

Introduction

In 1905, Karl Pearson defined the fourth central moment (i.e., the kurtosis[1]) of a distribution of a random variable to measure potential departures from normality and coined the terms "platokurtic," "mesokurtic," and "leptokurtic" to indicate kurtosis values smaller, equal, and larger than that of the normal distribution. Within the last decades, the interpretation of the kurtosis has vividly been debated. Some authors

[1]In this article, we define the kurtosis of a random variate, X, as $\beta_2 = \frac{E\left[(X-\mu)^4\right]}{E\left[(X-\mu)^2\right]^2} = \frac{\mu_4}{\sigma^4}$ and Pearson's adjusted excess kurtosis as $\delta_X = \frac{\kappa_4(X)}{\kappa_2(X)^2} = \frac{\mu_4}{\sigma^4} - 3$ with $\kappa_r(X)$ being the rth cumulant of X.

W. Wiedermann (✉)
Department of Educational, School, and Counseling Psychology College of Education, University of Missouri, Columbia, MO 65211, USA
e-mail: wiedermannw@missouri.edu

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_7

suggest that kurtosis is a measure of the peakedness of a distribution (e.g., Katz et al., 2013; Lee, Lee & Lee, 2013), while others suggested that the kurtosis is correctly interpreted as an indicator of tail heaviness (e.g., Ali 1974). The use of kurtosis as a measure of bimodality has also been discussed (e.g., Darlington, 1970). Livesay (2007) demonstrated that the kurtosis can also be used to detect outliers. Some researchers called for the provisional agreement that kurtosis is related to both peak and tails of a distribution (Ruppert, 1987; van Staden & Loots, 2009). Recently, Westfall (2014) analyzed whether kurtosis values convey useful information about the peakedness of a distribution and concluded that "... kurtosis should never be defined in terms of peakedness" (p. 194) and that "the relationship of peakedness with kurtosis is now officially over" (p. 194). Similar conclusions were already drawn by DeCarlo (1997, 1998).

Despite these interpretability issues, unambiguous consensus exists that the concept of kurtosis, together with measures of skew (e.g., the third central moment), are useful for determining whether a distribution deviates from the normal distribution. Nonzero skewness and/or kurtosis values deviating from 3 (equivalently to excess kurtosis values deviating from zero) suggest nonnormality. Note that the reverse corollary, i.e., skewness and excess kurtosis of zero, does not necessarily imply that data follow a normal distribution (van Staden & Loots, 2009). Thus, the line of research using information conveyed by higher order moments typically focuses on its potential to detect deviations from the normal distribution. Several significance tests have been proposed to infer whether higher moments statistically suggest deviations from the normal distribution (D'Agostino, 1970; Anscombe & Glynn, 1983, for an overview see Yap & Sim 2011). These tests are commonly applied to test the distributional assumptions of various parametric methods for statistical inference (Rochon & Kieser, 2011; Rochon, Gondan, & Kieser, 2012; Schucany & Ng, 2006).

In recent years, a second line of research has developed, *direction dependence research*, which analyzes whether indicators of higher moments can be used to address a fundamental issue associated with correlational and regression analyses in observational cross-sectional studies, i.e., the direction of observed effects (Bentler, 1983; Dodge & Rousson, 2000, 2001; Dodge & Yadegari, 2010; Muddapur, 2003; Shimizu & Kano, 2008; Shimizu et al., 2011; Sungur, 2005; von Eye & DeShon, 2008, 2012, Wiedermann et al., 2013, 2015; Wiedermann & von Eye, 2015a; Wiedermann & Hagmann, 2014). This line of research implicitly provokes a "paradigmatic shift" in which nonnormality of observed variables is no longer inevitable dismissed as a potential source of bias, but as a valuable source of information that can be used to gain deeper insights into the directional structures of variables. This proposition is based on distributional characteristics of nonnormal variables in the linear regression context. For example, in the bivariate linear regression setting, it is well known that second order moments (covariances, and correlations) cannot be used to decide whether two variables, X and Y, are related in the form $X \rightarrow Y$ (i.e., X is the cause and Y is the outcome) or whether the reverse flow of causality, $Y \rightarrow X$, (with Y being the cause and X being the outcome) is more likely to reflect the true data generating process (see, e.g., von Eye & DeShon,

2012). However, research on direction dependence revealed so-called asymmetric properties of the Pearson correlation coefficient which can help researchers make empirical statements about which of two competing regression models ($X \rightarrow Y$ or $Y \rightarrow X$) can be considered as a valid approximation of the true underlying model. These asymmetric properties are related to higher than second moments.

The present chapter (1) introduces asymmetric facets of the Pearson correlation coefficient which concern observed variables, (2) presents extensions of direction dependence methods using regression residuals, (3) demonstrates that kurtosis measures carry important information that can be used to arrive at conclusions concerning directionality in bivariate linear regression models, and (4) proposes significance tests to infer on the direction of effects. The performance of the significance tests is demonstrated using Monte Carlo simulations. An empirical data example is given for illustrative purposes.

Asymmetric Properties of the Pearson Correlation Coefficient

The Pearson correlation coefficient is commonly introduced in its symmetric form, i.e., $\rho_{XY} = cov\,(X, Y)/(\sigma_X \sigma_Y)$, where $cov(X, Y)$ denotes the covariance and σ_X and σ_Y are the standard deviations of X and Y. Because the covariance does not depend on variable order, we obtain $\rho_{XY} = \rho_{YX}$, which implies that, based on the Pearson correlation coefficient, no statements concerning cause and effect can be made. Similarly, in the bivariate linear regression case, the slope parameters obtained from the two competing regression models $X \rightarrow Y$ and $Y \rightarrow X$ will be identical when variables are standardized (e.g., von Eye & DeShon, 2012). Again, no decisions concerning directionality can be made because the two models fit the data equally well.

Direction dependence research investigates asymmetric facets of the Pearson correlation which result from the additive definition of the linear regression model. An outcome variable, Y, is assumed to emerge from the (additive) convolution of a predictor variable, X, and an error term. The error term is commonly assumed to be normally distributed, i.e., exhibiting zero skewness and zero excess kurtosis. When the true predictor is nonnormally distributed, the true outcome variable will be closer to the normal distribution due to the convolution of a non-normal variable (the predictor) and a normal variable (the error term). Dodge and Rousson (2000, 2001) presented an algebraic proof for this proposition and showed that the (absolute value of the) skewness of Y will always be smaller than the (absolute value of the) skewness of X. von Eye and DeShon (2008) as well as Dodge and Yadegari (2010) extended this relation to the fourth higher moment. Dodge and Yadegari (2010) showed that the Pearson correlation can be written as

$$\rho_{XY}^4 = \frac{\delta_Y}{\delta_X}$$

where δ_Y and δ_X denotes the excess kurtosis of Y and X. Note that ρ^4_{XY} cannot exceed the interval [0, 1] from which follows that the excess kurtosis of Y will always be smaller than the excess kurtosis of X provided that $\rho_{XY} \neq 0$ and $\delta_X \neq 0$.

von Eye and DeShon (2012) as well as Pornprasertmanit and Little (2012) discussed significance test to evaluate hypotheses which are compatible with direction dependence.

The observed variable-based direction dependence approach discussed so far exhibits two potential limitations. First, the approach is restricted to the bivariate setting which seriously hampers practical application. Second, decisions concerning directionality are made based on marginal properties of variables without estimating the corresponding linear regression models. This may entice researchers to make directional statements without asking whether the selected model is indeed capable of validly describing the relation between the observed variables. As an alternative, Wiedermann et al. (2013) and Wiedermann, Hagmann, and von Eye (2015) discussed residual-based direction dependence approaches which (1) can easily be extended to the multiple linear regression setting (see Wiedermann and von Eye, 2015b) and (2) base the decision concerning direction of effect on a synthesis of characteristics of both linear regression models. The latter approach is currently based solely on the third central moment. In the following section, we present an extension of the residual-based direction dependence methodology to the fourth central moment.

Direction Dependence Properties of Residuals

Wiedermann et al. (2013), von Eye and Wiedermann (2014), Wiedermann and Hagmann (2014), as well as Wiedermann, Hagmann and von Eye (2015) have discussed direction of dependence properties of regression residuals with respect to the third central moment in the bivariate regression setting. The multiple linear regression case is discussed by Wiedermann and von Eye (2015a). In the following paragraphs, we aim at extending the approach to the case of the fourth central moment, i.e., the excess kurtosis of the residuals of competing regression models.

In the bivariate data scenario, the following two regression models can be estimated based on the variables X and Y:

$$Y = \alpha_Y + \beta_Y X + \varepsilon_Y \tag{1}$$

and

$$X = \alpha_X + \beta_X Y + \varepsilon_X. \tag{2}$$

The subscripts denote the corresponding response variables, α_Y and α_X are the model intercepts, β_Y and β_X are the ordinary least squares (OLS) regression coefficients, and ε_Y and ε_X denote the OLS residuals. For the true model (i.e., the

true data generating process), residuals are assumed to be normally distributed with an expected value of zero and a variance σ_{ε}^2, homoscedastic, serially independent, and independent of the explanatory variable.

For the following discussion, let model (1) constitute the true model. Model (2) corresponds to the mis-specified regression model. Further, without loss of generality, we assume that the model intercepts are fixed at $\alpha_Y = \alpha_X = 0$. Inserting Eq. (1) into Eq. (2) and considering that $\beta_X \beta_Y$ equals the square of the Pearson correlation coefficient (ρ_{XY}^2), one obtains

$$\varepsilon_X = \left(1 - \rho_{XY}^2\right) X - \beta_X \varepsilon_Y. \tag{3}$$

for the regression residuals of the mis-specified model. The true predictor, X, and the true error term, ε_Y, are assumed to be stochastically independent. Thus, one obtains for the fourth-order cumulants $\kappa_4(\cdot)$ (see, for example, Kendall & Stuart 1979):

$$\kappa_4\left(\varepsilon_X\right) = \left(1 - \rho_{XY}^2\right)^4 \kappa_4(X) - \beta_X^4 \kappa_4\left(\varepsilon_Y\right). \tag{4}$$

Defining the excess kurtosis of ε_X and X in terms of higher order cumulants leads to

$$\delta_{\varepsilon_X} = \kappa_4\left(\varepsilon_X\right) / \sigma_{\varepsilon_X}^4 \tag{5}$$

and

$$\delta_X = \kappa_4(X) / \sigma_X^4. \tag{6}$$

Dividing Eq. (4) through the fourth power of the standard deviation of ε_X (i.e., $\sigma_{\varepsilon_X}^4$) and making use of Eqs. (5) and (6), one arrives at the following equation for the excess kurtosis of residuals of the mis-specified model:

$$\delta_{\varepsilon_X} = \left(\frac{\left(1 - \rho_{XY}^2\right) \sigma_X}{\sigma_{\varepsilon_X}}\right)^4 \delta_X - \left(\beta_X \frac{\sigma_{\varepsilon_Y}}{\sigma_{\varepsilon_X}}\right)^4 \delta_{\varepsilon_Y}. \tag{7}$$

The standard deviations of ε_X and ε_Y can be written as $\sigma_{\varepsilon_X} = \sqrt{1 - \rho_{XY}^2}\sigma_X$ and $\sigma_{\varepsilon_Y} = \sqrt{1 - \rho_{XY}^2}\sigma_Y$. Thus, Eq. (7) simplifies to

$$\delta_{\varepsilon_X} = \left(1 - \rho_{XY}^2\right)^2 \delta_X - \rho_{XY}^4 \delta_{\varepsilon_Y}. \tag{8}$$

Finally, in OLS regression, the excess kurtosis of the true error term is expected to be zero (i.e., $\delta_{\varepsilon_Y} = 0$) which leads to $\delta_{\varepsilon_X} = \left(1 - \rho_{XY}^2\right)^2 \delta_X$. From this expression, one can conclude that the excess kurtosis of the error term of the mis-specified model is, in fact, a weighted version of the excess kurtosis of the true predictor. To be more specific, we conclude that (1) the excess kurtosis of X and ε_X will always have the

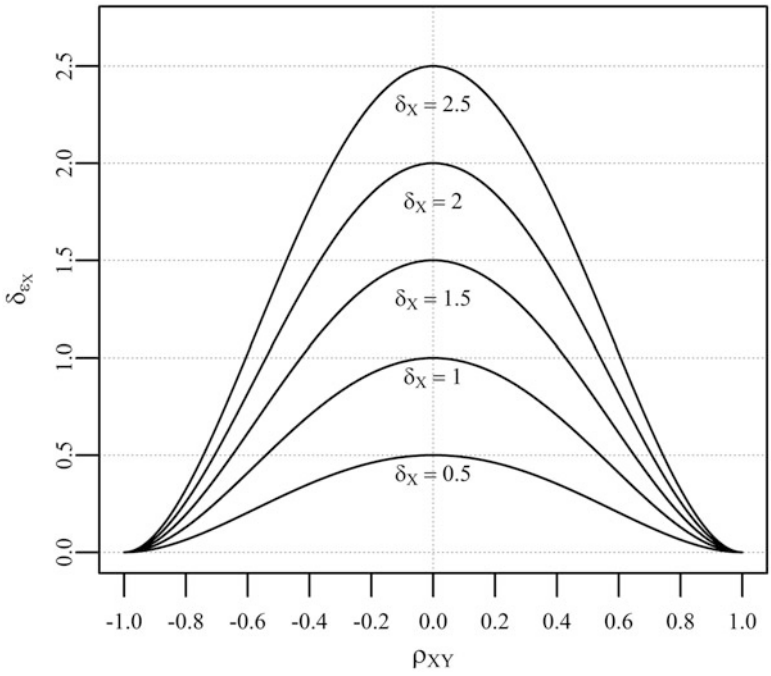


Fig. 1 Theoretical values of δ_{ε_X} as a function of the correlation between X and Y and the excess kurtosis of X

same sign; (2) the excess kurtosis of ε_X declines with the correlation between X and Y (i.e., the weighting term in Eq. (8) basically reflects the portion of unexplained variability of the bivariate relation), and, most importantly (3) the excess kurtosis of ε_X increases with the excess kurtosis of X (see Fig. 1).

Considering the third conclusion (stated above) in tandem with the routinely made assumption of normally distributed residuals of the correctly specified model, i.e., $\delta_{\varepsilon_Y} = 0$, it follows that the comparison of δ_{ε_Y} and δ_{ε_X} can be used to derive conclusions concerning the direction of effects. Assuming that observed variables show an excess kurtosis greater than zero, one can conclude that Y is the response variable and X the predictor when $\delta_{\varepsilon_X} > \delta_{\varepsilon_Y}$. Conversely, when $\delta_{\varepsilon_X} < \delta_{\varepsilon_Y}$ then X is more likely to be the response variable and Y is more likely to be on the predictor side. In addition to this rather descriptive decision rule, standard errors of excess kurtosis values have to be taken into account as well (von Eye & DeShon, 2012). In the next section, we present three different approaches to statistical inference on the equality of excess kurtosis values of regression residuals.

Statistical Inference on the Equality of Excess Kurtosis of Regression Residuals

This section discusses a combined Anscombe-Glynn normality test and two significance tests based on the difference of excess kurtosis values for determining the direction of effects in linear regression models. The combined normality test consists of two one-sample tests where the null hypotheses $H_0 : \delta_{\varepsilon_X} = 0$ and $H_0 : \delta_{\varepsilon_Y} = 0$ are evaluated separately.

Combined Anscombe-Glynn Test

Anscombe and Glynn (1983) suggested a transformation of the kurtosis estimate to more closely approximate normality. Let

$$b_2 = m_4/(m_2)^2 \tag{9}$$

where

$$m_k = \sum \left(X - \overline{X}\right)^k / n, \tag{10}$$

where k is a positive natural number, and $\overline{X}$ is the sample mean. Anscombe and Glynn's transformation involves the following computations:

$$E\left(b_2\right) = \frac{3\left(n-1\right)}{n+1} \tag{11}$$

$$var\left(b_2\right) = \frac{24n\left(n-2\right)\left(n-3\right)}{\left(n+1\right)^2\left(n+3\right)\left(n+5\right)} \tag{12}$$

$$x = \left(b_2 - E\left(b_2\right)\right) / \sqrt{var\left(b_2\right)} \tag{13}$$

$$\sqrt{b_2'} = \frac{6\left(n^2 - 5n + 2\right)}{\left(n+7\right)\left(n+9\right)} \sqrt{\frac{6\left(n+3\right)\left(n+5\right)}{n\left(n-2\right)\left(n-3\right)}} \tag{14}$$

$$a = 6 + \frac{8}{\sqrt{b_2'}}\left[\frac{2}{\sqrt{b_2'}} + \sqrt{\left(1 + \frac{4}{b_2'}\right)}\right] \tag{15}$$

and

$$z(\delta) = \left(\left(1 - \frac{2}{9a}\right) - \left(\frac{1 - \frac{2}{a}}{1 + x\sqrt{\frac{2}{a-4}}}\right)^{\frac{1}{3}}\right) \Big/ \sqrt{\frac{2}{9a}}, \tag{16}$$

where $z(\delta)$ approximately follows a standard normal distribution. The null hypothesis of zero excess kurtosis ($H_0 : \delta = 0$) is rejected against the two-sided alternative, $H_1 : \delta \neq 0$, when $|z(\delta)|$ exceeds the $1 - \alpha/2$ quantile of the standard normal distribution.

Recall that normality of the error term is assumed for the true model. This assumption, together with the theoretical result presented in Eq. (8), leads to the following decision rule for directional statements concerning the observed effect: When the null hypothesis $H_0 : \delta_{\varepsilon_Y} = 0$ is retained and, at the same time, the null hypothesis $H_0 : \delta_{\varepsilon_X} = 0$ is rejected, model (1) is more likely to reflect the true data generating process which implies that Y is the outcome and X is the predictor. Conversely, when the $H_0 : \delta_{\varepsilon_Y} = 0$ is rejected and $H_0 : \delta_{\varepsilon_X} = 0$ is retained, model (2) is more likely to reflect the data generating process. When both null hypotheses are retained/rejected, no distinct decision is possible based on the combined Anscombe-Glynn procedure.

Asymptotic Kurtosis Difference Test

The combined Anscombe-Glynn approach discussed above relies on separately testing the distributional properties of both error terms. For reasons of α protection, a test based on the difference of excess kurtosis values may be worthwhile. Wiedermann et al. (2015) proposed an asymptotic significance to evaluate the differences in skewness estimates based on the D'Agostino skewness z-values. In the following paragraphs, we extend the approach to the fourth central moment. Here, the difference of Anscombe-Glynn z-values (Anscombe & Glynn, 1983) can be used to make decisions upon the direction of effects. Let $z(\delta_{\varepsilon_X})$ and $z(\delta_{\varepsilon_Y})$ be the Anscombe-Glynn z-values corresponding to the kurtosis of ε_X and ε_Y. In the normal case,

$$z_\delta = \frac{z(\delta_{\varepsilon_X}) - z(\delta_{\varepsilon_Y})}{\sqrt{2 - 2\rho^4_{\varepsilon_X\varepsilon_Y}}}, \tag{17}$$

approximates a standard normal distribution, and the null hypothesis of excess kurtosis equality, $H_0 : \delta_{\varepsilon_X} = \delta_{\varepsilon_Y}$, is rejected against the two-sided alternative ($H_1 : \delta_{\varepsilon_X} \neq \delta_{\varepsilon_Y}$) when $|z_\delta|$ exceeds the $1-\alpha/2$ quantile of the standard normal distribution. Details of deriving the test statistic in Eq. (17) are given in the Appendix. When, for example, $\delta_X > 0$ and the null hypothesis $H_0 : \delta_{\varepsilon_X} = \delta_{\varepsilon_Y}$ is rejected against the one-sided alternative $H_1 : \delta_{\varepsilon_X} > \delta_{\varepsilon_Y}$, one can conclude that Y is the outcome and X is the explanatory variable.

Bootstrap Difference Test

The z_δ - test in Eq. (17) makes use of the Anscombe-Glynn z-values and, thus, relies on the assumption of normality of the true error term. The null hypothesis will be rejected when either one or both residual terms deviate from normality. In other words, normality for the true model is essential for best practice applications. This assumption can be relaxed for the following bootstrap difference test.

Again, let δ_{ε_Y} and δ_{ε_X} be the excess kurtosis values of residuals from the models $X \to Y$ and $Y \to X$ and let $\delta_d = \delta_{\varepsilon_X} - \delta_{\varepsilon_Y}$ be the difference in excess kurtosis values. When variables show excess kurtosis values larger than zero, we expect $\delta_d > 0$ when the true model states that $X \to Y$ and $\delta_d < 0$ when the true model states that $Y \to X$. A bootstrap p-value for testing the null hypothesis $H_0 : \delta_{\varepsilon_X} = \delta_{\varepsilon_Y}$ against the one-sided alternative hypothesis $H_1 : \delta_{\varepsilon_X} > \delta_{\varepsilon_Y}$ is obtained via randomly sampling pairs of residuals $(\varepsilon_Y, \varepsilon_X)$ with replacement from the original regression residuals. $\delta_d^{'} = \delta_{\varepsilon_X}^{'} - \delta_{\varepsilon_Y}^{'}$, the difference in excess kurtosis values for each of m resamples, leads to the bootstrap p-value:

$$p_\delta^{'} = m^{-1} \sum I\left(\delta_d^{'} < 0\right) \tag{18}$$

and the null hypothesis is then rejected against the one-sided alternative $H_1 : \delta_{\varepsilon_X} > \delta_{\varepsilon_Y}$ when $p_\gamma^{'}$ is smaller than the nominal significance level (e.g., 5 %). Note that in deriving $\delta_{\varepsilon_X} = \left(1 - \rho_{XY}^2\right)^2 \delta_X$ we assume that the excess kurtosis of the true error term is zero. However, no statement is made concerning symmetry of residuals corresponding to the true model. According to the well-known skewness-kurtosis inequality (Teuscher & Guiard, 1995), the skewness, thus, can vary from $-\sqrt{2}$ to $\sqrt{2}$ when $\delta_{\varepsilon_Y} = 0$.

Performance of the Direction Dependence Tests

To illustrate the performance of the three proposed direction dependence tests, a simulation study was performed using the R statistical environment (R Core Team, 2014). Two simulation experiments were implemented. First, predictors

were sampled from a standard normal distribution. Because the proposed direction dependence methods rely on the assumption of nonzero excess kurtosis of observed variables, this part of the simulation study was used to assess the Type I error robustness of the tests. Second, to systematically evaluate the power performance of the tests, predictors were sampled from a generalized error distribution (also known as the exponential power distribution) with zero mean and unit variance (Evans, Hastings, & Peacock 2000). All samples were generated according to the model $Y = \alpha_Y + \beta_Y X + \varepsilon_Y$, with $\beta_Y = \rho_{XY}/\sqrt{1-\rho_{XY}^2}$, $\rho_{XY} = 0,\ 0.2,\ 0.4,\ 0.6,$ and 0.8, and $\alpha_Y = 1$. The error term, ε_Y, was randomly sampled from the standard normal distribution. The excess kurtosis of X was set at $\delta_X = 0,\ 0.5,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,$ and 9. Sample sizes were $n = 50$, 75, 100, 125, 150, and 200. The simulation factors were fully crossed resulting in 11 (excess kurtosis of X) $\times$ 5 (correlation) $\times$ 6 (sample size) $= 330$ experimental conditions. For each condition, 5000 samples were generated and decisions concerning directionality were made based on the results of the combined Anscombe-Glynn test, the asymptotic kurtosis difference test, and the bootstrap difference test (applying 1000 resamples) following the decision rules described above. Decisions based on the difference of excess kurtosis values were performed one sided. We used Bradley's (1978) liberal criterion to determine in the Type I error robustness, i.e., given a nominal significance level of 0.05, a test is considered robust if the empirical Type I error rates do not exceed the interval [0.025; 0.075].

Table 1 gives the empirical Type I error rates of the three tests as a function of the sample size and the correlation of X and Y. The Type I error rates of the Anscombe-Glynn test are given in terms of model selection, i.e., the Type I error rates reflect the portion in which the null hypothesis $H_0 : \delta_{\varepsilon_Y} = 0$ is retained and, at the same time, the null hypothesis $H_0 : \delta_{\varepsilon_X} = 0$ is rejected. Overall, all tests were able to protect the nominal significance level across all experimental conditions. In other words, we can conclude that, under the given condition of normality, selecting the model $X \rightarrow Y$ is solely based on chance, within the nominal significance level, as expected.

Figure 2 gives the empirical power curves of the three direction dependence tests as a function of ρ_{XY} and δ_X for sample sizes $n = 50$, 100, 150, and 200. In general, power of tests increases with sample size due to increasing precision of parameter estimation. Further, power increases with excess kurtosis of the predictor and decreases with the correlation of X and Y. Both effects are in line with the theoretical results given in Eq. (8). Except for highly correlated variables, the combined Anscombe-Glynn test is more powerful than the tests based on the difference in excess kurtosis values. This can be explained by the fact that the latter tests consider sampling variability in both excess kurtosis values which decreases power. From Fig. 2, we conclude that all three tests are well suited to make statements concerning the direction of effects.

Table 1 Empirical Type I error rates of the three direction dependence tests using a nominal significance level of 5 %

Sample size	Correlation	Kurtosis difference	Bootstrap difference	Anscombe-Glynn
50	0	0.0468	0.0350	0.0528
50	0.2	0.0578	0.0412	0.0482
50	0.4	0.0528	0.0402	0.0534
50	0.6	0.0530	0.0384	0.0518
50	0.8	0.0560	0.0428	0.0460
75	0	0.0472	0.0424	0.0546
75	0.2	0.0510	0.0420	0.0498
75	0.4	0.0516	0.0492	0.0528
75	0.6	0.0546	0.0486	0.0492
75	0.8	0.0506	0.0416	0.0472
100	0	0.0526	0.0474	0.0548
100	0.2	0.0492	0.0500	0.0480
100	0.4	0.0542	0.0530	0.0534
100	0.6	0.0554	0.0534	0.0474
100	0.8	0.0604	0.0540	0.0420
125	0	0.0540	0.0530	0.0540
125	0.2	0.0518	0.0538	0.0484
125	0.4	0.0512	0.0582	0.0506
125	0.6	0.0536	0.0592	0.0474
125	0.8	0.0552	0.0622	0.0464
150	0	0.0550	0.0562	0.0560
150	0.2	0.0514	0.0532	0.0470
150	0.4	0.0550	0.0542	0.0574
150	0.6	0.0528	0.0568	0.0444
150	0.8	0.0542	0.0530	0.0476
175	0	0.0546	0.0554	0.0476
175	0.2	0.0624	0.0612	0.0558
175	0.4	0.0528	0.0542	0.0464
175	0.6	0.0544	0.0596	0.0508
175	0.8	0.0524	0.0556	0.0452
200	0	0.0518	0.0574	0.0496
200	0.2	0.0514	0.0548	0.0540
200	0.4	0.0496	0.0544	0.0506
200	0.6	0.0494	0.0576	0.0478
200	0.8	0.0516	0.0582	0.0452

Empirical Example: Relation Between Depression and PTSD

To illustrate the application of the proposed methodology, we use data from Bogat et al. (2003, 2004) on psychosocial development of women who are victims of intimate partner violence. von Eye and DeShon (2012) used the data to determine

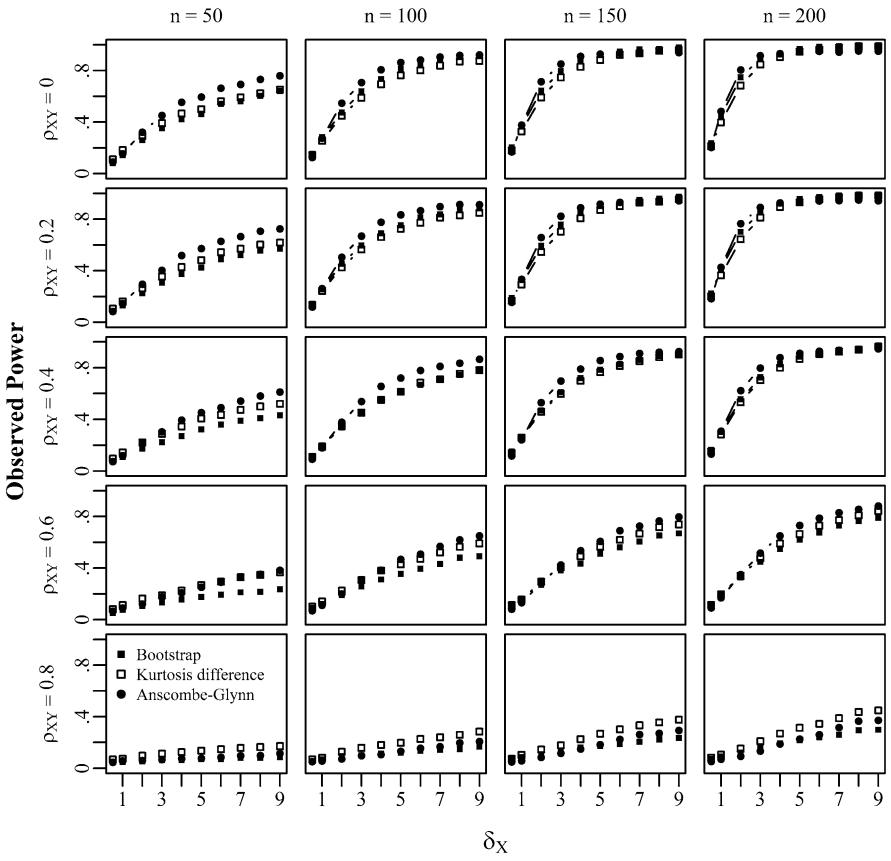


Fig. 2 Empirical power for the three direction dependence tests

the direction of effect of depression and post-traumatic stress disorder (PTSD) using observed-variable direction dependence methods and found that PTSD is more likely to be the cause of depression than vice versa. We reanalyze the relationship between depression and PTSD using fourth central moments of residuals of competing regression models. PTSD was measured using the PTSD scale for Battered Women (Saunders, 1994) which consists of seventeen items rated on an 8-point scale (higher scores indicate more pronounced PTSD symptomology; range = 0–119). Depression was measured using Beck's Depression Inventory (Beck, Ward, Mendelson, Mock & Erbaugh, 1961) which consists of 21 items (higher scores indicate more pronounced depression symptomology; range = 0–63). For the present reanalysis, we used cross-sectional data of 152 women who ever experienced domestic violence from the third wave of Bogat's et al. (2003, 2004) longitudinal study.

We observed an average depression score of 8.22 ($\widehat{\sigma}_{Depr} = 6.06$; excess kurtosis $\widehat{\delta}_{Depr} = 0.144$), an average PTSD score of 6.45 ($\widehat{\sigma}_{PTSD} = 11.16$; $\widehat{\delta}_{PTSD} = 5.988$) and a Pearson correlation of $\widehat{\rho}_{Depr,\,PTSD} = 0.327$. Both variables significantly deviate from normality according to the Shapiro-Wilk test (depression: $W = 0.942$, $p < .001$; PTSD: $W = 0.640$, $p < .001$). Moving to the linear regression models, we, first, regressed "Depression" on "PTSD" which constitutes the target model. The following regression diagnostic procedures were applied: (1) the linearity assumption was evaluated by including second and third order polynomials and testing the significance of the model fit change, (2) homoscedasticity was evaluated using the studentized Breusch-Pagan test (see Koenker, 1981), (3) the presence of outliers was evaluated using Bonferroni corrected tests of largest studentized residuals, and (4) potentially influential observations were searched using hat values and Cook's distances. For the target model, "$PTSD \rightarrow Depression$," we observed one conspicuous observation with a hat value larger than three times the average hat values and a Cook's distance of 0.402 (note that 97.5 % of the observations had a Cook's distance smaller than 0.06). We repeated the analysis after removing this conspicuous observation. Based on $n = 151$ women, we obtained $Depr = 6.910 + 0.215 \cdot PTSD$ ($t = 4.671$, $p < .001$) for the target model and $PTSD = 1.179 + 0.595 \cdot Depr$ (with, by necessity, $t = 4.671$, $p < .001$) for the alternative model. The linearity assumption was confirmed for both models, i.e., including higher order terms led to a non-significant change in model fit. For the target model, we obtained a nonsignificant Breusch-Pagan test ($\chi^2 = 0.880$, $df = 1$, $p = .348$) suggesting homoscedasticity. In contrast, the assumption of homoscedasticity was rejected for the alternative model ($\chi^2 = 10.453$, $df = 1$, $p = .001$). No outliers were detected for the target model based on largest studentized residuals (largest studentized residual: 3.488, Bonferroni adjusted $p = .097$). For the alternative model, studentized residuals suggested one potential outlier with a studentized residual of 4.213 ($p = .007$). Cook's distances and hat values were rather low for both models.

In the next step, we applied the proposed direction dependence tests to evaluate whether the target model or the alternative model is more likely to reflect the true causal flow. First, regression residuals were extracted from both models (see Fig. 3). For the model residuals of the target model, we observed an excess kurtosis of 0.721 and the Anscombe-Glynn test suggested retaining the null hypothesis of zero excess kurtosis ($z = 1.741$, $p = .082$). In contrast, for the residuals of the alternative model we observed an excess kurtosis of 3.099 and the Anscombe-Glynn test suggests rejecting the null hypothesis of zero excess kurtosis ($z = 4.028$, $p < .001$). Next, we asked whether the procedures based on the differences in excess kurtosis values reject the null hypotheses as well. The asymptotic kurtosis difference test failed to reject the null hypothesis of zero kurtosis difference according to the 5 % criterion ($z = 1.630$, $p = .052$). However, the bootstrap difference test gave a bootstrap p-value of 0.001 based on 5000 resamples and, thus, also rejected the null hypothesis. Overall, we can conclude that the initially selected target model

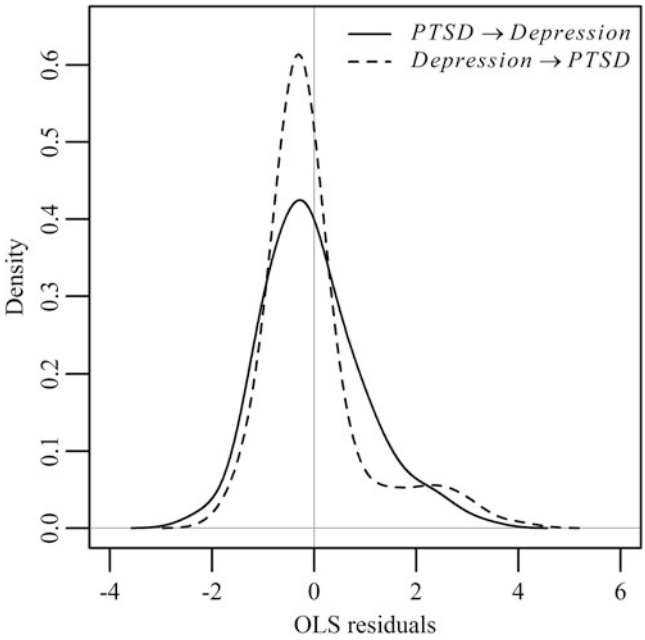


Fig. 3 Observed density of residuals from the two competing regression models

("*PTSD* → *Depression*") is more likely to reflect the true data generating process which is in line with results of von Eye and DeShon (2012).

Discussion

In the present chapter, we discussed using the fourth central moment to determine the direction of effects in the bivariate linear regression setting, and we proposed three significance procedures for statistical inference. Simulation results suggest good Type I error and power properties under various data scenarios. Note that comparing distributional characteristics of residuals obtained from competing regression models relies on the commonly met assumption of normality of the error term corresponding to the true model. In other words, the significance tests make use of a well-known assumption which should routinely be evaluated in practical applications. Normality is crucial in particular for the asymptotic kurtosis difference test which is based on the Anscombe-Glynn z-values. Violations of the assumptions can lead to distorted Type I error and power rates. However, the test statistic may serve as a starting point for the development of more robust tests. In contrast, for the bootstrap tests, the distributional assumption can be relaxed. Here, however, the true error term must exhibit an excess kurtosis of zero. Future studies

are needed to systematically quantify the robustness properties of the tests under violated distributional assumptions.

Wiedermann and Hagmann (2014) discuss direction dependence methods for the third central moment for nonnormal true error terms and show that $-\rho_{XY}^3 = \gamma_{\varepsilon_X}/\gamma_{\varepsilon_Y}$ with γ_{ε_X} and γ_{ε_Y} being the skewness of ε_X and ε_Y. In other words, when a priori considerations exist that justify the assumption of nonnormal true error terms, direction dependence methods can still be applied to test these hypotheses. From these results, two important implications arise.

First, from Eq. (8) it becomes obvious that a similar statement can be made concerning the fourth central moment, i.e., assuming a normally distributed predictor, one obtains $-\rho_{XY}^4 = \delta_{\varepsilon_X}/\delta_{\varepsilon_Y}$.

Second, the proposed direction dependence methods should not be confused with causal search algorithms (see, e.g., Scheines et al., 1998; Spirtes, Glymour, & Scheines 2000) which exhibit a stronger exploratory element, and are used to generate new hypotheses. In contrast, we characterize direction dependence tests as purely confirmatory in nature and assume that a valid regression model exists that includes proper distributional assumptions concerning the corresponding error term (the normality assumption may be sufficient in practice). Further, direction dependence methodology requires that the regression model can be validly interpreted. Thus, common regression diagnostics (e.g., Belsley, Kuh, & Welsch, 1980) are essential to avoid biased conclusions concerning the direction of effects.

Wiedermann and von Eye (2015b) proposed guidelines for confirmatory direction dependence analysis and showed that decision concerning the direction of effects have to be based on a careful synthesis of properties of both, the tentative model (e.g., $X \rightarrow Y$) and the alternative model ($Y \rightarrow X$). Important elements of the guidelines include (1) distributional characteristics of regression residuals (the proposed tests can be added to the already existing procedures for the systematic comparison of higher moments of residuals), (2) independence of the predictor and the error term (for further discussion see, e.g., Shimizu et al., 2011; Wiedermann & von Eye, 2015b), and (3) distributional characteristics of observed variables (such as the methodology proposed by von Eye and DeShon, 2012, and Pornprasertmanit and Little, 2012).

The presented approach addresses directional hypotheses in the bivariate linear regression setting. Extensions of direction dependency based on the third central moment to the multiple linear regression setting are discussed in Wiedermann and von Eye (2015a). Analogous extensions can be made for the fourth central moment. Consider the case of one response variable, Y, and two explanatory variables, X and Z, defining the true model $Y = \alpha_Y + \beta_{YX}X + \beta_{YZ}Z + \varepsilon_Y$. Further, suppose that Z is a priori known to be a predictor (e.g., due to logical order) which leads to the competing regression model $X = \alpha_X + \beta_{XY}Y + \beta_{XZ}Z + \varepsilon_X$. When true predictors are independent (which constitutes a common assumption in OLS regression), the fourth central moment of ε_X can be written as

$$\delta_{\varepsilon_X} = \left(\frac{\left(1 - \rho^2_{XY|Z}\right) \sigma_X}{\sigma_{\varepsilon_X}} \right)^4 \delta_X \tag{19}$$

where $\rho_{XY|Z}$ denotes the partial correlation of X and Y controlling for Z (the corresponding proof works in fashion analogous to the one presented in Wiedermann & von Eye, 2015a). From Eq. (19) it becomes evident that (when X and Z are independent) neither the distributional characteristic of Z, nor the correlation between Y and Z affect the excess kurtosis of ε_X. In addition, the excess kurtosis increases with δ_X which implies that comparing δ_{ε_Y} (again assumed to be zero) and δ_{ε_X} may help researchers to test hypotheses of directionality in the multivariate setting. Wiedermann and von Eye (2015a) showed that violations of the independence assumptions of predictors (so-called multicollinearity) do not affect the Type I error robustness of direction dependence tests using third central moments. Additional simulation experiments are being planned for the future to evaluate whether similar robustness properties hold for fourth central moments tests as well.

Note that the presented results are, in fact, not restricted to strictly bivariate applications. Multivariate data scenarios exist which can readily be analyzed using the proposed tests. For example, consider the case of a mediation model in which the predictor is randomized (e.g., intervention versus control group) and assumed to influence a mediator which, in turn, influences the outcome. When the distributional requirements of the tests hold (in this case nonnormality of the mediator), the proposed procedures can be applied to infer on directionality of the mediator-outcome path. For details concerning direction of effects in mediation analysis see Wiedermann and von Eye (2015c).

Further, the direction dependence principle is, of course, not restricted to OLS regression techniques and manifest variables. Concerning the former, direction dependence methodology may also be a valuable companion for structural equation modelling (SEM). Within SEM, an important distinction has to be made between residuals regularly obtained from structural analyses and the residuals necessary to perform direction dependence tests. Errors in SEM typically concern the variance-covariance matrix of variables (and are, thus, available on an aggregated level) while errors on the individual level are largely ignored. Direction dependence tests require so-called *individual case residuals* (ICR; Bollen & Arminger, 1991; Raykov et al., 2013; Raykov & Penev, 2014). Future studies are needed to analyze the performance of ICR-based direction dependence tests under various model specifications. Further, von Eye and Wiedermann (2014) demonstrate the application of observed variable-based and residuals-based direction dependence approaches in the context of latent variable models. Here, instead of manifest variables, component scores, factor scores, or, more generally, latent variable scores can be used to analyze direction dependence properties of variables.

The fact that nonnormality of observed data cannot solely be understood as a source of bias for parametric statistical inference, but also as valuable information

for testing structural relationships in variables has been pointed out by other authors as well. For example, Bentler (1983) suggested that higher than second moments can be used to solve the problem of model equivalence of structural equation models. Mooijaart (1985) suggested the use of third higher moments to make decisions concerning factor rotation in factor analysis for nonnormal variables. Bentler's (1983) proposition was also later adopted by, for example, Shimizu, Hyvärinen, Hoyer, and Kano (2006) and Shimizu and Kano (2008) which led to the development of linear non-Gaussian acyclic models (LiNGAM) and non-normal structural equation models (nnSEM). LiNGAM and nnSEM constitute other promising approaches to address directionality issues in cross-sectional studies. In contrast to the proposed methodology, these models assume non-normal error terms to test directional hypotheses. However, Shimizu and Kano (2008) note that comparatively large sample sizes are needed for model estimation. Dodge and Rousson's (2001) as well as Sungur's (2005) results concerning the properties of higher moments in the linear regression setting further led to the development of copula regression approaches for directional inference (Kim & Kim, 2013). In contrast to Dodge and Rousson's (2001) direction dependence approach, copula regressions focus on the joint distribution of variables.

Finally, the proposed residual-based direction dependence methodology as well as all other procedures (stated above) which are designed to make directional statement have one implicit assumption in common, i.e., observed nonnormality reflects an inherent characteristic of the variable of interest and does not result for other reasons such as outliers, ceiling/floor effects, or mixtures of normally distributed subpopulations (for a discussion see Bauer & Curran, 2003; Cudeck & Henly, 2003; Muthén, 2003; Rindskopf, 2003). Several examples of phenomena exist for which non-normality is likely to be an inherent characteristic. These include latent periods of infectious diseases and survival times after cancer diagnosis (in medicine), the concentration of elements in the earth crust and their radioactivity (in geology), income (in economics), reaction times (in psychology), species abundance (in ecology), lengths of spoken words and sentences (in linguistics), air pollution measured using the Pollutant Standard Index (PIS, in environmental sciences), and, more generally, various ability distributions (for an overview see Limpert, Stahel, & Abbt 2001). However, even when nonnormality can be considered an inherent property of the phenomena of interest, a priori theory is necessary to make directional claims. Direction dependence analysis is not intended to replace substantial considerations. Rather, it serves as an element in the systematic and careful synthesis of conclusions from various types of studies (such as observational, longitudinal prospective, and experimental studies; Cox, 2012).

Appendix: Deriving the Test Statistic for the Asymptotic Kurtosis Difference Test

In the following paragraphs, we first derive the test statistic of the kurtosis difference test based on the observed variables X and Y. We then discuss its application based on residuals from competing regression models.

Let $z(\delta_X)$ and $z(\delta_Y)$ be the standard normally distributed z-values of the variables X and Y obtained from the Anscombe-Glynn transformation (Anscombe & Glynn, 1983). For independent samples,

$$z_d = \frac{z\left(\delta_X\right) - z\left(\delta_Y\right)}{\sqrt{2}} \tag{A1}$$

follows a standard normal distribution as well. However, residuals obtained from competing regression models will be correlated. For correlated samples, the variance of differences in kurtosis values can be written as $\sigma^2_{\delta_X-\delta_Y} = \sigma^2_{\delta_X} + \sigma^2_{\delta_Y} - 2cov\left(\delta_X, \delta_Y\right)$, where $\sigma^2_{\delta_X}$ and $\sigma^2_{\delta_Y}$ denote the variances of δ_X and δ_Y and $cov(\delta_X, \delta_Y)$ is the covariance of δ_X and δ_Y. The variances of standard normally distributed quantities (such as $z(\delta_X)$ and $z(\delta_Y)$) equal one and the correlation between fourth central moments of two variables can be approximated by the fourth power of the correlation between X and Y, $\rho_{\delta_X\delta_Y} = cov\left(\delta_X, \delta_Y\right) / \left(\sigma_{\delta_X}\sigma_{\delta_Y}\right) \approx \rho^4_{XY}$. For a more detailed discussion of moments of moments and moments of product moment coefficients see, for example, Pearson and Young (1918), Wishart (1928), Pepper (1929), and Rider (1929). Thus, the variance of differences in excess kurtosis values can, in this special case, be re-written as

$$\begin{aligned}
\sigma^2_{\delta_X-\delta_Y} &= \sigma^2_{\delta_X} + \sigma^2_{\delta_Y} - 2cov\left(\delta_X, \delta_Y\right) \\
&= \sigma^2_{\delta_X} + \sigma^2_{\delta_Y} - 2\rho^4_{XY}\sqrt{\sigma^2_{\delta_X}\sigma^2_{\delta_Y}} \\
&= 2 - 2\rho^4_{XY}
\end{aligned} \tag{A2}$$

Inserting Eq. (A2) into Eq. (A1) results in

$$z_d = \frac{z\left(\delta_X\right) - z\left(\delta_Y\right)}{\sqrt{2 - 2\rho^4_{XY}}}. \tag{A3}$$

Finally, replacing X and Y by the residuals from competing regression models one obtains

$$z_\delta = \frac{z\left(\delta_{\varepsilon_X}\right) - z\left(\delta_{\varepsilon_Y}\right)}{\sqrt{2 - 2\rho^4_{\varepsilon_X\varepsilon_Y}}}. \tag{A4}$$

References

Ali, M. M. (1974). Stochastic ordering and kurtosis measure. *Journal of the American Statistical Association, 69*, 543–545.

Anscombe, F. J., & Glynn, W. J. (1983). Distribution of the kurtosis statistics b_2 for normal samples. *Biometrika, 70*, 227–234.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*, 338–363.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural equation modes: Specification and estimation of moment structures. *Psychometrika, 48*, 493–517.

Bogat, G. A., Levendosky, A. A., DeJonghe, E., Davidson, W. S., & von Eye, A. (2004). Pathways of suffering: The temporal effects of domestic violence on women's mental health. *Maltrattamento e abuso all'infanzia, 6*, 97–112.

Bogat, G. A., Levendosky, A. A., Theran, S., Von Eye, A., & Davidson, W. S. (2003). Predicting the psychosocial effects of interpersonal partner violence (ipv): How much does a woman's history of ipv matter? *Journal of Interpersonal Violence, 18*, 1271–1291.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology, 21*, 235–262.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods, 8*, 378–383.

Cox, D. R. (2012). Statistical causality: Some historical remarks. In C. Bezuini, P. Dawid, & L. Bernardinelli (Eds.), *Causality: Statistical perspectives and applications* (pp. 1–5). Chichester, West Sussex: Wiley.

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g_1. *Biometrika, 57*, 679–681.

Darlington, R. B. (1970). Is kurtosis really "peakedness"? *American Statistician, 24*, 19–22.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*, 186–205.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292–307.

Dodge, Y., & Rousson, V. (2000). Direction dependence in a regression line. *Communications in Statistics: Theory and Methods, 29*, 1957–1972.

Dodge, Y., & Rousson, V. (2001). On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician, 55*, 51–54.

Dodge, Y., & Yadegari, I. (2010). On direction of dependence. *Metrika, 72*, 139–150.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.

Katz, D. L., Elmore, J. G., Wild, D. M. G., & Lucan, S. C. (2013). *Jekel's epidemiology, biostatistics and preventive medicine* (4th ed.). Philadelphia: Elsevier.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics: Inference and relationship* (2nd ed.). London: Charles Griffin & Company.

Kim, D., & Kim, J. M. (2013). Analysis of directional dependence using asymmetric copula-based regression models. *Journal of Statistical Computation and Simulation, 84*, 1990–2010.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics, 17*, 107–112.

Lee, C. F., Lee, J. C., & Lee, A. C. (2013). *Statistics for business and financial economics* (3rd ed.). New York: Springer.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience, 51*, 341–352.

Livesay, J. H. (2007). Kurtosis provides a good omnibus test for outliers in small samples. *Clinical Biochemistry, 40*, 1032–1036.

Muddapur, M. V. (2003). On directional dependence in a regression line. *Communications in Statistics: Theory and Methods, 32*, 2053–2057.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran (2003). *Psychological Methods, 8*, 369–377.

Mooijaart, A. (1985). Factor analysis for non-normal variables. *Psychometrika, 50*, 323–342.

Pearson, K. (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. *Skew variation, a rejoinder. Biometrika, 4*(1-2), 169–212.

Pepper, J. (1929). Studies in the theory of sampling. *Biometrika, 21*, 231–258.

Pornprasertmanit, S., & Little, T. D. (2012). Determining directional dependency in causal associations. *International Journal of Behavioral Development, 36*, 313–322.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/

Raykov, T., & Penev, S. (2014). Latent growth curve model selection: The potential of individual case residuals. *Structural Equation Modeling, 21*, 20–30.

Raykov, T., Lee, C. L., Marcoulides, G. A., & Chang, C. (2013). A commentary on the relationship between model fit and saturated path models in structural equation modeling applications. *Educational and Psychological Measurement, 73*, 1054–1068.

Rider, P. R. (1929). Moments of moments. *Proceedings of the National Academy of Sciences, 15*, 430–434.

Rindskopf, D. (2003). Mixture or homogenous? Comment on Bauer and Curran (2003). *Psychological Methods, 8*, 364–368.

Rochon, J., & Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *British Journal of Mathematical and Statistical Psychology, 64*, 410–426.

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology, 12,* 81. Retrieved from www.biomedcentral.com/1471-2288/12/81

Ruppert, D. (1987). What is kurtosis? An influence function approach. *American Statistician, 41*, 1–5.

Saunders, D. G. (1994). Posttraumatic stress symptom profiles of battered women: A comparison of survivors in two settings. *Violence and Victims, 9*, 31–44.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research, 33*, 65–117.

Schucany, W. R., & Ng, H. K. T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student t. *Communications in Statistics: Theory and Methods, 35*, 2275–2286.

Shimizu, S., Hyvärinen, A., Hoyer, P. O., & Kano, Y. (2006). Finding a causal ordering via independent component analysis. *Computational Statistics & Data Analysis, 50*, 3278–3293.

Shimizu, S., & Kano, Y. (2008). Use of non-normality in structural equation modeling: Application to direction of causation. *Journal of Statistical Planning and Inference, 138*, 3483–3491.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., et al. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research, 12*, 1225–1248.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). Cambridge: MIT Press.

Sungur, E. A. (2005). A note on directional dependence in regression setting. *Communications in Statistics: Theory and Methods, 34*, 1957–1965.

Teuscher, F., & Guiard, V. (1995). Sharp inequalities between skewness and kurtosis for unimodal distributions. *Statistics & Probability Letters, 22*, 257–260.

van Staden, P. J., & Loots, M. T. (2009). *Teaching the concept of kurtosis in introductory statistics courses using Mathematica: Searching for platypuses and kangaroos beneath the cloth of table mountain*. Paper presented at the 7th southern right delta conference 2009.

von Eye, A., & DeShon, R. P. (2008). Characteristics of measures of directional dependence—A Monte Carlo study. *Interstat*. Retrieved March 13, 2013, from http://interstat.statjournals.net/YEAR/2008/articles/0802002.pdf

von Eye, A., & DeShon, R. P. (2012). Directional dependence in developmental research. *International Journal of Behavioral Development, 36*, 303–312.

von Eye, A., & Wiedermann, W. (2014). Direction of dependence in the latent variable context. *Educational and Psychological Measurement, 74*, 5–30.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *American Statistician, 68*, 191–195.

Wiedermann, W., Hagmann, M., Kossmeier, M., & von Eye, A. (2013). Resampling techniques to determine direction of effects in linear regression models. *Interstat*. Retrieved March 13, 2013, from http://interstat.statjournals.net/YEAR/2013/articles/1305002.pdf

Wiedermann, W., & Hagmann, M. (2014). Asymmetric properties of the Pearson correlation coefficient: Correlation as the negative association between linear regression residuals. *Communications in Statistics: Theory and Methods* (in press).

Wiedermann, W., Hagmann, M., & von Eye, A. (2015). Significance tests to determine the direction of effects in linear regression models. *British Journal of Mathematical and Statistical Psychology, 68*(1), 116–141.

Wiedermann, W., & von Eye, A. (2015a). Direction of effects in multiple linear regression models. *Multivariate Behavioral Research, 50*, 23–40.

Wiedermann, W., & von Eye, A. (2015b). Direction-dependence analysis: A confirmatory approach for testing directional theories. International Journal of Behavioral Development. DOI: 10.1177/0165025415582056

Wiedermann, W., & von Eye, A. (2015c). Direction of effects in mediation analysis. *Psychological Methods, 20*(2), 221–244.

Wishart, J. (1928). The generalised product moment distribution in samples from normal multivariate populations. *Biometrika, 20*, 32–52.

Yap, B. W., & Sim, C. H. (2011). Comparison of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*, 2141–2155.

Part III
Dyadic Data Modeling

Analyzing Dyadic Data with IRT Models

Rainer W. Alexandrowicz

Abstract Dyadic data frequently occur in social sciences and numerous techniques have been developed for their analysis. The most prominent methods involve using regression, path, and structural equation models. The present contribution extends these approaches by considering Item Response Theory (IRT) Models. Two pivotal dyadic data analysis models, the Actor-Partner Interdependence Model (APIM) and the Common Fate Model (CFM), are built using the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM). This approach combines the advantages of dyadic data analysis with a model for discrete data, thus allowing for categorical items while drawing inferences based on the estimated true scores on an interval scale.

Aims of This Contribution

This contribution presents a new approach to dyadic data analysis. It is organized as follows: After giving a short introduction to the basics of dyadic data (section "Dyadic Data") and the core principles of their statistical analysis (section "Modeling Dyadic Data"), the fundamentals of a new approach based on multidimensional Item Response Models (mIRT; e.g. Reckase 2009) are worked out. This approach combines the specific requirements of dyadic data analysis (i.e., taking into account the dependencies within a dyad) with the advantages and flexibility of discrete probability models for categorical data. The principles of mIRT will be introduced in section "Item Response Models" and exemplified for two important dyadic data models, including computational details, in section "Worked Examples".

R.W. Alexandrowicz (✉)
Applied Psychology and Methods Research Department, Institute for Psychology, Alps-Adria-University Klagenfurt, Universitaetsstr. 65, 9020 Klagenfurt, Austria
e-mail: rainer.alexandrowicz@aau.at

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_8

Dyadic Data

Dyadic data originate from responses of individuals sharing a common context, like partner or parent-child relationships. Apart from such natural (or voluntary) linkage, a dyad can also be established by membership to a common context, like in an experimental design, where two previously unacquainted individuals are assigned to each other to work on a common task. We cannot assume the responses of linked observations (e.g., parent and child) to be mutually independent. We have to act on the assumption that systematic variation arises due to both, individual and relational characteristics. Four kinds of nonindependence may be discerned: compositional nonindependence (the dyad members are linked due to preexisting common characteristics), partner effects (characteristics or behaviors of one partner necessarily affects those of the other partner, e.g. when resources have to be shared), mutual influence (due to some sort of feedback loop), and common fate (both dyad members are affected by common circumstances, like sharing the household or consanguinity, for example).

Another crucial distinction has to be made with regard to the identifiability of the members of a dyad (pair) under consideration: While, for example, the roles of parent and child allow for a clear distinction of individuals, monozygotic twins may not be uniquely allocated unless auxiliary variables are taken into account (e.g., the elder vs. the younger sibling). Hence, we have to differentiate between dyadic data models for distinguishable and indistinguishable members.

Further, we have to consider whether information is gained at the individual or at the dyad level: A dyad member's gender is usually a descriptor of the individual (except for studies deliberately focussing on same sex pairs, etc.), but the household income of a couple is identical for both members and therefore constitutes a dyad level variable. As a third category, we have to consider mixed variables, exhibiting variation on both the individual and the dyad level, like the respondents' age.

A comprehensive overview of dyadic data, models, analyses, and numerous references to original sources can be found in Kenny, Kashy, and Cook (2006).

Modeling Dyadic Data

The term "model" refers in the context of dyadic data to a substantive perspective, i.e. how measurements from dyad members are hypothesized to relate to each other, and will not necessarily determine the statistical model to be used for parameter estimation. It might, therefore, be helpful to differentiate between "dyadic models," focussing on substantive theory, and "statistical models." This distinction is not always clear-cut, because some dyadic models may correspond closely to a certain statistical model.

Several models for dyadic data have been proposed so far, two of which are outlined in this section, as they are common in the literature, and they are further pursued in the analyses presented here. These are the Actor-Partner Interdependence Model (APIM) and the Common Fate Model (CFM).

The Actor-Partner Interdependence Model

Basically, the APIM (Kenny et al. 2006, ch. 7) constitutes a regression model involving an independent variable X and a dependent variable Y, both available for both members of a dyad (A and B). Hence, we deal with four variables (or constructs, if more than one item is involved), X_A, X_B, Y_A, and Y_B. The regressions of the Y-variables on the X-variables are separately modeled for each dyad member (called *actor effects* in the dyadic context, a_{YX} and a'_{YX} in Fig. 1, top). In addition, each member's X may affect the other member's Y (called *partner effects*, p_{YX} and p'_{YX}). The magnitude of the partner effects relative to the actor effects expresses the extent of interdependence of dyad members. Furthermore, the two independent variables or constructs as well as the two dependent ones may exhibit a mutual relation (r_X and r_Y in Fig. 1, top).

If each of the four constructs involved (X_A, X_B, Y_A, and Y_B) is a single random variable fulfilling certain scale and distributional assumptions, the coefficients could be determined by means of Ordinary Least Squares (OLS) regression or path analysis (a comprehensive instruction can be found in Kenny et al. (2006), for example). However, such an approach ignores the measurement errors of the observed variables and becomes increasingly cumbersome for dyadic models that are more complex than those considered here.

The nesting of individuals within dyads constitutes a hierarchical data structure, which facilitates the application of multilevel models (cf. Hox 2010; for their specific application to dyadic data, see Campbell & Kashy 2002 or Kenny et al. 2006, ch. 4). Alternatively, the coefficients could be estimated by means of Structural Equation Models (SEM; e.g., Bollen 1989). In particular the SEM-approach allows for a sophisticated and flexible modeling of the hypothesized relationships and provides a highly differentiated assessment of model fit.

The Common Fate Model

The CFM (Campbell 1958; Kenny et al. 2006, pp. 409–412) also considers the relationship of two variables (X and Y) recorded for both members of a dyad (A and B). But instead of looking for mutual dyad members' influence (the partner effects in the APIM), we focus on the correlation of X and Y, assuming that they constitute a common background ("fate," hence the naming) for the individual expressions (X_A, X_B and Y_A, Y_B, respectively; cf. Fig. 1, bottom). As a special case, the CFM is

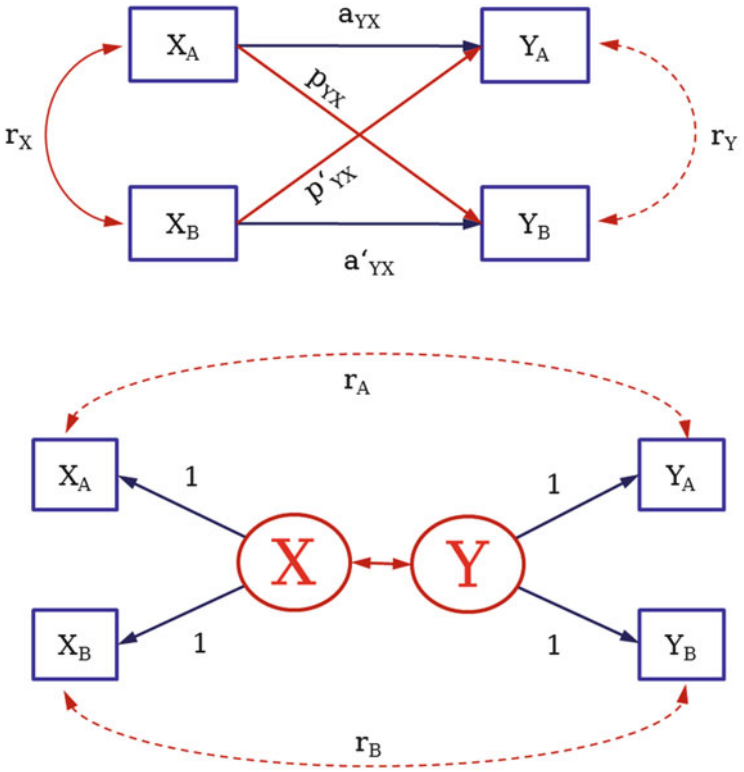


Fig. 1 *Top*: The Actor-Partner Interdependence Model (APIM). *Bottom*: The Common Fate Model (CFM). Notes: X: independent variable; Y: dependent variable; A, B: dyad members (e.g., actor/partner); a_{YX}, a'_{YX}: actor effects; p_{YX}, p'_{YX}: partner effects; r_X, r_Y: correlation of independent and dependent variables; r_A, r_B: correlation of X and Y for individuals A and B, respectively

particularly appropriate for designs, in which A and B express their assessment of a third person. This may be the case, for example, when both parents rate X and Y of their child, or a couple is asked to assess two characteristics X and Y of their marriage counselor.

The core principle of a CFM is that both X_A and X_B are affected by a latent variable X and, likewise, Y_A and Y_B can be traced back to a common latent variable Y. The latent correlation of X and Y reflects the substantial question of interest on the dyad-level. However, the two common fate constructs (X and Y) may not account for the entire observed covariance of the manifest variables X_A, X_B, Y_A, and Y_B, as individual characteristics could have an impact as well. Such individual level effects are expressed by the correlation coefficients r_A and r_B as indicated at the bottom of Fig. 1.

Because of the assumption of latent factors that underlie the manifest variables, the SEM approach is the most suitable technique for CFMs. Besides, there are also methods available not involving latent constructs, but based on simple regression

analysis (for an introduction, see Kenny et al. 2006, pp. 409–412). However, these should be considered outdated for the same reasons as in the APIM context.

Problem

SEM and Multi-Level Models prevail in current modeling approaches to dyadic data analysis. Either of these models requires certain scale and distributional prerequisites to be fulfilled—most prominently, in the standard case, interval scaled variables and (multivariate) normal distribution. Such assumptions are frequently made without hesitation. For example, Kenny et al. (2006) argue "Most scales developed and used in social science research are assumed to be measured on an interval scale" (p. 9), and "Throughout this book, we generally assume that outcome variables are measured at the interval level." (Kenny et al. 2006, p. 10).

In many cases, constructs are captured with scales comprising a reasonable number of items to be endorsed through ordered categories or by responding in a simple yes/no style. If such a set of items has undergone thorough statistical analysis, a (possibly weighted) sum of scores might fulfill the aforementioned scale assumptions—at the price of restricting the number of applicable instruments to those having been scrutinized accordingly. Sometimes, a sum score is even computed without bothering about properties of the involved items—dimensionality and scale assumptions remain conjectures then.

Loeys and Molenberghs (2013) have proposed Generalized Linear Mixed Models for dyadic data analysis with categorical data and Loeys, Cook, De Smet, Wietzker, and Buysse (2014) used Generalized Estimating Equations. McMahon, Puget, and Tortu (2006) have shown how to model binary data employing a Multi-Level approach. Furthermore, Log-Linear Models may be applied as well (e.g., von Eye & Mun 2013; see Kenny et al. 2006, pp. 131–135 for their specific application to dyadic designs). Log-Linear Models also allow for testing interaction effects and the assessment of model fit. However, we will take a slightly different approach to categorical data here, using Item Response Theory (IRT; de Ayala 2009; van der Linden & Hambleton 1997; Lord 1980).

Item Response Models

Before delving into the details of how dyadic data models may be expressed through IRT models, a brief introduction to IRT reviews some basic characteristics. Generally, Item Response Models link manifest responses to latent response probabilities,

expressed by model parameters, using a deliberately selected link function. Usually, the manifest responses are categorical, hence we deal with discrete probability models (an extension to quantitative responses has been developed by Müller 1987 but will not be further pursued here, as our intention is to model discrete data).

The Rasch Model and Some of Its Extensions

The most basic IRT model is the Rasch Model for dichotomous data (RM; Rasch 1960). It models the probability of a response $x_{vi} \in \{0, 1\}$ of individual v ($v = 1 \ldots n$) to item i ($i = 1 \ldots k$) with two parameters, θ_v, quantifying the ability of person v, and δ_i, quantifying the difficulty of item i. The link function is the logistic one. It yields the model equation

$$P(x_{vi}|\theta_v, \delta_i) = \frac{e^{x_{vi}(\theta_v - \delta_i)}}{1 + e^{\theta_v - \delta_i}}. \tag{1}$$

Note that a "positive" manifest response $x_{vi} = 1$ may represent the solution of an item during an ability test or the endorsement of a statement during a personality assessment. Hence the traditional term "ability" may also be understood as "proneness" (in the sense of "disposedness") to endorse a statement and "difficulty" as the "severity" or "particularity" of that statement.

The trace lines (or Item Characteristic Curves, ICC) of function (1) for selected δ_i across an arbitrary range of θ_v are parallel, which constitutes a distinct feature of the RM. The unweighted sum of scores x_{vi} per row v and per column i are the sufficient statistics for the person ability parameters θ_v and item difficulty parameters δ_i, respectively. Either parameter vector, $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, can be estimated independently of the distribution of the other one, which caused Georg Rasch to develop his infamous principle of Specific Objectivity (SO; Rasch 1961 1966, 1977). One decisive advantage of SO is that it allows for a rigorous assessment of model fit (for an overview, see Glas & Verhelst 1995, for example). If the model holds, all items measure the same latent trait (unidimensionality assumption).

Numerous extensions have been developed. For ordered polytomous data, the Eq. (1) is adopted to model the thresholds between adjacent categories while retaining all advantageous features inherited from the RM. Applying various restrictions, this yields the Partial Credit Model (PCM; Masters 1982; Wright & Stone 1979) or the Rating Scale Model (RSM; Andrich 1978 1982; Wright and Masters 1982). If, on the other hand, substantial considerations allow for decomposing the items into a well-defined set of p basic (cognitive) operations or technical features, their difficulty can be quantified by means of the Linear Logistic Test Model (LLTM; Fischer 1973 1995). For that purpose, a $k \times p$ weight (or design) matrix $\mathbf{A}$ is set up based on substantial theory, linking each item parameter δ_i to a hypothesized set of basic parameters ξ_j ($j = 1 \ldots p$; $p \leq k$), which represent cognitive operations or technical features.

Furthermore, additional item-specific parameters have been introduced (at the expense of losing SO). These parameters relax the rather restrictive assumption of parallel trace lines, which may be difficult to attain, especially when analyzing large item pools. A *discrimination parameter* α_i for each item (Birnbaum 1968) expresses the slope of an item's ICC at its inflection point along the θ-scale. It serves, roughly speaking, as a measure, of the degree to which the distinction (or "discrimination," hence the naming) between two individuals is clear-cut by this item. Thus, the trace lines' slopes are explicitly modelled rather than assumed parallel. In addition, an item specific *guessing parameter* may be employed, quantifying the probability of a positive response for arbitrary small values of the person ability parameters (technically, it defines the lower asymptote of an item's ICC; Birnbaum 1968).

A third line of development introduced a third kind of parameter for designs, where individuals (represented by the person ability parameter θ_v) respond to items (represented by the item difficulty parameters δ_i), and their responses are evaluated by raters. A rater's (r) leniency may be quantified by means of a rater parameter ψ_r (for details, see Linacre 1989).

Another important extension, crucial for modeling dyadic data, is introduced in the following section.

Multidimensional IRT Models

All IRT models sketched in section "The Rasch Model and Some of Its Extensions" share the assumption of unidimensionality, i.e. one single common latent trait being required for solving all items (or endorsing the respective statements) under consideration. In contrast, a set of items may also depend on several distinct latent traits, in fact in two ways: Either an item involves more than one trait (e.g., a math item embedded in a very complex instruction might require a certain amount of both language and math skills); such a case is referred to as *within item multidimensionality*. Or, one subset of items goes together with a latent trait θ_1 and a different subset of items with another latent trait θ_2; this case is called *between item multidimensionality*. In our application, we will refer to the latter case. The allocation of the $i = 1 \ldots k$ items to $\ell = 1 \ldots m$ latent traits is specified in scoring matrix $\mathbf{B} = (b_{i\ell})$, which is—as $\mathbf{A}$ before—set up based on theoretical reasoning prior to parameter estimation. As a consequence, each individual's ability *profile* (i.e., his or her location on each latent trait) is expressed through an individual's ability *vector* $\boldsymbol{\theta}_v = (\theta_{v\ell})$ of length m.

Now, combining such a person parameter decomposition with the item parameter decomposition sensu LLTM, as introduced in section "Item Response Models", leads to the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Adams, Wilson, & Wang 1997; Adams & Wu 2007). It follows the logistic structure of Model (1), in which both parameters are replaced by a product of a weight matrix (i.e., the scoring matrix $\mathbf{B}$ for person parameters and the design matrix $\mathbf{A}$ for item parameters), and the respective item and person parameter vector.

The model equation for an individual's response vector $\boldsymbol{x}_{vi}$ is

$$P(\mathbf{x}_v|\mathbf{A},\mathbf{B},\boldsymbol{\theta}_v,\boldsymbol{\xi}_i) = \frac{\exp\left[\mathbf{x}_v'(\mathbf{B}\boldsymbol{\theta}_v + \mathbf{A}\boldsymbol{\xi}_i)\right]}{\sum_{\mathbf{z}\in\Omega}\exp\left[\mathbf{z}'(\mathbf{B}\boldsymbol{\theta}_v + \mathbf{A}\boldsymbol{\xi}_i)\right]}, \tag{2}$$

where

$\mathbf{x}_v$ ... response vector of individual v

$\boldsymbol{\Omega}$ ... set of all possible response vectors

$\boldsymbol{\theta}_v$ ... vector of individuals$'$ parameters (ability profile)

$\boldsymbol{\xi}_i$ ... vector of items$'$ basic parameters

$\mathbf{B}$ ... scoring matrix

$\mathbf{A}$ ... design matrix.

Parameter estimation is usually accomplished with the Marginal Maximum Likelihood technique (cf. Baker & Kim 2004). This technique requires a distributional assumption regarding the person parameters. It is common practice to choose the normal, yielding for the unidimensional case

$$f(\theta|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2}. \tag{3}$$

Moreover, the MRCMLM allows for defining a correlational structure among the latent variables or regressing the latent variables onto each other and on a set of background variables Y, e.g., income or educational information. The latter leads to the so-called *background model*, which, in multivariate notation, shows as

$$\boldsymbol{\theta} = \mathbf{Z}'\boldsymbol{\gamma} + \epsilon, \tag{4}$$

with $\boldsymbol{\gamma}$ expressing the regression weights of $\boldsymbol{\theta}$ on $\mathbf{Z}$ and assuming $\epsilon \sim N(0;\sigma_\epsilon^2)$. Incorporating the background model (4) in the multivariate extension of (3) yields

$$f(\boldsymbol{\theta}|\mathbf{Z},\boldsymbol{\Gamma},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{m}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\ \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\Gamma}\mathbf{Z})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\Gamma}\mathbf{Z})} \tag{5}$$

with covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\theta_1} & & & & & & \\ \sigma_{\theta_2\theta_1} & \sigma^2_{\theta_2} & & & & & \\ \sigma_{\theta_3\theta_1} & & \sigma^2_{\theta_3} & & & & \\ \vdots & & & \ddots & & & \\ \sigma_{\theta_\ell\theta_1} & & & & \sigma^2_{\theta_\ell} & & \\ \vdots & & & & & \ddots & \\ \sigma_{\theta_d\theta_1} & \dots & \dots & \dots & \dots & \dots & \sigma^2_{\theta_m} \end{pmatrix}. \tag{6}$$

This covariance matrix (more precisely, its estimates) will prove eminently useful for the task of expressing dyadic models in terms of item response models. These coefficients allow for expressing correlations among the latent constructs. Moreover, we may also estimate directed relationships among the latent constructs, allowing for a SEM-like modeling approach, yet on a solid Rasch foundation. The core idea is to obtain the correct SSCP matrix of all exogenous ("independend") and endogenous ("dependent") variables and to find the desired regression coefficients by means of the Two-Stage Least Squares (TSLS) estimation approach (Gebhardt, in prep; for a delightful description of the TSLS history see Stock & Trebbi, 2003).

Expressing Dyadic Data Models in Terms of Item Response Models

This section outlines the central principle of how dyadic data models may be formulated in terms of IRT Models. We will consider two important dyadic models, the APIM and the CFM. In the graphical representations, boxes represent manifest variables (which are, in our case, categorical) and ellipses represent latent constructs. The core principles of all models to be introduced are to (1) assume a separate latent trait for each of the "dyadic variables" (i.e., the X_A, etc. in Fig. 1) and (2) model the postulated dyadic relationships in the latent domain, thus requiring a multidimensional model like the MRCMLM.

The APIM in Terms of an MRCMLM

Dyadic models as depicted in Fig. 1 assume relations among the X- and Y-measures of the dyad members A and B. While regression or path analysis assumes these constructs (i.e., X_A, Y_A, X_B, and Y_B) to be manifest, we may also model each of them as a separate latent construct. Of course, this could be achieved with a SEM

as well, but while the SEM (in its standard case) assumes the items' responses to lie on an interval scale, we want to model truly categorical responses (like "agree"—"partially agree"—"disagree") with a discrete probability model, such as the MRCMLM. Further, standard SEM applications use a linear link function of items and latent variables (although modifications for categorical variables have been developed as well, cf. Muthén 1984).

Each dyad (i.e., the pair A and B) forms a unit of observation v (usually a row in the data set). Hence the measures X_A, Y_A, X_B, and Y_B may be conceived as four latent dimensions of the dyad v and comply to one θ_ℓ of the MRCMLM as expressed in Eq. (2). We thus assert four latent constructs θ_1 to θ_4, constituting the four measures of interest (i.e., $\theta_1 = X_A$, and so on). Such a structure can be depicted as shown in Fig. 2. The double-headed arrows are based on the latent covariances [i.e., the elements of Matrix (6)] between the four constructs. Furthermore, the MRCMLM also allows for estimating a regression model of the latent constructs on the background variables [Eq. (4)] and on each other. The latter will be used to model the *directed* relationships as postulated in the APIM (and depicted by single-headed arrows in Fig. 1, top).

The CFM in Terms of an MRCMLM

Defining the CFM in terms of an MRCMLM is straightforward, as it already involves two latent constructs, θ_X and θ_Y, representing the variables of interest (cf. Fig. 3). The latent factor θ_X (representing X in Fig. 1, bottom) affects both dyad members' observed values, X_A and X_B, and, therefore, represents the common background (fate) of A and B. The same applies to the other latent variable of interest, θ_Y. The latent correlation $r_{\theta_X\theta_Y}$ is the central measure of interest. It

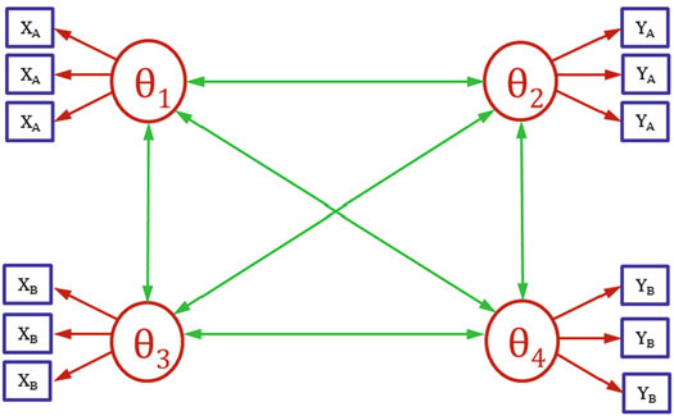


Fig. 2 The basic structure of an APIM expressed as an MRCMLM

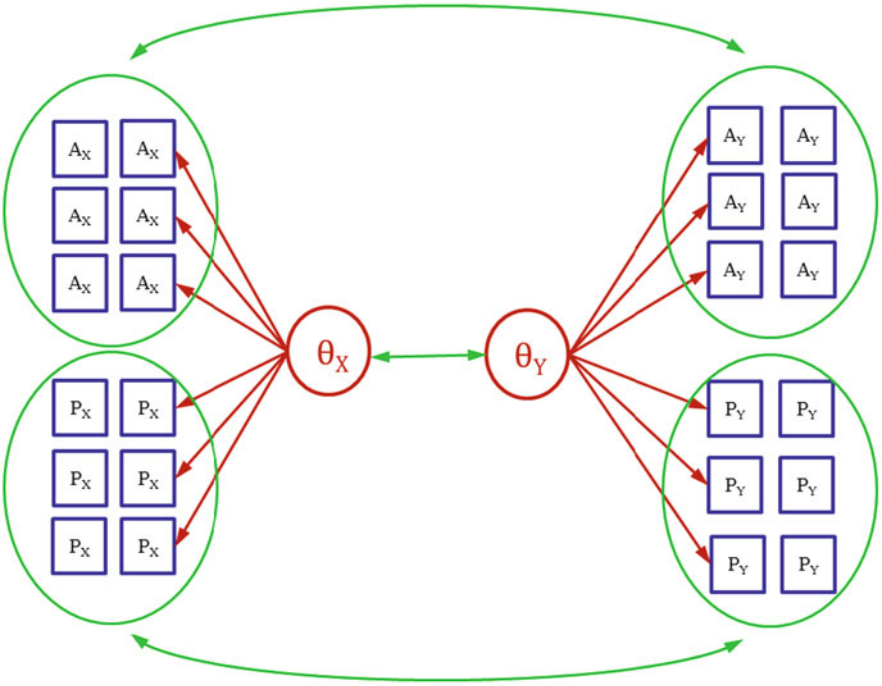


Fig. 3 The basic structure of a CFM expressed as an MRCMLM

expresses the association between θ_X (the latent representation of X) and θ_Y (the latent representation of Y) after taking the relationship between the actors into account. It is indicated with a double headed solid arrow in Fig. 3.

Not all covariation of X and Y variables may be explained by the dyad level correlation $r_{\theta_X\theta_Y}$, hence we further explore individual level correlations (termed r_A and r_B, respectively, indicated with dotted double headed arrows in Fig. 1). The calculation of these two coefficients requires additional reasoning. The latent factor θ_X reflects the common self-perception regarding the trait under consideration. Analoguously, the latent factor θ_Y reflects what is common in the other's perception. However, these two latent factors could miss certain aspects of one's self- or other's perception. Such omitted information is collected in the residuals, which will be used to determine the individual correlation coefficients r_A and r_B (see section "The CFM Approach").

Worked Examples

The proposed modeling approach shall be demonstrated in a psychological study focussing on selected personality traits of dyads of students and a parent. The question was whether the assessment of the respective other is influenced by the respective member's self-assessment.

The Study Framework

The theoretical scope of the study allows for different questions to be addressed. We will apply both the APIM and the CFM approaches with one data set, the design of which is outlined below. The respective research questions will be explained in the specific context of the model.

Instrument

The Gießen-Test (GT; Beckmann, Bräahler, & Richter 1990) is a self-assessment consisting of the following six scales (German original terms in brackets):

- social resonance (soziale Resonanz),
- dominance (Dominanz),
- control (Kontrolle),
- prevailing mood (Grundstimmung),
- responsiveness (Durchlässigkeit), and
- social power (soziale Potenz).

The test consists of 40 bipolar items and respondents have to indicate their preference on a 7-point scale of the form

I am rather (A) 3 2 1 0 1 2 3 rather (B)

with (A) and (B) representing opposite characteristics of a person. According to the manual, the construction of the GT involved exploratory factor analyses, resulting in six items per scale. The GT is capable of dyadic assessment because it can be employed for both self and partner assessment. For that purpose, three different forms of the questionnaire are available. For the self-assessment, the questions are formulated as

I think, I am rather patient 3 2 1 0 1 2 3 rather impatient.

The male/female partner assessment versions are worded

`I think, he/she is rather patient 3 2 1 0 1 2 3rather impatient.`

That way, four different versions exist and partnership assessment becomes feasible. Actor-self, Partner-self, Actor with respect to Partner, and Partner w.r.t Actor. Score sheets allow for comparing the four profiles. Note that there is also a dedicated partnership assessment version of the GT available, which uses only 5 out of the 6 scales. This version was not applied here, because it involves some indistinct scoring constants, not necessary for the present analysis.

Sample

The data set used for the present study has been simulated in a way that it reflects the characteristics of a smaller data set of psychology students. Hence, we will not draw substantial conclusions from the results obtained here. The students (first and third semester) were asked to fill out the questionnaire with respect to themselves and to a parent (preferably the mother). A total of 600 pairs has been simulated.

Method

Parameter estimation of the MRCMLM has been performed with the `ConQuest 3.0` software package (Adams, Wu, & Wilson 2012). To avoid estimation problems, the responses were dichotomized at the midpoint of the response scale (left vs. right direction). The online version of the instrument used in the present study comprised only six response categories per item (leaving out the middle category), thus fostering dichotomization.

The APIM Approach

Regarding the six personality traits of the GT, one could conceive of the following situation: A student's rating of the respective parent (θ_2) depends on the parent's status regarding that trait, reflected in their self ratings (θ_3). Therefore, we expect a strong coefficient p'_{YX} (see Fig. 4). In addition, the student's assessment might as well be influenced by his or her own *Selbstbild*, i.e. the way he or she perceives him- or herself with regard to the respective trait. For example, Sigmund Freud has keyed the term *Projektion* (projection), which describes, simply put, one's proneness to perceive one's own conflict-ridden, denied, or repressed emotions in others rather than in oneself (cf. Freud 1976; for more recent approaches see, for example, Baumeister, Dale, & Sommer 1998). Such a tendency would, if common, show in the regression coefficient a_{YX}, i.e. the actor effect in APIM terminology:

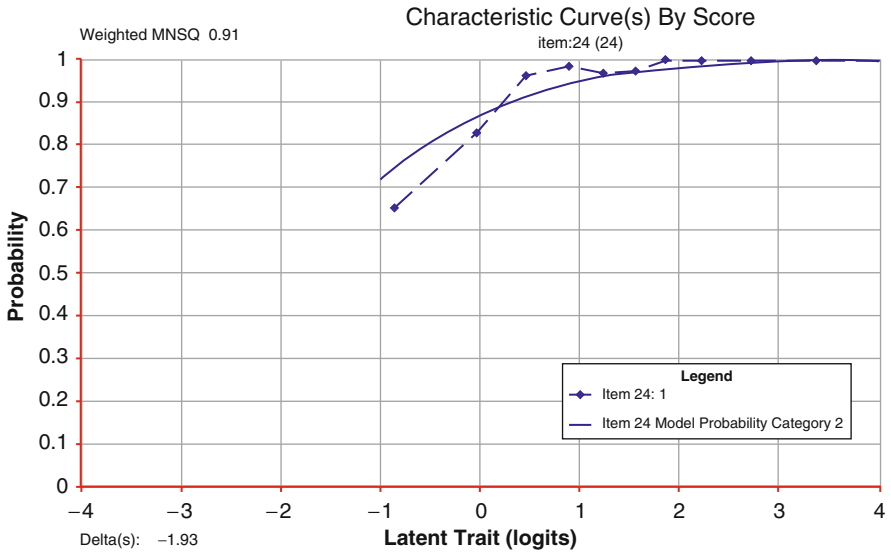


Fig. 4 Theoretical vs. empirical item characteristic curve (example plot for item 24). The dashed line represents the empirical ICC and the solid line the model derived ICC. Closeness of the two curves indicates a good model fit

The lower the student's self-rating, the higher his or her parent's rating on that scale, hence a substantial negative regression coefficient would arise. Analoguously, such an effect might as well appear in the parent's rating: His or her perception of the student (θ_4) would primarily depend on the student's trait, which should be expressed in the student's self-rating (θ_1), hence we expect a strong path p_{YX}. But the parent's *Selbstbild* might also influence this assessment—for example, because a parent might feel responsible for the offspring's development. Hence, a non-zero path a'_{YX} might appear as well. Finally, we have to consider that parents and children are prone to be similar with respect to personality traits as measured by the GT, reflected by the correlation $r_{\theta_1\theta_3}$ and which has to be corrected for.

We might therefore expect two strong (in terms of APIM) partner effects (p and p') as well as possible (but presumably weaker) actor effects (a and a'), reflecting the raters' involvement, like *Projektion*, for example. Moreover, a non-zero correlation of the two independent variables or the two dependent variables might occur as well. For each of the six GT scales, a separate APIM was estimated. For reasons of saving space, we will present only the results of the Social Resonance subscale.

Model Setup

The ConQuest 3.0 software accepts input via a command file. The ASCII-data file was named gt_res.dat. It comprised 25 columns (one ID and six items per subscale times four versions in the dyadic framework) with the responses coded numerically (1 to 6); missing data were coded with 9. Listing 1 shows the ConQuest command file for the APIM (a line by line explanation of these commands is given in Table 1 in Appendix "APIM Commands").

Listing 1 ConQuest Command Script for the APIM (Note that each command has to be terminated with a semicolon)

```
1  datafile gt_res.dat;
2  format responses 1-24;
3  codes 0,1;
4  recode (1 2 3 4 5 6) (0 0 0 1 1 1);
5  score (0,1)  (0,1) ()    ()    ()    !items (1-6);
6  score (0,1)  ()    (0,1) ()    ()    !items (7-12);
7  score (0,1)  ()    ()    (0,1) ()    !items (13-18);
8  score (0,1)  ()    ()    ()    (0,1) !items (19-24);
9  model item;
10 estimate ! storage=RAM, nodes=5, stderr=quick;
11 show parameters!table=3;
12 show parameters!table=2;
13 show ! estimate=mle;
14
15 structural /Dimension_2 on Dimension_1 Dimension_3;
16 structural /Dimension_4 on Dimension_1 Dimension_3;
```

These commands are stored in a file (named gt_res.cqc) and executed with the Run > Run all command from the menu bar (GUI version) or via submit gt_res.cqc; in the command line version.

Results

After submitting the command script to the program, a detailed output listing is available. The portions of this output relevant for building the APIM and assessing model fit will be described here.

Building the APIM from the Output One central part of the ConQuest output concerning the APIM is given in Listing 2. This section is produced by the structural commands in lines 15 and 16 of the command script.

In this output section we find the regression coefficients for the APIM, i.e., the single-headed arrows p_{YX}, a_{YX}, p'_{YX}, and a'_{YX} according to Fig. 1, top. These coefficients can be found in the columns headed Gamma (lines 25–27 and 45–47 in Listing 2). Another essential part of the APIM, the correlation of the two independent variables, (r_X in Fig. 1, top) can be found in the output section headed CONDITIONAL COVARIANCE/CORRELATION MATRIX (Listing 3).

Listing 2 Essential ConQuest Output for the APIM (Part 1a: Regression Coefficients)

```
STRUCTURAL MODEL
================================================================================

MODEL: dimension_2 on dimension_1 dimension_3

 ENDOGENOUS VARIABLES:
      dimension_2               (latent)

 EXOGENOUS VARIABLES:
      dimension_1               (observed)
      dimension_3               (observed)


EQUATION 1
--------------------
EQ1 N= 600 df=597
EQ1 R Squared  = 0.23744
EQ1 Multiple R = 0.48728

EQ1 Dependent Variable:        dimension_2
EQ1 Independent Variable(s):
EQ1                                                    Gamma    Beta     SE
--------------------------------------------------------------------------------
EQ1   exogenous                 Constant              -0.80118          0.134
EQ1   exogenous                 dimension_1           -0.13523          0.051
EQ1   exogenous                 dimension_3            0.87219          0.064
================================================================================

STRUCTURAL MODEL
================================================================================

   (...lines skipped...)

EQUATION 1
--------------------
EQ1 N= 600 df=597
EQ1 R Squared  = 0.50025
EQ1 Multiple R = 0.70728

EQ1 Dependent Variable:        dimension_4
EQ1 Independent Variable(s):
EQ1                                                    Gamma    Beta     SE
--------------------------------------------------------------------------------
EQ1   exogenous                 Constant              -0.54384          0.066
EQ1   exogenous                 dimension_1            0.36862          0.025
EQ1   exogenous                 dimension_3            0.49398          0.031
================================================================================
```

Listing 3 Essential ConQuest Output for the APIM (Part 1b: Correlation Coefficients)

```
CONDITIONAL COVARIANCE/CORRELATION MATRIX

                                Dimension
              ---------------------------------------------
Dimension         1          2           3           4

Dimension_1               0.053       0.332       0.808
Dimension_2     0.022                 0.948       0.449
Dimension_3     0.235      0.478                  0.685
Dimension_4     0.542      0.214       0.569
------------------------------------------------------------
```

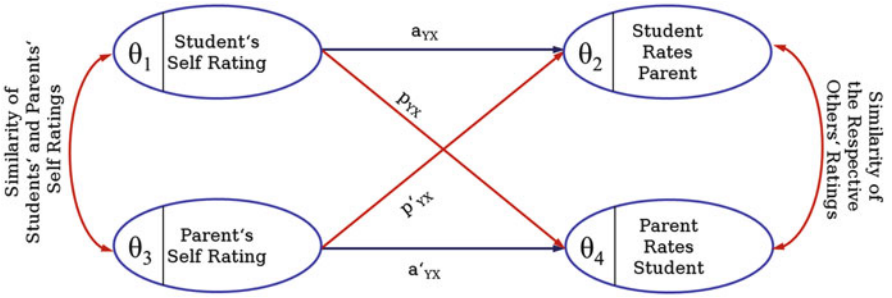


Fig. 5 The APIM approach of modeling other's assessment taking the *Selbstbild* into account

In this cut-out we find the estimated covariances (upper triangular matrix) and the correlations (lower triangular matrix) of the latent factors, i.e. the $\hat{\sigma}^2_{\theta_\ell\theta_{\ell'}}$ and the $\hat{r}_{\theta_\ell\theta_{\ell'}}$ for each pair θ_ℓ and $\theta_{\ell'}$. Hence, we may now draw the final diagram of the APIM (Fig. 5).

Each parameter estimate is of course accompanied by its respective standard error, facilitating the application of the Wald statistic. Listing 4 presents the example output for the regression models of our example.

Listing 4 Essential `ConQuest` Output for the APIM (Part 1c: Regression Coefficients)

```
EQUATION 1
--------------------
EQ1 N= 600 df=597
EQ1 R Squared  = 0.23744
EQ1 Multiple R = 0.48728

EQ1 Dependent Variable:        dimension_2
EQ1 Independent Variable(s):
EQ1                                                    Gamma      Beta     SE
------------------------------------------------------------------------------
EQ1   exogenous                 Constant              -0.80118        0.134
EQ1   exogenous                 dimension_1           -0.13523        0.051
EQ1   exogenous                 dimension_3            0.87219        0.064
==============================================================================

EQUATION 2
--------------------
EQ2 N= 600 df=597
EQ2 R Squared  = 0.50025
EQ2 Multiple R = 0.70728

EQ2 Dependent Variable:        dimension_4
EQ2 Independent Variable(s):
EQ2                                                    Gamma      Beta     SE
------------------------------------------------------------------------------
EQ2   exogenous                 Constant              -0.54384        0.066
EQ2   exogenous                 dimension_1            0.36862        0.025
EQ2   exogenous                 dimension_3            0.49398        0.031
==============================================================================
```

In order to obtain a test statistic for evaluating the null hypothesis that the parameter is zero, we have to divide the estimate by its standard error, yielding a standard normal variate. For example, to test the regression coefficient $\gamma_{\theta_2\theta_1}$ for

significance, we compute $-0.135/0.051 = -2.647$, the absolute value of which is larger than the 95 % quantile of the standard normal. Hence the coefficient is significantly different from zero (as are all coefficients of this model). However, because the sample size of the present data set has been fixed arbitrarily to 600, such a test is not informative in our case.

Interpretation We find two distinct partner effects ($p_{YX} = 0.463$ and $p'_{YX} = 0.833$). These results suggest effects of the personality of the target person (reflected in the students' and parents' self-description, θ_1 and θ_3) on the respective other's assessment. In contrast, the students' actor effect is close to zero ($a_{YX} = 0.031$), hence there is no evidence for the assumption of *Projektion* (as regards students) as has been hypothesized. However, the parents' actor effect is comparably large ($a_{YX} = 0.602$), which might be taken as an indicator for parental feelings of responsibility. The parameter $r_{\theta_2\theta_4} = 0.2$ shows that the ratings of the respective other are nearly uncorrelated, when taking the actor and partner effects into consideration.

Assessment of Model Fit As was noted above, the MRCMLM also allows for a multifaceted assessment of model fit. First of all, `ConQuest` supports the item mean square statistics Outfit (Unweighted Mean Square Statistic) and Infit (Weighted Mean Square Statistic). Basically, these are measures of discrepancy between observed and expected responses. Model fit is indicated by values close to one for either statistic (for details see Wright & Stone 1979). Listing 8 in Appendix "APIM Item Fit Indices" presents the model fit segment of the output. Generally, item fit is not convincing in our case, as many of the indexes lie outside the given confidence intervals (and, correspondingly, have t-values larger than 2).

Furthermore, the parameter estimates allow for expressing a reliability coefficient comparable to the one from classical test theory (cf. `ConQuest` manual, Wu, Adams, Wilson, & Haldane, 2007, p. 160). Listing 5 shows the original program output regarding this "Andrich-Reliability." It seems that all four latent constructs have low reliability, possibly a consequence of data dichotomization. Due to the artificial nature of the data, we will refrain from further interpretating this result.

Listing 5 Essential `ConQuest` Output for the APIM (Part 3: Scale Reliability)

```
RELIABILITY COEFFICIENTS
------------------------

Dimension: (Dimension_1)
MLE Person separation RELIABILITY:   0.346
------------------------

Dimension: (Dimension_2)
MLE Person separation RELIABILITY:   0.456
------------------------

Dimension: (Dimension_3)
MLE Person separation RELIABILITY:   0.393
------------------------

Dimension: (Dimension_4)
MLE Person separation RELIABILITY:   0.394
------------------------
```

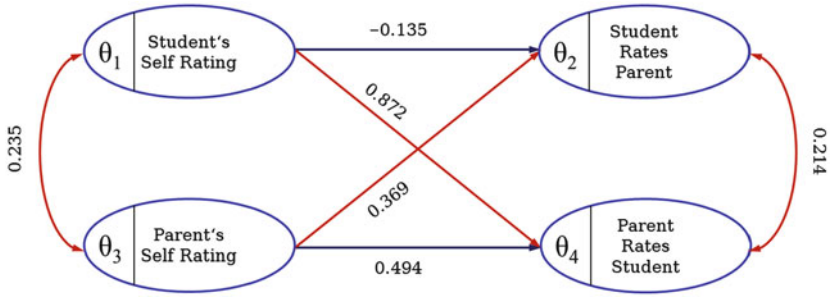


Fig. 6 The final APIM based on the MRCMLM

A measure of item fit is based on the comparison of observed and model derived ICC. The program delivers one such plot per item, one of which is shown in Fig. 6.

The horizontal axis shows the latent trait (θ) in the interval $[-4, +4]$ (covering the most frequently obtained values). The solid line is the expected probability of a positive response to item i, i.e. $P(X_{vi} = 1)$ according to Eq. (2) for $\theta_v \in [-4, +4]$, and the dotted line is the relative frequency of $X_{vi} = 1$ for all observed score groups (also called the *empirical ICC*). The closeness of the two lines is an expression of model fit.

The CFM Approach

We now discuss the correlation of the self-descriptions and the descriptions of the respective other more generally. We could assume that the complex processes within the family (here considering the dyad of two family members only) form the common background for developing a personality (θ_1 and θ_3) on the one hand and also establishing a common background for the perception of the family member (θ_2 and θ_4), on the other hand. Of course, individual components not captured by the correlation on the dyadic level may play an important role as well. These are expressed by the correlation coefficients of the residuals as described in section "The CFM in Terms of an MRCMLM". The MRCMLM directly delivers an estimation of the dyadic correlation coefficient with the (standardized) entries of the covariance matrix (6), while the individual level correlation coefficients require a little bit of craftsmanship.

Model Setup

The command script for the MRCMLM formulation of the CFM is similar to the previous script. However, the assignment of items to latent factors differs as we now

just use two latent factors for all self-description items on the one hand and the items expressing the partners' assessments, on the other hand (lines 5–8 in Listing 6).

Listing 6 `ConQuest` Command Script for the CFM

```
1  datafile gt_res.dat;
2  format responses 1-24;
3  codes 0,1;
4  recode (1 2 3 4 5 6) (0 0 0 1 1 1);
5  score (0,1)  (0,1)   ()   !items (1-6);
6  score (0,1)  ()   (0,1)   !items (7-12);
7  score (0,1)  (0,1)   ()   !items (13-18);
8  score (0,1)  ()   (0,1)   !items (19-24);
9  model item;
10 estimate ! storage=RAM, nodes=5, stderr=quick;
11 show parameters!table=3;
12 show parameters!table=2;
13 show ! estimate=mle;
14 show residuals ! estimates=wle >> resid.txt;
```

In line 14 of Command Script 6 the residuals are written into a file named `resid.txt`. These residuals are used to compute the individual level correlation coefficients r_A and r_B. In the present example, students and parents have responded to the same items. In order to take this mapping into account, we compute the correlation coefficients of the associated residuals (i.e., item 1 of student/self with item 1 of student → parent, etc.). The items within one block (e.g., all items regarding the self-rating of the student) are assumed to measure unidimensional. As a consequence, the residuals of each block represent the individual information not covered by the latent scales θ_X and θ_Y. We therefore compute the average of the correlation coefficients (cf. Monin & Oppenheimer 2005) across items per individual in order to obtain the desired coefficients r_A and r_B (cf. ellipses in Fig. 7; for technical details see Appendix "Extracting the Individual Level Correlation Coefficients").

Results

The essential output providing the CFM coefficients is given in Listing 7, where we find the estimated covariance (upper triangular matrix) and the correlation coefficient (lower triangular matrix) of the latent factors, i.e. the $\hat{\sigma}^2_{\theta_\ell\theta_{\ell'}}$ and the $\hat{r}_{\theta_\ell\theta_{\ell'}}$ for each pair θ_ℓ and $\theta_{\ell'}$. The bottom line contains the estimated variances of each latent variable, $\hat{\sigma}^2_{\theta_\ell}$, with the corresponding standard errors in brackets.

In this output we find that $\hat{r}_{\theta_1\theta_2} = 0.501$, indicating a medium sized correlation of the two latent constructs. Again, we may apply the Wald statistic to test whether this coefficient differs significantly from zero.

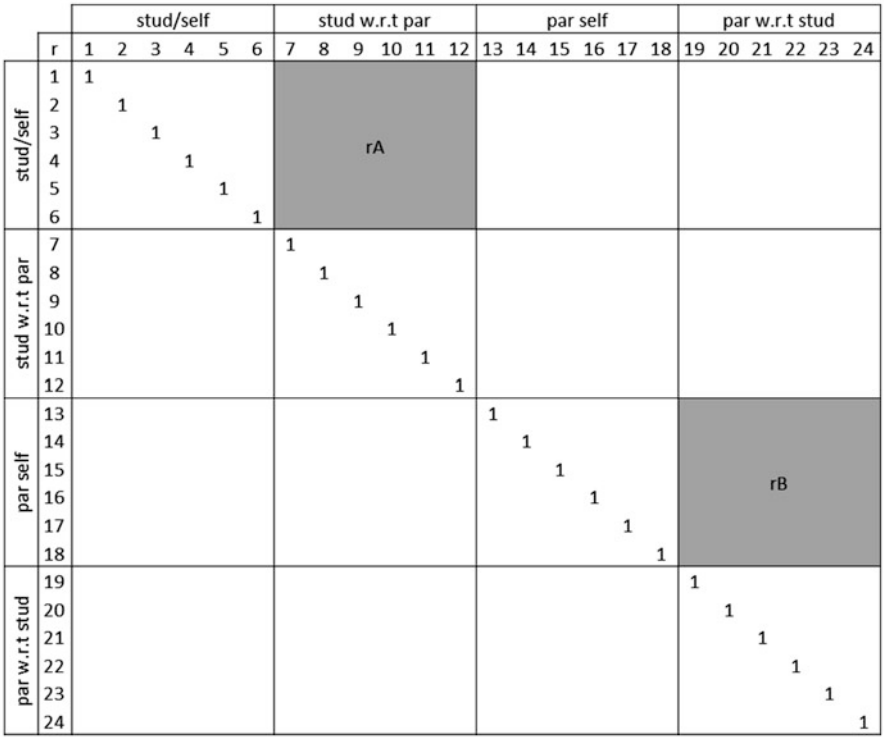


Fig. 7 Structure of the correlation matrix of the residuals

Listing 7 Essential `ConQuest` Output for the CFM (Part 1: Latent Correlation)

```
CONDITIONAL COVARIANCE/CORRELATION MATRIX

                                  Dimension
                          ------------------------------------
Dimension                      1                 2

Dimension_1                                    0.489
Dimension_2                0.501
------------------------------------------------------------
Variance                  0.799  ( 0.046)   1.196  ( 0.069)
------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Values below the diagonal are correlations and values above are covariances
============================================================
```

Next, we have to extract the residual correlations within individuals (student and parent, indicated by dashed double headed arrows in Fig. 7). For that purpose we have to evaluate the residuals stored in the external file, as has been done in line 14 of Listing 6. The necessary steps are explained in Listing 9 in Appendix "Extracting the Individual Level Correlation Coefficients", resulting in $r_{XY}^{(Student)} = -0.017$ and $r_{XY}^{(Parent)} = -0.010$.

Interpretation Interestingly, we find for Social Resonance only a dyadic correlation, while the two individual level correlation coefficients are in fact zero. From this, we could conclude that the common factors reflecting familial bonds seem to predominantly explain the agreement of self-description and other's assessment. Because of the nature of the data used for this analysis, we will again refrain from an in-depth interpretation of this evidence.

Assessment of Model Fit Model fit may be assessed in the same way as in the APIM. The item fit indices (Listing 10 in the Appendix) indicate that a few items do not fit and thus require further investigation. The Scale Reliability Coefficient for the self-rating latent scale was 0.401 and the value for the other's rating was 0.512. Both indicate slightly better fit than the scales constructed in the APIM. This could be an effect of scale length, as each latent factor comprises 12 items in the CFM, while there were only six items per scale in the APIM.

The ICC analysis would involve inspection of one plot per item, which is omitted here. Over all, acceptable model fit seems within range.

Discussion

The present contribution has demonstrated how two models of dyadic data analysis, the APIM and the CFM, can be cast in terms of a multidimensional Rasch Model, the MRCMLM. These two approaches have been conducted, yet many more could be conceived of. The common denominator is that the constructs of interest are not measured directly but rather with a set of variables each. These manifest variables are—as is often the case in social research—dichotomous or ordered categorical. Using the MRCMLM, a discrete probability model, we estimate a latent factor for each such construct. The relationships of these latent factors are then modeled in the latent domain.

Of course, one could argue that the SEM approach is readily applicable to ordinal or (ordered) categorical data as well by setting up an appropriate covariance matrix, using tetra- or polychorical correlation coefficient estimates. This argument definitely applies, but we must bear in mind that this extra step requires larger samples than the standard product-moment correlation coefficient for interval scaled variables (at least if standard maximum likelihood estimation is applied, which is usually the case; cf. Choi, Peters, & Mueller 2010). Alternatively, one may regard the category codings as valid quantifications of response categories assumed to be evenly spaced, hence assuming to work with coarsely categorized interval scaled variables in the sense of Bollen and Barb (1981). However, distributional issues may still arise then. If so, a Weighted ML Estimation Method is available, involving the estimation of the fourth moments. These require, for k items, the computation of a covariance matrix consisting of $(k^4 + 2k^3 + k^2)/4$ elements. Such a matrix would require a large number of observations to attain estimates with sufficient precision. Note that this argument also applies to data measured on an interval scale, when the

distributional assumptions are unclear. Hence, we may apply SEM by all means to (ordered) categorical data. However, the approach presented here treats such data in a much more natural manner, as it deliberately models the way, a latent response propensity θ is transformed into a response probability $P(x_{vi}|(\cdot))$. One advantage of this approach is that departures from the postulated link function can be detected, as has been exemplified in Fig. 6.

The MRCMLM applies the marginal maximum likelihood parameter estimation method, which assumes the person parameters to follow a certain distribution, usually the normal [cf. Eq. (5)]. Such an assumption may not necessarily hold (cf. Blanca, Arnau, López-Montiel, Bono, & Bendayan 2013; Micceri 1989), which might introduce an estimation bias. However, this assumption is a consequence of the applied estimation method, not of the model itself. Rasch Models not including a background structure as introduced in Eq. (4) support the conditional parameter estimation technique (Andersen 1970 1980), even in the multidimensional case (Andersen 1977). The CML estimation method conditions on the sufficient statistics of the incidental (in the sense of Neyman & Scott 1948) parameters and thus makes no distributional assumptions at all (for a comparison of MML and CML, see Adams & Wu, 2007, pp. 68–69). Moreover, the CML approach facilitates a model test (Andersen 1973), allowing for a rigid assessment of model fit.

When applying the APIM, researchers may be particularly interested in estimating two ratio parameters $k = p/a$ and $k' = p'/a'$ with p/p' representing the respective partner effects and a/a' representing the according actor effects (cf. Fig. 4). The ratios k and k' can be used to describe specific patterns in the APIM (e.g., $k = 1$ refers to a couple pattern, $k = -1$ refers to a contrast pattern, and $k = 0$ refers to an actor-only pattern). Kenny and Ledermann (2010) proposed a phantom variable approach to estimate k along with its standard error in the SEM context, thus allowing for a significance test of k as well. A merely descriptive value of k may be obtained with the estimated coefficients γ from the standard MRCMLM output.

One valuable option has not been incorporated in the presented examples: Each model could be enhanced with a background population model, thus controlling the latent variables for background variables (like age or socio-economic information, for example). Further extensions would consider non-distinguishable dyads or more complex designs (like the One-with-Many Design, taking more than two individuals into account).

While our examples have only dealt with dichotomous data, the full bandwith of IRT models for polytomous categorical data is readily available. Furthermore, one could drop the assumption of parallel trace lines and include a discrimination parameter in the model equation, thus explicitly capturing differing item characteristics within the items of a scale as well. Such extensions would allow for a wider range of items to be used.

Altogether, the presented approach provides a powerful framework for the complex requirements of dyadic data modeling, taking both scale and distributional requirements into account.

Acknowledgements I am indebted to Paul Czech for his assistance during data acquisition of the students' sample.

Technical Appendix

APIM Commands

Table 1 Description of `ConQuest` commands regarding the APIM

Line	Command
1	Where to read the data
2	Which columns contain the item response data
3	Valid codes for estimation (entries other than those listed here are treated as missing values)
4	Dichotomize codings: 1, 2, 3 = 0; 4, 5, 6 = 1
5	Matrix **B**: assign items 1–6 to first latent factor (A self)
6	Matrix **B**: assign items 7–12 to second latent factor (A w.r.t. B)
7	Matrix **B**: assign items 13–18 to third latent factor (B self)
8	Matrix **B**: assign items 19–24 to fourth latent factor (B w.r.t. A)
9	Estimate one item parameter (i.e. no thresholds required after dichotomization). A PCM would require `model item + item*step;` and the RSM `model item + step;`.
10	Estimation details
11–13	Output details
15–16	Regression coefficients as indicated by the APIM

APIM Item Fit Indices

Listing 8 Item Fit Indices for the APIM

```
-------------------------------------------------------------------------------------
   VARIABLES                          UNWEIGHTED FIT              WEIGHTED FIT
------------                        ------------------------   ------------------------
   item     ESTIMATE    ERROR^   MNSQ       CI         T    MNSQ       CI         T
-------------------------------------------------------------------------------------
 1   1      -1.452      0.083    1.23 ( 0.89, 1.11)  3.8   1.08 ( 0.79, 1.21)  0.8
 2   2       0.688      0.070    0.83 ( 0.89, 1.11) -3.1   0.87 ( 0.92, 1.08) -3.1
 3   3       0.137      0.072    1.08 ( 0.89, 1.11)  1.4   1.14 ( 0.90, 1.10)  2.6
 4   4       0.147      0.072    1.23 ( 0.89, 1.11)  3.7   1.23 ( 0.90, 1.10)  4.1
 5   5      -0.124      0.074    1.04 ( 0.89, 1.11)  0.7   1.08 ( 0.88, 1.12)  1.2
 6   6       0.604*     0.167    0.81 ( 0.89, 1.11) -3.5   0.84 ( 0.91, 1.09) -3.7

 7   7       5.097      0.112    4.68 ( 0.88, 1.12) 33.7   1.25 ( 0.63, 1.37)  1.3
 8   8      -2.283      0.095    0.51 ( 0.88, 1.12)-10.2   0.75 ( 0.83, 1.17) -3.1
 9   9      -0.872      0.086    1.19 ( 0.88, 1.12)  2.9   1.15 ( 0.88, 1.12)  2.2
10  10      -0.407      0.084    0.96 ( 0.88, 1.12) -0.7   0.97 ( 0.89, 1.11) -0.4
11  11      -0.243      0.084    1.05 ( 0.88, 1.12)  0.8   1.06 ( 0.89, 1.11)  1.1
12  12      -1.291*     0.208    0.72 ( 0.88, 1.12) -5.2   0.91 ( 0.87, 1.13) -1.4

13  13      -2.731      0.121    0.20 ( 0.87, 1.13)-18.4   0.70 ( 0.35, 1.65) -0.9
14  14       0.965      0.085    1.03 ( 0.87, 1.13)  0.4   1.03 ( 0.91, 1.09)  0.6
15  15      -0.776      0.103    0.53 ( 0.86, 1.14) -8.3   0.83 ( 0.76, 1.24) -1.4
16  16       1.097      0.084    0.96 ( 0.87, 1.13) -0.6   0.96 ( 0.91, 1.09) -1.0
17  17       1.189      0.084    1.18 ( 0.87, 1.13)  2.5   1.10 ( 0.91, 1.09)  2.1
18  18       0.257*     0.216    1.09 ( 0.87, 1.13)  1.4   1.05 ( 0.87, 1.13)  0.7

19  19       3.173      0.116    3.72 ( 0.87, 1.13) 24.2   1.25 ( 0.80, 1.20)  2.3
20  20      -0.171      0.096    1.03 ( 0.87, 1.13)  0.4   0.97 ( 0.89, 1.11) -0.5
21  21      -0.469      0.099    0.99 ( 0.87, 1.13) -0.1   1.02 ( 0.87, 1.13)  0.3
22  22      -0.484      0.099    0.84 ( 0.87, 1.13) -2.5   0.91 ( 0.87, 1.13) -1.5
23  23      -0.116      0.096    0.95 ( 0.87, 1.13) -0.7   1.00 ( 0.89, 1.11)  0.0
24  24      -1.934*     0.227    0.59 ( 0.87, 1.13) -7.0   0.86 ( 0.75, 1.25) -1.1
-------------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.997
Chi-square test of parameter equality =    5051.19,  df = 20,  Sig Level = 0.000
^ Quick standard errors have been used
=====================================================================================
```

`item`: Item number and label; as no label has been provided, the item number is repeated.

`ESTIMATE`: Item parameter estimate; in the dichotomous case, this is the item difficulty parameter [δ_i according to Eq. (1)]. To identify a latent scale, one item per latent dimension is fixed (indicated by an asterisk). By default, `ConQuest` sets the sum of the item parameters per latent dimension to zero (e.g.: $-1.452 + 0.688 + 0.137 + 0.147 + (-0.124) + \mathit{0.604} = 0$). This could be overridden with the command `set constraint=cases`, causing the mean of the latent variable to be fixed at zero.

`ERROR`: Standard error of item difficulty parameter.

`MNSQ`: Outfit (UNWEIGHTED FIT) and Infit (WEIGHTED FIT) Index.

`CI`: The 95 % confidence interval for the expected value (i.e., 1) of Infit and Outfit.

T: The *t*-statistic for the null hypothesis that the Outfit and Infit Index is 1. Values larger than 2 may be considered significant at the 95 % level (corresponds to MNSQ outside the CI).

Extracting the Individual Level Correlation Coefficients

To obtain the individual level correlation coefficients, we use the residuals stored in `resid.txt`. This file contains 600 lines and 25 columns. The first column is a numerical dyad identifier, followed by four groups of six columns each, comprising the residuals to the respective six items of student/self, student w.r.t parent, parent/self and parent w.r.t student. Any multi-purpose statistics software can be used to obtain the individual level correlation coefficients. We will resort to the R software (R Core Team 2014) for it is freely available (open source) and easy to use. The following script will perform the required steps:

Listing 9 R Script for Computing the CFM Individual Level Correlation Coefficients

```
1  d0 = read.table(file="resid.txt")
2  d0[d0==-99] = NA
3  colnames(d0) = c("id",paste("stud"   ,1:6,sep="")
4                       ,paste("studpar",1:6,sep="")
5                       ,paste("par"    ,1:6,sep="")
6                       ,paste("parstud",1:6,sep=""))
7
8  r0 = cor(d0[,-1],use="pair")
9
10 ra = r0[1:6,7:12]
11 rb = r0[13:18,19:24]
12
13 r2z = function(r) 0.5 * log( (1+r)/(1-r) )
14 z2r = function(z) (exp(2*z)-1) / (exp(2*z)+1)
15
16 z2r( mean(r2z(ra)) )
17 z2r( mean(r2z(rb)) )
```

The ten statements of Listing 9 perform the following operations:

- In line 1 of the script, we read the content of the file `resid.txt` and store it in a `data.frame` named `d0`.
- Then (line 2) we transform the missing values (`ConQuest` codes them with `-99` by default) to the R missing indicator `NA`.
- In lines 3–6, the columns obtain more informative variable names (the output file contains no header, therefore, `R` uses the generic names `V1` to `V25` by default). This step is merely cosmetic and may as well be omitted.
- Next (line 8), we compute the 25×25 correlation matrix of all residuals (omitting the `id` variable stored in column 1). A schematic view of this matrix is given in Fig. 8.

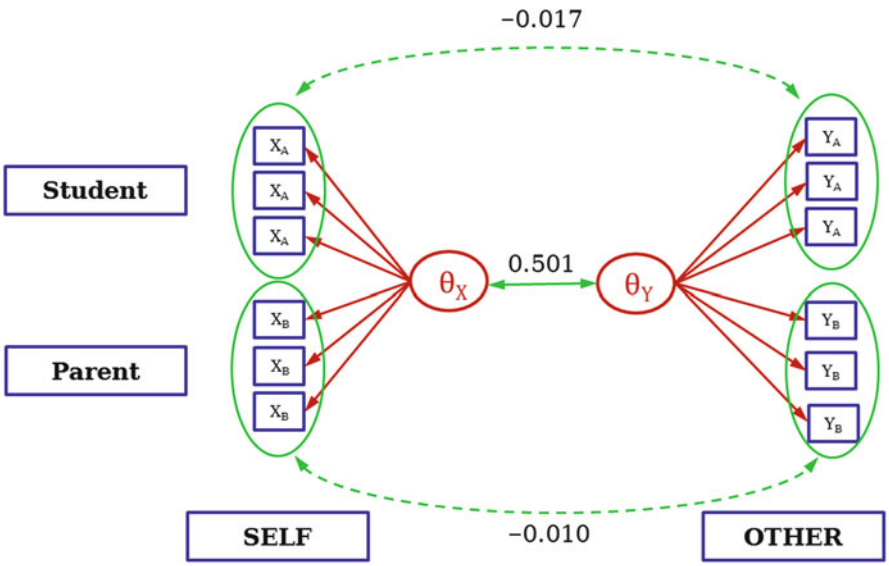


Fig. 8 The final CFM

- In line 10, we cut out blocks of correlation coefficients of the residuals of the students' self-description items with the columns covering the residuals of the students' assessments of the respective parents (rows 1–6/columns 7–12; grey shaded area termed `rA` in Fig. 8).
- Analoguously, in line 11, we cut out the correlation coefficients of the residuals of the parents' self-assessment items with the residuals of the items covering the parents' assessments of the respective students (rows 13–18/columns 19–24; grey shaded area termed `rB` in Fig. 8).
- In lines 13 and 14 we prepare two functions, transforming a correlation coefficient to a Fisher's Z-value (`r2z`) and backtransforming the latter into a correlation coefficient again (`z2r`). These functions could easily be enhanced to detect invalid input and issue a corresponding message.
- Finally (lines 16 and 17), we apply the Z-transformation to the two matrix parts, compute the mean and backtransform it to a valid correlation coefficient.

With these steps, we dispose of all required information to draw the complete CFM, depicted in Fig. 7.

CFM Item Fit Indices

Listing 10 Item Fit Indices for the CFM

```
-------------------------------------------------------------------------------------
   VARIABLES                          UNWEIGHTED FIT              WEIGHTED FIT
------------                      ------------------------   ------------------------
     item     ESTIMATE  ERROR^   MNSQ       CI        T     MNSQ       CI        T
-------------------------------------------------------------------------------------
 1   1        -1.036    0.093    1.08 ( 0.89, 1.11)  1.3   1.02 ( 0.80, 1.20)  0.2
 2   2         0.873    0.073    0.85 ( 0.89, 1.11) -2.7   0.89 ( 0.93, 1.07) -3.2
 3   3         0.389    0.076    1.12 ( 0.89, 1.11)  2.1   1.10 ( 0.91, 1.09)  2.1
 4   4         0.394    0.076    1.06 ( 0.89, 1.11)  1.0   1.08 ( 0.91, 1.09)  1.8
 5   5         0.139    0.079    0.97 ( 0.89, 1.11) -0.5   0.99 ( 0.89, 1.11) -0.2
 6   6         0.773    0.074    0.89 ( 0.89, 1.11) -2.0   0.90 ( 0.92, 1.08) -2.7
 7   7         4.485    0.140    1.64 ( 0.88, 1.12)  9.0   1.19 ( 0.57, 1.43)  0.9
 8   8        -1.587    0.103    0.63 ( 0.88, 1.12) -7.1   0.88 ( 0.84, 1.16) -1.5
 9   9        -0.503    0.088    0.98 ( 0.88, 1.12) -0.3   1.03 ( 0.90, 1.10)  0.5
10   10       -0.148    0.085    0.90 ( 0.88, 1.12) -1.7   0.92 ( 0.91, 1.09) -1.9
11   11       -0.023    0.084    1.07 ( 0.88, 1.12)  1.1   1.08 ( 0.92, 1.08)  1.8
12   12       -0.823    0.091    0.70 ( 0.88, 1.12) -5.7   0.80 ( 0.89, 1.11) -3.7
13   13       -3.056    0.116    0.39 ( 0.87, 1.13)-11.8   0.86 ( 0.28, 1.72) -0.3
14   14        0.693    0.081    1.07 ( 0.87, 1.13)  1.0   1.04 ( 0.91, 1.09)  0.8
15   15       -0.968    0.099    0.74 ( 0.86, 1.14) -4.1   0.92 ( 0.76, 1.24) -0.6
16   16        0.830    0.081    1.11 ( 0.87, 1.13)  1.5   1.06 ( 0.92, 1.08)  1.4
17   17        0.924    0.080    1.07 ( 0.87, 1.13)  1.0   1.05 ( 0.92, 1.08)  1.3
18   18        0.046*   0.283    1.04 ( 0.87, 1.13)  0.5   1.05 ( 0.87, 1.13)  0.8
19   19        2.933    0.116    3.88 ( 0.87, 1.13) 25.2   1.36 ( 0.79, 1.21)  3.1
20   20       -0.423    0.095    0.95 ( 0.87, 1.13) -0.8   0.93 ( 0.90, 1.10) -1.4
21   21       -0.712    0.098    1.09 ( 0.87, 1.13)  1.3   1.08 ( 0.88, 1.12)  1.3
22   22       -0.727    0.098    1.02 ( 0.87, 1.13)  0.4   1.01 ( 0.88, 1.12)  0.2
23   23       -0.371    0.095    0.99 ( 0.87, 1.13) -0.1   1.00 ( 0.90, 1.10)  0.1
24   24       -2.100*   0.333    0.80 ( 0.87, 1.13) -3.2   0.99 ( 0.76, 1.24) -0.1
-------------------------------------------------------------------------------------
An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability =  0.996
Chi-square test of parameter equality =    3697.27,  df = 22,  Sig Level = 0.000
^ Quick standard errors have been used
=====================================================================================
```

For an explanation of the column headings see Appendix “APIM Item Fit Indices”.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models. Extensions and applications* (pp. 57–75). New York, NY: Springer.

Adams, R. J., Wu, M. L., & Wilson, M. (2012). *Conquest 3.0 [Computer software]*. Melbourne: Australian Council for Educational Research (ACER).

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B, 32*, 283–301.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123–140.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69–81.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Andrich, D. (1982). An extension of the Rasch Model for ratings providing both location and dispersion parameters. *Psychometrika, 47*, 105–113.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory. Parameter estimation techniques*. New York, NY: Marcel Dekker.

Baumeister, R. R., Dale, K., & Sommer, K. L. (1998). Freudian defense mechanisms and empirical findings in modern social psychology: Reaction formation, projection, displacement, undoing, isolation, sublimation, and denial. *Journal of Personality, 66*, 1081–1124.

Beckmann, D., Bräahler, E., & Richter, H.-E. (1990). *Der Gießen-Test (GT). Ein Test für Individual- uind Gruppendiagnostik [The Gießen test (GT). A test for the assessment of individuals and groups]* (4th ed.). Bern: Hans Huber.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. E. Novick (Eds.), *Statistical theories of mental test scores with contributions by A. Birnbaum* (pp. 395–479). Reading, MA: Addison-Wesley.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*, 78–84.

Bollen, K. A. (1989). *Structural equations with latent variables*. Hoboken, NJ: Wiley.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232–239.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science, 3*, 14–25.

Campbell, L., & Kashy, D. A. (2002). Estimating actor, partner, and interaction effects for dyadic data using PROC MIXED and HLM: A guided tour. *Personal Relationship, 9*, 327–342.

Choi, J., Peters, M., & Mueller, R. O. (2010). Correlational analysis of ordinal data: From Pearson's r to Bayesian polychoric correlation. *Asia Pacific Educational Review, 11*, 459–466.

Gebhardt, E. C. (in preparation). *Latent Path Models within an IRT Framework.* Unpublished doctoral dissertation, University of Melbourne, Melbourne, Australia.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 131–155). New York, NY: Springer.

Freud, S. (1976). In J. Strachey (Ed.), *The complete psychological works of Sigmund Freud* (The standard edition). New York, NY: W. W. Norton & Company.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 69–95). New York, NY: Springer.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York, NY/Hove: Routledge.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford.

Kenny, D. A., & Ledermann, T. (2010). Detecting, measuring, and testing dyadic patterns in the actor-partner interdependence model. *Journal of Family Psychology, 24*, 359–366.

Linacre, J. M. (1989). *Multi-facet Rasch measurement*. Chicago, IL: Mesa Press.

Loeys, T., Cook, W., De Smet, O., Wietzker, A., & Buysse, A. (2014). The actor-partner interdependence model for categorical dyadic data: A user-friendly guide to GEE. *Personal Relationships, 21*, 225–241.

Loeys, T., & Molenberghs, G. (2013). Modeling actor and partner effects in dyadic data when outcomes are categorical. *Psychological Methods, 18*, 220–236.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
McMahon, J. M., Puget, E. R., & Tortu, S. (2006). A guide for multilevel modeling of dyadic data with binary outcomes using SAS PROC NLMIXED. *Computational Statistics & Data Analysis, 50*, 3663–3680.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.
Monin, B., & Oppenheimer, D. M. (2005). Correlated averages vs. averaged correlations: Demonstrating the warm glow heuristic beyond aggregations. *Social Cognition, 23*, 257–278.
Müller, H. (1987). A rasch model for continuous ratings. *Psychometrika, 52*, 165–181.
Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49*, 115–132.
Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1–32.
R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual], Vienna, Austria. Retrieved from http://www.R-project.org/
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Pædagogiske Institut.
Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Copenhagen: The Danish Institute of Educational Research.
Rasch, G. (1977). On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14*, 58–93.
Rasch, G. An informal report on the present state of a theory of objectivity in comparisons. In Proceedings of the NUFFIC International Summer Session in Science at “Het Oude Hof”, The Hague, 14–28, July, 1966. Retrieved July 22, 2015, from http://www.rasch.org/memo1966.pdf.
Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
Stock, J. H., & Trebbi, F. (2003). Who invented instrumental variable regression? *Journal of Economic Perspectives, 17*, 177–194.
van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.
von Eye, A., & Mun, E.-Y. (2013). *Log-linear modeling: Concepts, interpretation, and application*. Hoboken, NJ: Wiley.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.
Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest. Generalised item response modelling software*. Melbourne: ACER Press.

Longitudinal Analysis of Dyads Using Latent Variable Models: Current Practices and Constraints

Heather M. Foran and Sören Kliem

Abstract Interdependencies between dyads have long been recognized and taken into account in the analysis of partnership and marital data. However, most of the research that has examined dyadic influences is based on cross-sectional data or basic longitudinal models. When more complex longitudinal models are examined, several limitations and barriers arise. In this chapter, some of the practical issues with dyadic analyses of multi-time point samples will be discussed. In particular, we discuss (1) applications of latent growth curve mixture modeling trajectories of intimate partner relationship adjustment and (2) latent difference score modeling associations between relationship adjustment and depressive symptoms over time. A 4-year longitudinal sample of 237 families assessed over six time points will be used to illustrate these practical issues.

Why Are Intimate Relationships Important to Study?

Intimate relationships are among the most important contributors to well-being and life satisfaction. Although there are a number of important relationships, those with an intimate adult partner appear to play a particularly important role in psychological, physical, and economic well-being (e.g., Fincham & Beach 2010). Individuals in a satisfying intimate relationship or marriage experience better psychological health, economic security, and decreased risk for physical illnesses (Beach & Whisman 2013). Moreover, ending an intimate relationship through separation or divorce is one of the biggest risk factors for major depression and suicidality (e.g., Amato 2010; Sbarra, Law, & Portley 2011).

Accordingly, understanding what factors contribute to relationship satisfaction and long-term relationship success has been an area of research interest for many decades. This chapter focuses on longitudinal research concerning intimate couples

H.M. Foran (✉)
Ulm University, Ulm, Germany
e-mail: heather.foran@uni-ulm.de

S. Kliem
Criminological Research Institute of Lower Saxony, Hanover, Germany

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_9

and methodological approaches to these analyses (i.e., other dyads are not discussed, friendship pairs, parent–child dyads, or triads). Also not included in this chapter are daily diary studies as this represents a related yet distinct subset of longitudinal research with couples (e.g., Ferrer & Nesselroade 2003; Ferrer & Widaman 2008). In particular, this chapter focuses on longitudinal studies with multiple time points (i.e., in which it is possible to examine change over time) using a structural equation modeling framework.

History of Couple Longitudinal Research: Two Interdependencies

Repeated Measures

Prior to the mid-1990s, there were many studies examining couple processes but very few longitudinal studies of change. Typically, studies only measured relationship adjustment at one time point and measured other variables at a second time point. There was little use of repeated measures in research and very few samples of couples were followed over longer periods. In the seminal review of Karney and Bradbury (1995a), many of these methodological issues were highlighted. The authors reviewed 115 studies representing 68 separate samples in which marital processes were examined with two or more time points dating back as far as 1946. Of the 115 studies reviewed, 70 % used either *zero-order* correlations, *t*-tests, or analysis of variance (ANOVA). One-third of these studies examined the bivariate correlation between time 1 and a later time point and ignored any time point data in-between, if it was assessed. Another third of the studies ($n = 37$) used residualized change regression models in which a time-2-dependent (Y) variable was predicted by a time-1-independent (X) variable controlling for the time-1 Y variable. *t*-tests were used in 16 studies and 28 studies used ANOVAs (not repeated measures ANOVA). Of the 115 published studies, only 15 studies had longitudinal data in which at least three time points were assessed. However, none of these studies used growth curve analyses and frequently the same variable was not assessed at each time point.

Growth curve analyses were introduced to the couple research area with key papers in the 1990s (Barnett, Marshall, Raudenbush, & Brennan 1993; Karney & Bradbury 1995b; Raudenbush, Brennan, & Barnett 1995). Although these techniques were developed much earlier (e.g., Rogosa, Brandt, & Zimowski, 1982), they were not widely used by applied researchers until software programs that supported their use were introduced (Bryk, Raudenbush, & Congdon 1996; Raudenbush, Bryk, & Congdon 2000, HLM software) paralleling movement in the broader psychology field of longitudinal psychological research in which longitudinal change processes were given more in-depth consideration (see Collins & Sayer 2001; Little, Schnabel, & Baumert 2000).

Although growth curve modeling was possible with either structural equation modeling approaches or multi-level modeling approaches, much of the work in the couples field has been done from a multi-level modeling approach. This is likely due to the influence of applied methodology research papers in the couple field that described the multi-level modeling approach (via HLM software) (Karney & Bradbury 1995b) and early papers demonstrating this approach with couples (e.g., Barnett et al. 1993). Further, sample sizes tended to be small, and multi-level modeling could be used with small sample sizes (e.g., Maas & Hox 2005), whereas recommendations for structural equation modeling required larger sample sizes. Hence, applications using a MLM approach, particularly with HLM software, have dominated the analysis of dyads. We return to this issue in the next section where we will further discuss differences between MLM and SEM approaches, but before doing so, we introduce the second type of independence relevant for longitudinal dyadic analyses.

Nesting Within Couples

Around the same time as independence due to repeated measures, the issue of interdependence between members of a dyad began to be recognized as an important methodological problem for couple research. Kenny and colleagues introduced this issue to the broader couple research field with several papers and a commonly cited book ("Dyadic Data Analysis"; Kenny 1995; Kenny, Kashy, & Cook 2006; see also Atkins 2005). Analysis of dyads overcame some of the limitations related to analyzing members of a couple separately including a loss of degrees of freedom, biased standard errors (*F*- and *t*-tests), and incorrect *p*-values (vulnerability to both Type 1 and Type 2 errors) which can lead to biased estimates of the relationships in terms of correlations and regression weights, for example.

Several models were introduced as ways to handle dyadic analysis (actor partner independence model (APIM); Kenny & Cook 1999), mutual influence model (Kenny 1996), and the common fate model (Kenny & La Voie 1984). Of these, the APIM has been the most widely used in couple research (Ledermann & Kenny 2012). This is a model in which both actor effects and partner effects can be tested simultaneously and this has been applied in the couple research area to heterosexual couples in which they are considered "distinguishable dyads" due to gender. Other approaches were introduced and can be applied to analysis with homosexual couples (called "indistinguishable dyads").

Notably, much of the work using the APIM approach has been cross-sectional or across two time points. For example, Cook and Kenny (2005) applied the APIM model to a two-time point model of attachment. The APIM model can be applied through use of either multi-level models or structural equation modeling approaches in which the interdependencies between husband and wife scores can be modeled.

The mutual influence model differs from the APIM model in that it assumes that there is bidirectional causation in the outcome variable such that each member of

the couple directly influences the other member (Y1 ↔ Y2). To test this model in an SEM framework, one assumes that there are no partner effects (paths between partner 1 X and partner 2 Y variables) and that there are bidirectional paths between partner 1 and partner 2 Y1 and Y2 outcome variables. Kenny and colleagues describe this model as most plausible in the situation in which independent variables X1 and X2 are individual difference variables and outcome variables Y1 and Y2 are couple variables. In other words, X1 and X2 should show little within-partner correlations (e.g., a personality trait) whereas Y1 and Y2 should show a high within-partner correlation (e.g., relationship satisfaction). This model could be used for longitudinal data analysis as well, but has seldom been tested in couples research. We suspect there are several reasons why this model is often not used: it is less known, it is analytically more complex compared to the APIM model, and it is less applicable theoretically (i.e., partner effects occur often).

The common fate model is another alternative to dyadic analysis, but, similar to the mutual influence model, it is rarely used for couple longitudinal analysis. In this model, one or more latent factors are included and are indicated by each member's scores on some measured variable. There are various versions of the common fate model which vary in the number of latent variables and how the individual unique effects and dyadic effects are modeled (see Griffin & Gonzalez 1995; Kenny et al. 2006). The common fate model assumes that some common unmeasured factor explains both partner's scores on a measured variable and that unaccounted variance reflects each member's "uniqueness" or individual effects. This model has high applicability for understanding dyadic constructs but is rarely used in either cross-sectional or longitudinal models (see Ledermann & Kenny 2012). An advantage of the common fate model for longitudinal analysis is that it can result in a less complex longitudinal model compared to modeling growth curves of each partners as is the case in the APIM model.

MLM Versus SEM

Although the focus of the current article is on applications of the SEM approach to longitudinal dyadic analysis, we briefly note some of the differences between the MLM and SEM approaches (see also Kashy & Donnellan 2008). As mentioned above, the MLM approach to dyadic longitudinal analysis has been more extensively used and reviewed (see Atkins 2005; Karney & Bradbury 1995b; Raudenbush et al. 1995). It should be noted that longitudinal data analysis estimation via MLM can yield the same results as SEM growth modeling across a wide range of models if certain constraints are imposed (e.g., Bauer 2003; Curran 2003; Wu, Selig, & Little 2012). Regarding couple research data, the SEM and MLM approaches were compared using cross-sectional data with a sample of $N = 348$ couples (Wendorf 2002). Wendorf (2002) illustrated that one can obtain identical results with the MLM and SEM approaches to dyadic analyses with cross-sectional data if the SEM model is simplified to a MLM format (i.e., assumes no measurement error in the predictors

or covariates and constrains the error variances to be equal across all measurement points). However, as far as we are aware, there have been no direct comparisons with longitudinal dyadic data, although there are comparisons of longitudinal non-dyadic data (Chou, Bentler, & Pentz 1998; MacCallum, Kim, Malarkey, & Kiecolt-Glaser 1997).

Although SEM can be used to match the MLM modeling, there are several extensions that SEM affords (see Wendorf 2002). For example, SEM provides more flexibility in modeling choices, especially for analyzing types of relationship that cannot be modeled using MLM (Hox & Stoel 2005; Hoyle & Gottfredson 2015; Wu et al. 2012). With MLM one can only model one dependent variable (e.g., husbands' depressive symptoms) whereas with SEM one can model multiple (correlated) dependent variables (e.g., both husbands' and wives' depressive symptoms) and possible interrelationships simultaneously and account for their residual covariance. In other words, MLM models cannot address how trajectories of one variable relate to another over time (Kouros & Cummings 2011). In addition, MLM assumes no measurement error in predictors (exogenous variables in SEM terminology), whereas SEM allows measurement error to be modeled.

Using MLM software, on the other hand, has several other benefits such as: (a) including additional levels of nesting (e.g., individuals nested in groups), (b) including time-varying (with random effects) or time-invariant covariates to the model, and (c) handling non-continuous dependent variables is straightforward (Hox & Stoel 2005; Wu et al. 2012). Furthermore, MLM can handle designs with a large number of unequal intervals between assessment points (Mehta & West 2000).

Dearth of Longitudinal Dyadic Peer-Reviewed Method Papers

In addition, methodological articles in which the latent variable approach was applied to couple longitudinal research have been scarce and this may also partially explain the less frequent use in longitudinal couple research. There were some early applications in which growth curve modeling of couples was conducted with structural equation modeling (Kurdek 2005). Kurdek (2005) modeled both husband and wife growth curves simultaneously over four time points representing 4 years. The authors predicted the intercepts and slopes of the husband and wife growth curves using time-1 latent variables. In total, the authors tested four separate models with different time-1 latent variables (psychological distress, marital satisfaction, attributions, or social support). The authors modeled the error covariances between adjacent time points and between spouses at each time point. In addition, the authors tested gender differences by comparing model fit ($\triangle\chi^2$) between constrained models in which intercept and slope effects were equal across gender versus freely estimated.

In methodological journals, dyadic analysis is rarely addressed. In *Structural Equation Modeling*: *A Multidisciplinary Journal* through 2013, only five papers were found that addressed dyadic analysis and only two of these discussed

longitudinal data (Newsom 2002; Peugh, DiLillo, & Panuzio 2013). One paper was on state-space modeling of dyadic daily data (Song & Ferrer 2009), one compared SEM and HLM with cross-sectional data (Wendorf 2002), and one discussed APIM mediation with cross-sectional data (Ledermann, Macho, & Kenny 2011).

In *Psychological Methods*, only one paper has been published about analysis of distinguishable dyads (Loeys & Molenberghs 2013). In this paper, the authors apply the APIM model to cross-sectional data using a categorical outcome. Similarly, in *Multivariate Behavioral Research*, there are no papers that have been published that address longitudinal dyadic analysis (although there are several papers that address momentary data or daily diary data; Ferrer, Steele, & Hsieh 2012; Song & Ferrer 2012; Steele & Ferrer 2011). Thus, there is a need for more methodological papers which focus particularly on longitudinal dyadic analyses from a latent variable framework.

Practical Examples of Two Couple Research Questions

To illustrate contemporary issues that arise in longitudinal analysis of couples from a SEM framework, we narrow our discussion to two common research questions in the couple field. First, we address the basic question of how relationship adjustment changes over time using latent growth mixture modeling (LGMM). Next, we examine the association between relationship adjustment with depressive symptoms using a recent extension of latent difference score (LDS) modeling (Grimm, An, & McArdle 2012). Although there are many other approaches in a latent variable framework that could be applied (e.g., traditional parallel process growth curve models), we have selected these two approaches to illustrate the importance of attending to one's match with the theoretical models of change.

Example 1: How Does Relationship Adjustment Change Over Time?

Early research into the longitudinal course of relationship satisfaction consistently reported declines in satisfaction over time (e.g., Karney & Bradbury 1997; Kurdek 1998). These findings were based on analyses of means and did not take into account different trajectories that may exist for subgroups. Recently, we analyzed the trajectories of relationship satisfaction in two samples of parents of young children using LGMM to determine whether different trajectories may exist. LGMM was used to identify latent trajectory groups of relationship adjustment. This approach allows one to identify subpopulation trajectories rather than assuming population homogeneity in trajectories. Furthermore, this approach allows for within-class variability and more flexibility in modeling patterns within classes that is limited with other types of person-centered approaches (e.g., traditional latent class analysis, taxometric analysis). Consistent across both the German

and American prospective samples ($N = 242$ and 453 families, respectively), two distinct longitudinal latent classes were detected (see Foran, Hahlweg, Kliem, & O'Leary 2013; Foran, O'Leary, & Slep 2013). Approximately 90 % of men and women could be classified as showing high relationship satisfaction and a stable or increasing trajectory. The remaining 10 % were initially more distressed and tended to show a decline in relationship satisfaction over time.

Independently, another group of researchers in the United States found similar results using mixture modeling techniques among newlywed samples ($N = 251$ couples), although the distressed groups were larger among newlyweds (Lavner, Bradbury, & Karney 2012). Taken together, there is growing evidence in contrast to the earlier research which only examined means and suggests that relationship satisfaction is relatively stable for the majority of couples and that only a small subgroup experiences significant decline in satisfaction over time.

Although the LGMM approach has proven fruitful in application to understanding relationship adjustment trajectories, there have been certain practical limitations in the application of this approach to dyadic data (e.g., small sample sizes may limit the number of reliable classes that can be detected). This approach provides an elegant approach to dealing with longitudinal interdependencies (via growth curve modeling), but the best approach to dyadic interdependencies is selected based on the theoretical conceptualization of relationship adjustment. To date, researchers who have examined relationship satisfaction in the context of a latent mixture growth curve model have either averaged couple relationship satisfaction or modeled each partner's scores separately.

The rationale for selection of a dyadic or individual model requires careful consideration. Averaging men's and women's scores across relationship satisfaction often simplifies the model but causes loss of an important source of variance. The dyadic model (dual growth mixture model, DGMM) takes into account the shared variance between partners in determining the latent classes. An advantage of the DGMM approach is that when one partner's data are missing, this could be estimated based on the other partner's responses, resulting in reduction of lost data. However, this may not be the best match to the research question. In the case of trajectories of relationship adjustment, one may be interested in men's or women's individual variance or in modeling who is more distressed in the relationship. This depends on the conceptualization of relationship adjustment. Clinically, it only takes one partner who reports relationship distress to indicate a problem and one partner who wants to end the relationship. This would suggest that modeling the *worse* score may yield relevant trajectory information.

We illustrate these differences empirically using a sample of $N = 242$ couples, followed over 4 years, in which we have previously examined latent class growth curves separately for men and women (Foran, Hahlweg et al. 2013). Specifically, we apply LGMM to five models: (1a) men only and (1b) women only models (described previously in Foran, Hahlweg et al. 2013) and three new models (2) DGMM, (3) average scores of relationship satisfaction, and (4) worse score. The model structure for the single growth curve models (models 1a, 1b, 3, and 4) is shown in Fig. 1. This model is similar to a traditional latent growth curve model (i.e., includes continuous

latent intercept and slope variables); a latent categorical variable is labeled with "c" in the model and is used to represent the latent trajectory classes (see Duncan, Duncan, Strycker, Okut, & Li 2002, for more details). In addition, a covariate is included in the model. The DGMM structure (model 2) is illustrated in Fig. 2. This is similar to Fig. 1 but includes growth curves for both men and women (rather than only 1 growth curve as in models 1a, 1b, 3, and 4). The residual variances of men's and women's relationship adjustment within each time point are free to covary as shown in Fig. 2.

Participants and Procedure

Participants were recruited from daycare centers in Braunschweig, Germany (see Heinrichs, Bertram, Kuschel, & Hahlweg 2005 for more detail on the recruitment process) to participate in a randomized control trial of a universal primary parenting prevention program (the Triple-P positive parenting program; see Sanders 2012 for more detail). Briefly, 17 kindergartens were selected to recruit a sample representative of a range of socioeconomic statuses using the social index of their living area via the objective Kita Social Index. Parents, fluent in German, were eligible to participate if they had a child 2½–6 years old attending daycare.

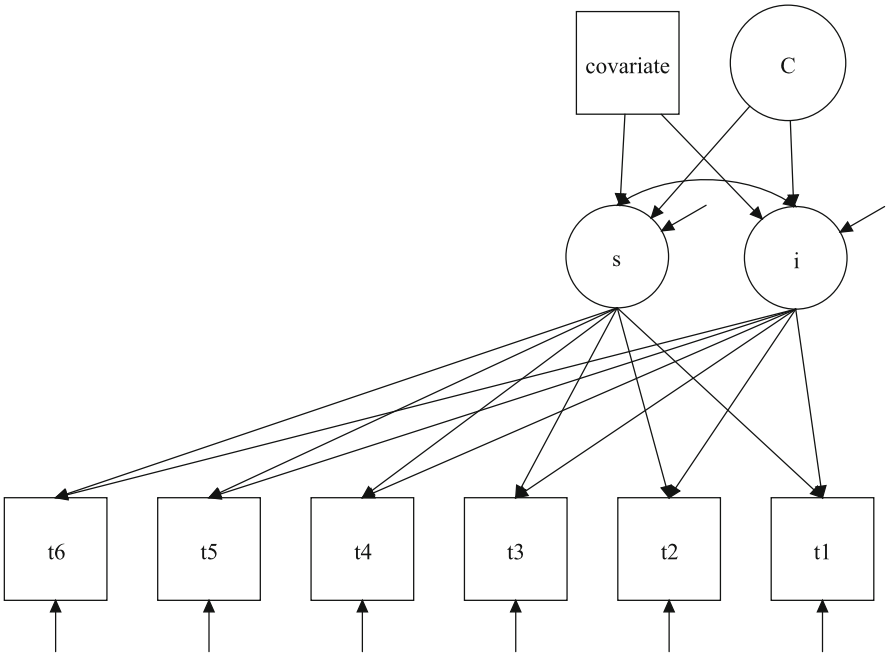


Fig. 1 Latent growth curve mixture model (LGMM) of relationship adjustment

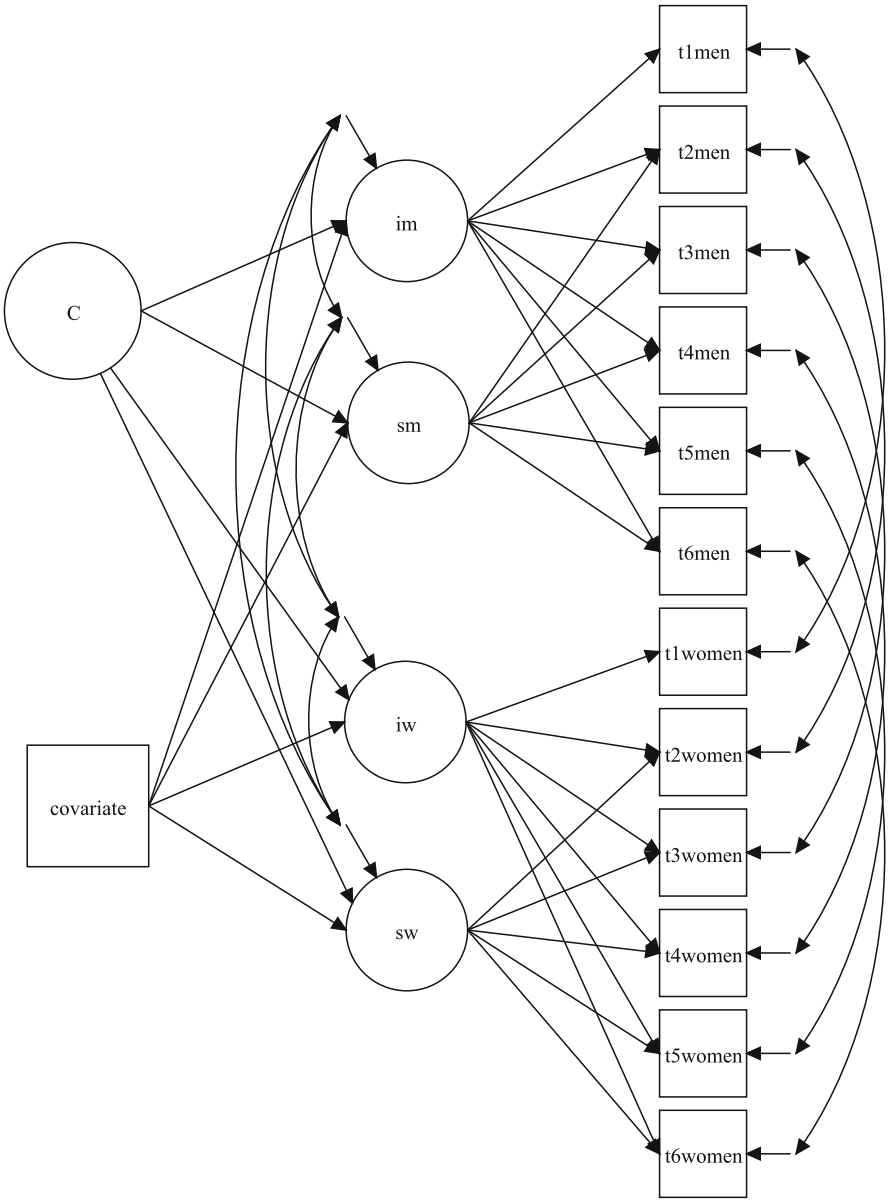


Fig. 2 Dual growth mixture model or relationship adjustment for men and women

The population response rate was 31 % ($N = 280$) of those invited to participate (Heinrichs et al. 2005), similar to other international prevention trials (Sanders 2012). Only parents who were in a committed relationship or married were eligible for the current study ($N = 242$). Although both partners were invited to participate,

more female partners agreed to participate ($N = 242$ women) than male partners ($N = 205$ men). Thus, $N = 205$ couples participated and $n = 37$ women without their partners. Due to missing data on the relationship satisfaction measure, the analytical sample consisted of $n = 237$ women and $n = 205$ men.

Participants were assessed six times over the course of the 4-year study (time 1, post-intervention (approximately 6 months following the initial assessment), and four additional times every 12 months after the time-1 assessment). Participant retention was excellent across all follow-ups (follow-up 1 = 99.2 %; follow-up 2 = 98.2 %; follow-up 3 = 96.3 %; follow-up 4 = 95 %). To account for the small amount of missing data across time, full information maximum likelihood estimation was used for all analyses. This study was approved by the university IRB board and informed consent was provided.

The mean age of the sample was 38.7 (6.0) years for men and 35.4 (4.7) years for women. The target child was 4.5 years old on average ($SD = 0.98$). The majority of the sample was married (88 %) and reported middle income (53 %, 1500–3000 Euros per month after taxes); 34 % reported income greater than 3000 Euros per month and 11 % of the sample reported income of less than 1500 Euros per month. Employment characteristics were as follows: full-time salaried position or self-employed 84.8 % men, 15.7 % women; part-time or paid by the hour = 7.4 % men, 46.3 % women; stay-at-home parent = 0.5 % men, 30.6 % women; unemployed = 3.2 % men, 2.9 % women; other = 4.2 % men, 4.5 % women.

In Germany, there are three levels of secondary education (high, middle, and low). Over half of the men and women (57 % and 58 %, respectively) had completed the high level (typically indicative of individuals who attend college); 20 % of the men and 33 % of the women completed the middle level (typically indicative of individuals who obtain some specialized training other than a bachelor's degree) and 14 % of the men and 10 % of the women reported the low level (typically indicative of individuals who do not complete high school). The number of children living in the household was $M = 2.1$ ($SD = .86$) on average.

Measures

Relationship Adjustment. The 7-item Abbreviated Dyadic Adjustment Scale (Sharpley & Rogers 1984) was used to assess relationship adjustment (3 items assessing topics of disagreements between partners, 3 items assessing frequency of positive exchanges, and 1 item assessing overall happiness). Items are scored on a Likert scale from 0 to 5, with higher scores indicating more relationship adjustment ($\alpha = .82$). Means and standard deviations of relationship adjustment for men, women, averaged across gender, and based on the worse score are presented in Table 1.

Analytical Strategy: Example 1

LGMM with Mplus 7 (Muthén & Muthén 2012) was used to examine the trajectories of relationship adjustment. To consider whether participation in the parenting program may have impacted relationship adjustment trajectory, we examined treatment group ($n = 133$ prevention group; $n = 109$ control group) as a predictor in all analyses. Prior to examining latent growth mixture trajectories for relationship adjustment, the unconditional model of growth (i.e., one class solution) was reviewed to provide an overall picture of average growth for this sample and overall model fit. Fit was evaluated based on chi-square values, Comparative Fit Index (CFI > .90), Tucker Lewis Index (TLI > .90), Root-mean square error of approximation (RMSEA < .06), and Standardized Root Mean Square Residuals (SRMR < .06).

Next, LGMM was used to detect subgroup trajectories of relationship adjustment. This approach allows for individual variability within classes; derived classes were free to differ on latent intercept and slope growth factors. Similar to previous studies, Akaike information criterion (AIC) and Bayesian information criterion (BIC) with lower values representing a better model fit were used to determine the number of classes that best fit the data. In addition, only class solutions that had an adequate amount of cases per class (>20) were retained.

Results: Example 1

To examine trajectories of relationship adjustment, we first examined change in mean scores over time. Intercept factor loadings were fixed at 1; slope factor loadings were 0 at the initial assessment, at .5 at post, and 1, 2, 3, 4 chronologically at each year follow-up. The model was a good fit for men ($N = 205$, $\chi^2(20) = 21.74$, $p = .35$, RMSEA = .02, CFI = 1.00, TLI = 1.0, SRMR = .04), for women ($N = 237$, $\chi^2(20) = 38.92$, $p = .01$, RMSEA = .06, CFI = .98, TLI = .97, SRMR = .03), and in the dual growth curve model ($N = 237$, $\chi^2(66) = 99.97$, $p = .00$, RMSEA = .06, CFI = .98, TLI = .98, SRMR = .04). Consistent across models, the slope was not significant for men ($bs = .10$–$.13$, $p > .05$) but was significant for women, such that their relationship adjustment increased over time ($bs = .50$–$.51$, $p < .05$). There was significant variance in both the intercept and slope factors for men and women, indicating there were individual differences in initial levels and rates of growth. In all models, intervention assignment was a covariate of intercepts and slopes. It did not significantly predict intercepts or men's slope but predicted slope for women such that those who received the parenting intervention showed less declines in relationship adjustment over time (e.g., Model women alone: slope $b = -.39$, $se = .15$, $t = -2.58$, $p = .01$).

Table 1 Descriptive statistics

	Relationship adjustment		Average score relationship adjustment	Worse score relationship adjustment	Depressive symptoms	
	Men	Women			Men	Women
	M (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)	*M* (SD)
T1	23.30 (4.97)	22.85 (5.11)	23.08 (4.73)	21.91 (5.01)	4.89 (5.53)	5.36 (5.71)
T2	23.21 (4.63)	23.46 (4.91)	23.35 (4.38)	22.20 (4.68)	4.55 (5.02)	5.14 (6.25)
T3	23.58 (4.72)	23.34 (5.13)	23.37 (4.78)	22.71 (5.01)	4.57 (5.36)	4.46 (5.87)
T4	23.56 (5.14)	22.95 (5.13)	23.18 (4.87)	22.50 (5.26)	4.18 (5.42)	4.34 (6.10)
T5	23.77 (5.23)	23.13 (5.14)	23.34 (5.14)	22.55 (5.69)	3.73 (4.40)	4.44 (5.88)
T6	23.23 (5.60)	22.99 (5.50)	23.00 (5.48)	22.22 (6.02)	4.32 (5.27)	4.34 (6.20)

$N = 205$ men. $N = 237$ women

Latent Growth Mixture Modeling

Next, LGMM was used to test the different approaches for grouping the relationship adjustment in comparison with the men only and women only models (see Figs. 1 and 2). Intervention assignment was included in all models as a covariate (see Figs. 1 and 2, "covariate"). Consistent with Foran and colleagues (2013), the two class solution was the best fit for all models tested based on our criteria (lowest AIC and BIC; more than 20 cases per class), and thus, only results from the two class model will be presented. The results of these models are presented in Table 2. See also Figs. 1 and 2 for the graphical representation of the models. Class 1 represented satisfied couples whose relationship adjustment remained high over time (labeled "*non-distressed*"). For all models, the slope was positive but it was only statistically significant for the women only model (1b), the DGMM (2), and in the worse score model (4). This represented between 87 and 93 % of participants across models (Model 1a: 89 %; Model 1b: 90 %; Model 2: 87 %; Model 3: 93 %; Model 4: 89 %). The second class represented couples with more relationship distress and a tendency to decline over time. Slopes for men and women in the distressed group were statistically significant in the men only, women only, and worse score models; they were not statistically significant in the DGMM or averaged models. This represented between 7 and 13 % of couples across models.

Overlap in classification of distressed couples. To further understand the differences across models, the correlations among the distressed class latent probabilities for each model are presented in Table 3. Although there was a high degree of overlap across models, the DGMM and men's models showed only a correlation of .33. The DGMM and women's models showed the highest correlation ($r = .89$), suggesting that the DGMM approach is more reflective of women's relationship adjustment than men's (and this was not explained by differences in missing data for men).

In addition, we examined the overlap between classes by seeing which models had higher false positives or false negatives. False positives were defined as a case

Table 2 Latent growth mixture models of relationship adjustment using different approaches

		Latent class prob.		Men non-distressed class		Men distressed class		Women non-distressed class		Women distressed class	
		Non-distressed	Distressed	I	S	I	S	I	S	I	S
1a	Men only	.97	.85	23.53	.27	21.90	−1.81***	—	—	—	—
1b	Women only	.97	.85	—	—	—	—	23.56	.58**	19.88	−1.45**
2	DGMM	.97	.89	23.43	.26	22.03	−.73	23.97	.70**	19.45	−.92
3	Averaged[a]	.97	.88	23.36	.38	21.32	−1.81	—	—	—	—
4	Worse score[a]	.07	.90	22.28	.61**	19.64	−1.64***	—	—	—	—

I = Intercept, S = mean slope, Latent class prob. = Average latent class probabilities for most likely class membership

[a]Indicates couple data. Unstandardized coefficients

Entropy was .83 for model 1a, .86 for Model 3, and .85 for all other models. N = 237 for all models except 1a ($n = 205$). Percentage in distressed class: Model 1a: 11 %, Model 1b: 10 %, Model 2: 13 %, Model 3: 7 %, Model 4: 11 %

*p < .01, ***p < .001

Table 3 Correlations between latent class probabilities in the distressed groups

	Women $N = 24$	DGMM $N = 31$	Average $N = 17$	Worse $N = 25$
Men	.41	.33	.71	.68
Women	–	.89	.83	.83
DGMM		–	.68	.75
Average			–	.82

being present in one model but not in any other models. There were $n = 45$ cases which were classified as distressed in any model. False positives were only detected for the DGMM (26 %) and men only model (27 %). To understand this false-positive rate for the DGMM, we compared the cases in which they were classified as distressed in the DGMM to all other cases classified as distressed. There should be no statistically significant differences among these distressed groups. However, the false positives in the DGMM were significantly higher on men's relationship adjustment than those classified as distressed across other groups, suggesting that they were indeed false positives (or at least not representative of men's relationship adjustment).

To explore the false-negative rate, we examined the detection of the $n = 31$ cases which were classified as distressed across two or more models. The false-negative rate (cases which were present in other models but rejected in that particular model) was highest for the men only model (54 %), followed by the average score model (45 %), dual score model (26 %), the women only model (23 %), and the worse score model (19 %).

Example 1: Discussion

Our goal in example 1 was to highlight the utility of the LGMM approach for differentiating distressed and non-distressed couples over time and highlight some of the different approaches for handling dyadic information. Although results were similar across one gender only, averaged score, DGMM, and worse score approaches, they were not identical. The DGMM model for example appeared to be more reflective of women's relationship adjustment (and this was not explained by missing data of men). In contrast, the worse score was more consistent in classifying "distressed" couples with the men and women only models. These differences highlight the importance of careful selection of dyadic model and consideration of what best maps the researchers construct and theory. If one takes the perspective that one distressed partner is sufficient to cause the relationship to be "distressed," then a worse score model may be a good option. This approach also corresponds to clinical models of relationship distress in which the DSM-5 diagnosis of an intimate partner relationship problem is defined based on one partner's clinical cut-off score on relationship adjustment measures (Foran, Whisman, & Beach 2015).

An interesting follow-up to this study is to determine whether the different ways of modeling the relationship adjustment and their respective latent classes differentially predict outcomes such as divorce. Some previous work has suggested that women are often the "emotional barometer" of relationships, which could lead to their report holding more weight (Gottman 1990); this could be evaluated further to determine which approach yields the most predictive validity.

In sum, LGMM is a methodology that fits well with theories of relationship adjustment, but more research on differences in handling the relationship adjustment scores is needed. Although this approach is useful for examining one variable such as relationship adjustment, there are constraints in integrating these types of models in multiple variable growth curve models. One approach is to use the latent class assignments or latent class probabilities to predict other outcomes, however some caution against using the probability information in such a way and concerns have also been expressed about the replicability of latent classes across studies (e.g., Bauer & Curran 2003; Nagin & Tremblay 2005).

An interesting integration that has rarely been used to-date and has not at all been used with dyadic data is combining LGMM with latent difference score modeling (LDS; also known as latent change score modeling; McArdle 2001; McArdle & Hamagami 2001). LDS modeling allows one to simultaneously examine change processes of two or more variables over time. In the next section, we apply LDS modeling to examine relationship adjustment and depressive symptoms over time. We then return to this issue of integration of the approaches illustrated in examples 1 and 2 at the end of the chapter.

Example 2: How Do Relationship Adjustment and Depressive Symptoms Relate Over Time?

Understanding the link between relationship adjustment and depressive symptoms (or depressive disorders) has been an active research question in the couples literature. Among women who had never experienced a depressive episode and who had a negative relationship event, 38 % developed a major depressive episode within the next 4 weeks after the event (Christian-Herman, O'Leary, & Avery-Leaf 2001) and this rate is significantly higher than incidence rates of approximately 2 % reported in epidemiological studies. Additional support for the role of relationship problems in depression onset comes from intervention studies that have demonstrated that treating relationship problem leads to reductions in depressive symptoms (see Beach 2001; Cohen, O'Leary, & Foran 2010), relationship distress moderates individual psychotherapy/psychopharmacological depression treatment outcome (Denton et al. 2010), and relationship distress predicts depression relapse (Hooley & Teasdale 1989).

In 2001, Whisman conducted a review of the association both cross-sectionally and longitudinally. Although there were numerous cross-sectional studies estab-

lishing the link, there was little longitudinal research at the time of the review. Of the six studies reviewed that examined the link between marital distress and depressive symptoms, three studies used residualized change analyses (regressions), two studies used structural equation modeling, and one study used HLM analyses. Although beyond the scope of this chapter, there were other studies which examined diagnostic depression (e.g., Gotlib, Lewinsohm, & Seeley 1998) and marital distress and depression in the context of treatment (e.g., Hooley & Teasdale 1989).

Another meta-analysis examined the link between relationship functioning and well-being broadly defined (Proulx, Helms, & Buehler 2007). Twenty-seven multiple time point studies were included in the meta-analysis. Unfortunately, the number of studies which specifically examined relationship adjustment and depressive symptoms was not reported, nor was the study sample sizes, number of time points, whether the analysis was dyadic or involved modeling each spouse data separately, or the type of statistical approach used for the analyses. Interestingly, the authors did find that longitudinal studies more recently found smaller effect sizes for well-being and relationship functioning compared to earlier studies. Although we can only speculate, this could be related to the different methods used in earlier studies compared to more recent studies.

Some of the best methodological work on trajectories of marital distress and depressive symptoms has been done by Karney, Bradbury and colleagues at UCLA. The authors collected two newlywed samples and followed them over 4 years with assessments at eight time points, using MLM (via HLM software) to examine marital trajectories. Attrition was much lower than typical in other studies (7 %; N = 60 initial sample; $N = 54$ analytical sample, see Karney & Bradbury 1997; 21 % study 2 N = 172, Davila, Karney, Hall, & Bradbury 2003). The authors found evidence that depressive symptoms and relationship distress covary over time (e.g., Davila, Bradbury, Cohan, & Tochluk, 1997; Davila et al. 2003; Karney 2001).

More recently, Kouros, Papp, and Cummings (2008) analyzed the association between depressive symptoms and relationship distress using three different methods with the same sample in three separate publications. The sample included $N = 296$ parents of 8–18-year olds followed over three time points (2 years). In the first paper (Kouros et al. 2008), the authors used multivariate HLM analyses to examine the reciprocal associations between depressive symptoms and marital distress. Results replicated the earlier findings of Davila et al. (2003) in which bidirectional within-person associations were also found with HLM analyses in a 4-year newlywed sample of 164 couples.

In a second reanalysis of these data, Kouros and Cummings (2010) examined dual growth curves of depressive symptoms for husbands and wives. The authors applied LDS models to look at dynamic coupling between spouses' depressive symptoms over the three time points (McArdle & Hamagami 2001). The authors then tested whether the growth curves of depressive symptoms were different for low or high maritally satisfied couples by conducting multi-group analyses. The martially distressed group included $n = 118$ couples and the martially satisfied group included $n = 178$ couples based on whether either partner reported scores below or above 100 on the Marital Adjustment Scale and the first time point,

respectively. The authors found that husbands' depressive symptoms were linked with changes in wives' depressive symptoms for the martially distressed group but not for the martially satisfied group.

In the third paper with this sample, Kouros and Cummings (2011) again applied LDS to understand depressive symptoms association with marital distress. This study differed from the 2010 study in that the two modeled growth curves were marital distress and depressive symptoms (rather than two growth curves for depressive symptoms of each partner). Thus, analyses in this last study were conducted separately by gender. To consider, cross-partner effects, the authors ran two additional models in which wives' marital satisfaction and husbands' depressive symptoms and husbands' marital satisfaction and wives' depressive symptoms were examined controlling for average scores on each variable for each spouse. Results suggested that women's marital satisfaction level predicted their depressive symptoms level, rather than depressive symptom change over time (Kouros & Cummings 2011). For men, marital satisfaction level predicted change in depressive symptoms.

Based on the current state of the literature, there are several relevant future directions. As noted, by Kouros and Cummings (2011), "methodological approaches based on HLM were limited in testing theoretical notions of how depressive symptoms and marital satisfaction simultaneously change and simultaneously predict change in each other over time." There is need for new longitudinal studies which take advantage of advances in latent growth curve modeling to test change processes. As far as we are aware, the Kouros & Cummings 2011 is the only paper that has used LDS modeling to address this research question.

A recent extension of LDS modeling in which previous latent changes in one variable predict subsequent latent changes in another variable (Grimm et al. 2012; see Fig. 3) has yet to be used and this approach may map more closely with theory than other approaches used. This extension differs from traditional LDS modeling in that previous change instead of previous level is used to predict future change in the other variable. The extension of Grimm et al. (2012) is shown in Fig. 3. The main changes from the traditional bivariate latent change score model are the additional paths between previous latent change and subsequent latent change (indicated by ϕ in Fig. 3) and the additional paths from previous latent change of X to subsequent latent change of Y (the coupling parameter ξ in Fig. 3).

There are many well developed theories of how relationship adjustment and depressive symptoms influence each other (see Beach 2001), but often there is a gap between the theory and the methodology used to test it and timing is not explicitly clarified. Based on the marital model of depression (Beach, Sandeen, & O'Leary 1990), one would expect that when relationship adjustment declines this leads to a simultaneous increase in depressive symptoms. This would be shown by levels of relationship adjustment and depressive symptoms covarying within time and slopes covarying across time. Thus, to detect change effects, shorter time periods (such as days or weeks) rather than months or years would be needed (e.g., Whitton, Stanley, Markman, & Baucom 2008). The most appropriate time frame to detect the theorized associations is often not given enough consideration in the literature.

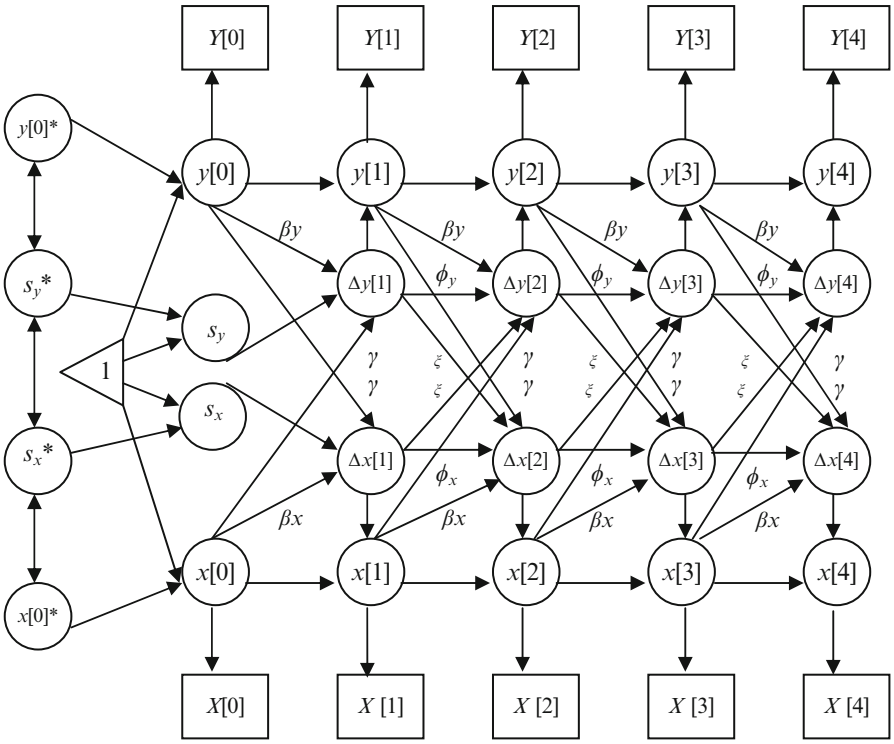


Fig. 3 "Changes to Changes" bivariate latent change score model extension (Grimm et al. 2012). See Grimm et al., 2012 for the full diagram with residual variances, covariances and paths from slopes to latent changes in *y* and *x*

In contrast, the effect of depressive symptoms on relationship adjustment change may be more delayed. The depressive behaviors of one partner over a period time may not result in immediate declines in relationship adjustment, but rather later declines in relationship adjustment. Depressive behaviors of one partner can be aversive to other partner, especially when attempts to improve the depressed partners' mood are ineffective over time. This type of association would be detected with LDS models of previous change in depressive symptoms predicting subsequent change in relationship adjustment. Notably, daily relationship satisfaction may change in response to daily changes in mood, but relationship adjustment (a broader construct evaluating the relationship overall) would be less susceptible to daily changes in mood or brief increases in depressive symptoms.

Thus, different time frames would be needed to test the bidirectional associations between relationship adjustment and depressive symptoms. Although it is beyond the scope of this chapter to fully test these differences, we would like to present an example of how longer term associations (over 4 years) between relationship adjustment and depressive symptoms can be tested with LDS modeling and how

this approach helps differentiate types of change processes that can be tested. We have chosen to briefly illustrate this new extension of LDS modeling since it matches particularly well with theory and to encourage researchers in this area to consider these approaches in their future work.

Method: Example 2

The same sample as in example 1 is used for the analyses of example 2. Measures included relationship adjustment (described above in example 1) and a measure of depressive symptoms. Depressive symptoms were assessed with the widely used 14-item depression subscale of the Depression Anxiety Stress Scale (Lovibond & Lovibond 1993) scored from *never* = 1 to *very often* = 4 (αs = .93 for men and women). Means and standard deviations for relationship adjustment and depressive symptoms are presented in Table 1.

Analytical Strategy: Example 2

Univariate LDS models have been reviewed in many places previously (see Grimm et al. 2012; McArdle 2009), and thus, we present only the results from the second step of our analyses, the bivariate LDS models. Eight bivariate LDS models were tested (see Grimm et al. 2012, for specific details on these models as well as sample syntax for setting these models up in Mplus). The first four models imply testing the traditional LDS models (Model 1: no coupling, Model 2: level of relationship adjustment to latent change in subsequent depressive symptoms, Model 3: level of depressive symptoms to subsequent latent change in relationship adjustment, and Model 4: bidirectional coupling model with paths in both directions). The next four models test the "changes to changes" components. This includes an additional LDS variable for each growth curve that indicates the change from time $t-2$ to $t-1$. The "no coupling" model (Model 5) yields a regression coefficient for each growth curve that describes the effect of previous change in that variable on subsequent changes (e.g., previous change in depressive symptoms predicting subsequent change in depressive symptoms). The next two models test the unidirectional coupling parameters in which in one model (Model 6) previous change in relationship adjustment predicts subsequent change in depressive symptoms and in the next model (Model 7) previous change in depressive symptoms predicts subsequent change in relationship adjustment. The final model (Model 8) is the full model in which both directions of change on changes are included. Note that Model 8 is similar to the Grimm et al. (2012) model shown in Fig. 3 but we had also included covariates in our models (intervention assignment, initial scores of spouses' relationship adjustment and depressive symptoms).

Model fit was compared using AIC, BIC, and sample size adjusted BIC with lower values indicating better model fit. In addition, given that all the bivariate LDS models were nested, model fit can also be compared using the difference in −2log-likelihood related to change in parameters. Full information maximum likelihood (FIML) estimation with robust standard errors was used to account for missing data and adjust for non-normality in the data.

Similar to Kouros and Cummings (2011), we choose to analyze men's and women's parallel growth curves for relationship adjustment and depressive symptoms separately. Time-1 levels of partner's relationship adjustment and depressive symptoms were included as covariates. Although there are some applications in which four process growth curves have begun to be used (Hoppman, Gerstorf, & Hibbert 2011), our sample size was not adequate for such analyses, an issue we return to more fully in the discussion, as this represents one of the constraints in modeling dyadic data with repeated measures.

Results: Example 2

Results of the bivariate LDS models for women indicated that levels of relationship adjustment and depressive symptoms did not predict change in the other variable; this is consistent with earlier analyses with these data in which latent class of relationship distress or initial levels of relationship distress did not predict slope of depressive symptoms for women (Foran, Hahlweg et al. 2013). Thus, we focus on only results for men in the following section.

Results of the bivariate LDS models for men are provided in Table 4. Of the eight models tested, the bidirectional coupling changes on changes model was the best fit in terms of the lowest AIC, BIC, and adjusted BIC values (Model 8). A Satorra–Bentler scale chi-square difference tests indicated that this model fit significantly better than all other models except that the difference between model 7 and 8 was not statistically significant (Satorra–Bentler scale: $\triangle\chi^2 = 1.49$, $p = .23$). Thus, for parsimony, model 7 was the selected model. Further, the additional path of change in relationship adjustment to change in depressive symptoms that differentiated models 7 and 8 was not statistically significant.

Model 7 parameter estimates are presented in Table 5. Higher previous levels of depressive symptoms (parameter β), lower previous levels of relationship adjustment (coupling parameter γ), and more previous decreases in depressive symptoms (parameter ϕ) lead to more subsequent decreases in depressive symptoms. Changes in relationship adjustment were accounted for by previous changes in relationship adjustment (parameter ϕ) as well as previous changes in depressive symptoms (coupling parameter ξ). In other words, of most interest, the results show that previous changes in depressive symptoms predicted subsequent changes in relationship adjustment such that if depressive symptoms increased, then subsequent relationship adjustment would decrease, consistent with theoretical expectations.

Table 4 Bivariate LDS score modeling of relationship adjustment and depressive symptoms over 4 years

	No coupling	Relationship adjustment → Δ Depression	Depression → Δ Relationship adjustment	Bidirectional Coupling	No coupling change on change but prior changes on change	Δ Relationship adjustment → Δ Depression	Δ Depression → Δ Relationship adjustment[a]	Bidirectional Coupling Changes on Changes
Model number	1	2	3	4	5	6	7	8
Parameters	39	40	41	41	43	44	44	45
−2 log-likelihood	11,952.60	11,952.56	11,952.42	11,952.38	11,945.98	11,939.12	11,934.48	11,927.72
AIC	12,030.59	12,032.56	12,032.42	12,034.39	12,031.98	12,027.19	12,022.47	12,017.71
BIC	12,159.82	12,165.09	12,164.95	12,170.23	12,174.45	12,172.97	12,168.25	12,166.80
Adjusted BIC	12,036.24	12,038.36	12,038.22	12,040.33	12,038.22	12,033.57	12,028.85	12,024.23

[a]Selected model. N = 203. Dual Growth curves of Relationship Adjustment and Depressive Symptoms for men; Controlling for pre values for women

Table 5 Bivariate LDS score Model 7 parameters for men's relationship adjustment and depressive symptoms

	Change relationship adjustment	Change in depressive symptoms
	Parameter estimate (SE)	Parameter estimate (SE)
Mean intercept	8.33 (1.59)***	6.60 (2.17)**
Mean slope	−7.14 (5.36)	−.70 (1.21)
β level to change (within same variable)	.02 (.01)	−.25 (.12)*
γ coupling parameter (across variables)	1.17 (.65)	.27 (.10)**
ϕ change on change (within same variable)	−.65 (.24)**	1.74 (.63)**
ξ change on change coupling parameter (across variables)	−3.34 (1.26)**	−

Unstandardized coefficient

General Discussion

The results of example 2 highlight the utility of this extension LDS modeling for understanding change processes of relationship adjustment and depressive symptoms, particularly for men. Including the additional parameters in which previous changes in depressive symptoms were modeled to predict subsequent changes in relationship adjustment allowed us to test our theoretical expectation of time-lagged effects of depressive symptom change on relationship adjustment change. We did not find support for time-lagged effects for women. Women may be more reactive to bidirectional changes in relationship adjustment and depressive symptoms and shorter lags may be needed to see these effects.

This represents an extension of LDS modeling to test theoretical change processes. The most important consideration is that the approach selected is a match with the purported change processes being tested. At least in the couple research field, and we expect in many other fields, this particular consideration does not receive enough attention. An important related issue is timing of measurement in relation to change. Many longitudinal designs select equal interval time frames that may not provide the appropriate window to a change process. In our example of depressive symptoms and relationship adjustment, shorter time frames may be needed to fully capture some of the purported effects of changes in relationship adjustment and satisfaction on mood symptoms (e.g., Whitton et al. 2008), and this may help explain some of the discrepant findings across studies for men and women.

Across examples 1 and 2, various constraints in handling the dyadic nature of the data were encountered. In example 1, we illustrated how results may vary

depending on the way that the dyadic data are aggregated in a growth mixture model. In example 2, the focus was on illustrating a two-growth process model, which did not allow us to model partners' growths simultaneously. Partner scores were incorporated into the model (similar to a basic APIM model), but dual processes for men and women were not explored in the same model. One alternative is to test a four-growth process LDS model. As far as we are aware, there has only been a limited number of applications of a four-variable model for traditional latent growth models (see Hoppman et al. 2011) and little work in the context of LDS (Gerstorf, Hoppmann, Kadlec, & McArdle 2009). Constraints of application in longitudinal dyadic studies include the sample size needed and difficulties in the interpretation of a four-growth curve LDS model.

Many other future options for better integration of methods exist in longitudinal analysis of couples. LGMM could be combined with LDS modeling within the same model, but there are few applications that integrate these two approaches. The simplest way to integrate these two would be to proceed in two steps in which the latent classes for the variable of interest were derived and then could be integrated as subsequent predictors of the second step in which LDS modeling is used to identify change relations of two other variables. However, one would have to show that the derived latent classes provided more information than would be the case were the full growth model included in the model, and this may not be the case in many situations. Thus, in many cases it may be better to use the LGMM or exploratory growth curve analyses to describe development processes as a separate approach (e.g., Grimm, Steele, Ram, & Nesselroade 2013). Bivariate LDS then provide a good fit for testing theorized interrelations between variables over time.

In sum, new longitudinal models proposed over the last two decades offer a robust set of tools for dyadic analyses. However, constraints in terms of longitudinal sample sizes, numbers of growth curves that can easily be tested simultaneously, model complexity, and timing of assessments remain. In addition, although methods for accounting for dyadic independence exist and have been widely applied in the couple field, modeling barriers as well as theoretical controversies on whether to consider a variable dyadic or individual still need more in-depth consideration.

Acknowledgement Preparation of this chapter was supported by a grant from the German Research Foundation awarded to Heather Foran (FO788/1-2).

References

Amato, P. R. (2010). Research on divorce: Continuing trends and new developments. *Journal of Marriage and the Family, 72*, 650–666. doi:10.1111/j.1741-3737.2010.00723.x.

Atkins, D. C. (2005). Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. *Journal of Family Psychology, 19*, 98–110.

Barnett, R. C., Marshall, N. L., Raudenbush, S. W., & Brennan, R. T. (1993). Gender and the relationship between job experiences and psychological distress: A study of dual-earner couples. *Journal of Personality & Social Psychology, 64*, 794–806.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*, 135–167.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes (with discussion). *Psychological Methods, 8*, 338–363.

Beach, S. R. H. (2001). Marital therapy for co-occurring marital discord and depression. In S. R. H. Beach (Ed.), *Marital and family processes in depression: A scientific foundation for clinical practice* (pp. 205–224). Washington, DC: American Psychological Association.

Beach, S. R. H., Sandeen, E. E., & O'Leary, K. D. (1990). *Depression in marriage: A model for etiology and treatment*. New York: Guilford.

Beach, S. R. H., & Whisman, M. A. (2013). Relationship distress: Impact on mental illness, physical health, children and family economics. In H. M. Foran, S. R. H. Beach, A. M. S. Slep, R. E. Heyman, M. Z. Wamboldt, N. Kaslow, & D. Reiss (Eds.), *Family problems and family violence: Reliable assessment and the ICD-11*. New York: Springer.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (1996). *HLM 4 for Windows [Computer software]*. Chicago: Scientific Software International.

Chou, C., Bentler, P. M., & Pentz, M. (1998). A comparison of two statistical approaches to study growth curves: The multilevel model and latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*, 247–266.

Christian-Herman, J. L., O'Leary, K. D., & Avery-Leaf, S. (2001). The impact of severe negative events in marriage on depression. *Journal of Social and Clinical Psychology, 20*, 24–40.

Cohen, S., O'Leary, K. D., & Foran, H. M. (2010). A randomized clinical trial of a brief, problem-focused couple therapy for depression. *Behavior Therapy, 41*, 433–446. doi:10.1016/j.beth.2009.111.004.

Collins, L. M., & Sayer, A. G. (2001). *New methods for the analysis of change*. Washington, DC: American Psychological Association.

Cook, W. L., & Kenny, D. A. (2005). The actor-partner interdependence model: A model of bidirectional effects in developmental studies. *International Journal of Behavioral Development, 29*, 101–109.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*(4), 529–569.

Davila, J., Bradbury, T. N., Cohan, C. L., & Tochluk, S. (1997). Marital functioning and depressive symptoms: evidence for a stress generation model. *Journal of Personality & Social Psychology, 73*, 849–861.

Davila, J., Karney, B. R., Hall, T. W., & Bradbury, T. N. (2003). Depressive symptoms and marital satisfaction: Within-subject associations and the moderating effects of gender and neuroticism. *Journal of Family Psychology, 17*, 557–570.

Denton, W. H., Carmody, T. J., Rush, A. J., Thase, M. E., Trivedi, M. H., Arnow, B. A., et al. (2010). Dyadic discord at baseline is associated with lack of remission in the acute treatment of chronic depression. *Psychological Medicine, 40*, 415–424.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Okut, H., & Li, F. (2002). *Growth mixture modeling of adolescent alcohol use data*. Eugene, OR: Oregon Research Institute.

Ferrer, E., & Nesselroade, J. R. (2003). Modeling affective processes in dyadic relations via dynamic factor analysis. *Emotion, 3*, 344–360.

Ferrer, E., Steele, J., & Hsieh, F. (2012). Analyzing dynamics of affective dyadic interactions using patterns of intra- and inter-individual variability. *Multivariate Behavioral Research, 47*, 136–171.

Ferrer, E., & Widaman, K. F. (2008). Dynamic factor analysis of dyadic affective processes with inter-group differences. In N. A. Card, J. P. Selig, & T. D. Little (Eds.), *Modeling dyadic and interdependent data in the developmental and behavioral sciences* (pp. 107–137). Hillsdale, NJ: Psychology Press.

Fincham, F. D., & Beach, S. R. H. (2010). Marriage in the new millennium: A decade in review. *Journal of Marriage and Family, 72*, 630–649.

Foran, H. M., Hahlweg, K., Kliem, S., & O'Leary, K. D. (2013). Longitudinal patterns of relationship adjustment among German parents. *Journal of Family Psychology, 27*, 838–843. doi:10.1037/a0034183.

Foran, H. M., O'Leary, K. D., & Slep, A. M. S. (2013). *Course and predictors of relationship distress among parents with young children*. Unpublished manuscript, Technical University of Braunschweig, Braunschweig, Germany.

Foran, H. M., Whisman, M. A., & Beach, S. R. H. (2015). Intimate partner relationship distress in the DSM-5. *Family Process, 54*, 48.

Gerstorf, D., Hoppmann, C. A., Kadlec, K. M., & McArdle, J. J. (2009). Memory and depressive symptoms are dynamically linked among married couples: Longitudinal evidence from the AHEAD study. *Developmental Psychopathology, 45*, 1595–1610.

Gotlib, I. H., Lewinsohm, P. M., & Seeley, J. R. (1998). Consequences of depression during adolescence: Marital status and marital functioning in early adulthood. *Journal of Abnormal Psychology, 4*, 686–690.

Gottman, J. M. (1990). How marriages change. In G. R. Patterson (Ed.), *New directions in family research: Depression and aggression* (pp. 75–101). Hillsdale, NJ: Lawrence Erlbaum.

Griffin, D., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in exchangeable case. *Psychological Bulletin, 118*, 430–439.

Grimm, K. J., An, Y., & McArdle, J. J. (2012). Recent changes leading to subsequent changes: Extensions of multivariate latent difference score models. *Structural Equation Modeling, 19*, 268–292.

Grimm, K. J., Steele, J. S., Ram, N., & Nesselroade, J. R. (2013). Exploratory latent growth models in the structural equation modeling framework. *Structural Equation Modeling, 20*, 568–591.

Heinrichs, N., Bertram, H., Kuschel, A., & Hahlweg, K. (2005). Parent recruitment and retention in a universal prevention program for child behavior and emotional problems: Barriers to research and program participation. *Prevention Science, 6*, 275–286. doi:10.1007/s11121-005-0006-1.

Hooley, J. M., & Teasdale, J. D. (1989). Predictors of relapse in unipolar depressives: Expressed emotion, marital distress, and perceived criticism. *Journal of Abnormal Psychology, 98*, 229–235.

Hoppman, C. A., Gerstorf, D., & Hibbert, A. (2011). Spousal associations between functional limitation and depressive symptom trajectories: Longitudinal findings from the study of Assets and Health Dynamics among the Oldest Old (AHEAD). *Health Psychology, 30*, 153–162.

Hox, J., & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1296–1305). Chichester: John Wiley & Sons.

Hoyle, R. H., & Gottfredson, N. C. (2015). Sample size considerations in prevention research applications of multilevel modeling and structural equation modeling. *Prevention Science,* 1–10. doi: 10.1007/s11121-014-0489-8

Karney, B. R. (2001). Depressive symptoms and marital satisfaction in the early years of marriage: Narrowing the gap between theory and research. In S. R. H. Beach (Ed.), *Marital and family processes in depression: A scientific foundation for clinical practice* (pp. 45–68). Washington, DC: American Psychological Association.

Karney, B. R., & Bradbury, T. N. (1995a). The longitudinal course of marital quality and stability: A review of theory, method, and research. *Psychological Bulletin, 118*, 3–34.

Karney, B. R., & Bradbury, T. N. (1995b). Assessing longitudinal change in marriage: An introduction to the analysis of growth curves. *Journal of Marriage and the Family, 57*, 1091–1108.

Karney, B. R., & Bradbury, T. N. (1997). Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology, 72*, 1075–1092.

Kashy, D. A., & Donnellan, M. B. (2008). Comparing MLM and SEM approaches to analyzing developmental dyadic data: Growth curve models of hostility in families. In N. A. Card, J. P. Selig, & T. D. Little (Eds.), *Modeling dyadic and interdependent data in the developmental and behavioral sciences* (pp. 165–190). New York: Routledge.

Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships, 2*, 67–75.

Kenny, D. A. (1996). Models of non-independence in dyadic research. *Journal of Social and Personal Relationships, 13*, 279–294.

Kenny, D. A., & Cook, W. L. (1999). Partner effects in relationship research: Conceptual issues, analytic difficulties, and illustrations. *Personal Relationships, 6*, 433–448. doi:10.1111/j.1475-6811.1999.tb00202.x.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford.

Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 18, pp. 141–182). New York: Academic.

Kouros, C. D., & Cummings, E. M. (2010). Longitudinal associations between husbands' and wives' depressive symptoms. *Journal of Marriage and Family, 72*, 135–147.

Kouros, C. D., & Cummings, E. M. (2011). Transactional relations between marital functioning and depressive symptoms. *American Journal of Orthopsychiatry, 81*, 128–138.

Kouros, C. D., Papp, L. M., & Cummings, E. M. (2008). Interrelations and moderators of longitudinal links between marital satisfaction and depressive symptoms among couples in established relationships. *Journal of Family Psychology, 22*, 667–677.

Kurdek, L. A. (1998). The nature and predictors of the trajectory of change in marital quality over the first 4 years of marriage for first-married husbands and wives. *Journal of Family Psychology, 12*, 494–510. doi:10.1037/0893-3200.12.4.494.

Kurdek, L. A. (2005). Gender and marital satisfaction in early marriage: A growth curve approach. *Journal of Marriage & Family, 67*, 68–84.

Lavner, J. A., Bradbury, T. N., & Karney, B. R. (2012). Incremental change or initial differences? Testing two models of marital deterioration. *Journal of Family Psychology, 26*, 606–616. doi:10.1037/a0029052.

Ledermann, T., & Kenny, D. A. (2012). The common fate model for dyadic data: Variations of a theoretically important but underutilized model. *Journal of Family Psychology, 26*(1), 140–148.

Ledermann, T., Macho, S., & Kenny, D. A. (2011). Assessing mediation in dyadic data using the actor-partner interdependence model. *Structural Equation Modeling: A Multidisciplinary Journal, 18*, 585–612.

Little, T. D., Schnabel, K. U., & Baumert, J. (2000). *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples*. Mahwah, NJ: Lawrence Erlbaum Associates.

Loeys, T., & Molenberghs, G. (2013). Modeling actor and partner effects in dyadic data when outcomes are categorical. *Psychological Methods, 18*, 220–236.

Lovibond, S. H., & Lovibond, P. F. (1993). *Manual for the Depression Anxiety Stress Scales (DASS). Kensington, NSW, Australia: University of New South Wales*. Mahwah, NJ: Erlbaum.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*, 86–92.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research, 32*, 215–253.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605.

McArdle, J. J., & Hamagami, F. (2001). Linear dynamic analyses of incomplete longitudinal data. In L. Collins & A. Sayer (Eds.), *Methods for the analysis of change* (pp. 137–176). Washington, DC: APA Press.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*, 23–43.

Muthén, B. O., & Muthén, L. K. (2012). *Mplus 7 base program*. Los Angeles: Muthén & Muthén.

Nagin, D. S., & Tremblay, R. E. (2005). Developmental trajectory groups: Facts or a useful statistical fiction? *Criminology: An Interdisciplinary Journal, 43*, 873–904.

Newsom, J. T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling, 9*, 431–447.

Peugh, J. L., DiLillo, D., & Panuzio, J. (2013). Analyzing mixed-dyadic data using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 314–337.

Proulx, C. M., Helms, H. M., & Buehler, C. (2007). Marital quality and personal well-being: A meta-analysis. *Journal of Marriage & Family, 69*, 576–593.

Raudenbush, S. W., Brennan, R. T., & Barnett, R. C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *Journal of Family Psychology, 9*, 161–174.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2000). *HLM 5 for Windows [Computer software]*. Skokie, IL: Scientific Software International.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90*, 726–748.

Sanders, M. R. (2012). Development, evaluation, and multinational dissemination of the Triple P-Positive Parenting Program. *Annual Review of Clinical Psychology, 8*, 345–379. doi:10.1146/annurev-clinpsy-032511-143104.

Sbarra, D. A., Law, R. W., & Portley, R. M. (2011). Divorce and death: A meta-analysis and research agenda for clinical, social, and health psychology. *Perspectives on Psychological Science, 6*, 454–474. doi:10.1177/1745691611414724.

Sharpley, C. F., & Rogers, H. J. (1984). Preliminary validation of the Abbreviated Spanier Dyadic Adjustment Scale: Some psychometric data regarding a screening test of marital adjustment. *Educational and Psychological Measurement, 44*, 1045–1049. doi:10.1177/0013164484444029.

Song, H., & Ferrer, E. (2009). State-space modeling of dynamic psychological processes via the Kalman Smoother algorithm: Rationale, finite sample properties, and applications. *Structural Equation Modeling, 16*, 338–363.

Song, H., & Ferrer, E. (2012). Bayesian estimation of random coefficient dynamic factor models. *Multivariate Behavioral Research, 47*, 26–60.

Steele, J., & Ferrer, E. (2011). Latent differential equation modeling of self-regulatory and coregulatory affective processes. *Multivariate Behavioral Research, 46*, 956–984.

Wendorf, C. A. (2002). Comparisons of structural equation modeling and hierarchical linear modeling approaches to couples' data. *Structural Equation Modeling, 9*, 126–140.

Whisman, M. A. (2001). The association between depression and marital dissatisfaction. In S. R. H. Beach (Ed.), *Marital and family processes in depression* (pp. 3–24). Washington, DC: American Psychological Association.

Whitton, S. W., Stanley, S. M., Markman, H. J., & Baucom, B. R. (2008). Women's weekly relationship functioning and depressive symptoms. *Personal Relationships, 15*, 533–550.

Wu, W., Selig, J. P., & Little, T. D. (2012). Longitudinal models. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 387–410). New York: OUP.

Can Psychometric Measurement Models Inform Behavior Genetic Models? A Bayesian Model Comparison Approach

Ting Wang, Phillip K. Wood, and Andrew C. Heath

As methodologists have increasingly noted, the role of psychometrics in operationalizing a construct is often overlooked when evaluating research claims (Borsboom 2006). In a related vein, others have noted that psychological research appears to move away from assessment and interpretation of a single a priori statistical model to a more nuanced comparison of models which assess the trade-off between a model's parsimony and complexity in explaining behavior (e.g., Rodgers 2010). The genetic factor model is one such statistical model often used to estimate the relative contributions of genetic and environmental components of observed behavior in genetically informative designs (Heath, Neale, Hewitt, Eaves, & Fulker 1989; Martin & Eaves 1977; Neale & Cardon 1992). Mathematically, the genetic factor model decomposes observed phenotypic variability into additive genetic (A), common (C), and unique (E) environmental components and is, for that reason, often referred to as the ACE model.

Recently, Franić et al. (2013) discussed how the genetic factor model can be used in the service of psychometrics by informing researchers about the different patterns of dimensionality and factor structure associated with genetic and environmental components of the ACE model. They note that adjudication of dimensionality is obviously not possible based on phenotypic factor analysis which does not take into account the genetically informative nature of the data. In their paper, Franić et al. propose conducting a Cholesky decomposition for the genetic and

Electronic supplementary material: The online version of this chapter (doi: 10.1007/978-3-319-20585-4_10) contains supplementary material, which is available to authorized users.

T. Wang • P.K. Wood (✉) • A.C. Heath
Department of Psychological Sciences, University of Missouri, Columbia, MO, USA
e-mail: phillipkwood@gmail.com

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_10

environmental components of the ACE model and decide on dimensionality of each of the environmental and genetic components of the model. They then rotate this solution to a more substantively meaningful form using Promax rotation.

Absent a strong a priori rationale for the factor structure of the environmental and genetic components of the ACE model, this approach appears reasonable and reflects the general practice of behavior genetic models which involve several items (e.g., Heath, Eaves, & Martin 1989; Heath, Jardin, Eaves, & Martin 1989), repeated measurements across successive occasions (e.g., Chang, Lichtenstein, Asherson, & Larsson 2013; Roberson-Nay et al. 2013), and multivariate studies of simultaneously measured variables in which some rationally defined order or priority exists across the manifest variables (e.g., Ludeke, Johnson, & Bouchard 2013). It is well appreciated that such Cholesky decompositions are not unique and that models consisting of other triangular orderings, models with common factors and residual subfactors, or autoregressive factors may fit such data equally well (Loehlin 1996). Rotation of initial Cholesky factorization to more conceptually meaningful form such as simple structure is also a reasonable procedure (e.g., Carey & DiLalla 1994).

Although the strategy outlined by Franić et al. is quite promising, the present paper proposes four reasons why a more fine-grained Bayesian psychometric approach may prove useful. First, for reasons discussed below, multifactor ACE models sometimes encounter empirical under-identification problems. Second, in some research contexts (such as, for example, the genetic analysis of body mass index data considered below where a variety of ages are considered but for which any one individual is assessed at multiple, but not all, measurement occasions), Cholesky factorization across all measurement occasions is not mathematically possible. Third, rotation of the identified solution to simple structure and the original Cholesky decomposition may obscure the psychometric measurement model underlying the construct of interest. Finally, there is reason to believe that Bayesian estimation may be preferable to ML or eigenvalue decomposition. This is particularly the case when sample sizes are small (Boomsma 1982; Chou, Bentler, & Satorra 1991; Hoogland & Boomsma 1998; Hu, Bentler, & Kano 1992; Lee & Song 2004). Additionally, Carey, Goldsmith, Tellegen, and Gottesman (1978) speculate that discrepant estimates of genetic and environmental effects in personality and psychiatric traits may be due to over-extraction of factors or to factors which describe weak effects which limit the generalizability of exploratory factor loadings in the ACE model. Again, these concerns are not meant to criticize the general approach outlined by Franić et al., but instead to highlight that refining the set of candidate psychometric measurement models provide researchers with models which may not be immediately obvious in some situations or estimable in other contexts.

Empirical Under-Identification

The issue of empirical under-identification is not unique to the estimation of genetic models and can occur when researchers attempt to fit a factor model which is more complex than the true model which generated the data, when small sample sizes are examined, and when the factor loadings of the model describe weak or non-existent effects (Kenny, Kashy, & Bolger 1998; Kenny & Milan 2013). Within genetic factor models, the problem of empirical under-identification manifests itself in convergence failures or improper solutions (such as negative variance estimates or estimation correlations which exceed one; see e.g., Phillips & Matheny 1997). Rietveld, Posthuma, Dolan, and Boomsma (2003) discuss the identification issue as it bears on the statistical power of a given genetic model, noting that a given behavior genetic model is mathematically identified if and only if the null space of the Jacobian is zero (i.e., has full column rank). This is, however, only a necessary but not sufficient condition for a specific model within the context of a particular data set.

As Kenny and Milan (2013) note, researchers who encounter empirical under-identification problems usually make post-hoc changes to the model such as redacting individual parameters thought superfluous or adding indicator variables to improve the resolution of the factor structure or instrumental variables which help resolve erroneously specified directions of causality in the model. Researchers using genetic models often constrain parameters of the model to equality or set other parameters to zero (Henderson 1982). Other strategies have included reducing the number of factors considered due to the presumed lack of statistical power associated with the sample (e.g., Martin, Scourfield, & McGuffin 2002). Rietveld et al. (2003) have noted that this state of affairs can be somewhat confusing given that at times researchers have claimed particular genetic models are over-parameterized and not identified while others have investigated the model and found this not to be the case.

Measurement Models

That notions of strictly parallel, tau-equivalent, and congeneric measurement models can be expressed as structural equation models has been noted since Lord, Novick, and Birnbaum's (1968) classical test theory text. In the case of measurement equivalence across a set of manifest variables, strictly parallel measurement requires that both error variances and factor loadings are identical for all variables. Tau equivalence, by contrast, assumes only that the loadings are identical and congeneric measurement permits the factor loadings and error variances across items to be different. Mathematically identified exploratory factor models correspond to a congeneric measurement model, while the tau equivalent model constitutes a more parsimonious model because the loadings across manifest variables are constrained to equality.

In other cases, however, a behavioral or genetic component may be poorly represented by a single congeneric factor, requiring more complex measurement alternatives. Although in many situations multiple oblique or orthogonal factors may be appropriate, measurement models which are intermediary between the one- and two-factor models may be appropriate in other situations. The random intercept model (Maydeu-Olivares & Coffman 2006) is one such model, consisting of both a freely estimated factor and an orthogonal general random intercept factor. The interpretational status of the random intercept factor depends on the particular constructs under investigation: Maydeu-Olivares and Coffman, for example, interpreted the random intercept factor they found in questionnaire data as a general response bias method factor and interpreted the remaining congeneric factor as the construct of interest. When the manifest variables under consideration consist of repeated measurements of the same variable, the factor pattern of the random intercept factor model corresponds to those which would be observed under the free basis growth curve model of Meredith and Tisak (1990). The random intercept factor model differs from the free basis model only in that the random intercept model estimates separate intercepts for each manifest variable and assumes that the latent variables of interest have a zero mean, while the growth curve model assumes that such intercepts are constrained to zero and mean levels of the manifest variables are explained by estimated latent variable means. Taken together, the tau equivalent, congeneric, and random intercept factor models constitute a more fine-grained set of measurement models which are simpler (in the case of the tau-equivalent model) or intermediate models between the dimensions considered under traditional factor analytic models. It is hoped that such a process will result in a "right-sizing" of the statistical model which will result in models which are easier to fit and may well be more generalizable across replications.

Specifically, we speculate that the standard single-factor model may be an over-complex measurement model when effects are relatively week. Specifically, estimation of the distinct individual loadings of the common factor model assumes a congeneric measurement model for a particular genetic or environmental component while the tau-equivalent measurement model which constrains loadings to be equal across variables may be more appropriate. Mathematical derivations (Davis-Stober 2011) also support the idea that predictor weights in the general linear model fail to replicate across samples because of just such over-complexity. This effect is found to be especially true when the sample size is small ($N < 150$) and the effect size of interest is moderate or small (R^2 is smaller than 0.6). Since the measurement model of factor analysis is a type of regression as well (although admittedly one in which the predictor variable for all observed variables is missing), it seems reasonable that similar difficulties in generalizability would be found. Although the sample sizes for behavior genetic studies are frequently quite large, in some contexts (such as the assessment of multiple cohorts of twins measured prospectively), the sample sizes associated with the data in some contexts may be rather small and comparable to the values considered by Davis-Stober. Because phenotypical behavior is frequently thought to entail expression of multiple genes, with each gene exhibiting only a small unique effect (Joseph & Ratner 2013; Turkheimer 2000), the effect sizes of

interest may well fall into the "moderate to small" criterion considered by Davis-Stober. In any event, exploration of genetic and environmental components under a tau equivalent model may provide a useful parsimonious comparison to the estimates from the congeneric factor model.

Robustness to Small Sample Size and/or Small Experimental Effects

Finally, as reviewed above, there is some reason to believe that exploration of more parsimonious measurement models using Bayesian estimation may be preferable to congeneric ML estimates when the effect of interest is small or when measurement is based on relatively few observations. If, for example, the additive genetic components of a model consist of a random intercept factor model, but the remaining environmental components are congeneric factor models, a researcher who fits a random intercept or congeneric factor model to all components will likely find that the resulting model is not empirically identified under maximum likelihood (ML) estimation. Even assuming congeneric measurement across all components, this predicament would also occur under triangular factorization if some components consist of multiple factors while others are well-represented by single factors. As another example, assuming a tau equivalent factor model may be appropriately parsimonious when summarizing effects which appear to be small across all manifest variables. As described below, we propose that Bayesian models which compare measurement models for the individual components of the genetic factor model may inform researchers of the relative explanatory power of different measurement models across genetic and environmental components (Lee 2007; Lindley 1977).

We will now present the formal definitions of the three measurement models we wish to consider in the genetic factor model, the tau equivalent, congeneric (i.e., standard factor), and random intercept factor models. We will then describe how such measurement models can be estimated and compared using a Bayesian conjugate approach. This approach will then be illustrated using simulated and real-world data.

Psychometric Models: Tau-Equivalent and Congeneric Factor Models

The standard factor model for N individuals measured across k variables in which j latent variables are assessed can be represented in matrix notation as follows (using Sörbom's 1974 notation but with the small adaptation that models are presented so that rows of observed scores correspond to individuals and columns correspond to variables):

$$Y = \alpha + \eta\Lambda + \varepsilon$$

The matrix Y contains N rows of individuals, and k variables (which can consist of repeated measurements of a single variable or different manifest variables at a single occasion). α represents an N by k column scalar matrix of intercepts. η is an N by j matrix of values on the latent variable(s) of the model, and Λ is a j by k matrix of factor loadings. ε is an N by k matrix of errors under the assumption that each column of ε is *i.i.d.* across the N rows. When only one factor is present, the variance/covariance matrix is constrained to unity to mathematically identify the model. Identification of multiple orthogonal factors via triangular decomposition was discussed above. The variance/covariance matrix associated with the matrix of errors of predictions, ε, is usually referred to as Ψ and is most often specified as a diagonal, freely estimated matrix. When all possible factor loadings are freely estimated the resulting measurement model is referred to as a congeneric factor model and is the standard measurement model used in the ACE model.

As noted above, the tau-equivalent factor model (Lord et al. 1968, pp. 47–50) assumes that factor loadings in λ are equal. Mathematically, this model is equivalent to the random intercept component employed in some hierarchical linear models, except that in these models, the variance of the factor is assumed to be freely estimated and the factor loadings in Λ are fixed to 1.

Complex Alternative Models: Random Intercept Model

In the random intercept model (RI) (Maydeu-Olivares & Coffman 2006), two orthogonal factors are estimated, with one factor consisting of freely estimated parameters as in the single-factor congeneric model, and the remaining factor's loadings constrained to equality (or equivalently, to unity with a freely estimated factor variance). As noted earlier, in terms of the number of estimated parameters, the RI model is more complex than the single-factor congeneric model (by estimating a single loading across all manifest variables on the second factor), but more parsimonious than the orthogonal two-factor model (which has k-2 more degrees of freedom than the RI model due to the k-1 freely estimated loadings on the second factor). The RI model also differs from the usual multifactor orthogonal models (such as Cholesky or other triangular decomposition) in that each manifest variable is assumed to load on both factors.

Specifically, using the factor model notation defined above each row vector of Λ can now be written as $\Lambda_0, \Lambda_1, \Lambda_2, \ldots, \Lambda_k$. Λ_0 represents the random intercept factor and all Λ_0 are constrained to 1 with the variance associated with the intercept factor freely estimated or, equivalently, with all Λ_0 constrained to equality and the intercept variance constrained to unity. Λ_1 through Λ_k are defined as before for the multifactor congeneric measurement model. For those more accustomed to path diagram representations, Fig. 1 shows the random intercept model for the case of

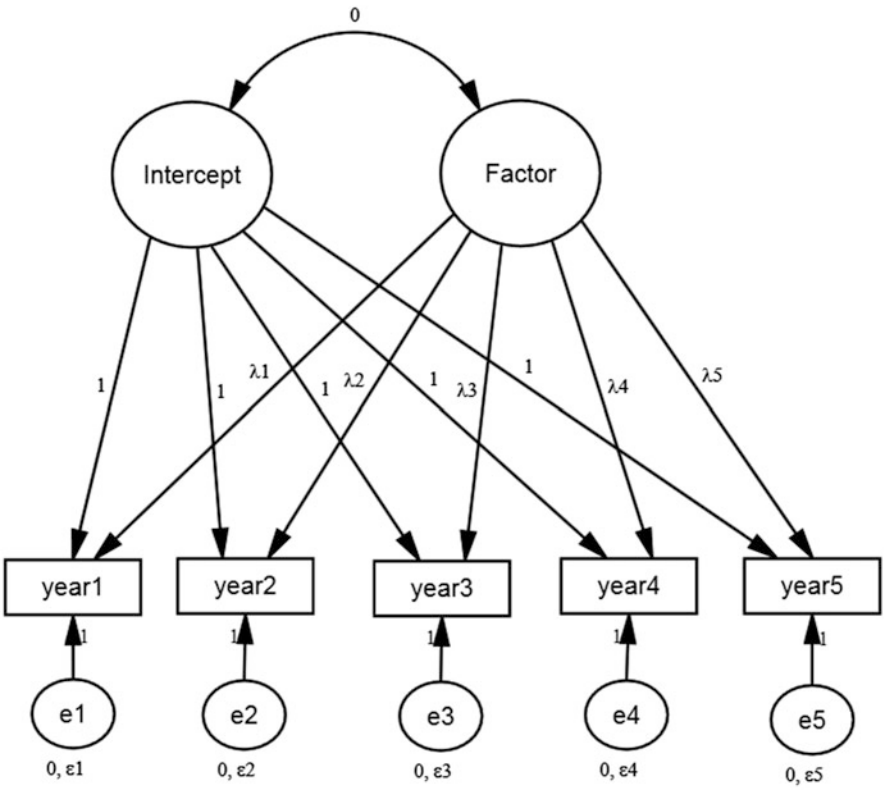


Fig. 1 RI model path diagram

five time measurements. In this Figure λ_1–λ_5 indicate loadings in Λ_1. The random intercept's loadings are represented by the λ_{RI} which are fixed to equality over the measurement occasions. Given that the ACE model frequently assumes unit variances, we chose this strategy to identify the random intercept factor.

The interpretational status of the random intercept factor depends on the particular research situation. As Maydeu-Olivares and Coffman (2006) note, one source of such variability in cross-sectional data may be due to response format, such as systematic negative (or positive) wording of items or a general method factor associated with response. In their analysis of optimistic orientation, Maydeu-Olivares and Coffman found that the random intercept factor resulted in better fit to the data than the traditional one-factor model and was also a parsimonious alternative to a two-factor simple structure model. They interpreted the random intercept factor as a general endorsement or acquiescence factor or, more generally, as a method factor associated with the Likert assessment format. Within the context of longitudinal data, however, the RI model is identical in structure to the free basis growth curve model (Meredith & Tisak 1990) except that, in the growth

curve model, mean level information in the manifest variables is used to estimate factor means for both factors while in the random intercept model, factor means are assumed to be zero and individual manifest variable intercepts are estimated. As such, the RI model could represent such a growth process, but the statistical model relies only on the variance/covariance matrix for the identification of such change patterns. As such, when a single group of monozygous and dizygous twins is analyzed (as for the female twin data considered below), the random intercept factor model loadings are identical to those associated with the reference group considered in Dolan, Molenaar, and Boomsma's (1989 1992) multigroup structured means genetic factor model. An explication of an approach to the structuring of mean effects models for genetic data involves a survey of several articles by Dolan and colleagues as well as consideration of additional psychometric models and is the object of a companion article.

Genetic Factor Model in Factor Analysis Notation

As described in Heath et al. (Heath, Eaves & Martin 1989; Heath, Jardin et al. 1989; Heath, Neale et al. 1989), the genetic factor model for twin data is an extension of the factor model described above, except that η is an *n*6* matrix, with distinct η_A, η_C, and η_E representing the additive genetic, common environmental and unique environmental components for each member of the twin pairs under consideration. Variances across all latent variables are fixed to unity and three additional constraints are placed across the three factors associated with on the ACE structural model: For monozygotic twins, the correlation between genetic components across twins is fixed to 1; for dizygotic twins, this correlation is fixed to 0.5. Finally, the correlation between common factors across both twins is constrained to 1.

Random Intercept Factor Model Applied to ACE Model

One general model for assessment of the psychometric properties of the ACE model occurs when all the three components of the ACE model are modeled as random intercept factors. We therefore differentiate six factors for the resulting genetic model in which we subscript intercept factors to indicate their status as random intercept factors. Accordingly, the terms A, $A_{Intercept}$, C, $C_{Intercept}$, E, and $E_{Intercept}$ denote the congeneric and tau-equivalent components of the genetic factor model for the additive genetic, common environmental, and unique environmental effects respectively. Matrices of the resulting genetic factor model consist of the observed scores of *Y* as an *n by 2k* matrix for k measurement occasions. The column scalar matrix of α has dimensions *n by 2k* matrix, η is an *n by 12* matrix of factor values, and λ is a patterned *12 by 2k* matrix of factor loadings. This random intercept genetic factor model is the same as the traditional genetic factor model except that each

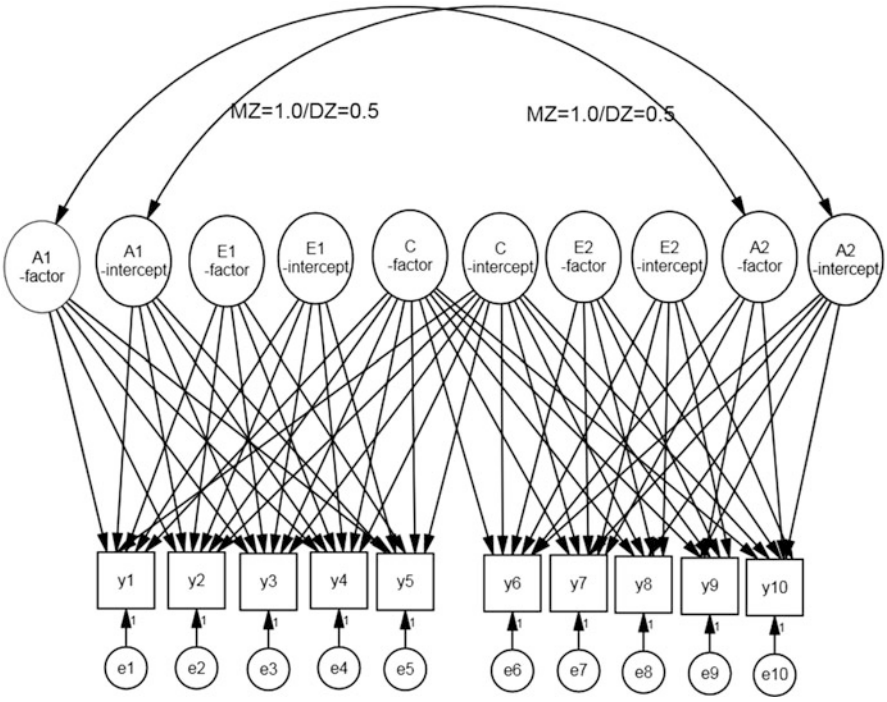


Fig. 2 Proposed model path diagram

genetic and environmental component is represented by two, rather than one latent variable due to the addition of a random intercept model. Model constraints for this model are identical to those for the genetic factor model described above, with the A, $A_{Intercept}$, C, $C_{Intercept}$, E, and $E_{Intercept}$ components assumed uncorrelated. The proposed full model path diagram is shown in Fig. 2.

Bayesian Estimation

Basic Principles and Concepts of Bayesian SEM

As noted above, different measurement models may be appropriate across the genetic and environmental components of the model. Some components may be modeled best as tau equivalent, for example, while the congeneric or random intercept factor models may be most appropriate for other components. If this is the case, researchers attempting to estimate the full RI measurement model for all components are "over-factoring" the data (Rindskopf 1984; Sato 1987) and are likely to find that the model is empirically under-identified due to the non-

uniqueness of the solution space (Savalei & Kolenikov 2008). Such estimation difficulties are not present, however, in Bayesian approaches and parameters which are zero or very close to zero are simply estimated as any other parameters in the model (e.g., Lee 2007). The methodological benefit of this approach, however, is that prior distributions may exclude improper values by definition. (For example, variances estimated in the Bayesian approach using the inverse Gamma distribution can never take negative values, thereby preventing one type of improper solution.) In addition, because the matrix of parameter estimates does not need to be inverted, locally degenerate solutions are not encountered during the process of estimation (Shi & Lee 1998). This permits researchers to compare the relative fit of models with different measurement models across components.

Several excellent treatments of Bayesian inference and use of the Gibbs Sampler are available in both systematic (Gelman, Carlin, Stern, Dunson, & Vehtari 2013) and didactic presentations. For Winbugs applications of the Gibbs Sampler, Eaves et al. (2005) present a Bayesian genetic IRT analysis of questionnaire items and Zhang, Hamagami, Wang, Nesselroade, and Grimm (2007) present Winbugs specifications of growth models. Muthén (2010) presents a similar discussion of example analyses and technical aspects using Mplus. In the interests of space, we will not repeat these presentations, but will limit our discussion to those topics which deal with the basic logic of Bayesian SEM and those technical aspects of estimation which proved most important to the estimation of the random intercept genetic factor model.

Bayes' Theorem

Let M be an arbitrary structural equation model consisting of both parameter specifications of the model with a vector of unknown parameters θ. For brevity of presentation, we will take M to represent both the structural equations representing the model as well as any (possibly informative) prior beliefs of the researcher about these parameters expressed via an appropriate probability distribution. Let Y again be the observed data defined as in Equation 1 above. Based on a well-known identity in probability (Gelman et al. 2013), the posterior probability density function associated with θ given the observed data and structural model may be defined as:

$$p\left\langle \theta | Y, M \right\rangle = p\left(Y | \theta, M\right) p\left(\theta\right)$$

$p(\theta|Y, M)$ represents the posterior density function of the researcher's beliefs about the parameters of the model. $p(Y|\theta, M)$ can be regarded as the likelihood function. The posterior density function incorporates the sample information and the prior density function $p(\theta)$ (Lee & Song 2004).

Gibbs Sampler

The joint analytic form of the posterior distribution poses difficulties to a formal evaluation of the density (Lee 1980). As a result, data augmentation procedures involving Markov Chain Monte Carlo (MCMC) methods such as the Gibbs Sampler are used to obtain the posterior distribution of $p(\theta|Y)$. Such techniques involve a successive iterative approach to generating estimates of the posterior distributions of the parameters and also provide some indication of the reasonableness of the distributional assumptions of the model. Let η be the set of latent variables in the model. The rationale is that adding latent variables η could turn the conditional distribution $p(\theta|Y,\eta)$ and $p(\eta|Y,\theta)$ into simpler form. Given a sample $\{\theta^{(t)},\eta^{(t)}\}$ draws from $p(\theta,\eta|Y)$, an iteration

$$\theta^{(t+1)} p\left(\theta \,|Y,\eta^{(t)}\right)$$

$$\eta^{(t+1)} p\left(\eta \,\middle|Y,\theta^{(t+1)}\right)$$

samples a new state $\left\{\theta^{(t+1)},\eta^{(t+1)}\right\}$. In the end, we could get enough samples in the chain and observe the posterior distribution of θ (Geman & Geman 1984). At convergence, different chains generated with different starting values are merged together (after discarding a number of iterations during the beginning phases of each, which are treated as burn-in iterations). If successive observations are highly positively correlated (as was frequently found in several of the genetic factor models we considered) values are taken only from successive intervals (such as every 20th iteration), a process known as "thinning" (Gelman et al. 2013).

Data from the MCMC iterations used in estimation can also be plotted as a diagnostic of whether the parameter of interest appears to take the form assumed by the distributions chosen by the researcher to represent beliefs about the parameter, a method known as Posterior Predictive Checking (PPC, Gelman et al. 2013). As described below, in the data sets considered in this paper, PPC of the estimated posterior distributions alerted us to the fact that the Gibbs Sampler was prone to produce multi-modal posterior distributions symmetric about zero for random intercept factor loadings, particularly if the size of the effect was modest. We discuss this issue and solutions below.

Model Fit

In addition to providing posterior distributions about the parameters of interest, the Bayesian approach also permits the researcher to evaluate the fit of the structural model based on its likelihood. Although several approaches to assessing model fit can be taken (Gelman et al. 2013; Lee 2007) we will discuss three here. The BIC (Schwarz 1978) is popular within structural modeling because it penalizes models for their complexity (expressed as the number of parameters in the model).

The Deviance Information Criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde 2002) is a Bayesian generalization of information criteria such as the AIC and BIC which penalizes models based on the effective number of parameters in the model. For both the DIC and BIC, smaller values are considered better-fitting. Posterior predictive checking of the likelihood of the model as a whole also permits the researcher to estimate the Posterior Predictive p-value, an estimate based on the PPC of the likelihood ratio chi-square statistic for the model (Meng 1994). This represents a rough estimate of the probability that the data could have been generated under the candidate model. The proposed model may be considered as plausible if the PP p-value estimate is not far from 0.5 (acceptable range 0.3–0.7). Meng (1994) notes that the PP p-value is not suitable for comparing different models but is a reliable index of stand-alone model fit.

Model Comparison: Bayes Factors

In addition to providing stand-alone measures of model fit, it is also possible to assess the relative fit of candidate structural models for the data. In addition to simply comparing the incremental fit, the Bayesian approach also permits the researcher to assess the relative informative power associated with increases in model complexity. Most generally, this comparison is made using the Bayes factor, which we now introduce in some greater detail given the need to understand its basic logic and the fact that its estimation is the object of ongoing study. From the Bayes theorem comparing the odds ratio associated with the comparison of a base model, M_0 with a more complex model, M_1, we can obtain:

$$\frac{p\left(M_1 \mid Y\right)}{p\left(M_0 \mid Y\right)} = \frac{p\left(Y \mid M_1\right) p\left(M_1\right)}{p\left(Y \mid M_0\right) p\left(M_0\right)}$$

which permits us to define the Bayes factor as

$$B_{10} = \frac{p\left(Y \mid M_1\right)}{p\left(Y \mid M_0\right)}$$

Thus we see that posterior odds = Bayes factor*prior odds (Lee 2007). Larger Bayes factors mean stronger evidence for M_1 relative to M_0. $p\langle Y|\, M_1\rangle, p\langle Y|\, M_0\rangle$ is obtained by integrating $p\langle Y|\, \theta, M_1\rangle, p\langle Y|\, \theta, M_0\rangle$ over the parameter space, respectively. It is, however, often difficult to obtain Bayes factor analytically using a path sampling approach (Gelman & Meng 1998) and, for that reason, another easy and quick way to calculate Bayes factor is by using BIC (Muthén & Asparouhov 2011):

$$BF = \frac{p\left(M_1\right)}{p\left(M_0\right)} = \frac{\exp\left(-0.5 BIC_{M_1}\right)}{\exp\left(-0.5 BIC_{M_0}\right)}$$

Although there is some dispute about the validity of calculating Bayes factor by using BIC (Gelman et al. 2013), in the simulation study presented below, we found that this criterion worked well in practice. Generally speaking, Bayes factors less than 3 represent minimal support for the alternative model, values between 3 and 20 positive support for the alternative model, values between 20 and 150 strong support, and values larger than 150 decisive support.

Simulated Data Example

We will now illustrate our general approach of fitting a general random intercept genetic factor model and assessing the relative fit of more parsimonious measurement models for the genetic and environmental components using simulated data (generated from SAS). We generated simulated twin data for 1000 hypothetical twin pairs using the following factor loadings: Across all variables, $A_{Intercept}$, $C_{Intercept}$, and $E_{Intercept} = 0.4$, 0.4, and 0.3, respectively. A factor loadings were zero. C factor loadings were chosen as 0.4, −0.4, 0.3, −0.3 and 0.2 across the five variables. E factor loadings were chosen as −0.3, −0.3, 0.3, 0.3, and 0.3.

To demonstrate the ability of the procedure to correctly arrive at a more parsimonious model and to highlight the empirical under-identification issues associated with more parsimonious models under ML estimation, we chose to simulate data in which an intercept model was appropriate for the additive genetic component, but for which RI models were appropriate for the shared and unique environmental components. 1000 replication data sets were generated to investigate the sampling behavior of the approach using SAS. Models were estimated using Mplus (Muthén and Muthén 1998–2010). The Gibbs Sampler iteration number was set at 5000 to allow a generous amount of iterations for the MCMC chains in the Bayesian analyses. By default, the first half of these iterations was used as a burned-in phase. Initial inspection of the MCMC chains revealed marked auto-correlation across iterations of the Gibbs Sampler, and so a thinning value of 50 was chosen for the analyses which appeared to remedy the auto-correlation problem (Albert & Chib 1993). All 1000 replications met the convergence criteria by Bayesian estimation (PSR close to 1 for each parameter) (Muthén & Asparouhov 2011). The PP-p value associated with the general RI model had a mean of 0.53, standard deviation 0.25 across replications, indicating good fit. In addition, the genetic slope factor loadings are all non-significant. Under ML estimation, however, all 1000 samples failed to converge which we take as evidence that they were not empirically identified. When the correct model is fit to the data, however, ML models did converge. There was little difference between the Bayesian and ML estimates under the correct model, with bias estimates not exceeding 2 % across the estimated loadings (See Table S1 in supplemental materials accompanying the manuscript.)

Table 1 Percentages of replications with Bayes factors > 20 favoring column model over row model across 1000 simulated samples

Model	Two-factor model	RI model	ACE model	True model
Two-factor model	NA	91.6 %	47.4 %	99.8 %
RI model	7.70 %	NA	10.4 %	99.4 %
ACE model	52.0 %	89.3 %	NA	96.3 %
True model	0.20 %	0.60 %	3.50 %	NA

Table 2 Summary of parameter bias in traditional ACE genetic factor model

	Bayesian estimation	ML estimation
Parameter	Bias Mean (S.D.)	Bias Mean (S.D.)
A loadings	45.8 % (0.106)	45.5 % (0.110)
C loadings	−39.7 % (0.071)	−39.1 % (0.073)
E loadings	−27.7 % (0.379)	−27.7 % (0.377)

Model Comparison

Table 1 presents proportions of model comparisons across replications in the simulated data which exceed criteria for strong support in comparisons of the true model, the traditional ACE model, and a freely estimated two-factor solution across all genetic and environmental components. As can be seen from the fifth column of the table, the true model is preferred over the competing two-factor, random intercept, and traditional ACE models in 96.3–99.8 % of the cases. The two-factor model is preferred over the traditional ACE model in only 52 % of the cases, and the random intercept model is preferred over the traditional ACE model in 89.3 % of the cases. This high latter percentage is unsurprising, given that the random intercept model differs from the true model only in that the A factor is redacted from the RI model to produce the true model. Taken together, model comparisons based on the simulated data reveal that Bayesian estimation appears able to correctly identify the correct model and, even when the model under consideration is slightly over-complex, the factorial complexity of the genetic and environmental components in these data is detected.

Genetic Factor Model

When these data were analyzed with the (mis-specified) traditional genetic factor model in which all three components are assumed to have a congeneric measurement model (i.e., have only A, C, and E factors), all 1000 replication yielded a zero PP-p value, indicating poor model fit. The bias summary associated with the ACE model is presented in Table 2. Results suggest that, for the simulated data considered here, failure to correctly include intercept components for the common unique environmental effects introduces substantial bias in the estimated additive effects of the model.

Two-Factor Model

As noted above, the RI measurement model is a two-factor model, but one with considerably more parsimony than the traditional two-factor congeneric model. The question therefore remains as to whether the Bayes factor can also correctly reward the greater parsimony of the RI model relative to the more complex traditional two-factor model. When the traditional two-factor model is estimated from the data, the PP-p value has mean of 0.50, with standard deviation 0.23, indicating the high degree of model fit found for the (true) RI model. To secure a mathematically identified solution for the two-factor model, the first loading associated with each of the A2, C2, and E2 factors was set to zero. Bias estimates for the two-factor model are of necessity quite pronounced, given that the Cholesky form of the two-factor model represents an affine rotation of the true structure of the data. If calculated as a percent bias relative to the true model, bias estimates of the two-factor Cholesky model averaged 37.2 %, with bias across the particular types of loading ranging from 12.5 to 97.5 % (See Table S2 in the supplemental materials accompanying the manuscript.)

Alternatively, if the approach outlined by Franić et al. (2013) is followed, the correct dimensionality of the genetic and environmental components is identified as a two-factor solution. However, the structure of the random intercept model is not correctly specified due to the fact that the resulting decomposition is triangular in nature. Even if the two-dimensional factor structure is rotated via an affine transformation to a form most closely resembling the true factor structure, two of the recovered loadings still deviate by approximately .05 due to sampling variability. Since it is difficult to judge empirically in real-world applications whether such variation represents sampling variability or a true multifactor structure in which factor loadings of one factor are unequal to each other, we believe it reasonable to directly compare the two-factor and random intercept models as outlined here.

Other Alternative Models: Bayesian Estimation

In addition to these selected model comparisons, we also compared the true model with all other combinations of the three possible measurement models (tau, congeneric, and random intercept) for each component of the genetic factor model. Model fit indices for the models are shown in Table 3 as well as the Bayes factor comparing the true model to each candidate. Although it would be possible to compare all of these candidate models using Bayes factors, the evaluation of such a matrix of pairwise comparisons would be both tedious and liable to substantial experiment-wise error given the number of contrasts. If, however, researchers compare the relative fit of the random intercept model to models which redact intercept or factor models from the genetic model, a relatively proscribed set of model comparisons results. Well-fitting parsimonious models can then be compared to the random intercept model in an attempt to identify a more parsimonious model. As can be seen in Table 3, when Bayes factors are calculated relative to the random intercept

Table 3 Candidate models' PP-p value and percent of Bayes factors strongly preferring the RI and true models

Model	PP-p value (S.D.)	BIC (S.D.)	BF Perc. Over 20 (RI) (%)	BF Perc. Over 20 (True) (%)
RI model	0.53 (0.25)	49,327.21 (234.73)	NA	99.4
True model	0.52 (0.26)	49,249.37 (215.04)	0.06	NA
RI without $A_{Intercept}$ model	0.52 (0.25)	49,279.5 (215.83)	4.00	99.3
RI without $C_{Intercept}$ model	0.07 (0.11)	51,798.18 (2371.75)	95.3	99.8
RI without $E_{Intercept}$ model	0.00 (0.00)	49,892.57 (712.31)	96.0	99.3
$ACEE_{Intercept}$ model	0.07 (0.11)	50,167.56 (1492.76)	79.6	100
$ACC_{Intercept}E$ model	0.00 (0.00)	49,515.99 (220.87)	92.8	100
$AA_{Intercept}CE$ model	0.00 (0.00)	50,304.43 (1602.06)	96.3	100

model, only the true model and the random intercept model without the $A_{Intercept}$ factor were not significantly worse fitting than the RI model, as shown in the fourth column. When the true model is considered as a base model, the evidence strongly supporting the true model is found between 99.3 and 100 % of the replications.

Summary Remarks for Simulation Study

Under ML estimation, estimating an over-complex RI measurement model for all three components results in empirical under-identification. When the random intercept is present for the genetic component but the data are analyzed using the traditional ACE model, estimates of heritability of the genetic component are over-estimated under both ML and Bayesian estimation. When the true model is known, however, ML and Bayesian parameter estimates appeared similar. Because of the empirical under-identification problems in ML estimation, comparison of candidate measurement models was only possible under the Bayesian approach. For these data, the correct model was identified using the Bayes factor. Significantly, the random intercept measurement model was also found to be a parsimonious alternative to the traditional two-factor model.

Care must be taken in conducting Bayesian analyses, however. Even with the simulated data under consideration, large thinning values were necessary to reduce autocorrelation across iterations of the Gibbs Sampler and bimodality was observed in some of the PPC plots which indicated possibly misleading estimates and confidence intervals for the Bayesian approach. Once identified, however, these bimodality issues were successfully addressed. In the next section, the general RI genetic factor model and its more parsimonious alternatives are considered in an empirical data example. In addition to the didactic value of a real-world example, use of a real-world example also permits exploration of the effects of the non-normality and unmodeled causal effects on model fit, comparison, and parameter estimation.

Empirical Study

The genetic and environmental effects on body mass index (BMI) have been investigated across several studies. Allison et al. (1996), in a study of Japanese, Finnish, and American twins, reported that additive genetic effects appeared more pronounced at early ages, that the genetic effects did not appear due to shared environmental effects during this time, and that heritability coefficients ranged between .5 and .7 for the data sets considered. Elks et al. (2012), in a review of 88 estimates of the heritability of BMI across twin studies, found heritability estimates ranging from .47 to .90. It is worth noting that most of these estimates (61) were based on AE models (i.e., a model with no common environmental effects), while 15 were based on the traditional ACE model. (The remainder were based on direct comparisons of within and between twin correlations or the non-additive genetic model.) Estimates of the genetic heritability of BMI using the ACE model were generally .12 higher than estimates from the AE model. Readers are referred to Elkes et al. for a discussion of the genome-wide association studies investigating the loci associated with BMI.

The BMI data we wish to analyze are taken from the Missouri Adolescent Female Twin Study (MOAFTS), a genetic-epidemiological, prospective twin-family study of alcohol use in young females. (For full details, including response rates, see Waldron, Bucholz, Lynskey, Madden, & Heath 2013.) Using a cohort sequential design, twins were aged 13, 15, 17, and 19 when first enrolled in the study. In analyses presented here, we exclude African-American twins, because of small numbers but significant mean differences in BMI distribution. A total of 3416 Missouri female adolescent twins (85 % participation rate, approximately 55 % MZ and 45 % DZ) were interviewed from 1995 to 2012 with a telephone version of the Child Semi-Structured Assessment for the Genetics of Alcoholism. In this study, we only concentrated on the body mass index (BMI) variable. Observations from twin pairs with at least five measurement occasions were selected for this longitudinal analysis. Descriptive statistics by age groups are listed in Table 4. Since all observed variables are positively skewed, even after fitting the model, we transformed the data by taking the log of the original data. The following analyses were based on the transformed data.

Bayesian Model Comparison

As in the simulation study, two-factor, RI, and simpler alternatives were considered for the BMI data. Table 5 presents the Bayes factor (relative to the final model), PP-p value, and DIC for each reduced model as well as the two-factor model. A model consisting of a RI model for the additive genetic effect, a tau equivalent model for the unique environmental effect, and no common environmental effect was chosen as the final model based on its Bayes factor relative to the RI model

Table 4 Descriptive statistics of body mass index by age group

	N	Twin 1			Twin 2		
Age	Twin pairs	Mean	S.D.	Skewness	Mean	S.D.	Skewness
13	58	19.9	2.684	0.82	20.26	3.244	1.323
14	71	20.75	3.109	1.061	20.1	3.019	1.465
15	150	21.1	3.222	1.542	21	3.188	1.819
16	110	21.05	3.04	1.437	21.16	3.419	1.842
17	188	21.74	3.113	1.222	21.58	3.731	2.199
18	160	21.97	3.359	1.156	22	3.646	1.434
19	89	23.09	4.443	1.81	22.7	3.523	1.351
20	117	23.07	4.334	1.458	22.53	3.653	1.315
21	31	23.28	4.925	1.763	23.22	5.379	2.387
22	80	23.6	4.833	2.298	23.38	4.736	1.793
23	86	24.84	5.068	1.059	24.23	4.587	1.372
24	68	23.89	4.782	1.221	23.83	4.705	0.997
25	65	24.95	5.538	1.777	24.84	5.35	1.437
>25	113	26.54	5.804	1.125	25.88	6.08	1.152

($1.84*10^{19}$). Although such a choice of models may seem somewhat unusual, it is a choice consonant with other research on BMI during young adulthood; Elks et al. (2012) report 26 studies of BMI spanning both young and older samples compared to nine studies reporting the traditional ACE model. Although such a contrast does not ensure correctness via democratic vote, it does speak to the fact that a decision to redact the common environmental component is not without precedent.

ML Model Comparison

The model comparison results using ML estimation were similar to their Bayesian counterparts and are shown in Table S3 in the supplemental materials for the manuscript. Model fit index such as RMSEA and CFI were very similar across the different models. Moreover, chi-square test cannot be used to compare all models given that they are not nested models. However, based on examination of the BIC values, the $AA_{Intercept}C_{Intercept}EE_{Intercept}$ model demonstrated the best fit (BIC = −5335.9), with the $AA_{Intercept}C_{Intercept}E_{Intercept}$ model showing a value only slightly larger than this (BIC = −5304.5). The model chosen under Bayesian estimation, $AA_{Intercept}E_{Intercept}$ (BIC = 5165.7) was larger than these other two models but still lower than the other models considered. On examination of the

It should be noted that when all Bayesian models which included a common environmental effect failed to find environmental effects greater than zero, regardless of whether a tau equivalent, congeneric or random intercept model was used to model the component.

Table 5 PP-p and BIC values for candidate models and Bayes factor of body mass index data

Model	PP-p	BIC	Bayes factor	Bayes factor vs. $AA_{Intercept},E_{Intercept}$
$AA_{Intercept},E_{Intercept}$	0.291	−5157.05	1.84E + 19	1
Two-factor	0.401	−4729.84	3.14E−74	1.71E−93
RI full model	0.301	−5068.33	1.00E + 00	5.43E−20
$A,C,C_{Intercept},E,E_{Intercept}$	0.309	−5075.19	3.09E + 01	1.67E−18
$A,A_{Intercept},C,E,E_{Intercept}$	0.317	−4864.42	5.27E−45	2.86E−64
$A,A_{Intercept},C,C_{Intercept},E$	0.333	−4840.35	3.12E−50	1.70E−69
$A,A_{Intercept},C_{Intercept},E_{Intercept}$	0.299	−4918.28	2.61E−33	1.42E−52
$A,C,C_{Intercept},E$	0.289	−5055.06	1.31E−03	7.13E−23
$A,A_{Intercept},C,E$	0.293	−4910.09	4.34E−35	2.36E−54
A, C, E (Genetic Factor Model)	0.285	−5042.8	2.85E−06	1.55E−25
$A_{Intercept},C,C_{Intercept},E,E_{Intercept}$	0.289	−5130.3	2.85E + 13	1.55E−06
$A,A_{Intercept}, C_{Intercept},E,E_{Intercept}$	0.323	−5131.53	5.28E + 13	2.87E−06
$A,A_{Intercept},C,C_{Intercept}, E_{Intercept}$	0.285	−5067.28	5.90E−01	3.21E−20
$A_{Intercept}, C_{Intercept},E,E_{Intercept}$	0.133	−5130.5	3.16E + 13	1.72E−06
$A_{Intercept},C,C_{Intercept}, E_{Intercept}$	0.275	−5136.93	7.87E + 14	4.28E−05
$A_{Intercept}C_{Intercept}E_{Intercept}$	0.122	−5000.67	2.03E−15	1.10E−34
AE	0.124	−4963.992	2.03E−15	1.19E−42

ML estimates, the $C_{Intercept}$ and E factor loadings were, although significant, modest in magnitude (all λ's < .05). Because of the advantages of the Bayesian estimation approach to model comparison and because the additional factors, if present, appeared to represent modest effects, we chose to report ML and Bayesian estimates for this model.

Parameter Estimation

Bayesian parameter estimates based on the final model are shown in Table 6. (Corresponding ML parameter estimates for this model were almost identical in value.) Consistent with Allison et al.'s (1996) finding, the genetic intercept appears to explain more variability than the genetic factor in early years, especially from ages 13 through 18. During later years (from ages 21 through 26 and later), the genetic factor appears to explain roughly the same proportion of variability as the intercept. The pattern of loadings for the genetic factor appears to be roughly nonlinear and suggests systematic differences in the genetic component associated with BMI during the adolescent, young adult, and adult years.

Also consistent with the majority of the twin studies reviewed by Elks et al. (2012), common environmental effect was either not statistically significant (based on Bayesian estimates). Given that dropping CI gave a similar model fit index

Table 6 Bayesian parameter estimates for body mass index data

Age	λ[a]	Std. Dev.[a]	$p\|H_0 = 0$	Intercept	Std. Dev.	$p\|H_0 = 0$
	A					
13	−3.40	1.70	0.023	3.00	0.01	0.00
14	−0.30	1.10	0.408	3.02	0.01	0.00
15	1.60	1.00	0.063	3.04	0.01	0.00
16	0.50	1.00	0.285	3.05	0.01	0.00
17	1.80	1.00	0.034	3.07	0.01	0.00
18	3.40	1.00	0	3.08	0.01	0.00
19	5.50	1.30	0	3.11	0.01	0.00
20	5.10	1.10	0	3.12	0.01	0.00
21	8.70	1.80	0	3.14	0.01	0.00
22	8.10	1.20	0	3.14	0.01	0.00
23	10.20	1.30	0	3.17	0.01	0.00
24	9.40	1.50	0	3.18	0.01	0.00
25	9.90	1.40	0	3.18	0.01	0.00
>26	13.00	1.40	0	3.22	0.01	0.00
	$A_{intercept}$					
All Ages	12.90	0.50	0			
	$E_{intercept}$					
All Ages	4.10	0.30	0			

[a]Values in columns multiplied by 100 for ease of presentation

(PP-p value is 0.293) and the Bayes factor is 2.81 favoring the model without CI, we conclude that dropping the CI factor from the model seems reasonable and we note that inclusion of the effect does not seem to affect other parameters and explains at most a minimal amount of variability.

The proportion of variability explained by genetic and environmental effects by age is shown in Table 7. For these data, heritability estimates for the final model (shown in the column labeled "Additive" under the Heading "Final Model") ranged from 0.72 to 0.82 with an average of 0.77 across years, which compares favorably with the 0.75 median estimate from Elks et al.'s (2012) meta-analysis. In contrast to the heritability estimates based on the traditional ACE model and models used in Elks et al.'s study, heritability does not appear to be more pronounced in younger ages than in older ages. Heritability estimates from the traditional ACE model (shown in the same column under the heading "ACE") for these data are somewhat lower (mean = .68, range 0.49–0.82 across years) and appear to be slightly lower for twins older than 21. The difference in average heritability between the final and traditional ACE model of .09 is similar to the 0.12 increase noted by Elks et al. when models are fit which do not include an environmental effect. It is also worth noting that statistically significant common environmental effects using the single-factor ACE model were only found for ages 21 through 25 and, even for these, the proportion of variability in BMI explained was on average 7 %. The discrepancy between the final model and the traditional one-factor ACE model does not appear

to be due entirely, however, to the estimation of a common environmental effect, because when a two-factor ACE model is estimated from these data, the average heritability across ages is 0.59, and none of the factor loadings associated with the common environmental effects is statistically significant.

A comparison of unique environmental effects of the final model with the one- and two-factor traditional ACE models reveal that the average unique environmental effect was slightly smaller for the final model (0.06) than for either the one- or two-factor ACE models (.09 and .12, respectively).

Summary Remarks of Empirical Study

Taken together, estimates from the final model under Bayesian estimation produce estimates of heritability consonant with the Elks et al. (2012) review and replicate the conclusion made by many researchers that common environmental effects in BMI appear to be negligible. The pattern of differential common environmental effects found under a one-factor ACE model is not replicated by either the random intercept model selected as most reasonable or by a freely estimated two-factor model. Although, as Visscher, Gordon, and Neale (2008) note, small sample studies may be underpowered to detect a statistically significant common environmental effect, the existence of such differential effects were not found using the Bayesian model comparison procedure outlined here and, even if thought to exist, their magnitude appears to be confined to older ages and to be minimal in comparison to the magnitude of heritability coefficients during these ages. For these data, the proposed model comparison approach appears to yield a model which is both parsimonious and reasonably similar to the larger literature on the magnitude of environmental and genetic effects.

Discussion

The measurement model which researchers choose to operationalize environmental and genetic components of behavior genetic models has important implications for the estimation and interpretations of such models. When psychometric alternatives to the traditional factor model such as the tau equivalent and random intercept models are considered, substantially different estimates of the relative salience of genetic and environmental contributions are obtained. Comparison of candidate measurement models seems warranted given that the psychometric complexity of the true model is largely unknown to the researcher prior to analysis and, even if it were, such exploration can inform the researcher about possible alternate estimates for genetic and environmental components that a reasonable skeptic might raise. Consideration of overly complex genetic models, however, is often prevented in maximum likelihood estimation because such models are not empirically identified,

Table 7 Estimated proportion of variability explained in BMI by final, traditional ACE, and two-factor models

Source	Additive genetic			Common Env.			Unique Env.		
Age	Final model	ACE	Two-factor	Final model[a]	ACE	Two-factor	Final model	ACE	Two-factor
13	0.72	0.79	0.69	0	0.05	0.00	0.07	0.01	0.11
14	0.78	0.81	0.69	0	0.01	0.00	0.08	0.01	0.10
15	0.77	0.82	0.68	0	0.00	0.00	0.08	0.01	0.06
16	0.82	0.81	0.70	0	0.01	0.00	0.08	0.05	0.06
17	0.77	0.77	0.66	0	0.00	0.00	0.08	0.07	0.07
18	0.76	0.66	0.58	0	0.01	0.00	0.07	0.15	0.20
19	0.73	0.57	0.54	0	0.03	0.00	0.06	0.25	0.17
20	0.80	0.77	0.73	0	0.01	0.00	0.07	0.11	0.09
21	0.81	0.69	0.50	0	0.10	0.00	0.06	0.05	0.13
22	0.79	0.59	0.55	0	0.12	0.00	0.06	0.16	0.11
23	0.79	0.57	0.53	0	0.12	0.00	0.05	0.18	0.18
24	0.77	0.68	0.54	0	0.11	0.00	0.05	0.05	0.08
25	0.74	0.49	0.49	0	0.17	0.00	0.05	0.14	0.16
>25	0.76	0.51	0.45	0	0.24	0.01	0.04	0.03	0.11

[a]Fixed to zero

which leaves open the question of whether complex alternatives were simply not numerically obtained given the software or whether they are empirically unidentified due to being over-complex. Such model estimation was, however, possible under Bayesian estimation and estimated parameters for the final model appeared largely similar to corresponding estimates from ML estimation for both the simulated and real-world data.

Given that traditional behavior genetic models often involve the assessment of only different numbers of freely estimated latent variables, this paper seeks to highlight the fact that a greater number of models are possible, given parsimonious patterning of the factor loadings involved. The extension of such models to convey mean effects makes it possible for the researcher to specify the patterning of such variance components as growth curve models. As noted above, a great variety of models are possible under the model comparison procedure described above, which may prompt some researchers to wonder how best to limit the specification and search of models to a more tractable number in practice. In the fortunate cases where the researcher is in the position of having some knowledge concerning the functional form of growth over time, it would be possible to specify nonlinear constraints on the estimated factor loadings so that the underlying estimated curve corresponds to a parametric growth model such as the logistic or Gompertz curve (Grimm & Ram 2009). Increasingly, however, it appears that the patterns of growth observed over time in empirical data do not follow such tidy mathematical specifications, leading some to adopt the nonparametric growth curve as a reference curve for characterizing the form of growth over time. For example, Ram and Grimm (2009), in a study of longitudinal finite mixture models, advocate for initial specification of a free curve growth model as a model of functional change over time which can then serve as a reference form for the identification of finite mixtures. In general, however, adoption of such a "nonparametric" growth curve model raises questions concerning the interpretability of the identified curves.

It is quite possible that one reason for the failure of identified patterns of growth over time to follow a parametric form is due to the fact that Alessandri, Caprara, and Tisak (2012) point out that the presence of a single, but nonparametric pattern of loadings for growth data may indicate the presence of several, rather than one source of stability of time which may include environmental effects, age-related effects or turning points. Given that the genetic and environmental effects identified through behavior genetic models are genetically multi-determined, it is probably more reasonable to expect that identified effects should probably exhibit such composite patterns over the lifespan.

As such, it is important to recall that the proposed model comparison approach to fitting genetic and environmental effects is no panacea and that behavior genetic growth curve models, as with any latent variable model, are subject to the "naming problem" in that the latent variables identified may not represent the constructs initially intended. Although it is possible to attempt a remedy of this by modeling one of the factors of the model according to some agreed upon parametric form and to identify "residual factors" which would model additional covariation due to the extraneous effects, the fact that there is at least in the context of much behavior

genetic models, little agreement as to the functional form of growth over time makes such an approach untenable. In the end analysis, probably the best remedy for this ambiguity lies in the identification of such possible confounding effects and data collection strategies designed to provide a less ambiguous portrait of change over time.

Model Support

Rather than basing model comparisons in terms of probability statements common under the frequentist approach, the Bayesian approach permitted adjudication between candidate models based on a quantification of the relative support for a particular model relative to other candidate measurement models. Operationally, measurement models which include random intercept components permit the researcher to consider the model of tau equivalence as a parsimonious alternative to the single-factor model usually considered in genetic factor models. The possibility that the manifest variables of the study constitute equivalently scaled measures would seem an attractive one to researchers, especially if the manifest variables in the model constitute longitudinal assessments. More generally, inclusion of a random intercept model makes more fine-grained comparisons of models intermediary between those usually considered by researchers which are based on factor dimensionality. For example, in both simulated and real-world data, factor models with random intercept components were estimated and selected which were more complex than the single-factor model but yet more parsimonious than the freely estimated two-factor model. The question of whether the random intercept model or multi-factor measurement model better describe the data also has important implications for the investigation of multigroup invariance. Equality constraints have sometimes been used across Cholesky factors to test for invariance across groups (e.g., Loehlin & Martin 2013). Different hypotheses about equality constraints of factor loadings and component variability are implied under the RI model, however, suggesting that different conclusions about partial invariance may be made under the RI and Cholesky factor models.

Model Support Varies as a Function of Study Design

When such comparisons are considered across studies in an area, such model comparisons provide statements of what measurement models seem reasonable based on characteristics of the study. When the statistical power of the data is low (i.e., when effect size is small or small sample sizes are analyzed), researchers are more likely to find the tau equivalent measurement and its associated standard error are a parsimonious summary. In the body mass data, such intercept factors seemed sufficient to explain variability due to common and unique environmental

effects. Congeneric measurement models, by contrast, provide information that markedly different effect sizes across manifest variables have been found. Similarly, multivariate studies of relatively few manifest variables are similarly less likely than studies with several manifest variables to recover random intercept models or multiple factors due to the lower power associated with smaller samples of the multivariate space. Accordingly, the ability to identify method variability, response set, or a complex measurement model such as that underlying a growth process varies as a function of study design.

Limitations

Although use of Bayesian estimation for genetic modeling has promise, it is not without its difficulties. The bimodality of estimated factor loadings across MCMC replications was one difficulty most often encountered when estimated factor loadings were modest and successive MCMC iterations varied between small positive and equally well-fitting small negative values. This problem is equivalent to the reflection problem in factor analysis in general (i.e., that a factor model with loadings multiplied by -1 fits the data as well as the original factor loadings). In Bayesian analysis, researchers can detect the resulting bimodal posterior distribution using different starting values across chains. If the estimated loading is not far away from zero, the convergence criteria or K–S test are unlikely to detect such bimodality. Such bimodality can, however, be remedied by constraining one or more such marginal loadings to be positive across those parameters which appear to exhibit bimodality (Congdon & Congdon 2003). Although such a remedy is appropriate in many situations (Erosheva & Curtis 2013), it should be noted that it is not a universal solution and requires further research (Chan & Jeliazkov 2009).

Future Directions

Use of the model comparison approach outlined here can be readily extended to a greater variety of genetic models for twin and family data. Although the models considered here assumed that the manifest variables were continuously measured variables, extensions of the models presented here to genetic factor models using categorical data (Cho, Wood, & Heath 2009) would appear straightforward, subject to additional identification requirements of the latent response variable approach required for categorical data. Developments in both behavior genetic modeling and Bayesian statistics have also extended structural models using generalized linear mixed models, enabling researchers to specify random effects for variables with other known distributions such as Poisson or other exponential link functions (e.g., Bolker et al. 2009). Additionally, given that multilevel behavior genetic models have also been proposed for genetic data (e.g., Guo & Wang 2002), modeling a random intercept term within a factor model provides the researcher with the

ability to assess the relative fit of such models within a factor analytic framework, with the added benefit that the multilevel models can be modeled as special cases of the general genetic models considered here, as they result when factor loadings are constrained to fixed values. Finally, as mentioned above, the random intercept model can also be extended to the case of estimation of growth curve models, although the psychometric measurement alternatives are slightly more complex in those situations.

Conclusion

The exploration of more fine-grained model comparisons motivated by psychometric models for the environmental and genetic components of behavior genetic models appears promising when Bayesian estimation is considered. Bayesian models appear less susceptible to problems of empirical under-identification frequently encountered under ML estimation. The tau equivalent and random intercept models in particular appear to be two parsimonious alternatives to the factor components usually considered under a Cholesky decomposition or other exploratory factor approaches. Although care must be taken to assure that estimation difficulties related to multi-modality and serial correlation in the MCMC estimation procedure are identified and remedied, use of the Bayes factor appears to be a promising means for assessing the relative support of candidate psychometric behavior genetic models.

Electronic supplementary material

Below is the link to the electronic supplementary material.Mplus Program for Fitting Bayesian One-Factor ACE model (DOCX 21 kb)Mplus Program for Fitting Final Bayesian Random Intercept Model for Simulated Data (DOCX 25 kb)

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 422*, 669–679.

Alessandri, G., Caprara, G. V., & Tisak, J. (2012). A unified latent curve, latent state-trait analysis of the developmental trajectories and correlates of positive orientation. *Multivariate Behavioral Research, 47*, 341–368. doi:10.1080/00273171.2012.673954.

Allison, D. B., Kaprio, J., Korkeila, M., Koskenvuo, M., Neale, M. C., & Hayakawa, K. (1996). The heritability of body mass index among an international sample of monozygotic twins reared apart. *International Journal of Obesity, 20*, 501–506.

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, M., et al. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–135. doi:10.1016/j.tree.2008.10.008.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (pp. 149–173). Amsterdam: North-Holland.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*, 425–440.

Carey, G., & DiLalla, D. L. (1994). Genetics, personality, and psychopathology. *Journal of Abnormal Psychology, 103*, 32–43.

Carey, G., Goldsmith, H. H., Tellegen, A., & Gottesman, I. I. (1978). Genetics and personality inventories: The limits of replication with twin data. *Behavior Genetics, 8*(4), 299–313.

Chan, J. C. C., & Jeliazkov, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation, 1*(1), 101–120.

Chang, Z., Lichtenstein, P., Asherson, P. J., & Larsson, H. (2013). Developmental twin study of attention problems: High heritabilities throughout development. *JAMA Psychiatry, 70*(3), 311–318.

Cho, S. B., Wood, P. K., & Heath, A. (2009). Decomposing group differences of latent means of ordered categorical variables with the genetic factor model. *Behavior Genetics, 39*, 101–122.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347–357.

Congdon, P., & Congdon, P. (2003). *Applied Bayesian modelling* (Vol. 394). New York: Wiley.

Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika, 76*(4), 650–669.

Dolan, C., Molenaar, P., & Boomsma, D. (1989). LISREL analysis of twin data with structured means. *Behavior Genetics, 19*(1), 51–62.

Dolan, C. V., Molenaar, P. C., & Boomsma, D. I. (1992). Decomposition of multivariate phenotypic means in multigroup genetic covariance structure analysis. *Behavior Genetics, 22*(3), 319–335.

Dolan, C. V., Molenaar, P. C. M., & Boomsma, D. I. (1994). Simultaneous genetic analysis of means and covariance structure: Pearson-Lawley selection rules. *Behavior Genetics, 24*, 17–24.

Eaves, L., Erkanli, A., Silberg, J., Angold, A., Maes, H. H., & Foley, D. (2005). Application of Bayesian inference using Gibbs sampling to item-response theory modelling of multi-symptom genetic data. *Behavior Genetics, 35*(6), 765–780. doi:10.1007/s10519-005-7284-z.

Elks, C. E., Den Hoed, M., Zhao, J. H., Sharp, S. J., Wareham, N. J., Loos, R. J., et al. (2012). Variability in the heritability of body mass index: A systematic review and meta-regression. *Frontiers in Endocrinology, 3*, 29. doi:10.3389/fendo.2012.00029.

Erosheva, E. A., & Curtis, S. M. (2013). Dealing with rotational invariance in Bayesian confirmatory factor analysis. Technical report #589, Seattle, WA: Department of Statistics, University of Washington. http://www.stat.washington.edu/research/reports/2011/tr589.pdf

Franić, S., Dolan, C. V., Borsboom, D., Hudziak, J. J., van Beijsterveldt, C. E. M., & Boomsma, D. I. (2013). Can genetics help psychometrics? Improving dimensionality assessment through genetic factor modeling. *Psychological Methods, 18*, 406. doi:10.1037/a0032755.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., & Vehtari, A. (2013). *Bayesian data analysis* (3rd ed.). New York: CRC press.

Gelman, A., & Meng, X. L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science, 13*(2), 163–185.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(6), 721–741. doi:10.1109/TPAMI.1984.4767596.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in Mplus and SAS. *Structural Equation Modeling, 16*(4), 676–701. doi:10.1080/10705510903206055.

Guo, G., & Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. *Behavior Genetics, 32*(1), 37–49.

Heath, A. C., Eaves, L. J., & Martin, N. G. (1989). The genetic structure of personality III. Multivariate genetic item analysis of the EPQ scales. *Personality and Individual Differences, 10*(8), 877–888.

Heath, A. C., Jardin, R., Eaves, L. J., & Martin, N. G. (1989). The genetic structure of personality II: Genetic item analysis of the EPQ. *Personality & Individual Differences, 10*, 615–624.

Heath, A., Neale, M., Hewitt, J., Eaves, L., & Fulker, D. (1989). Testing structural equation models for twin data using LISREL. *Behavior Genetics, 19*(1), 9–35.

Henderson, N. D. (1982). Human behavior genetics. *Annual Review of Psychology, 33*, 403–440.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329–367.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351–362.

Joseph, J., & Ratner, C. (2013). The fruitless search for genes in psychiatry and psychology: Time to re-examine a paradigm. In S. Krimsky & J. Gruber (Eds.), *Genetic explanations: Sense and nonsense* (pp. 94–106). Cambridge, MA: Harvard University Press.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindsey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). Boston: McGraw-Hill.

Kenny, D. A., & Milan, S. (2013). Identification: A non-technical discussion of a technical issue. In R. Hoyle, D. Kaplan, G. Marcoulides, & S. West (Eds.), *Handbook of structural equation modeling* (pp. 145–163). New York: Guilford.

Lee, S. Y. (1980). Estimation of covariance structure models with parameters subject to functional restraints. *Psychometrika, 45*(3), 309–324.

Lee, S. Y. (2007). *Structural equation modelling: A Bayesian approach*. New York: John Wiley.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653–686.

Lindley, D. V. (1977). A problem in forensic science. *Biometrika, 64*(2), 207–213.

Loehlin, J. C. (1996). The Cholesky approach: A cautionary note. *Behavior Genetics, 26*, 65–69.

Loehlin, J. C., & Martin, N. G. (2013). General and supplementary factors of personality in genetic and environmental correlation matrices. *Personality and Individual Differences, 54*, 761–766.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Ludeke, S., Johnson, W., & Bouchard, T. J. (2013). "Obedience to traditional authority:" A heritable factor underlying authoritarianism, conservatism and religiousness. *Personality and Individual Differences, 55*, 375–380.

Martin, N. G., & Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity, 38*(1), 79–95.

Martin, N., Scourfield, J., & McGuffin, P. (2002). Observer effects and heritability of childhood attention-deficit hyperactivity disorder symptoms. *British Journal of Psychiatry, 180*, 260–265.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11*(4), 344.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*(3), 1142–1160.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*(1), 107–122.

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. Retrieved from https://www.statmodel.com/download/IntroBayesVersion%203.pdf

Muthén, B., & Asparouhov, T. (2011). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. doi:10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Neale, M. C., & Cardon, L. R. (1992). *Methodology for genetic studies of twins and families*. New York: Springer.

Phillips, K., & Matheny, A. P. (1997). Evidence for genetic influence on both cross-situation and situation-specific components of behavior. *Journal of Personality and Social Psychology, 21*(1), 129–138.

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*(6), 565–576. doi:10.1177/0165025409343765.

Rietveld, M. J., Posthuma, I. D., Dolan, C. V., & Boomsma, D. I. (2003). ADHD: Sibling interaction or dominance: An evaluation of statistical power. *Behavior Genetics, 33*(3), 247–255.

Rindskopf, D. (1984). Structural equation models. *Sociological Methods & Research, 13*(1), 109–119.

Roberson-Nay, R., Moruzzi, S., Ogliari, A., Pezzica, E., Tambs, K., Kendler, K. S., et al. (2013). Evidence for distinct genetic effects associated with response to 35% CO_2. *Depression and Anxiety, 30*(3), 259–266. doi:10.1002/da.22038.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1–12. doi:10.1037/a0018326.

Sato, M. (1987). Pragmatic treatment of improper solutions in factor analysis. *Annals of the Institute of Statistical Mathematics, 39*(1), 443–455.

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods, 13*(2), 150–170.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.

Shi, J. Q., & Lee, S. Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51*(2), 233–252.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64*, 583–639.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science, 9*(5), 160–164.

Visscher, P. M., Gordon, S., & Neale, M. C. (2008). Power of the classical twin design revisited: II Detection of common environmental variance. *Twin Research and Human Genetics, 11*, 48–54.

Waldron, M., Bucholz, K. K., Lynskey, M. T., Madden, P. A. F., & Heath, A. C. (2013). Alcoholism and timing of separation in parents: Findings in a Midwestern birth cohort. *Journal of Studies on Alcohol and Drugs, 74*, 337–348.

Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development, 31*(4), 374–383. doi:10.1177/016502540707776.

Part IV
Item-Response-Modeling

Item Response Models for Dependent Data: Quasi-exact Tests for the Investigation of Some Preconditions for Measuring Change

Ingrid Koller, Wolfgang Wiedermann, and Judith Glück

Abstract The Rasch model has several advantages for the psychometric investigation of item quality (e.g., specific objectivity). One approach to testing model fit uses quasi-exact tests which are well suited to test the validity of the Rasch model when sample sizes are rather small. Application of these tests is not restricted to Rasch modeling. In this chapter, we show that these tests can be used to test preconditions for measuring change such as measurement invariance, unidimensionality, and local independence across time points. For example, if items are unidimensional across time points (i.e., all items measure the same latent construct across time) and groups (e.g., control and training groups), it follows that there are no significant interindividual differences within groups and over time. All individuals in a group change in the same direction. On the other hand, significant results across time but not within groups suggest group differences in change, such as training effects. In this chapter, we first give an introduction to quasi-exact tests. Then, we demonstrate the applicability of three test statistics for the investigation of preconditions for measuring change using empirical power analysis and an empirical example concerning spatial ability.

Introduction

The Rasch model (Rasch 1960; see also Fischer & Molenaar 1995) is commonly applied for the psychometric investigation of items. If a data set conforms to the Rasch model, several positive mathematical properties hold for the data (e.g., Fischer & Molenaar 1995; Koller & Hatzinger 2013): (a) *Unidimensionality*: all items of a test measure the same latent construct. (b) *Local independence*: holding

I. Koller (✉) • J. Glück
Department for Psychology, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
e-mail: ingrid.koller@aau.at; judith.glueck@aau.at

W. Wiedermann
Department of Educational, School, and Counseling Psychology College of Education, University of Missouri, Columbia, MO 65211, USA
e-mail: wiedermannw@missouri.edu

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_11

ability constant, an item's probability of being solved does not depend on other items. (c) Parallel and strictly increasing *item characteristic curves* (ICCs): the probability to solve an item strictly increases with person ability, which results in non-overlapping ICCs with the same discrimination for all items. (d) *Specific objectivity*: it is irrelevant which items are used to compare two individuals, and it is irrelevant which individual is used to compare two items. (e) *Measurement invariance* (an important aspect of specific objectivity): subgroups of individuals show the same conditional probabilities of solving items. Testing the assumption of measurement invariance across levels of external variables (e.g., gender) is commonly known as the investigation of differential item functioning (DIF; e.g., Holland & Wainer 1993). (f) If the Rasch model holds for a data set, an individual's raw score contains all information necessary to characterize that individual's ability, and at the same time, the number of individuals who have solved an item (i.e., the sum score) contains all information necessary to determine that item's difficulty. This last property is known as the property of *sufficient statistics* and constitutes the central part of the quasi-exact tests described in this chapter.

The mathematical properties of the Rasch model are important not only for scaling items but also for investigating preconditions for measuring change (e.g., Ponocny 2002) or other cases of dependent data. In this chapter, we focus on preconditions for item response models designed to measure change (see, e.g., Fischer 1974 1989 1995a 1995b; Fischer & Ponocny-Seliger 1998; Formann & Spiel 1989; Glück & Spiel 1997 2007), namely (1) unidimensionality of items across time points, (2) unidimensionality of change, and (3) response independence within items over time. In the following section, we review statistical tests that are commonly used to test these preconditions.

Unidimensionality Between Time Points (The Person Side)

An important question in the analysis of change is whether change occurs for all individuals in the same direction or whether it is necessary to model individual change over time. If individuals change in different directions within groups, but only group differences are modeled, change parameters can be biased. In other words, if the correlation of scores or latent abilities between time points is lower than expected under the Rasch model, the assumption of unidimensionality across time points is violated. There exist different approaches to the investigation of unidimensionality across time points. For example, the mixed Rasch model (e.g., Rost 1990) can be used to detect latent classes within which the Rasch model holds across time points (e.g., Glück & Spiel 1997). Another option is to check unidimensionality on the level of subscales using the Martin-Löf test (e.g., Glück & Spiel 1997; see Eq. (3) below), or to apply the recently proposed likelihood ratio test by Gittler and Fischer (2011). Alternative modeling approaches to address these unidimensionality issues include multidimensional item response models (Adams, Wilson, & Wang 1997) or log-linear representations of these models (e.g., Meiser 1996).

Unidimensionality of Change (The Item Side)

Unidimensionality of change on the item side is another important precondition for drawing valid conclusions about change over time. If this precondition is fulfilled, all items of a test show the same magnitude, and direction of change. Thus, it is irrelevant which items are used to assess change, a property known as specific objectivity of change. Again, violations of this assumption can lead to distorted results, e.g., suggesting no change even though the latent construct of interest changes across time. Because the two preconditions of unidimensionality between time points and unidimensionality of change are not independent of each other, it is also possible to investigate the assumption of unidimensionality of change using the Martin-Löf test or multidimensional item response models, as well as testing the respective model (a single change parameter for all items) against a maximum model that estimates a change parameter for each item separately (see, e.g., Fischer 1976; Glück & Spiel 1997 2007), or using various model tests to assess measurement invariance (Cho, Athay, & Preacher 2013).

Response Independence Between Time Points

The third precondition is response independence across time, which means local independence of items. When the same items are used across time points, the probability of items becoming response dependent increases, for example due to practice effects. Violations of this precondition lead to inflated correlations of the same item across time points. Whether this kind of violation is considered or ignored in the analysis of change is the decision of the researcher. In any case, if violations of response independence are present, researchers should pay attention to the magnitude of the correlations within items over time because highly correlated items impede the assessment of change effects (non-change-sensitive item). A straightforward approach to solve this problem is to use different (but unidimensional) items at each time point (e.g., Embretson 1991).

Again, the Martin-Löf test and multidimensional item response models can be used to determine whether the precondition of response independence is fulfilled. It is also possible to investigate response independence using methods assessing measurement invariance. For example, Andersen's (1973) likelihood ratio test can be used to evaluate the assumption of locally independent items by splitting the data according to an item of interest (for details see Formann 1981; Koller, Alexandrowicz, & Hatzinger 2012).

As discussed above, several methods for the investigation of preconditions have been proposed; a comprehensive overview is given by Fischer and Molenaar (1995). However, all of these methods have the serious drawback that large numbers of

participants and/or items are required (e.g., Fischer 1981; Gittler & Fischer, 2011; Fischer & Molenaar 1995; Glück & Spiel 1997; Ponocny 2001). Another option for the examination of model fit is provided by quasi-exact tests (e.g., Koller, Maier, & Hatzinger 2005; Koller & Hatzinger 2013; Ponocny 2001) which can assess fit of the Rasch model even when a sample is small. These goodness of fit tests are based on the assumption of sufficient statistics and can also be used for the investigation of the three preconditions described above.

The aim of this chapter is to give an overview of quasi-exact tests and to illustrate that these tests can be used to determine whether the three preconditions for valid conclusions concerning change are met. In addition, empirical power analyses and an empirical example are given.

Quasi-exact Tests

Quasi-exact tests for the Rasch model (Koller et al. 2012; Koller & Hatzinger 2013; Ponocny 2001) can be considered a generalization of Fisher's exact test. The idea is based on the mathematical property of sufficient statistics, which implies that all possible matrices with the same margins will have the same parameter estimates. With this property, an exact test can be algorithmically described as follows (a more detailed description is given in Koller & Hatzinger 2013): (1) Consider an observed $r \times c$ matrix $\mathbf{A}_0$ ($r =$ rows, $c =$ columns). (2) All possible matrices with the same margins as $\mathbf{A}_0$ have to be generated, that is, $\mathbf{A}_1, \ldots, \mathbf{A}_s, \ldots, \mathbf{A}_S$. (3) A test statistic T_0 is calculated for $\mathbf{A}_0$ and for all the generated matrices, that is, $T_1, \ldots, T_s, \ldots, T_S$. (4) The p-value of the model test is defined as the relative frequency of the T's which show the same or a more extreme value compared to T_0.

Due to computational limitations, computing all possible matrices with given margins is not always practical. Several authors have addressed this problem by simulating matrices. For example, Verhelst (2008) introduced a Markov Chain Monte Carlo simulation algorithm which is implemented in the package RaschSampler (Verhelst, Hatzinger, & Mair 2007) for the open-source software R (R Core Team 2014). A general description of the simulation algorithm is given by Koller and Hatzinger (2013), and the detailed theoretical background is given in Verhelst (2008) and Verhelst et al. (2007).

Several authors have developed various quasi-exact tests for the mathematical properties mentioned above (e.g., Koller et al. 2012; Koller & Hatzinger 2013; Ponocny 1996 2001; Verhelst et al. 2007). Many of these tests are implemented in the R package eRm (Mair, Hatzinger, & Maier 2014). In this chapter, we focus on three test statistics. Note that other preconditions can also be investigated with quasi-exact tests. Examples are given in Ponocny (2002).

Unidimensionality Between Time Points: The Statistic T_{md}

To investigate the "person side" of unidimensionality, Koller and Hatzinger (2013; see also Koller et al., 2012) proposed the test statistic T_{md}. To calculate this statistic, the set of items ($i = 1, \ldots, k$) is divided into two subsets that represent time point 1 (t_1) and time point 2 (t_2). If the assumption of unidimensionality holds for the observed data, the two raw scores $r_v^{(t1)}$ and $r_v^{(t2)}$ are expected to be positively associated. Low correlations between time points indicate multidimensionality issues between time points. The test statistic can be written as

$$T_{md}(\mathbf{A}) = Cor\left(r_v^{(t1)},\ r_v^{(t2)}\right) \quad \text{where} \quad \left(r_v^{(t.)}\right) = \sum\nolimits_{i \in t} x_{vt.} \tag{1}$$

The model test statistic is given in Eq. (2) and is defined as the relative frequency of T_s (1, ..., s, ..., *nsim*), where *nsim* is the number of simulated matrices which has the same correlation as T_0 or a smaller correlation (the number is denoted with $d = 1$, ..., s, ..., *nsim*). If more than two time points are involved, each combination must be investigated separately, i.e., for three points in time, t_1 vs. t_2, t_1 vs. t_3, and t_2 vs. t_3.

$$p = \frac{1}{nsim}\sum\nolimits_{s=1}^{nsim} d_s \quad \text{where} \quad d_s = \begin{cases} 1, \text{ if } T_s(\mathbf{A}_s) \leq T_0(\mathbf{A}_0) \\ 0, \text{ else} \end{cases} \tag{2}$$

A nonsignificant result suggests that the assumption of unidimensionality across time points holds. Thus, there is either no change or homogeneous change for all individuals. Hypotheses about different types of unidimensionality can be investigated this way as well. When the null hypothesis of unidimensionality is rejected, further analyses are required to test whether there is multidimensionality on the person side (across time points) or whether one or more items change in a specific way (see section "Unidimensionality of Change"). In the first case, the assessment of multidimensionality on the person side, subgroup comparisons (e.g., control vs. experimental group) can be performed. When the correlation between raw scores is lower than expected for the aggregate of the data set, but nonsignificant results are observed within groups, then unidimensionality holds within but not across the groups, which suggests group-specific change. However, if the results are significant within groups, this may result from person-specific changes or item-specific changes. In these cases, models for the assessment of group-specific change and models which assume unidimensionality across time points cannot be used for the assessment of change.

It is also possible to investigate the precondition of unidimensionality on the item level using the test statistics T_{1m} or T_{2m}. Details can be found in Koller et al. (2012) and Koller and Hatzinger (2013).

In the following section, simulation results on the type-I error and power performance of the proposed test statistic T_{md} are reported.

Empirical Power Analysis

Multidimensional data were simulated using the multidimensional random coefficient multinomial logit model (MRCMLM; Adams et al. 1997) implemented in the function sim.xdim() in eRm. Simulations were carried out with sample sizes $n = 30$, 50, 200, and 500 and three test lengths $k = 10$ (5 items at $t_1 + 5$ items at t_2), $k = 20$ (10 items at $t_1 + 10$ items at t_2), $k = 40$ (20 items at $t_1 + 20$ items at t_2). Item parameters at each time point (i.e., each dimension) were drawn from a uniform distribution with a mean of zero and a range of [−2, 2], and person parameters from a bivariate standard normal distribution also with a range of [−2, 2]. The latent correlations between the two time points were $\rho(\theta_{t1}, \theta_{t2}) = 0$, 0.3, 0.5, and 0.8. For each of the 48 combinations (4 sample sizes × 3 test lengths × 4 correlations = 48), 1000 simulations were carried out.

To examine type-I error rates ($\alpha = 5$ %), data sets were generated so that all item parameters were the same at both time points. Thus, for the second time point the weights were fixed at zero ($D_1 = 1$, $D_2 = 0$). In this scenario the assumption of unidimensionality holds. Thus, the null hypothesis of the test statistic in Eq. (2) is expected to be rejected according to the nominal significance level of 5 %.

In the case of multidimensionality two different models of multidimensionality were defined following Adams et al. (1997):

1. All items at t_2 are influenced by an additional dimension, implying multidimensionality within time point 2 (see Fig. 1, left panel). Different weights were used so that the effect of the second dimension on the items at t_2 increases whereas the effect of the first dimension decreases in decrements of 0.2 ($D_1 = 0.8$, $D_2 = 0.2$; $D_1 = 0.6$, $D_2 = 0.4$; $D_1 = 0.4$, $D_2 = 0.6$; $D_1 = 0.2$, $D_2 = 0.8$).

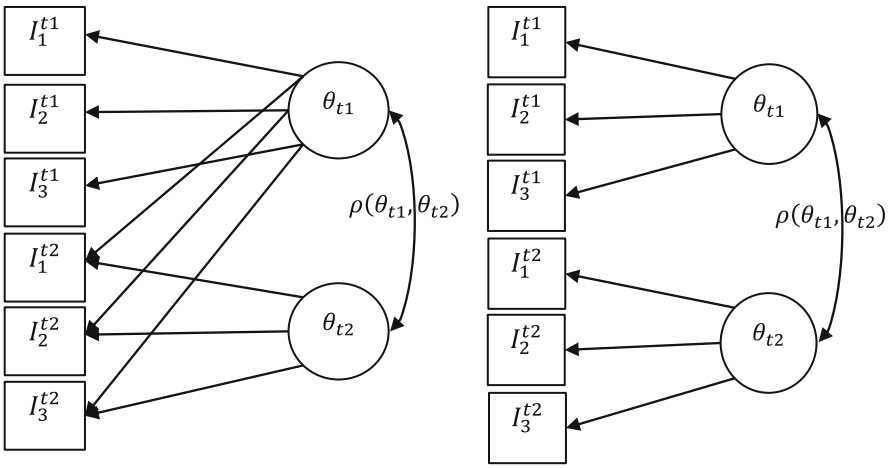


Fig. 1 Within-time point multidimensionality (*right panel*) and between-time point multidimensionality (*left panel*) according to Adams et al. (1997)

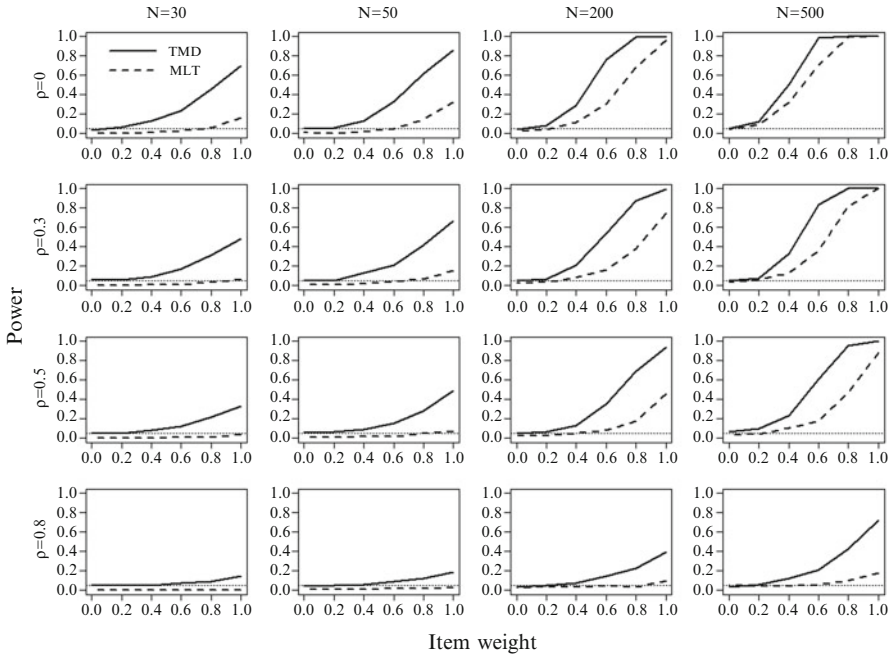


Fig. 2 Results for the test length $k = 5 + 5 = 10$; *solid lines* represent the T_{md} and the *dashed lines* represent the *MLT*. The *x*-axis gives the weights of the second dimension and the *y*-axis gives the probability of detecting a model violation (empirical power). The *dotted line* indicates the nominal significance level of 5 %

2. The items at t_2 measure another dimension, implying multidimensionality across time points (see Fig. 1 right panel; $D_1 = 0$, $D_2 = 1$).

In addition, we compared the performance of the quasi-exact test with that of the Martin-Löf test (*MLT*; as described in Fischer & Molenaar 1995). The *MLT* is a popular likelihood ratio test of the unidimensionality assumption (see, e.g., Verhelst 2001). As in T_{md}, the data set is split into two subgroups of items (or, as in the current case, two time points). Then the item parameters for the overall sample and for both subsamples are estimated and compared. The *MLT* can be written as

$$MLT\,(\mathbf{A}) = 2ln\left(\frac{\prod_w\prod_u\left(\frac{n_{\{wu\}}}{n}\right)^{n_{\{wu\}}}}{\prod_r\left(\frac{n_r}{n}\right)^{n_r}} \times \frac{L_c^{(1)} \times L_c^{(2)}}{L_c^{(0)}}\right), \tag{3}$$

where $w = 1, \ldots, k_1$ is the raw score for the first subset of items, $u = 1, \ldots, k_2$ is the raw score for the second subset of items, $r = 1, \ldots, k$ is the raw score for the overall data matrix (i.e., $r = w + u$), n_r, n_w, and n_u are the frequencies of the raw scores r, w, and u, n is the number of observations, $L_c^{(0)}$ is the conditional likelihood for the item parameters estimated for the overall data matrix, and $L_c^{(1)}$ and $L_c^{(2)}$ are

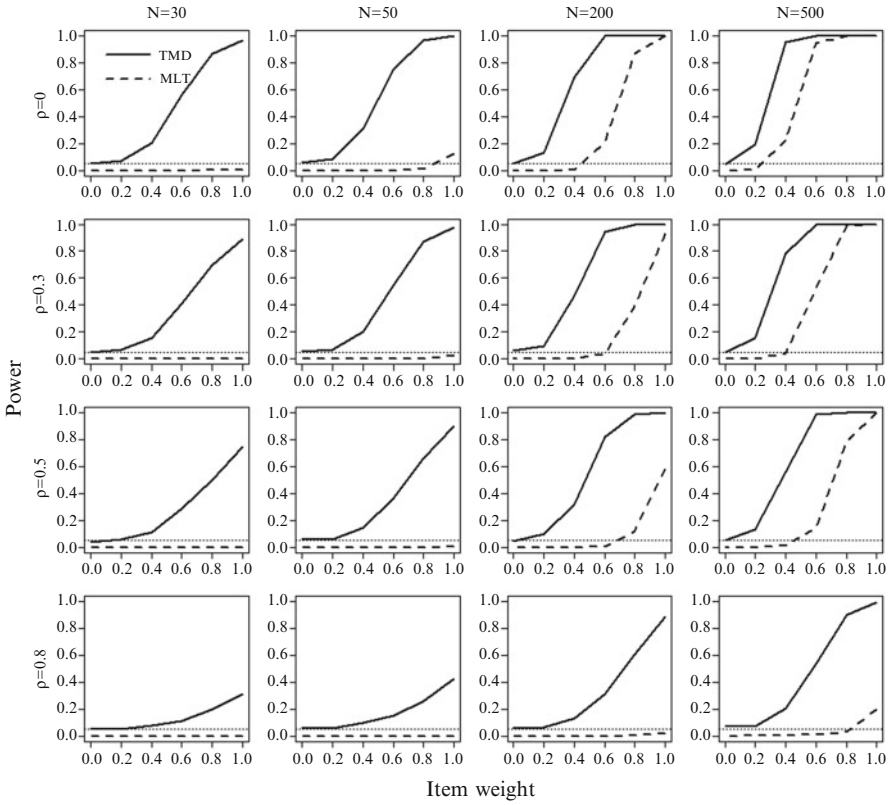


Fig. 3 Results for the test length $k = 10 + 10 = 20$; *solid lines* represent the T_{md} and the *dashed lines* represent the *MLT*. The *x*-axis gives the weights of the second dimension and the *y*-axis gives the probability of detecting a model violation (empirical power). The *dotted line* indicates the nominal significance level of 5 %

the conditional likelihoods for the item parameters estimated for the two subsets. The *MLT* statistic is asymptotically distributed as χ^2 with $k_1 + k_2 - 1$ degrees of freedom.

Results are given in Figs. 2, 3, and 4. The solid lines represent the results for T_{md} and the dashed lines represent the results for *MLT*. On the *x*-axis, the weights of the second dimension are displayed, e.g., a value of 0.8 means that the second dimension has a weight of 0.8 and the first dimension of 0.2.

Overall, the T_{md} was able to protect the nominal significance level of 5 % in all scenarios. The *MLT* showed deflated type-I error rates. In the shortest test-length scenario ($k = 5 + 5 = 10$), the type-I error rates for the MLT increase with sample size and were around 5 % at $n = 200$. This result is in line with results given in Futschek (2014), Verguts and DeBoeck (2001), and Verhelst (2001).

The cases of $D_2 > 0$ depict the empirical power of the test. In general, the power of both test statistics increased with sample size and decreased with the magnitude

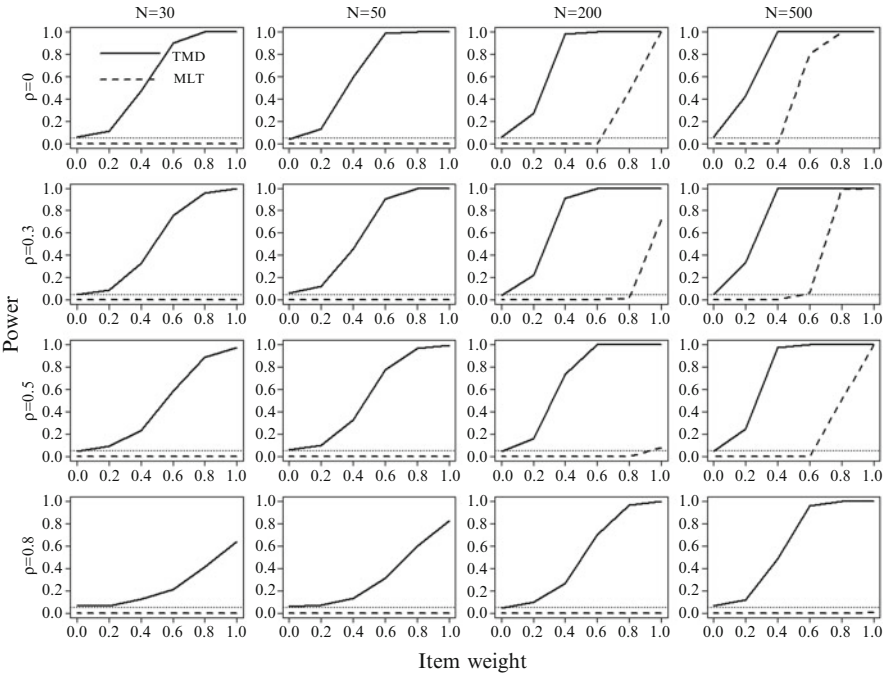


Fig. 4 Results for the test length $k = 20 + 20 = 40$; *solid lines* represent the T_{md} and the *dashed lines* represent the *MLT*. The *x*-axis gives the weights of the second dimension and the *y*-axis gives the probability of detecting a model violation (empirical power). The *dotted line* indicates the nominal significance level of 5 %

of correlation between the latent traits. In addition, however, there were several differences in performance. First, the power of T_{md} increased with test length, whereas the power of the *MLT* decreased with test length. This can be explained as follows. For constant sample size, longer tests provide more information for generating permutation matrices in the T_{md} approach, but increase the magnitude of estimation errors in the MLT. Second, the T_{md} detected within multidimensionality (range 0.2–0.8) and between multidimensionality (1.0) even in small samples. In addition, the T_{md} detected multidimensionality in large samples and longer tests even if the latent traits were highly correlated. In sum, T_{md} outperformed the *MLT* regardless of sample size.

Unidimensionality of Change: The Statistic T_4

To test item-specific change or measurement invariance, the statistic T_4 (Ponocny 2001; see also, Koller & Hatzinger 2013) can be used. This statistic can be used to evaluate whether one or more items are easier or more difficult in a predefined

group of individuals (or, when measuring change, at t_i) than in another group (or at t_j). The test statistic can be written as

$$T_4(\mathbf{A}) = \sum\nolimits_{v \in G_g} x_{vi}, \tag{4}$$

where the x_{vi} of an item in the predefined group of individuals ($g = 1, \ldots, G$) are summed up over all points in time. When an item within the predefined group of individuals is more difficult than expected under the Rasch model (i.e., the number of correct responses is smaller than expected), the model test is given in Eq. (2). To test the assumption that an item is easier than expected, the model test is

$$p = \frac{1}{nsim} \sum\nolimits_{s=1}^{nsim} d_s \quad \text{where} \quad d_s = \begin{cases} 1, \text{ if } T_s(\boldsymbol{A}_s) \geq T_0(\boldsymbol{A}_0) \\ 0, \text{ else} \end{cases} \tag{5}$$

According to Cho et al. (2013), four different violations of measurement invariance across time are of interest: (1) The item parameters differ across person groups (see above). (2) The item parameters differ across time points. (3) The item parameters differ across person groups within a time point. (4) The item parameters differ across time points within a person group. These four potential violations can be analyzed using T_4 by rearranging the data matrix, for example, using the same individuals at t_2 as "virtual" individuals at t_1 and using the time point as splitting variable. Further tests are discussed in Koller and Hatzinger (2013), Koller et al. (2012), and Ponocny (2001).

An empirical power analysis for T_4 (Koller, Maier, & Hatzinger 2015) showed that the type-I error rates of T_4 were far below the nominal level of 5 %. Thus, the test tends to be rather conservative, which may lead to increased type-II error rates. Severe model violations can still be detected with samples of sizes $n = 50$ or $n = 100$, and $n = 200$ seems sufficient to detect weaker violations regardless of the shape of the ability distribution.

Response Independence Between Time Points: The Statistic T_2

The statistic T_2 (Ponocny 2001; see also, Koller & Hatzinger 2013) is well suited to test the assumption of response independence between items. T_2 tests for increased dispersion of raw scores for a set of items. The test statistic can be written as

$$T_2(\mathbf{A}) = Var\left(r_v^{(I)}\right) \quad \text{with} \quad \left(r_v^{(I)}\right) = \sum\nolimits_{i \in I} x_{vi}. \tag{6}$$

$Var(r_v^{(I)})$ is the variance of a subscale I with a minimum of at least two items, which can also mean the same item at t_1 and t_2. According to the variance addition theorem, the variance of a scale that consists of two subscales t_1 and t_2 is defined

Table 1 Results of the empirical power analysis for T_2

N	*30*				*50*			
Items	*RM*	$\bar{r} = .47$	$\bar{r} = .71$	$\bar{r} = .86$	*RM*	$\bar{r} = .47$	$\bar{r} = .71$	$\bar{r} = .86$
$k = 5$	.025	.335	.811	.936	.026	.562	.966	.988
$k = 10$	.013	.372	.860	.965	.025	.655	.974	.995
$k = 20$	.020	.390	.872	.965	.029	.649	.968	.996
$k = 40$	.016	.398	.879	.976	.032	.632	.978	.996
N	*200*				*500*			
	RM	$\bar{r} = .47$	$\bar{r} = .71$	$\bar{r} = .86$	*RM*	$\bar{r} = .47$	$\bar{r} = .71$	$\bar{r} = .86$
$k = 5$	.025	.980	1.00	1.00	.042	.999	1.00	1.00
$k = 10$	.035	.994	.999	1.00	.043	.999	1.00	1.00
$k = 20$	.042	.986	.999	1.00	.032	.999	1.00	1.00
$k = 40$	.034	.993	1.00	1.00	.030	.999	1.00	1.00

Note. The column RM (Rasch model) shows the empirical error rates (e.g., .025 is 2.5 %) and the columns $\bar{r}$ show the average between-item correlations

as $Var\left(r_v^{(t1)}\right) + Var\left(r_v^{(t2)}\right) + 2 \times Cov\left(r_v^{(t1,t2)}\right)$ and is expected to increase with the covariance of t_1 and t_2. The model test, given in Eq. (5), tests whether the items are more highly correlated than assumed under the Rasch model.

As for the previously described statistics, the following section describes the results of a simulation study on the type-I error rates and empirical power performance.

Empirical Power Analysis

Violations of the assumption of response independence of two items were simulated as by Marais and Andrich (2008; cf. Andrich & Kreiner 2010). Simulations with 1000 samples each were performed for each combination of sample sizes $n = 30, 50$, 200, and 500 and test lengths $k = 5$, 10, 20, and 40. Item and person parameters were drawn from a standard normal distribution with a range of $[-2, 2]$. For each sample, two moderately difficult items showed violations of the response independence assumption, namely for $k = 5$: (i = 2, i = 3), $k = 10$: (i = 5, i = 6), $k = 20$: (i = 10, i = 11), and $k = 40$: (i = 20, i = 21). Four different weights were used to simulate different magnitudes of violations ($d = 0$, 1, 2, and 3). A weight of zero represents the case of no violation, i.e., the type-I error scenarios. Cases of $d > 0$ correspond to average manifest correlations between items of $\bar{r}_1 = .474$, $\bar{r}_2 = .710$, $\bar{r}_3 = .855$.

Simulation results are displayed in Table 1. In general, the empirical error rates of T_2 (column RM) increased with sample size. The rejection rates were generally far below the nominal level of 5 %. In other words, T_2 tends to be rather conservative, which results in increased type-II error rates. However, small departures from the response independence assumption can be detected even in very small samples. In addition, power increases with sample size and magnitude of violations.

In sum, the power analyses suggest that quasi-exact tests are well suited for small samples. Next, an illustration is given of these tests for the assessment of the three preconditions concerning measurement of change using data from a spatial ability training study.

An Empirical Example: A Spatial Ability Training Study

The data set was collected in the project "Educating Spatial Ability with Augmented Reality" (Kaufmann, Steinbügl, Dünser, & Glück 2005). The study compared the effects of a spatial ability training intervention in an augmented reality (AR)-based three-dimensional setting to a two-dimensional intervention and a no-training control condition. Training effects were measured by a battery of paper-pencil spatial ability tests including the Mental Cutting Test (MCT, CEEB College Entrance Examination Board 1939). Each item of the MCT consists of a perspective drawing of a solid figure that is cut by a plane. Participants are asked to imagine the shape of the cross section and select the correct solution out of five alternatives. The original test consists of 25 items, but, in the current study, a 15-items short version was used.

The sample consisted of 317 high school students, 213 of whom completed both pretest and posttest (51.6 % males; age: $M = 17.0$, $SD = 1.1$, $min = 14.4$, $max = 20.5$). As no differences between the three-dimensional and the two-dimensional training were found, we focus on comparing the control group (CG; $n = 123$) to the training group (TG; $n = 90$) that participated in six weekly training sessions. TG participants were trained either with a computer-aided design software presented two dimensionally on the computer screen or three dimensionally using Augmented Reality (e.g., Kaufmann 2004 2006; Kaufmann & Schmalstieg, 2003). In the following analyses, we use the data from this study to illustrate the usefulness of quasi-exact tests to test preconditions for measuring change.

General results of the training study are reported by Dünser (2005) and Kaufmann et al. (2005); specific results for the MCT can be found in Koller (2010), where quasi-exact tests were used to investigate whether the Rasch model fits the items of the MCT. The Rasch model held at the first and second time point for eight of the 15 items. These items were used here to test the three preconditions for measuring change.

Additionally, it is not possible to apply quasi-exact tests to data including missing values. However, a systematic analysis of the performance of missing value imputation algorithms in the context of quasi-exact tests is clearly beyond the scope of the present chapter. Thus, only the data from 183 individuals who had no missing values across the eight items were used for the analyses. The sample consisted of 81 females and 102 males; 79 of the participants were in the training group.

Results

First, the assumption of unidimensionality across time points was assessed using T_{md}. Analyses were performed separately for the two groups CG and TG. Results suggest that the assumption of unidimensionality holds across time for both groups ($p_{CG} = .789$; $p_{TG} = .998$). For example, $p_{CG} = .789$ means that 78.9 % of the simulated matrices showed the same or a more extreme violation of the unidimensionality assumption. In addition, we analyzed whether the data were sufficiently unidimensional across time for females and males. Again, these results were not significant ($p_{male} = .972$; $p_{female} = .977$). Thus, the test measures the same latent dimension across time points.

Even a high correlation between raw scores over time still allows for changes in the difficulty of individual items; for example, one item might have become easier while another became more difficult from t_1 to t_2. Thus, second, the assumption of measurement invariance over time was assessed by splitting the data set according to time point in general and per group and analyzing whether the items at t_2 were significantly easier than expected under the Rasch model (one test for each item). In this analysis, a low p-value for an item suggests that the item was indeed easier at t_2 than at t_1. On the other hand, a very high p-value, for example .90, would imply that 90 % of the simulated matrices showed the same or a lower number of correct answers on the item (i.e., a higher item difficulty) than in the observed matrix. In such cases, we additionally used the statistic T_4 to assess whether the item was more difficult at t_2 than assumed under the Rasch model.

The results, given in Table 2, suggest that the first item was easier than expected under the Rasch model in the whole sample and three of the four subgroup analyses. Thus, this item was significantly easier at t_2 than at t_1 for females, for males, and for the training group. For the control group the result was nonsignificant, which suggests that the effect was largely due to the training. On the other hand, item 7 was more difficult at t_2 than at t_1 in the total sample, though not clearly in any of

Table 2 Results concerning measurement invariance across time points

Items/groups	Overall	Females	Males	Control group	Training group
1	.003	.054	.014	.141	.008
2	.796	.703	.797	.175	.981 (.035)
3	.856	.462	.953 (.090)	.894	.711
4	.706	.452	.865	.899	.444
5	.149	.115	.430	.031	.725
6	.451	.575	.438	.455	.555
7	.963 (.051)	.943 (.113)	.915 (.140)	.964 (.089)	.884
8	.667	.967 (.066)	.192	.891	.395

Note. The table shows the p-values for the hypothesis that the item was easier than expected at t_2. If an item had a p-value above .90, we tested whether the item is more difficult than expected at t_2. These results are given in parentheses

Table 3 Results for the investigation of the assumption of response independency

p-value	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Overall	.001	.004	.668	.043	.246	.170	.001	.005
CG	.017	.061	.518	.062	.220	.253	.020	.225
TG	.001	.019	.749	.192	.355	.439	.070	.007
Females	.001	.113	.272	.144	.996	.254	.235	.016
Males	.009	.006	.924	.181	.045	.331	.001	.051

the subgroups. This suggests that training and practice effects did not affected the difficulty of this item. This may imply that this item measures other performance components than the others. These violations of measurement invariance should be considered in the analysis of change and modeling of change effects (e.g., specific change parameters for item groups).

Third, the assumption of response independence over time was investigated for each item separately. The same splitting variables were used as before (i.e., CG vs. TG, females vs. males). As explained earlier, the test statistic T_2, given in Eq. (5), tests whether item responses at t_1 and t_2 are more highly correlated than assumed under the Rasch model. The results in Table 3 suggest several significant response dependencies, most consistently, for items 1, 2, 7, and 8. This type of dependency is typical when the same items were presented at consecutive points in time, and the time interval between two assessments is short. Together with the significant results of the previous analysis, the present results suggest that most participants who solved Item 1 or Item 7 at t_1 also solved these items at t_2, but of those few participants whose response did change, the majority moved from not solving to solving for Item 1 and from solving to not solving for Item 7. Interestingly, the items in the middle of the test (items 3, 4, 5, and 6) showed no significant response dependencies. It may be very interesting to assess the change parameters for both groups of items separately and to compare the results.

In sum, the analyses showed that, on the level of correlations between the raw scores, the assumption of unidimensionality holds for the data set. However, when individual items were inspected, the analyses showed violations of the assumptions of measurement invariance and response independence for some items. These results suggest concrete alternatives for the modeling of change in mental rotation test performance. Although the requirements of classical Rasch-family models for measuring change are not fulfilled, other models can be used to model the changes in this data set. For example, researchers may model several change parameters for the items showing violations of measurement invariance and only one change parameter for the middle group of items. Several item response models are available for measuring change in this way, for example, explanatory item response models (e.g., Cho et al. 2013; Stevenson, Hickendorff, Resing, Heiser, & DeBoeck 2013) or multidimensional Rasch models (e.g., Koller, Carstensen, Wiedermann, & vonEye 2014; Wang, Wilson, & Adams 1998).

Conclusion

In this chapter, we hope to have shown that (1) quasi-exact tests are very well suited to evaluate Rasch model conformity even in small samples, that (2) they can also be used to test important preconditions of item response theory models for measuring change, and (3) yield additional information for model selection.

The empirical power results presented here suggest good performances of all proposed tests. For example, T_{md} is an excellent statistic for the investigation of between-time point and within-time point multidimensionality in small as well as larger samples. Another advantage of T_{md} is the possibility to detect group-specific multidimensionality, even when latent traits are highly correlated. In the simulation, T_{md} outperformed the *MLT* in all cases.

Of course, further studies are needed to systematically evaluate the behavior of these test statistics under various conditions (see also Koller et al., 2014). For example, future studies should evaluate the behavior of the tests in cases of varying item discrimination and/or multiple model violations. In addition, item position effects which violate one of the underlying mathematical properties of the Rasch model should be investigated in more detail. Furthermore, for T_{md}, further simulation scenarios are needed in which not all items show violations of the unidimensionality assumption, and in which more than two dimensions influence the probability of solving an item.

The current permutation algorithm does not allow missing values. Thus, researchers have to decide a priori whether cases with missing values are removed from the data set or whether missing value imputation methods are carried out prior to the psychometric analysis. First promising results of applying quasi-exact tests to dichotomous data including missing values and trichotomous items are given in Verhelst and Gruber (2013).

Acknowledgement This research was partly funded by the Austrian Research Fund, grant nr. P 16803-N12.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123–140.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181–192.

CEEB College Entrance Examination Board. (1939). *Special aptitude test in spatial relations*. New York, NY: CEEB.

Cho, S.-J., Athay, M., & Preacher, K. J. (2013). Measuring change for a multidimensional test using a generalized explanatory longitudinal item response model. *British Journal of Mathematical and Statistical Psychology, 66*(2), 353–381.

Dünser, A. (2005). *Trainierbarkeit der Raumvorstellung mit Augmented Reality* [*Trainability of spatial ability with Augmented Reality*]. Unpublished doctoral thesis, University of Vienna, Austria.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495–515.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen*. Bern: Huber.

Fischer, G. H. (1976). In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York, NY: John Wiley.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59–77.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*(4), 599–624.

Fischer, G. H. (1995a). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York, NY: Springer.

Fischer, G. H. (1995b). Some neglected problems in IRT. *Psychometrika, 60*(4), 459–487.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. Handbook of the usage of LPCM-WIN 1.0*. Groningen: ProGAMMA.

Formann, A. K. (1981). Über die Verwendung von Items als Teilungskriterium für Modellkontrollen im Modell von Rasch [The application of items as split criterion for goodness of fit tests for the Rasch model]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 28*(4), 541–560.

Formann, A. K., & Spiel, C. (1989). Measuring change by means of a hybrid variant of the linear logistic model with relaxed assumptions. *Applied Psychological Measurement, 13*(1), 91–103.

Futschek, K. (2014). Actual type-I- and type-II-risk of four different model tests of the Rasch model. *Psychological Test and Assessment Modeling, 56*(2), 168–177.

Gittler, G., & Fischer, G. (2011). IRT-based measurement of short-term changes of ability, with an application to assessing the "Mozart Effect". *Journal of Educational and Behavioral Statistics, 36*(1), 33–75.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches. *Methods of Psychological Research Online, 2*(1). Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/.

Glück, J., & Spiel, C. (2007). Using item response models to analyze change: Advantages and limitations. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 349–361). Oxford: Oxford University Press.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, MI: Erlbaum.

Kaufmann, H., & Schmalstieg, D. (2003). Mathematics and geometry education with collaborative augmented reality. Computer & Graphics, 27(3), 339–345.

Kaufmann, H. (2004). *Geometry education with augmented reality.* Unpublished doctoral thesis, Technical University of Vienna, Austria. Retrieved from http://www.ims.tuwien.ac.at

Kaufmann, H. (2006, August). *The potential of augmented reality in dynamic geometry education.* Paper presented at the 12th International Conference on Geometry and Graphics (ICGG), Salvador, Brazil.

Kaufmann, H., Steinbügl, K., Dünser, A., & Glück, J. (2005). General training of spatial abilities by geometry education in augmented reality. *Annual Review of Cyber Therapy and Telemedicine: A Decade of VR, 3*, 65–76.

Koller, I. (2010). *Item response models in practice: Testing the assumptions in small samples and comparing different models for repeated measurements.* Unpublished doctoral thesis, University of Klagenfurt, Austria.

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012). *Das Rasch Modell in der Praxis: Eine Einführung mit eRm [The Rasch model in practical applications: An introduction using eRm]*. Wien: facultaswuv, UTB.

Koller, I., & Hatzinger, R. (2013). Nonparametric tests for the Rasch model: Explanation, development, and application of quasi-exact tests for small samples. *Interstat,* 1–16. Retrieved from http://interstat.statjournals.net/INDEX/Nov13.html

Koller, I., Maier, M. J., & Hatzinger, R. (2015). An Empirical power analysis of quasi-exact tests for the rasch model: *Measurement invariance in small Samples. Methodology, 11*(2), 45–55.

Mair, P., Hatzinger, R., & Maier, M. J. (2014). *eRm: Extended Rasch Modeling.* [Computer software]. R package version 0.15-3. Retrieved from http://CRAN.R-project.org/package=eRm

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200–215.

Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. *Psychometrika, 61*(4), 629–645.

Ponocny, I. (1996). *Kombinatorische Modelltests für das Rasch-Modell.* [Combinatorial goodness-of-fit tests for the Rasch model.] Unpublished doctoral thesis, University of Vienna, Austria.

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika, 66*(3), 437–460.

Ponocny, I. (2002). On the applicability of some IRT models for repeated measurement designs: Conditions, consequences, and goodness-of-fit tests. *Methods of Psychological Research Online, 7*(1), 22–40.

R Core Team. (2014). *R: A language and environment for statistical computing.* [Computer software] R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Kopenhagen: Danish Institute for Educational Research.

Rost, J. (1990). An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282.

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & DeBoeck, P. A. L. (2013). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence, 41*(3), 157–168.

Verguts, T., & DeBoeck, P. (2001). Some Mantel-Haenszel tests of Rasch model assumptions. *British Journal of Mathematical and Statistical Psychology, 54*(1), 21–37.

Verhelst, N. D. (2001). Testing the unidimensionality assumption of the Rasch model. *Methods of Psychological Research Online, 6*(3), 231–271.

Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika, 74*(4), 705–728.

Verhelst, N. D., & Gruber, K. (2013). *The PCM2-sampler*. Paper presented at the Psychoco 2013 (International Workshop on Psychometric Computing), Zurich, Switzerland. Retrieved from http://eeecon.uibk.ac.at/psychoco/2013/

Verhelst, N. D., Hatzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software, 20*(4). Retrieved from http://www.jstatsoft.org.

Wang, W.-C., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with multidimensional Rasch model. *Journal of Outcome Measurement, 2*(3), 240–265.

Measuring Competencies across the Lifespan - Challenges of Linking Test Scores

Steffi Pohl, Kerstin Haberkorn, and Claus H. Carstensen

Abstract The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole lifespan. Competencies are assessed via tests and competence scores are estimated based on models of Item Response Theory (IRT). IRT allows a comparison of test scores—and, thus, the investigation of change across time and differences between cohorts—even when the respective competence is measured with different items. As in NEPS for most of the competencies retest effects are assumed, linking is done via additional link studies in which the tests for two age groups are administered to a separate sample of participants. However, in order to be able to link the test results of two different measurement occasions, certain assumptions, such as, that the measures are invariant across samples and that the tests measure the same construct, need to hold. These are challenging assumptions regarding the linking of competencies across the whole lifespan. Before linking reading tests in NEPS for different age cohorts in secondary school as well as in adulthood, we, thus, investigated unidimensionality of the items for different cohorts as well as measurement invariance across samples. Our results show that the tests for different age groups do measure a unidimensional construct within the same sample. However, measurement invariance of the same test across different samples does not hold for all age groups. Thus, the same test exhibits a different measurement model in different samples. Based on our results, linking may well be justified within secondary school, while linking test scores in secondary school with those in adult age is threatened by differences in the measurement model. Possible reasons for these results are discussed and implications for the design of longitudinal studies as well as for possible analyses strategies are drawn.

S. Pohl (✉)
Faculty of Education and Psychology, Free University, Habelschwerdter Allee 45, 14195 Berlin, Germany
e-mail: steffi.pohl@fu-berlin.de

K. Haberkorn • C.H. Carstensen
Otto-Friedrich-University, Bamberg, Germany

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_12

Large-scale assessments generally aim at drawing inferences about individuals' knowledge, competencies, and skills (Popham 2000). Thus, international large-scale assessments such as the Program for International Student Assessment (PISA; e.g., OECD 2013), the Third International Mathematics and Science Study (TIMSS; e.g., Mullis, Martin, Foy, & Arora 2012), or the Progress in International Reading Literacy Study (PIRLS; e.g., Mullis, Martin, Foy, & Drucker 2012) aim at accurately measuring competencies, such as reading comprehension or mathematical literacy, of participants. As most of these studies have a cross-sequential design, with a new sample being drawn at every cycle, an investigation of competence development and factors influencing this development is limited. This is different in longitudinal studies, such as the National Educational Panel Study (NEPS, see Blossfeld, von Maurice, & Schneider 2011), where due to the repeated measurement of competencies, competence development may be investigated. Specifically, the NEPS is the only study so far considering competence development across the whole lifespan, from newborns to adults. It does, thus, provide a rich data pool for the investigation of competence development. In order to investigate competence development, competence scores need to be linked across test administrations and test forms. While linking has so far been performed in studies across smaller age ranges, it has not been investigated whether assumptions necessary for linking test scores hold in studies across such a long age span as in NEPS. In this study, we investigated whether it is possible to link test scores for reading competence across the lifespan. In the following sections, we discuss the necessity of linking, describe different link designs, delineate the assumptions of linking, and discuss their plausibility in longitudinal studies. We then present the National Educational Panel Study and derive specific research questions.

Linking of Test Scores

Necessity of Linking Test Scores

It is often not feasible to administer the same test to the participants across time or age, but tests need to be adapted in difficulty and content to the respective age group. Thus, a direct comparison of competence scores from different tests is not possible, since differences in competence values across different tests represent both, differences in competence *and* difference in test items. In longitudinal studies, it is a major aim to investigate competence development over time or to compare the competencies of different age cohorts. In order to be able to compare competence scores across time or cohorts from different tests assessing the same dimension, the test scores need to be linked.

As described by von Davier, Carstensen, and von Davier (2008) and Kolen and Brennan (2004) linking means to establish a common scale for different measurement instruments that are intended to measure the same construct. *Vertical* linking allows for placing the competence scores of different test forms for different age groups on the same scale, thus allowing for a comparison of these test scores. IRT provides means to develop vertical scales encompassing different test versions. In order to obtain a common scale, certain test designs and analyses methods are necessary.

Link Designs

For linking of test scores, some common information or overlap between different test administrations (say grades) to be linked is needed. This can be achieved by various linking designs (for an overview see, e.g., Kolen & Brennan 2004; Reckase 2009, or von Davier et al. 2008). Overlap can be achieved by collecting common observations in (a) a common-person design, (b) a common-item design, or (c) a scaling-test design. In a common-person design, a sample of subjects takes the two test forms to be linked. Because of the single group completing both tests, differences in the scores on these tests can be attributed to differences in the test forms. In a common-item design, two samples of different populations take two tests and the link is established by a set of common items within both tests (anchor items). This design is also called the nonequivalent group anchor test (NEAT) design (e.g., Reckase 2009; von Davier et al. 2008). Assuming invariance of item functioning (i.e., no item drift), the common items may be used as anchors for establishing a common scale between the test versions. In vertical linking, the common-item design with overlapping items is often used across adjacent grades. The scaling-test design can be seen as a special form of the common-item design. Whereas in the common-item design, anchor items are usually administered across adjacent grades, in a scaling-test design, a common test, appropriate to all levels of ability, is implemented in each grade in addition to grade-specific items. Consequently, all students of a study deal with the same test and additionally answer items specifically constructed for their age group. There are different challenges associated with each of these designs which have to be considered (Kolen & Brennan 2004). For instance, in the common-item or scaling-test design, one has to assume that there are no retest effects. Otherwise, item drift might occur and the measurement model would change. There is no such threat in the common-person design; instead this design requires drawing an additional sample, which is less economic, and the challenge arises from an adequate sampling strategy. Note that it is also possible to combine the different designs to build more complex data collection designs (see also Dorans, Pommerich, & Holland 2007; von Davier, Holland, & Thayer 2004).

Coherence of Measurement

Assumptions for Linking

In order to establish a link between test forms that allows one to depict change across time or cohorts, certain assumptions need to be fulfilled (e.g., Camilli, Yamamoto, & Wang 1993; Doran & Cohen 2005; Hoover 1984; Linn 1993; Mislevy 1992; Tong & Kolen 2007): the construct to be measured needs to be the same across (a) samples and (b) tests. This implies that (a) measurement invariance of the same items in different samples holds and that (b) the items of two different tests form a unidimensional construct. Violations of these assumptions may lead to errors in linking (Monseur & Berezner 2007; Monseur, Sibberns, & Hastedt 2008). As a consequence, change scores do not only represent competence development but also changes in the test instrument and inferences on competence development or cohort differences will be biased.

Plausibility of Assumptions in Empirical Studies

Some researchers have stated that the assumption of measuring the same construct are hardly met in applications (e.g., Martineau 2006; Reckase & Martineau 2004; Wang & Jiao 2009). For instance, Wu (2010) reported that "In general, the further the grades are apart the less reliable the vertical scaling across grades is found to be" (p. 23). We draw on studies assessing competencies that incorporated longitudinal or multi-cohort designs for collecting evidence on whether and how coherent measurement of competencies may be obtained. We first reviewed studies that Kristen, Römmer, Müller, and Kalter (2005) found in a systematic stocktaking of the most important longitudinal studies on educational pathways in selected countries in Europe and North America. Kristen et al. identified a number of longitudinal large-scale studies in education. These usually considered competence assessment across some part of the lifespan. Only a few of them included competence assessment in their design and for those who did hardly any information on vertical scaling and on tests of assumptions of linking was available. For those that did assess competencies, results on the coherence of measurement were ambivalent. Additionally to the studies reviewed in Kristen et al., we collected information on measurement coherence from small-scale studies or multi-cohort studies.

Evidence Supporting Coherence of Measurement

There are some studies that did find evidence for the coherence of measurement. One of them is the National Education Longitudinal Study of 1988 (NELS: 88; Rock, Pollack, Owings, & Hafner 1991), a very prominent longitudinal study on

competencies in the USA. In this study, students were followed in intervals of 2 years from eighth grade to 24–25 years. For the three waves of data collection in school in 8th, 10th, and 12th grade, students' reading, math, social studies, and science competencies were assessed (Rock, Pollack, & Quinn 1995). In order to link the test forms of the competence tests across age, a common-item design was used. Half of the items (in reading) to three-quarters of the items (in math) from one measurement occasion were also used in the following assessment. The authors reported that measurement invariance was found across measurement occasions.

Another longitudinal study for which a coherent measurement was supported is the Early Childhood Longitudinal Study (ECLS; Pollack, Atkins-Burnett, Najarian, & Rock 2005) in the USA. It consists of a birth cohort (ECLS-B), with measurements starting with 9-month-old children which are followed up to first grade, and two kindergarten cohorts (ECLS-K and ECLS-K: 2011), one ranging from fourth to eighth grade and the second following children from kindergarten till fifth grade. In the kindergarten cohorts reading, math, and scientific competencies were assessed and linking was performed using a common-item design. Analyzing differential functioning of the items in the ECLS-K study across time, Pollack and colleagues (2005) found measurement invariance of the common items across measurement occasions. Thus, in this study measurement invariance across the wide span from kindergarten to secondary school could be assured.

Besides these large-scale studies, there is some evidence on the coherence of competence measures across age from other studies. Wang and Jiao (2009), for example, investigated the equivalence of the factorial structure of the Stanford Reading Comprehension Test (Stanford Achievement Test Series, Tenth Edition, 2004) across eight samples in grades 3–10. They found that on subtest level the measurement models were invariant across grades. While Wang and Jiao investigated measurement invariance only on subtest level, in a longitudinal study, Wang, Jiao, and Zahng (2013) investigated measurement invariance of the Measures of Academic Progress (MAP) for mathematic and reading competence on item level. The authors found that measurement invariance could be assured across fifth to seventh grade.

Evidence Questioning Coherence of Measurement

However, there is also evidence that the competence assessed changes across time or cohorts. This is the case in the BiKS-3-10 study on Educational Processes, Competence Development, and Selection Decisions at Preschool and Elementary School Age (von Maurice et al. 2007), a longitudinal study on competence development and educational progress from kindergarten to primary school. Linking between testing waves was done via a common-item design. Robitzsch, Dörfler, Pfost, and Artelt (2011) investigated measurement invariance of the common items in reading competence tests from three measurement occasions in Grade 3 and Grade 4. The authors found considerable item drift across measurement occasions, threatening the interpretation of change scores as indicators of competence development.

Also some cross-sectional large-scale studies, specifically the National Assessment of Educational Progress (NAEP) and a German study evaluating the National Educational Standards (NES), found evidence for measurement non-invariance across age. The NAEP, the largest representative educational assessment in the USA, explores achievement of students in various domains, among others mathematics and reading, every 2 years in Grades 4, 8, and 12 (Jones & Olkin 2004). After the first waves of assessments, measurement invariance of anchor items across grades was checked and threats to measurement invariance were reported on a significant number of mathematics and history items (Haertel 1991; McClellan, Donoghue, Gladkova, & Xu 2005), whereas the reading test functioned rather well across grade levels. Altogether, Haertel questioned the usefulness of cross-age scales for the NAEP regarding the costs in terms of constraints on the framework. He even concluded that comparing students separated by 4–8 years is "largely meaningless" (p. 14). As a consequence in the following assessments cross-age comparisons were discouraged (Thissen 2012). Threats to measurement invariance were also found in the evaluation of the German National Educational Standards (NES; Klieme et al. 2003; Rupp & Vock 2007) by the Institute for Educational Progress. In the domain of language assessment, Böhme and Robitzsch (2009) analyzed reading tests which were administered in pilot and calibration studies facilitating a cross-sectional setting in Grade 3 and 4 of elementary school. For evaluating the item parameter drift, the authors evaluated the variance of differential item functioning (DIF) between the two grades. DIF occurs when items function differently for different groups, that is, when estimated item difficulties differ between subgroups after controlling for overall group differences on the latent trait. Based on the classification scheme of Penfield and Algina (2006), the results indicated a medium DIF variance indicating that some items considerably favored third or fourth graders.

In addition to the above-mentioned large-scale studies, we also reviewed small longitudinal studies. As such in a study on science competence development, Carstensen, Lankes, and Steffensky (2012) found that measurement invariance was not warranted for common science items across three measurement occasions in fifth- to sixth-year-old children. In an U.S. American study, Tong and Kolen (2007) investigated the performance of various vertical linking methods in simulation studies as well as in empirical data. The analyses of the empirical data were based on the assessments of the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie 2003) in the four different domains vocabulary, mathematics, language, and reading covering Grade 3 through Grade 8 via a scaling- and an anchor-test design. Tong and Kolen found that the scaling designs in the empirical studies produced scales with dissimilar properties, especially for tests that tended to be less homogeneous in content across grades and for tests that included testlet-based items such as the reading test.

Summary of Previous Findings on Coherence of Measurement

The results from previous longitudinal or multi-cohort studies show that the assumption of measuring the same latent variable across different age groups is not a trivial one. Indeed, results from some studies such as NELS or ECLS-B confirmed measurement invariance across age, but other studies such as NAEP or BiKS report challenges in creating a common scale. Even in studies with a short age span such as in the NES study or the study by Carstensen et al. (2012), measurement invariance is not always fully warranted. The issue of coherence of measurement is even more prevalent in the NEPS covering such a broad age span.

The National Educational Panel Study: Competence Development Across the Lifespan

The German National Educational Panel Study (NEPS, see Blossfeld et al. 2011) is a current longitudinal study on competence development in Germany. A particular strength of the NEPS is that it considers competence development and educational pathways across the whole lifespan. NEPS incorporates a multi-cohort sequence design (see Fig. 1) that incorporates around 60,000 target persons in six different starting cohorts (newborns, children in kindergarten, students in fifth grade, students in ninth grade, university students, and adults). In order to provide information on educational processes already at an early stage of the study, the six starting cohorts simultaneously started in 2010[1] at different important educational stages and are followed concurrently in their development over time. By regarding different cohorts that overlap at some point in the design, it is possible to investigate educational processes across the whole lifespan without following the same participants across their whole life.[2] Competencies as well as a variety of data on conditions for and consequences of individual educational careers are assessed. Information is gained from the target persons as well as their parents, teachers, or other educators. For many cohorts, different competence domains are repeatedly measured every 2 years, allowing researchers to explore the evolvement of these competencies. Based on the data a wide range of research questions regarding the development of competencies as well as the interaction between competence development and context factors with respect to individual educational careers may be investigated (see, e.g., Blossfeld et al. 2011).

[1] Newborns started 2012 and the adult sample was pursued from the former ALWA study.

[2] This is possible if measurement invariance for the instruments for comparisons between cohorts can be assumed. One may also investigate and account for cohort effects with this design.

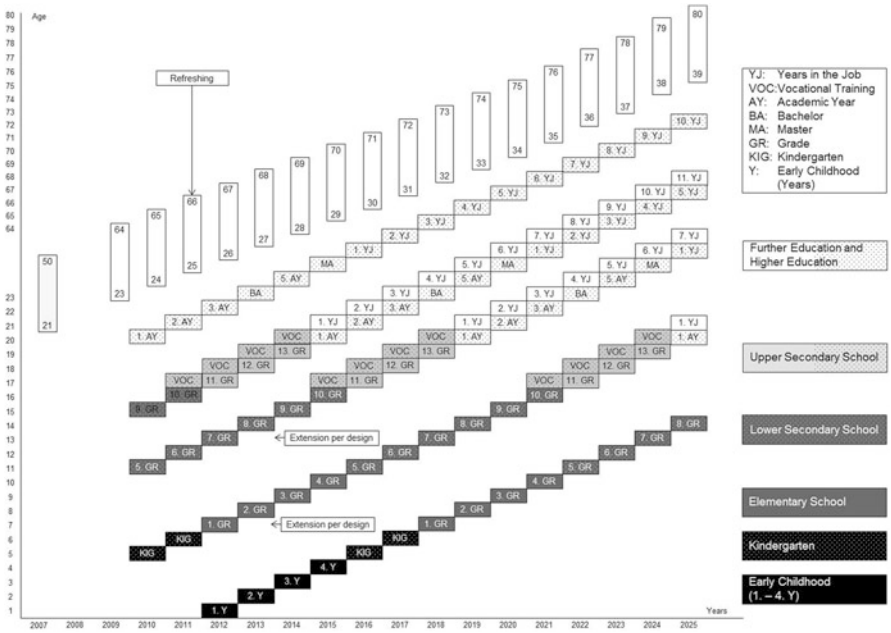


Fig. 1 The multi-cohort sequence design of the NEPS (©LIfBi)

Competence Assessment in the NEPS

The framework for assessing competencies in the NEPS employs a number of different domains (Artelt, Weinert, & Carstensen 2013). These include, among others, reading competence (Gehrer, Zimmermann, Artelt, & Weinert 2013), mathematical competence (Neumann et al. 2013), scientific literacy (Hahn et al. 2013), and information and communication technologies (ICT) literacy (Senkbeil, Ihme, & Wittwer 2013). The NEPS aims at assessing these domain-specific competencies coherently across the lifespan in order to appropriately describe the competencies' developmental progress over time. Therefore, competence models have been specified comprising a consistent structure of each domain across ages and cohorts (Weinert et al. 2011). In order to facilitate a coherent competence assessment, the same conceptual framework has been applied for the tests of different age groups. For reading competence, for instance, the same cognitive processes and text types are used in the tests across different age groups. According to the competence models, new tests are developed and evaluated in pilot studies with the most appropriate items being used in the final main study tests in the NEPS. The newly developed test instruments require the participants to respond to tasks with different response formats. The responses to these tasks are scaled using models of Item Response Theory. In the NEPS, reading, mathematical, scientific, and ICT competence are scaled using the Rasch (Rasch 1960) or the Partial Credit

Model (Masters 1982) (for the scaling model in the NEPS see Pohl & Carstensen 2012). Here, we focus on the measurement of reading competence from Grade 5 to adulthood.

Linking in the NEPS

Linking in the NEPS includes linking of test scores within cohorts over measurement occasions as well as across cohorts. Linking test scores within cohorts is obviously needed to enable the analysis of change over time within each cohort of the NEPS. For example, a question might be how reading competence of students develops between fifth grade (in 2010) and ninth grade (in 2014). However, with the multi-cohort-sequence design of the NEPS, comparisons between cohorts are also intended. As an example, a question might focus on how much ninth graders in 2010 differ in their reading competence from fifth graders, at the same measurement occasion.

In the NEPS different linking strategies are employed. Since retest effects are expected for reading and science items, neither a common-item nor a scaling-test design are applicable as the same items would need to be presented twice to the participants. Instead, common person designs were employed to obtain linking information. In the NEPS, link samples are additionally drawn randomly from the older of the two age groups, that is, students in the link sample typically take the on-grade and below-grade test. In the domain of mathematical competence retest effects are not expected and both common-item designs as well as common-person designs are implemented (Pohl & Carstensen 2013).

Coherence of Measurement in NEPS

Coherence of measurement is a special challenge in the NEPS as the NEPS, in contrast to many other educational studies, follows the development of persons across the whole lifespan. In constructing the test instruments, a great deal of effort was put on a coherent assessment of competencies over the lifespan (see, e.g., Gehrer et al. 2013; Neumann et al. 2013, or Hahn et al. 2013). For (almost) all age groups the same conceptual framework, the same cognitive demands, as well as the same item formats were used for test construction. However, while the assumption of comparable test scores seems to be very plausible for cohorts that are similar in age and in educational or institutional setting, such as linking Grade 5 to Grade 7 students, it is still questionable across very different cohorts, such as Grade 9 students and adults. Adults differ from Grade 9 students not only in age (with a rather large age gap between both samples) but also in institutional settings (Grade 9 students being in school and used to tests and adults mainly being in labor market). This poses a challenge on the comparability of test results across time and cohorts.

Possible threats to the assumptions of linking (Camilli et al. 1993; Hoover 1984; Tong & Kolen 2007) were addressed in the NEPS. In the NEPS fixed item position within a test and rotation of test position within a testlet were used to control for position effects. The possible mismatch of item difficulty to person ability was evaluated in pilot studies within test development and was assured for the main samples. Note that they do, however, not necessarily need to hold for link samples. From the construction point of view, in the NEPS effort is invested to assure coherent measurement of the same latent variable across age groups. Whether this proves successful needs to be tested empirically.

Research Questions

The NEPS is the first study that aims to measure competencies across *the whole lifespan*. So far, there has been no empirical evidence whether and how coherent measurement may be obtained across such a wide age range. The aim of the present study was to investigate whether the construction of coherent instruments for the measurement of competencies across the lifespan is possible and as such was successful in the NEPS. Here, we focused on reading competence and investigated whether it is possible to measure reading competence *coherently* from fifth grade to adulthood. Specifically, we asked whether the assumption holds that the measured reading competence is the same for different age cohorts and measurement occasions. Methodologically phrased, we investigated whether the assumptions for vertical scaling are met, that is: (1) Is competence measurement on reading invariant across studies and age groups? and (2) Is reading competence in NEPS unidimensional across tests for different age groups? Additionally, we explored item and test characteristics related to the coherence of measurement. Only if a competence measurement is coherent and an adequate link between measurements can be established, we may investigate development and change of the competencies (which is one of the main aims in longitudinal studies) as well as compare competencies across different cohorts (on which the multi-cohort sequence design relies).

Method

Sample and Design

Sample

In the present study we analyzed data from four main studies (in Grade 5, Grade 7, Grade 9, and on Adults) and three corresponding link studies of the NEPS. The three link studies are designed to link the measurements of the main studies

between Grade 5 and Grade 7 (G5–G7), between Grade 7 and Grade 9 (G7–G9), and between Grade 9 and Adults (G9–AD). Thus, the studies considered in this paper allow for linking reading competence measures from Grade 5 to adults. The main study in Grade 5, Grade 9, and on Adults took place in the first assessment wave of the NEPS (starting in 2010). The participants in these studies comprised different starting cohorts. The second competence assessment of the fifth graders of 2010 took place in 2012 in Grade 7. As the main studies of Grade 5 and Grade 7 comprised the same starting cohort, most of the students in Grade 5 also participated in the assessment in Grade 7. The link studies were administered parallel to the last of the main studies that are to be linked. Thus, the link study G9–AD took place in the first wave of the NEPS in 2010, while the link studies G5–G7 and G7–G9 were carried out in 2012 (when the main study in G7 took place). The participants in the link studies were always drawn from the older of the two populations, e.g., for linking Grade 9 students to adults, the link study was performed on an adult sample.

The main studies had sample sizes between 5000 (in Grade 5 and Adults) up to about 14,000 (in Grade 9) participants, whereas the link samples were considerably smaller with 500–600 participants (see Table 1). In all main studies, the participants constituted representative samples of German inhabitants at different ages (Aßmann et al. 2011). For the link study G9–AD, adults were representatively drawn from the 16 German federal states, while the link studies G5–G7 and G7–G9 were conducted in only four federal states: Lower Saxony, Bremen, North Rhine-Westphalia, and Saxony. Although no representative sample of the whole country could be drawn for two of the link studies, representative samples were drawn from the four federal states and we did not expect large differences in populations. However, it is to note that participants in the main studies agreed to take part in a longitudinal study, while participants in the link study were only recruited for one assessment. This may result in different participation processes and, thus, in different populations.

Looking at demographic characteristics (Table 1), the link studies and the respective main studies seem to be rather similar. Comparing the main study in G7 with the link study G5–G7, a relatively equal distribution of male and female students and similar percentages of school type and migration background were found when missing values were not taken into account. The average age in the link study G5–G7 was almost identical to that in the main study G7. Based on the design, students in the main study G5 were about 2 years younger than in the corresponding link study G5–G7. They were, however, similar in many of the other demographic characteristics.

The link study G7–G9 and the corresponding main studies in G7 and G9 featured similar properties regarding gender and migration background. However, the link study G7–G9 and the main study G9 slightly differed in age, with the participants in the main study being on average about half a year older. Participants in the different studies also differed in school type. There were more students in the highest academic track in the main study in G7 than in the link study; the lowest number of students in the highest academic track was found in the main study in G9. Thus, the link study G7–G9 and the respective main study in G9 may have been drawn from different populations.

Table 1 Description of the samples in main and link studies

	Main study G5	Link study G5–G7	Main study G7	Link study G7–G9	Main study G9	Link study G9–AD	Main study Adults
N	5193	608	6186	534	13,897	502	5335
Gender (rel. freq.)							
Male	51.6 %	48.6 %	51.7 %	51.1 %	50.2 %	43.5 %	49.9 %
Female	48.4 %	51.4 %	48.3 %	48.9 %	49.8 %	56.5 %	50.1 %
Age							
Mean	10.9	12.9	13.0	15.3	15.7	45.2	47.6
(SD)	(0.5)	(0.6)	(0.5)	(0.7)	(0.6)	(12.7)	(10.9)
Migration background (rel. freq.)							
No	68.0 %	69.7 %	66.6 %	71.5 %	70.5 %	83.7 %	80.3 %
Yes	25.1 %	26.0 %	22.0 %	23.2 %	25.0 %	15.5 %	14.6 %
No information	6.9 %	4.3 %	11.3 %	5.2 %	4.5 %	0.8 %	5.2 %
School type/degree (rel. freq.)							
Lower school type	54.3 %	56.0 %	53.0 %	59.1 %	65.0 %	66.0 %	54.6 %
High school type	45.4 %	44.0 %	47.0 %	40.9 %	35.0 %	34.0 %	45.4 %

Migration background either the person itself or one of its parents is born in a foreign country; *School type/degree* refers to the school type in the school cohort samples and to the school degree in the Adults' samples; high school type: at least grammar school/A-level degree, lower school type: other school types/a lower school degree

Adults in the main study and the corresponding link study G9–AD had a similar age distribution and a similar percentage of persons with migration background. Slight differences occurred on the variables gender and school degree. These differences possibly reflect differences in participation between the two studies. The Grade 9 students in the main study and adults in the link sample G9–AD were by design drawn from different populations and they differed in some of the background variables. Note that in the school cohorts the dichotomous variable school type/degree refers to the school type participants attend at the moment. For the school cohorts the variable differentiates between students attending grammar school (German: Gymnasium) and students with a lower school type. Since (most of the) participants in the Adult sample did not attend school any more, the respective variable refers to the highest school degree achieved so far, distinguishing between an A-level degree (German: Abitur) and a lower school degree. In the current study, the Grade 9 sample and the link study sample differed in the distribution of school

type/degree, and additionally in the variables gender, and migration background. Moreover, as expected by design, the link study sample was substantially older than students in Grade 9.

In summary, while the link study G5–G7 shows similar demographic properties as the corresponding main studies G5 and G7, there are some differences in the samples between the link study G7–G9 and the corresponding main studies as well as the link study G9–AD and its corresponding main studies.

Design

A common-person link design was used to link the reading competence scores of different age groups. We describe the design exemplary for linking the Grade 9 test to the adult reading test. The link sample was always drawn from the older population of the two main studies to be linked. Thus, the link study G9–AD was conducted on adults. In the main studies, one test constructed for this age group was administered, while the link sample completed the tests of the two adjacent years. Regarding the link between Grade 9 and Adults, the ninth graders and the adults in the main studies received only the Grade 9 test or the Adults test, respectively. In the link study, both tests were administered to the participants. The two tests in the link studies were given in randomized order to balance position effects. The same link design was applied for linking competence scores of Grade 5 students to Grade 7 students and of Grade 7 students to Grade 9 students.

Whereas in the first testing wave (here main studies in Grade 5, Grade 9, and Adults), reading competence was measured using a single test form for all students, in later waves (here Grade 7) longitudinal multi-stage testing using information from the previous testing wave for routing to test forms of different difficulty was applied in order to enhance test targeting, motivation, and measurement precision (see Pohl 2014). Thus, the test in Grade 7 consisted of two test forms that differ in mean difficulty. 61.9 % of the students in Grade 7 took part in the previous competence testing wave in Grade 5, so competence scores from the previous wave were available for these students. Additionally, 2357 (38.1 %) new students were recruited in Grade 7 to enlarge the sample size. Students with an ability estimate in Grade 5 below the median were assigned to an easy test form in Grade 7 ($N = 1771$), students with an ability estimate equal or greater than the median were assigned to the difficult test form ($N = 2058$) (see Fig. 2). Students with no available competence score from Grade 5 ($N = 2357$) were assigned to the difficult test form, since pilot studies had shown that the difficult test form targets a wider ability range than the easy test form. Altogether, 1771 students in the main study in Grade 7 took the easy test form, and 4415 subjects took the difficult test form. The assignment to the test forms was different in the corresponding link studies (G5–G7 and G7–G9). As these were cross-sectional samples, no preliminary information about the student's competencies was available and the two test forms of the G7 reading competence test were administered randomly to the participants of the

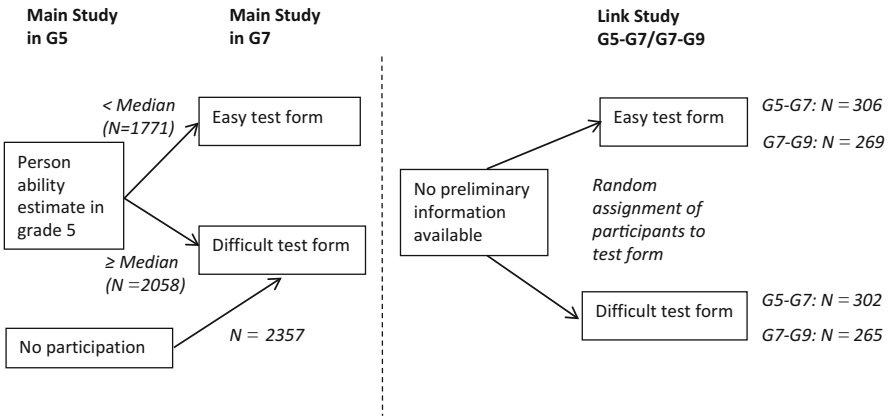


Fig. 2 Allocation of the Grade 7 test forms to the examinees in the main study and the link studies

link studies. Note that the different assignment of test forms results in different population characteristics between the main study and the link studies, conditional on the test form.

Measures and Procedures

In the NEPS, reading competence tests were developed that aim at measuring reading competence coherently across the lifespan (Weinert et al. 2011)—using the same conceptual framework across age (Gehrer et al. 2013). The NEPS framework on reading competence embodies different text functions and different cognitive requirements. Tests across all ages consist of five texts each with a different text function: (1) information texts, (2) commenting or argumenting texts, (3) literary texts, (4) instruction texts, and (5) advertising texts. The specific questions focusing on the texts' content can be classified into three types of items according to their cognitive requirement: (a) finding information in texts, (b) drawing text-related conclusions, and (c) reflecting and assessing. The three types of items are not intended to primarily differ in difficulty, but qualitatively (Gehrer et al. 2013). Most of the items are multiple-choice (MC) items with one option out of four being correct. Furthermore, complex multiple choice (CMC) items and matching (MA) items are included in the tests. CMC items consist of several subtasks with two response options, MA items include a number of responses which have to be matched to a given set of statements. Subtasks of CMC and MA items were aggregated to one polytomous variable per item and given (partial) credit scores (see Haberkorn, Pohl, Carstensen, & Wiegand 2015; Pohl & Carstensen 2012).

As described above, in Grade 7 two test forms were administered which differed in difficulty, and students were assigned to either the difficult or the easy test form.

Each test form comprised five texts, and the two test forms had three out of five texts (plus the respective items) in common which enabled a link between the test forms. The three common texts were presented on the same positions in both test forms.

In the main and link studies the reading competence test was administered with other competence tests and questionnaire items assessing further information of the examinees. The reading test featured a paper-and-pencil format and participants had 30 min to complete the test. While the test was presented to the students in Grade 5, 7, and 9 in a group setting at school with a group size of up to 25 subjects, the adults took the test individually at their homes.

Analyses

We scaled the data within the framework of Item Response Theory (IRT). As described above, the reading test included simple MC, complex MC, and matching items. The complex MC and the matching items consisted of a set of subtasks that were aggregated to a polytomous variable in the final scaling model in the NEPS. In accordance with the scaling procedure for competence data in the NEPS (Pohl & Carstensen 2012 2013), we used the Partial Credit model (Masters 1982) for scaling the data. The models were fitted to the data using ConQuest (Wu, Adams, Wilson, & Haldane 2007). Missing responses were ignored in the estimation of the parameters (see Pohl, Gräfe, & Rose 2014).

We evaluated both assumptions of measurement invariance. First, we investigated the dimensionality of the tests. For this, we used the link samples, which took the reading tests of two adjacent age groups, and specified (a) a two-dimensional model—each test form of a specific age group forming one dimension and (b) a unidimensional model across both tests. For the Grade 7 test, both test forms were included in the analyses and the information of test form assignment was included in the model. The dimensionality of the test was assessed by comparison of the AIC and BIC of the two models and by evaluating the latent correlation between the dimensions of the two test forms (estimated in the two-dimensional model). If the model comparison supports a unidimensional model and the correlation between the test forms is close to one, the assumption that the tests of adjacent age groups measuring the same construct within one population is supported.

Second, we investigated whether the tests measure the same construct in the different studies. For this purpose, we applied a multi-facet Rasch model and evaluated differential item functioning (DIF) by comparing estimated item difficulties between main study and link study. Note that for the tests, where main study and link study are drawn from the same population, the test of DIF is mainly a test for equivalence of the samples drawn. DIF of items that were administered to different populations in the main study and the link study is mainly a test of measurement coherence across age groups and settings. For the Grade 7 test, DIF was investigated separately for the easy and the difficult test form. Although the participants in the main study G7 and the link study attend the same grade, differences in populations are present,

as the assignment to the different test forms differed between main study and link study. The populations may especially differ in person abilities and as a consequence possibly also in test taking strategies.

In subsequent analyses we investigated whether there is a relationship between DIF and test as well as item characteristics as possible explanations for measurement variance. We specifically considered the competence domain assessed, item difficulty, text functions, cognitive requirements, and response format.

Results

Dimensionality

Using the link studies we investigated whether the reading tests of adjacent years do measure a unidimensional construct. The results showed that in all three studies the fit indices supported a two-dimensional over a unidimensional model (see Table 2). It is, however, to note that the differences in AIC and BIC were rather small compared to sample size and test length, so that statistical inferences will not be without ambiguity (Alexandrowicz 2008). The latent correlations between the test forms of two adjacent age groups were very high (see Table 2), indicating that within the same sample, the different tests measure the same construct.

Measurement Invariance

In the following the results on measurement invariance are presented by reporting the DIF between main study and link study for each of the three links. Afterwards, the relationships of item and test characteristics with DIF are described.

Table 2 Fit indices of the uni- and the multidimensional models in the link studies

Link study	Model	AIC	BIC	Latent correlation
G5–G7	Unidimensional	32,782.13	33,267.25	
	Two-dimensional	32,756.85	33,250.79	0.93
G7–G9	Unidimensional	27,013.45	27,462.89	
	Two-dimensional	26,993.14	27,451.14	0.95
G9–AD	Unidimensional	25,257.80	25,586.85	
	Two-dimensional	25,241.78	25,579.26	0.95

Linking Grade 5 to Grade 7

The absolute differences in the estimated item parameters of the Grade 5 test in the main study of Grade 5 and the link study G5–G7 are presented at the top of Fig. 3. Note that the test was administered to Grade 5 students in the main study and to Grade 7 students in the link study and, thus, allows one to describe differences in item functioning across age groups. As can be seen in the Figure, the differences in item difficulties between the two studies were negligible, ranging from −0.794 to 0.504 logits. For only one item DIF exceeded 0.6 logits. Overall, the measurement model of the Grade 5 test in the main study on Grade 5 students seems to be similar to that in the link study on Grade 7 students. In Fig. 3 also DIF for the items of the easy and the difficult Grade 7 test is shown. Although the main study and link study were both sampled from the population of Grade 7 students, the assignment to test forms differed between main study and link study. In the main study the assignment was based on ability estimates from the previous testing waves, resulting in subgroups with rather homogenous ability scores and a good test targeting.[3] In contrast, random assignment was performed in the link study, resulting in heterogeneous subgroups and a test targeting that was less tailored to the ability level of the subgroups. As in the main study of Grade 7 the students newly recruited in Grade 7 all received the difficult test (regardless of their ability), the competence distribution for the students receiving the difficult test should be more similar between main study and link study than for the easy test form. This is also reflected in the results of measurement invariance of the test forms across samples (Fig. 3). DIF was smaller for the difficult test form than for the easy test form. DIF values ranged from −0.606 to 0.480 logits in the difficult test form and from −0.664 to 0.750 logits in the easy test form. Only one item in the difficult test form and four items in the easy test form showed DIF greater than 0.6.

Linking Grade 7 to Grade 9

The results on measurement invariance linking the tests in Grade 7 to the test in Grade 9 are presented in Fig. 4. There was no considerable DIF for the items of the Grade 9 test. For all items absolute differences in estimated item difficulty were less than 0.5. As for the Grade 9 test, the samples of the main study and the link study were both drawn from the population of ninth graders, these results support the comparability of the samples.

Considering the link across different age groups, there was noticeable DIF for both test forms in Grade 7 across samples. DIF values ranged from −0.846 to 0.792 logits in the easy test form and from −0.746 to 0.914 in the difficult test form. There was also a non-negligible amount of items with rather large DIF, especially in the

[3]Test targeting is good, when the item difficulties of the test items well fit to the ability levels of the specific target group. A good test targeting enhances reliability of the ability measurement.

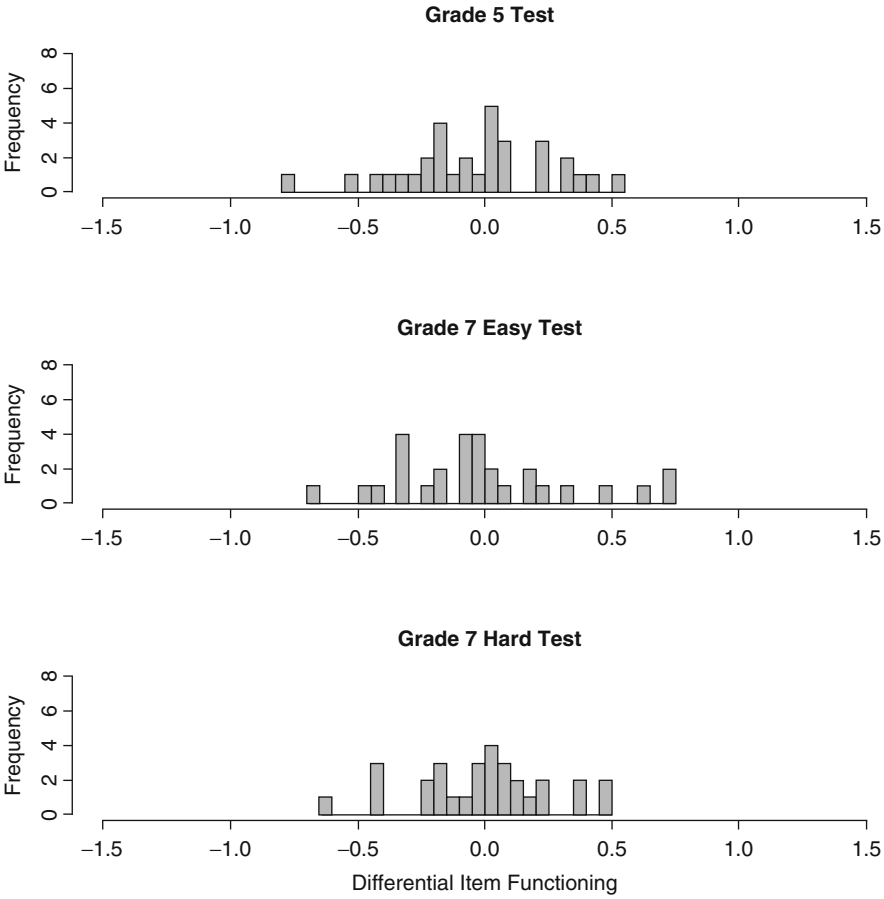


Fig. 3 DIF of items linking Grade 5 to Grade 7

easy test form. For seven items in the easy and five items in the difficult test form DIF exceeded 0.6. The results indicate that the two test forms function differently in the different populations.

Linking Grade 9 to Adults

Figure 5 shows the differences in estimated item difficulties for linking the Grade 9 test to the adult test. For the adult test, estimated item difficulties were very similar across the main study and the link study, indicating similarity of both samples. No DIF value exceeded an absolute value of 0.4 logits (range from −0.300 to 0.392 logits). This was different for the Grade 9 test, where main sample and link sample were drawn from different populations. DIF values were large, ranging from −1.298

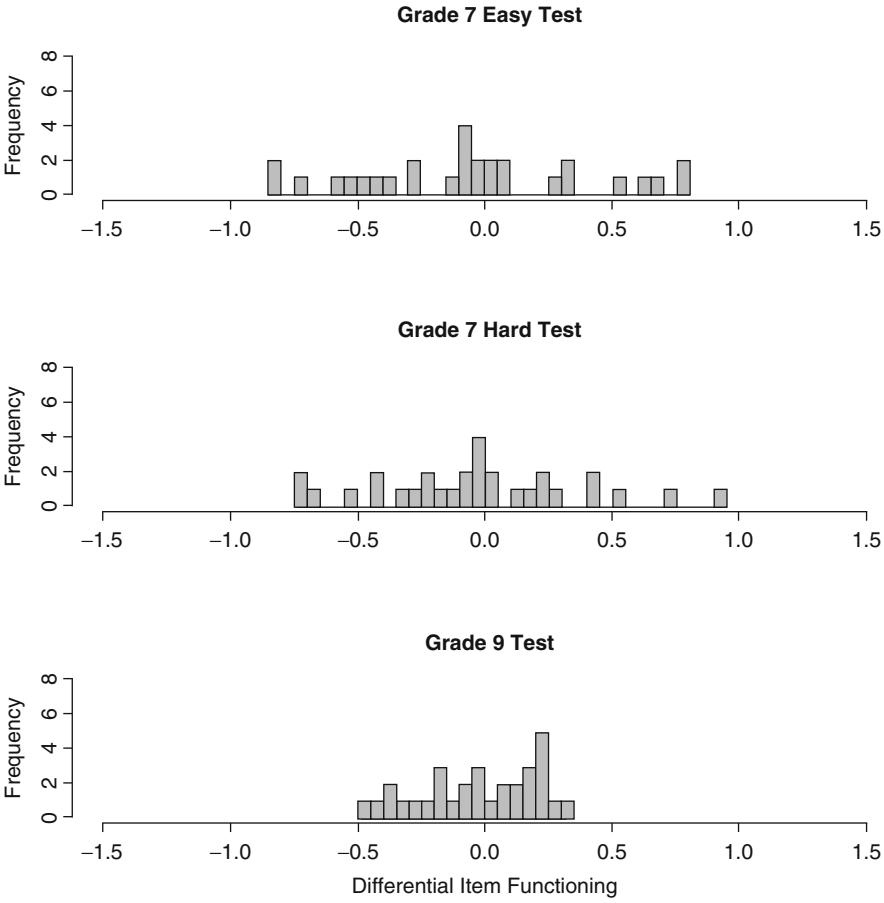


Fig. 4 DIF of items linking Grade 7 to Grade 9

to 1.394 logits. Nine items exhibited absolute DIF values greater than 0.6 logits with four of them even exceeding differences of 1 logit. Thus, as a considerable number of items showed large DIF indicating great differences in the measurement of ninth graders (in the main study) and adults (in the link study). The same test seems to assess a different construct in the different populations. Note that here the main study and link study differ not only by a large age difference (ninth graders aged 16 to adults of age 21–78), but also in educational and occupational setting (school vs. mainly work), test setting (group testing in Grade 9 vs. individual testing at home for Adults), and most probably also in competence level. These differences between the populations seem to challenge the coherence of measurement.

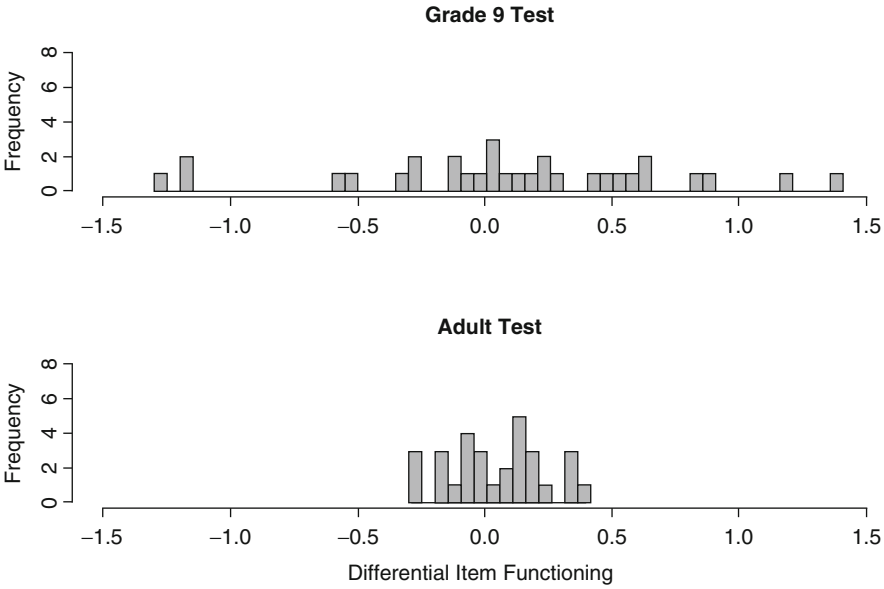


Fig. 5 DIF of items linking Grade 9 to Adults

Subsequent Analyses

In subsequent analyses, we investigated the impact of item and test characteristics on the amount of DIF. Regarding item characteristics, we investigated whether DIF is related to item difficulty, text functions, cognitive requirements, and response format. We found no considerable relationship between DIF and text functions, cognitive requirements, or response format. Concerning text functions, the mean absolute DIF across all items and studies ranged from 0.24 (SD across studies [SDacross] being 0.08 and average SD within studies [SDwithin] being 0.18) for literary texts to 0.32 (SDacross = 0.18, SDwithin = 0.23) for commenting texts. Regarding cognitive requirements, mean absolute DIF values across all items and studies were 0.33 (SDacross = 0.13, SDwithin = 0.26) for finding information in the text, 0.23 (SDacross = 0.10, SDwithin = 0.19) for drawing text-related conclusions, and 0.28 (SDacross = 0.11, SDwithin = 0.20) for reflecting and assessing. Complex MC items were affected with slightly lower absolute DIF values (M = 0.23, SDacross = 0.08, SDwithin = 0.20), than MA items (M = 0.28, SDacross = 0.13, SDwithin = 0.19) and simple MC items (M = 0.29, SDacross = 0.13, SDwithin = 0.22). The impact of the different text functions, cognitive requirements, and response formats was moderate and very similar for the different studies.

There was a strong relationship of DIF with item difficulty. Table 3 shows the correlation of item difficulty with both, absolute DIF value and DIF value. Note that DIF was calculated as the differences in estimated item difficulty in the link study

Table 3 Correlation of the value (DIF) and the absolute value (DIFabs) of differential item functioning and item difficulty (β) across studies and tests

Link	Test	$cor(\beta,DIFabs)$	$cor(\beta,DIF)$
G5–G7	G5	0.27	−0.26
	G7 easy	0.31	0.23
	G7 difficult	−0.48	−0.21
G7–G9	G7 easy	−0.33	0.30
	G7 difficult	0.25	0.05
	G9	−0.27	−0.25
G9–AD	G9	−0.03	−0.60
	AD	−0.00	−0.26

minus the estimated item difficulty in the main study. Thus, positive DIF values indicate that an item is easier in the main study than in the link study and negative values that the item is easier in the link study than in the main study. The correlation of item difficulty with absolute DIF indicates for which items DIF occurs; when positive, DIF tends to occur in difficult items; when negative, DIF tends to occur in easy items; and when zero, it tends to occur in both easy and difficult items. The sign of the correlation of item difficulty with DIF indicates in which direction DIF occurs. Positive values indicate that easy items are easier in the link study than in the main study and difficult items are more difficult in the link study than in the main study. The results for the various studies (see Table 3) suggest a very heterogeneous picture with different correlation patterns for different studies. We had no theory on that and investigated the patterns exploratorily as possible explanations for DIF. We focus on the studies with large DIF, that is, the G7 easy and hard tests in the G7–G9 link and the G9 test in the G9–AD link. For the easy G7 test form in the link G7–G9, DIF mainly occurred for easy items ($cor(\beta, DIFabs) = -0.33$) with items being more difficult in the main study (on seventh graders) than in the link study (on ninth graders) ($cor(\beta, DIF) = 0.30$). For difficult items, DIF hardly occurred ($cor(\beta, DIFabs) = -0.33$). This is different for the respective difficult test form of the same study. For the G7 difficult test form, DIF mainly occurred for difficult items ($cor(\beta, DIFabs) = 0.25$). There was no relationship of item difficulty and the direction of DIF ($cor(\beta, DIF) = 0.05$). Another pattern was found for the G9 test linking the G9 test to the adult test. Here DIF occurred for easy and for difficult items ($cor(\beta, DIFabs) = -0.03$) with the easy items being more difficult in the link study (with an adult sample) and the difficult items being more difficult in the main study (with a sample of ninth graders) ($cor(\beta,DIF) = -0.60$). As the size and direction of DIF for different item difficulties varied a lot across studies, it is difficult to find an explanation. The results on measurement invariance for the Grade 7 easy test linking G7–G9 seem to be affected by test targeting, with DIF occurring mainly for items with a low targeting (i.e., that are either too difficult or too easy for the respective sample). In fact, the Grade 7 easy test that was completed by ninth graders in the link study, yielded considerable ceiling effects for some items. About eight out of 29 items had a probability to be solved above 95 %. When these items were excluded from the DIF analyses, the relative amount of DIF could be reduced. This is not

necessarily true in the other studies, as in most studies (e.g., studies in Grade 9 and on Adults) item difficulty is rather low, but DIF occurs on items of all difficulties. Thus, although there does not seem to be a clear pattern, there are indications that measurement variance is related to item difficulty and test targeting.

On test level, we evaluated whether similar results of measurement invariance can be found for assessing other competencies. This facilitates the drawing of conclusions concerning the extent to which the results depend on the specific test or are rather population specific. The main and link studies linking the Grade 9 test to the adult test were also used to link mathematical competence. For mathematical competence, we found similar results on dimensionality and measurement invariance as for reading competence. There was hardly any DIF on the adult test, which was administered to the same population in the main study and the link study; there was, however, large DIF for items of the Grade 9 test (also see Pohl & Carstensen 2013). Similar coherence across competence domains was found in the school cohorts, such as for linking ICT literacy from Grade 6 to Grade 9. In the Grade 9 ICT test that was administered to ninth graders in the main and the link study almost no DIF occurred (analogous to the results of reading competence linking Grade 7 to Grade 9). In contrast, some DIF was present in the Grade 6 test that was administered to sixth graders in the main study and to ninth graders in the link study with four out of 30 items exceeding DIF values of 0.6 logits (there was also DIF present in the respective analyses comparing measurement models between Grade 7 and Grade 9 on reading). In summary, different competencies that were assessed in the NEPS such as reading, mathematical competence, or ICT literacy showed similar patterns of measurement (in)variance across specific age spans.

Discussion

In the present study, we investigated whether it is possible to coherently measure reading competence across the lifespan within the NEPS. We specifically asked whether the reading tests for different cohorts measure the same construct and whether each reading test measures the same construct across different samples. The results on dimensionality showed that within the same population tests for different age groups did measure the same construct. Thus, the tests were well constructed to assess the same construct coherently across the test forms. However, when the same test was administered to samples drawn from very different populations, the measurement models differed between samples, that is, measurement invariance did not fully hold. The more different the populations were, the larger DIF was found. Differences in populations are indicated by differences in age (e.g., linking Grade 5 and Grade 7), differences in educational and occupations settings (e.g., students in school and adults at work), differences in test settings (e.g., group testing in school, individual testing at home), and differences in competence levels (e.g., differences in the assignment to test forms in Grade 7). Only for linking Grade 5 to Grade 7,

which were similar in educational setting and test setting, an adequate amount of measurement invariance could be assured. On test level, item difficulty and test targeting seem to play a role for results on measurement invariance.

The differences in item functioning for different populations may to some extent occur due to differences in test-taking behavior. This can, for example, be evaluated by missing values. While samples from similar populations in our study showed rather similar missing item patterns, samples from different populations differed in their missing item patterns. Adults in the main study and in the link study, for example, showed a very similar missing pattern on the amount of omitted and not reached items as well as non-valid responses. Students in Grade 9 and adults differed immensely in their missingness patterns. The adult sample reached fewer items, omitted more items, and produced more invalid responses than Grade 9 students. We also found greater correlations between the number of omissions and item difficulty for the student populations (correlations ranging from 0.23 to 0.55) than for adults ($cor = 0.12$). The greater age and competence level, the higher these correlations were in school. This may indicate that students in school, especially the older they are and the more competence they gained, apply a different test-taking strategy than adults. This is also corroborated by the finding that the number of omissions of adults is greater with lower competence levels (correlation of reading competence and number of omissions being −.26), while it is hardly related in the students' samples (correlations ranging from −.07 for Grade 9 students to −.12 for Grade 5 students). Especially older and more competent students seem to use some test-taking strategy omitting difficult items. This fits well in the research on testwiseness (e.g., Diamond & Evans 1972; Gibb 1964; Millman, Bishop, & Ebel 1965) and test motivation (e.g., Wise & DeMars 2005 2006), which also reports on omission of items and quitting on the test (e.g., Schmitt, Chan, Sacco, McFarland, & Jennings 1999; Zerpa, Hachey, van Barnfield, & Simon 2011). Investigating differences in test-taking behavior may help explain the results on differences in measurement invariance across different populations in further research.

There are some implications for large-scale studies that can be drawn from our results. As our results show, although within the same sample, adjacent test forms may assess a unidimensional construct, the measurement model may differ for different populations. This is especially the case when differences between populations increase. Thus, for planning a longitudinal study that requires linking of test forms, the differences in the populations to be linked should be kept to a minimum. This means that linking should be performed across smaller age ranges. In NEPS, linking between Grade 9 and Adults did not prove successful, but possibly linking Grade 9 students to Grade 12 students and students in the school cohort to younger adults might facilitate appropriate linking. Similarities in linked samples also include similarities in test settings. Mode effect studies may help assessing the effect of individual vs. group testing and, thus, accounting for it. This is done in NEPS in other age cohorts (Kröhne & Martens 2011). As it might be that DIF between different populations occurs due to differences in test-taking strategies, a more thorough instruction on how to take the test may help prevent from

measurement variation. This issue can be approached in an even more sophisticated way, by computerized testing, where more control over item skipping and response time is possible.

In our study we focused on the prerequisites for linking. These results are very relevant for the NEPS, since they are the basis for choosing the actual linking models within the age cohorts and, if possible, across age cohorts as well. One of the outcomes of the NEPS will be an empirical answer to the question, whether and for which domains it is possible to construct a common scale across the lifespan. As far as the results presented here indicate, it will be feasible to construct common scales within some age limits.

In order to establish a common scale, one has to make assumptions about item drift. If one assumes that observed item drift is not due to any systematic reason like a shift in constructs, a link may be based on items that did not show DIF, assuming partial measurement invariance. One has to rely on the assumption that the items chosen for linking are not confounded by item drift. This, however, cannot be empirically tested. This assumption may be more plausible for the tests in the school cohorts, that is, linking Grade 5 to Grade 7 and Grade 7 to Grade 9. As the DIF on the Grade 9 test linking G9–AD is very large and the populations differ a lot, it may be less plausible here. Further link studies, e.g., linking G9–G12, G12 to university students or tertiary students to younger adults may and will give more evidence to investigate whether linking across age cohorts will be possible.

After having evaluated the plausibility of different linking assumptions, the question is how to link different test forms. From research we know that different decisions in the scaling process typically lead to somewhat different vertical scales (Camilli et al. 1993; Loyd & Hoover 1980; Williams, Pommerich, & Thissen 1998; Yen 1986). No consensus exists in the literature as to which set of procedures produces the vertical scale that most adequately captures the nature of development (Kolen & Brennan 2004). It rather seems that the optimal linking model depends on the degree of violation of the assumptions made in a linking model given its particular design and sample sizes. In any way, within the NEPS different linking analyses, preferably linking with restrictions on item difficulty on an item level and as an alternative, linking with restrictions on item difficulty on the test level, will be explored to quantify the impact on the linking results. One of the crucial questions will be to decide which items are considered "undrifted" and thus will contribute to the link and which items are considered to show item drift and will be excluded from establishing the link. Consequently, a thorough evaluation of the linking model applied to a particular study is needed. In order to quantify the degree of linkability, linking errors will be computed and compared. A possible solution for linking approaches for the NEPS might be to distinguish *strict linking* from linking of tests that might be considered as *connected* in a less stringent way. A strict link requires most items to be invariant over time resulting in small linking errors only, whereas connected tests may allow item drifts to occur more frequently and the link error might thus be larger. From a substantive NEPS point of view, to have connected test forms across different age cohorts may have the potential for relevant cohort comparisons in the NEPS, whereas following the competence development

of students longitudinally over subsequent years will require a strict link assuming measurement invariance and small linking errors. The investigation of which linking models are appropriate in NEPS falls in the scope of further research.

Acknowledgement This research used data from the National Educational Panel Study (NEPS). From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

This research is based on the dedicated work of professors and research assistants within the NEPS. We especially thank Karin Gehrer, Stefan Zimmermann, Cordula Artelt, and Sabine Weinert for developing the tests on reading competence, that are the basis of our research, and Maike Krannich, Michael Wenzler, Theresa Rohm, and Odin Jost for their valuable assistance in analyzing the data. Our thanks also go to the staff of the NEPS administration of surveys and to the methods group.

References

Alexandrowicz, R. (2008). Wieviel ist “ein bisserl”? Ein neuer zugang zum BIC im rahmen von Latent-Class-Analysen [How much is “a bit”? A new approach to the BIC within the framework of Latent Class Analyses]. In J. Reinecke & C. Tarnai (Eds.), *Klassifikationsanalysen in theorie und anwendung* (pp. 141–165). Münster: Waxmann.

Artelt, C., Weinert, S., & Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. *Journal for Educational Research Online, 5*, 5–14.

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., et al. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. *Zeitschrift für Erziehungswissenschaft, 14*, 51–65.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft, 14*, 5–17.

Böhme, K., & Robitzsch, A. (2009). Methodische aspekte der erfassung der lesekompetenz [Methodological aspects of reading assessment]. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss, & G. Walther (Eds.), *Bildungsstandards Deutsch und mathematik. Leistungsmessung in der grundschule* (pp. 250–289). Weinheim: Beltz.

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*, 379–388.

Carstensen, C. H., Lankes, E. M., & Steffensky, M. (2012). Modellierung von längsschnittlichen daten am beispiel einer quasi-experimentellen studie zur erfassung von naturwissenschaftlichen kompetenzen im kindergartenalter [Modeling of longitudinal data illustrated on a quasi-experimental study of the assessment of scientific competencies in preschool children]. In W. Kempf & R. Langeheine (Eds.), *Item-response-modelle in der sozialwissenschaftlichen forschung* (pp. 109–126). Berlin: Regener.

Diamond, J. J., & Evans, W. J. (1972). An investigation of the cognitive correlates of testwiseness. *Journal of Educational Measurement, 9*, 145–150.

Doran, H. C., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. W. Lissitz (Ed.), *Value-added models in education: Theory and applications* (pp. 80–104). Maple Grove, MN: JAM Press.

Dorans, N. J., Pommerich, M., & Holland, P. (Eds.). (2007). *Linking and aligning scores and scales*. New York, NY: Springer.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online, 5*, 50–79.

Gibb, B. G. (1964). *Testwiseness as secondary cue response* (Doctoral dissertation). Stanford University, Ann Arbor, Michigan: University Microfilms, 1964. No. 64-7643.

Haberkorn, K., Pohl, S., Carstensen, C., & Wiegand, E. (in press). Scoring of complex multiple choice items in NEPS competence tests. In H. -P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), Methodological issues in longitudinal surveys. Springer.

Haertel, E. (1991). *Report on TRP analyses of issues concerning within-age versus across-age scales for the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., et al. (2013). Assessing science literacy over the lifespan – A description of the NEPS science framework and the test development. *Journal for Educational Research Online, 5*, 110–138.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice, 3*, 8–14.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa Tests: Guide to research and development*. Chicago, IL: Riverside Publishing.

Jones, L. V., & Olkin, I. (Eds.). (2004). *The Nation's Report Card: Evolution and perspectives*. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (Eds.). (2003). *The development of National Educational Standards. An expertise* (Vol. 1). Berlin: BMBF.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Kristen, C., Römmer, A., Müller, W., & Kalter, F. (2005). *Longitudinal studies for education reports – European and North American examples,* Report commissioned by the Federal Ministry of Education and Research. Bonn, Berlin: Federal Ministry of Education and Research (BMBF).

Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14*, 169–186.

Linn, R. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

Martineau, J. (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Psychological Statistics, 31*, 35–62.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

McClellan, C. A., Donoghue, J. R., Gladkova, L., & Xu, X. (2005). *Cross-grade scales in NAEP: Research and real-life experience.* Presentation at the conference Longitudinal Modeling of Student Achievement, Maryland Assessment Research Center for Education Success, University of Maryland, College Park, MD.

Millman, J., Bishop, D. H., & Ebel, R. (1965). An analysis of test wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*, 323–335.

Monseur, C., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (pp. 113–122). Hamburg: IEA-ETS Research Institute.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online, 5*, 80–109.

OECD. (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving, and financial literacy*. Paris: OECD Publishing.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295–312.

Pohl, S. (2014). Longitudinal multi-stage testing. *Journal of Educational Measurement, 50*, 447–468.

Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report: Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study—Many questions, some answers, and further challenges. *Journal of Educational Research Online, 5*, 189–216.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests-Evaluating different approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74*, 423–452.

Pollack, J. M., Atkins-Burnett, S., Najarian, M., & Rock, D. A. (2005). *Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS–K), Psychometric report for the fifth grade*. Washington, DC: National Center for Education Statistics. U.S. Department of Education.

Popham, W. J. (2000). *Educational measurement*. Boston, MA: Allyn and Bacon.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reckase, M. D., & Martineau, J. A. (2004). *Growth as a multidimensional process.* Paper presented at the Annual Meeting of the Society for Multivariate Experimental Psychology, Naples, FL.

Robitzsch, A., Dörfler, T., Pfost, M., & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe [Relevance of item selection and model selection for assessing the development of competencies: The development of reading competence in primary school students]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43*, 213–227.

Rock, D. A., Pollack, J. M., Owings, J., & Hafner, A. (1991). *Psychometric report for the NELS: 88 base year test battery*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report of the NELS: 88 base year through second follow-up*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Rupp, A. A., & Vock, M. (2007). National educational standards in Germany: Methodological challenges for developing and calibrating standards-based tests. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Making it comparable: Standards in science education* (pp. 173–198). Münster: Waxmann.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41–54.

Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal for Educational Research Online, 5*, 139–161.

Thissen, D. (2012). *Validity issues involved in cross-grade statements about NAEP results*. Washington, DC: American Institutes for Research, NAEP Validity Studies Panel.

Tong, Y., & Kolen, M. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*, 227–253.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2008). Linking competencies in horizontal, vertical and longitudinal settings and measuring growth. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 121–149). New York, NY: Hogrefe & Huber.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.

von Maurice, J., Artelt, C., Blossfeld, H. -P., Faust, G., Rossbach, H. -G., & Weinert, S. (2007). Bildungsprozesse, kompetenzentwicklung und formation von selektionsentscheidungen im vor- und grundschulalter: Überblick über die erhebungen in den längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden projektjahren [Educational processes, competence development and formation of selection decisions in preschool and primary school age: An overview of the first two years of data collection in the longitudinal studies BiKS-3-8 and BiKS-8-12]. Bamberg: Otto-Friedrich-Universität.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*, 760–777.

Wang, S., Jiao, H., & Zahng, L. (2013). Validation of longitudinal achievement constructs of vertically scaled computerized adaptive tests: A multiple-indicator, latent-growth modeling approach. *International Journal of Quantitative Research in Education, 1*, 383–407.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, 14*, 67–86.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*, 93–107.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort moderated model. *Journal of Educational Measurement, 43*, 19–38.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*, 15–27.

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *Conquest 2.0 [Computer Software]*. Camberwell, VIC: ACER Press.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299–325.

Zerpa, C., Hachey, K., van Barnfield, C., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. *Sage Open, 1*, 1–9.

Mixed Rasch Models for Analyzing the Stability of Response Styles Across Time: An Illustration with the Beck Depression Inventory (BDI-II)

Ferdinand Keller and Ingrid Koller

Abstract Questionnaires for clinical studies are often evaluated in cross-sectional settings and on the basis of classical test theory. Some of them, like the BDI-II which is one of the most widely used self-report instruments for assessing depression severity, are considered to have very good psychometric properties. However, these properties are rarely evaluated in longitudinal designs, and even less with models of item response theory (IRT). In addition, analyses of self-report questionnaires with IRT models provided evidence of two major response styles: the tendency to prefer extreme response categories, and the tendency to prefer the middle categories. Rasch models, in particular their extension to the so-called mixed Rasch model, are well suited to address these questions. They allow one to determine latent classes with different response styles and to analyze qualitative aspects of change such as the consistency of response styles across time. In this chapter first, an introduction to response styles and an overview of the mixed Rasch model, especially in the context of measuring change, are given and second, a practical example is elaborated using a sample of in-patients from a psychosomatic clinic that were assessed with the BDI-II at the beginning and at the end of in-patient treatment. The presence of two response styles is confirmed for the admission data, whereas for the discharge data the Rasch model seems sufficient. A combined analysis of both time points reveals three classes, one of which is a low symptom class and the other two reflect, again, the two response styles; these two classes remain quite stable over time.

F. Keller (✉)
Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital of Ulm, Steinhoevelstr. 5, 89075 Ulm, Germany
e-mail: ferdinand.keller@uniklinik-ulm.de

I. Koller
Department for Psychology, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_13

Measurement of Change in Clinical Psychology and Response Styles

Measurement of change in clinical psychology and psychiatry is of major importance for the evaluation of treatment approaches that are suited best for patient groups (i.e., comparing different types of psychotherapy and/or psychopharmacological treatment) as well as for monitoring improvement on an individual level (e.g., is there clinically significant progress across the treatment sessions, or is it indicated to modify the treatment approach?).

Unlike in achievement research (e.g., measurement of educational trajectories) where sophisticated statistical models are applied to assess the psychometric properties of items and to investigate change across time, treatment evaluation studies in the clinical realm mostly rely on a few, well-established outcome instruments that have high clinical face validity but whose psychometric properties in designs with repeated measurement are nonetheless rarely tested. Although this facilitates the comparison of study results, the measurement properties in longitudinal designs are largely unknown, except the test-retest-reliability based on the classical test theory approach.

A further threat for reliability and validity of self-report measures using Likert-type response scales are response styles which denote the tendency of an individual to respond to items irrespective of content. Plieninger and Meiser (2014) and Wetzel, Carstensen, and Böhnke (2013) give an overview on research regarding different types, in particular the extreme response style (ERS), i.e., the tendency to prefer the extreme response categories, and midpoint responding (MRS), i.e., the tendency to choose the middle categories (other response styles are, e.g., acquiescence and its opposite, disacquiescence). The authors conclude that past research suggested that response styles may be conceptualized as trait-like constructs that are stable across content domains and time. However, Weijters, Geuens, and Schillewaert (2010) question the results of previous studies on stability over time because of several methodological problems that arise with longitudinal designs, in particular possible memory effects and the usage of the same items which makes it impossible to distinguish between common variance due to response style and due to content.

In this chapter, we focus on the assessment of depression with a self-report instrument, the Beck Depression Inventory in its revised version (BDI-II; Beck, Steer, & Brown 1996; German version: Hautzinger, Keller, & Kühner 2006). The BDI-II is one of the most widely used self-report instruments to assess severity of depression in treatment studies as well as in psychodiagnostics. The psychometric properties are considered to be very good and extensive factor analytic studies have been done on cross-sectional samples (e.g., Brouwer, Meijer, & Zevalkink 2013a; Bühler, Keller, & Läge 2014; Ward 2006).

The BDI-II is used to address the presence of ERS and MRS in a clinical context. Moreover, the stability or change of these (potential) response styles across two time points (admission and discharge in a psychosomatic hospital) and the impact on the measurement of depression severity are examined. To our knowledge, neither issue

has been addressed before in the literature. Furthermore, the assessment of stability is confounded by the clinical intervention (treatment of the patients during their hospital stay) and thus more complicated than in studies where relatively stable traits (personality or achievement) are analyzed. Relations to basic variables which are available for this sample (gender, age, as well as diagnostic subgroups) will be assessed, too. Our method of choice is the mixed Rasch model (MRM; Rost 1990; Rost & von Davier 1995) which is an item response theory model (IRT) that is well suited to identify subgroups of patients that differ in response style, and offers the possibility to assess qualitative change across time (e.g., Glück & Spiel 1997). In the next sections the MRM and its application in the context of assessing different response styles and measuring change are described. After that the empirical example with the BDI-II is elaborated using the MRM approach.

The Mixed Rasch Model

The MRM is a generalization of the Rasch model (RM; Rasch 1960) to a discrete mixture distribution model which makes it possible to extract latent classes of individuals within which the RM holds. Between the extracted classes the RM has not to fit the data and, therefore the order of item difficulty and the range of item difficulties are allowed to vary. Thus, different response scale category usage can exist and therefore RM properties, e.g., measurement invariance, are not given between latent classes (e.g., Baghaei & Carstensen 2013; Embretson 2010; Meiser, Hein-Eggers, Rompe, & Rudinger 1995; Rost, Carstensen, & von Davier 1999; Rost & von Davier 1995). In summary, the MRM combines the unidimensional Rasch model with latent class analysis (LCA; e.g., Meiser et al. 1995; Meiser 2010; Rost 1991). But contrary to LCA, where within classes no person ability variation is assumed, MRM allows the quantification within classes, which means that individuals can differ in ability (e.g., Rost 2004; Spiel & Glück 2008).

In addition to the MRM for two-categorical items, extensions for items with polytomous response formats exist, for example, the mixed partial credit model (PCM) and the mixed rating scale model (RSM; e.g., Von Davier & Rost 1995). Because the applied example in this chapter is based on a polytomous response format the equation for the mixed PCM and one restriction, the mixed RSM, are shown. The restriction to the MRM is straightforward and is explained in Rost and von Davier (1995).

The mixed PCM defines the probability for a person $v = 1, \ldots, n$ to pass the threshold $l = 1, \ldots, m$ (with $s = 0, \ldots, m$ categories) of an item $i = 1, \ldots, k$ given the person ability θ_v in class $c = 1, \ldots, C$ and the item difficulty β_{ilc} with

$$P\left(x_{vilc} = l \middle| \theta_{vc}, \beta_{ilc}\right) = \sum_{c=1}^{C} \pi_c \frac{\exp\left(l\theta_{vc} - \beta_{ilc}\right)}{\sum_{s=0}^{m} \exp\left(s\theta_{vc} - \beta_{isc}\right)},$$

where π_c is the probability to belonging in latent class c (class size parameter) and the item difficulty $\beta_{ilc} = \sum_{l=1}^{m} \tau_{ilc}$, with the normalization $\sum_{i=1}^{k} \sum_{l=1}^{m} \tau_{ilc} = 0$, and $\beta_{i0c} = 0$ within all classes (see also, Rost 1991 or Wetzel et al. 2013). Furthermore, the mixed RSM results from the restriction $\tau_{ilc} = \beta_{ic} + \tau_{sc}$ where the same distances between thresholds are assumed for all items within all classes.

MRM fit will be tested in two ways. First, to test whether the estimated model fits the data, it has to be compared with the saturated model (i.e., the model with the maximum of estimable parameters) by a likelihood ratio test or Pearson chi-square test (see, e.g., Spiel & Glück 2008). Second, the estimated models (e.g., two-class and three-class solution) have to be compared using information criteria, such as, the Akaike information criterion (AIC; Akaike 1974), Bayesian Information Criterion (BIC; Schwarz 1978), or Consistent Akaike Information Criterion (CAIC; Bozdogan 1987). Based on the literature (Baghaei & Carstensen 2013; Wetzel et al. 2013) and simulation studies for the evaluation of performance of information criteria (Preinerstorfer & Formann 2012) BIC and CAIC should be preferred. A qualitative goodness of fit check is the comparison of the average membership probability of different individuals. If it is possible to assign individuals with high probability to one class, the MRM describes the data or response patterns well (see, Spiel & Glück 2008).

Assessment of Response Styles with the MRM

Several studies exist where the MRM was applied to various types of data for the detection of response styles in achievement tests (e.g., Baghaei & Carstensen 2013; Spiel & Glück 2008) and in personality questionnaires (e.g., Eid & Zickar 2010; Gollwitzer, Eid, & Jürgensen 2005; Rost et al. 1999; Rost, Carstensen, & von Davier 1997). All studies showed the suitability of the MRM for the identification and better understanding of different response styles. For example, Wetzel et al. (2013) analyzed several PISA 2006 attitude scales and the subscales of the NEO-PI-R with mixed PCM and further combined the respective latent response classes by means of a second order latent class analysis (c.f. Keller & Kempf 1997). The authors found that for 77 % of the participants a response style (ERS or MRS) occurred consistently across traits.

Furthermore, Wetzel et al. (2013) state that testing the consistency of response styles with the MRM requires that participants only differ in their response style but not in the trait that is being assessed or other factors that might influence the choice of a response category. Thus, the authors recommend estimating a constrained PCM where item locations are fixed to be equal across classes.

Assessment of Response Styles Across Time

In addition to the assessment of response styles in general, it can be of interest to determine whether class membership and, therefore, response style change over time. It is also possible to investigate this kind of question with MRMs (Glück & Spiel 1997 2010; Spiel & Glück 1998). With this exploratory approach it is possible to assess qualitative change across time. Research questions could be whether the class membership is constant over time or whether changes in membership are constant over time (e.g., those associated with class one at time point one change primarily to class two at time point two). For applications of the MRM in the case of dependent data, see, e.g., Glück and Spiel (1997 2010), Meiser et al. (1995), Meiser, Stern, and Langeheine (1998), and Rost (2004).

Technically, the data matrix has to be rearranged before analysis depending on research question. Two examples can be seen in Fig. 1 (see Rost 2004). Further possibilities for longitudinal data are conceivable (see, e.g., Meiser et al. 1998), but not of interest for our study and thus not discussed in this chapter.

If the data matrix is rearranged as shown in the left panel of Fig. 1 (long-format), one gets twice (or t times) as many participants, and change can be analyzed in one step (e.g., Glück & Spiel 1997). Thus, each time point can be seen as an independent subgroup of individuals. The individuals starting from t_2 are called virtual individuals. With this approach the item parameters are estimated in one step

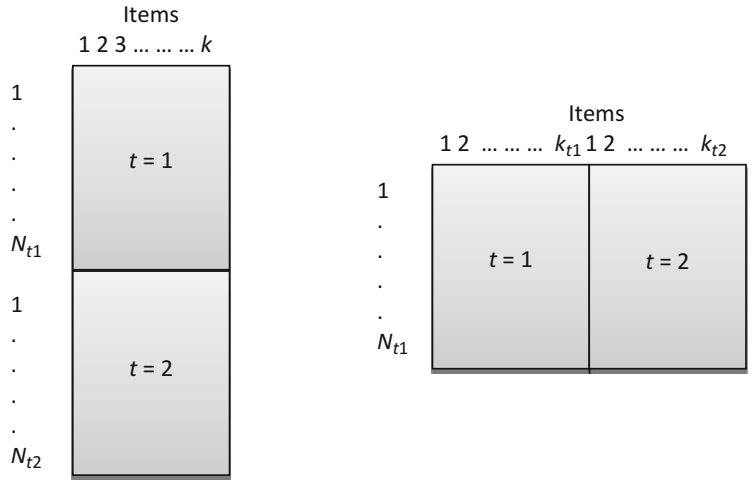


Fig. 1 Two possible ways to rearrange the data matrices for MRM in longitudinal studies. *Left panel* (long format): Data matrix with virtual persons at t_2. With this rearranging twice (or t) as many persons are available for analysis of dependent data. The MRM analysis can be performed in one step for all time points but the instrument must contain the same items across time. *Right panel* (wide format): Data matrix with virtual items at t_2. The MRM analysis can be performed also in one step

and it can be seen whether the individuals are staying within or moving between classes. However, there is one restriction, that is, the tests must contain the same items at all time points. In addition it must be taken into account that the assumption of local independence on the person side is violated.

It is also possible to rearrange the data as shown in the right panel in Fig. 1 (wide-format), and to analyze the time points as one long test. Again, in this approach the item parameters are estimated in one step, but, in addition, classes of participants are identified whose items at, e.g., t_2 reflect different magnitudes of change and different types of change (see, Glück & Spiel 1997). This approach, however, hides one major drawback. Due to the prolonged test, the sample size must be increased for a sufficiently accurate estimation of item parameters.

In the previous sections, the application of the MRM when assessing response styles and the procedure for the investigation of qualitative change in dependent data were described. In the next section we show the assessment of response styles using the MRM in a clinical context. First, the sample, the BDI-II and the procedure are described and second the results are given and discussed.

Assessment of Response Styles with the BDI-II

Sample

The sample consisted of in-patients from a clinic for psychosomatic disorders ($N = 1164$); they completed the BDI-II at admission within the routine diagnostic procedure and also at discharge. The mean age in the sample was 45.2 years ($SD = 10.8$; range: 19–72) and 64.7 % of the patients were female. The mean BDI-II total score at admission was 21.4 ($SD = 10.6$) and at discharge 9.1 ($SD = 8.1$). Eight hundred and two patients (68.9 %) were diagnosed with a primary affective disorder (ICD-10: chapter F3) as their main diagnosis; when taking F3 as a comorbid diagnosis, 1001 patients (86.0 %) fulfill the criteria of a depression. The most frequent comorbid disorder was substance-related disorders (ICD-10: F1; $n = 254$ (21.8 %)), and within a range of 15–19 % were somatoform disorders, anxiety-related disorders and post-traumatic stress disorder (PTSD), eating disorders, and personality disorders (see Table 3).

Description of the BDI-II

The BDI-II consists of 21 items that assess a wide range of depressive symptoms (e.g., sadness, suicidal thoughts and wishes, concentration difficulty, or loss of energy). Each item has four categories numbered from 0 to 3 that are formulated in a symptom-specific way (e.g., item 9 “suicidal thoughts and wishes” has the four

response options: 0 = "I don't have any thoughts of killing myself," 1 = "I have thoughts of killing myself, but I would not carry them out", 2 = "I would like to kill myself," and 3 = "I would kill myself if I had the chance"). The total score of these items reflect the severity of depression. In 1996, a minor revision of the BDI was carried out to meet the criteria of the DSM-IV (American Psychiatric Association 1994) and resulted in the BDI-II (Beck et al. 1996). Symptom scores from 14 to 19 indicate a mild depression, 20 to 28 a moderate, and above 28 a severe depression (Beck et al. 1996).

Procedure

The software program WINMIRA v1.45 (Von Davier 2001) was used to estimate the MRMs. We restricted ourselves for this data example to the mixed PCM, since it has been found in several samples that the fit of the RSM was worse than the fit of the PCM (Keller 2012), which supports the theoretical assumption that the BDI-II with its symptom- and category-specific text requires no restrictions on the category thresholds. The number of latent classes was successively increased from the PCM (1-RM) up to a PCM with three latent classes (3-RM) and parsimony of the models was evaluated using BIC and CAIC, as described above. Participants are then assigned to their most probable class and frequency tables are used to explore relations between time points and to the demographic variables. To compare the identified latent classes and to test the fit of the PCM, MRM analyses are performed, first, for the two time points separately, and then for the virtual sample (long-format, see Fig. 1, left panel) as suggested by Glück and Spiel (1997) and Rost (2004). Additionally, to test the model fit of the final solution (critical $\alpha = 5\,\%$), 500 re-simulations were carried out and the Pearson X^2 test-statistic was calculated (see Langeheine, van de Pol, & Pannekoek 1996); according to the recommendation in the WINMIRA output, only the p-value of the empirical probability distribution is reported.

An MRM analysis of the virtual items (wide-format, see Fig. 1, right panel) in one step was omitted, since it runs into several problems: (a) the number of estimated parameters gets in misbalance with our sample size (e.g., for two latent classes almost 500 parameters have to be estimated); (b) the dimensionality of item parameters could be tested, in particular the interesting question whether the items at t_1 and the items at t_2 are homogeneous, but the result would be valid only for this special split of items (t_1 vs. t_2). There is no analogue to the MRM for determining person heterogeneity (where two or more groups (latent classes) are built to achieve maximum person heterogeneity between classes) for the detection of maximum item heterogeneity (Rost 2004).

Following Wetzel et al. (2013), a constrained PCM is also estimated where the item locations are fixed to be equal across classes. The constrained PCM delivers homogeneous latent classes which only differ in the distribution of the threshold parameters (Wetzel et al. 2013) that is in response style. Consequently, the authors

compare the unconstrained PCM with the constrained PCM and use only those subscales for which the constrained PCM (i.e., ensuring trait homogeneity between the latent classes) shows a better fit in BIC and CAIC than the unconstrained PCM.

Results

Mixed PCM Estimated Separately for the Two Time Points

The likelihood, number of parameters, and the information criteria for the PCM and the two-class and the three-class solution are displayed in Table 1. For the admission data, there is a clear minimum in BIC and CAIC for the solution with two latent classes (Modelfit$_{2Class}$: empirical $p = .046$). The first class consists of 64.3 % of the individuals, and the thresholds (see Fig. 2) suggest that this class prefers to use the middle categories. The estimated thresholds for the second class (35.7 %) are closer together; that is, it is more difficult for them to "leave" category zero and also not very difficult to endorse the highest category: they prefer the extreme categories. Item 9 (suicidal thoughts) has a high threshold in both classes, because acute suicidality is an exclusion criteria in a psychosomatic clinic and thus, the frequencies

Table 1 Model fit for the PCM at admission, at discharge (both estimated separately), and for the virtual sample (long format)

Partial credit models	Log-Lik.	# of parameters	BIC	CAIC	Reliability	Class sizes (%)
Admission						
1-RM	−26,227.83	125	53,338.1	53,463.1	.91	100
2-RM	−25,640.07	249	**53,038.0**	**53,287.0**	.91/.92	64/36
3-RM	−25,344.08	373	53,321.4	53,694.4	.89/.92/.90	43/32/25
2-RM constr.	−25,803.55	228	53,216.7	53,444.7	.90/.90	61/39
Discharge						
1-RM	−17,714.72	125	36,311.9	**36,436.9**	.84	100
2-RM	−17,261.26	249	**36,280.4**	36,529.4	.73/.87	63/37
2-RM constr.	−17,335.96	228	36,281.5	36,509.5	.75/.87	64/36
Long format						
1-RM	−45,483.88	125	91,936.9	92,061.9	.90	100
2-RM	−44,153.86	249	90,238.2	90,487.2	.90/.78	54/46
3-RM	−43,496.61	373	**89,885.0**	**90,258.0**	.77/.90/.91	41/38/21
4-RM[a]						
3-RM constr.	−44,015.01	331	90,596.2	90,927.2	.90/.72/.89	41/33/26

Note. constr. = constrained, i.e., item locations set equal across classes

[a]Several attempts to estimate a four-class solution resulted always in non-convergent solutions and the fourth class consists of almost no person (class sizes <0.1 %); the other classes remain the same

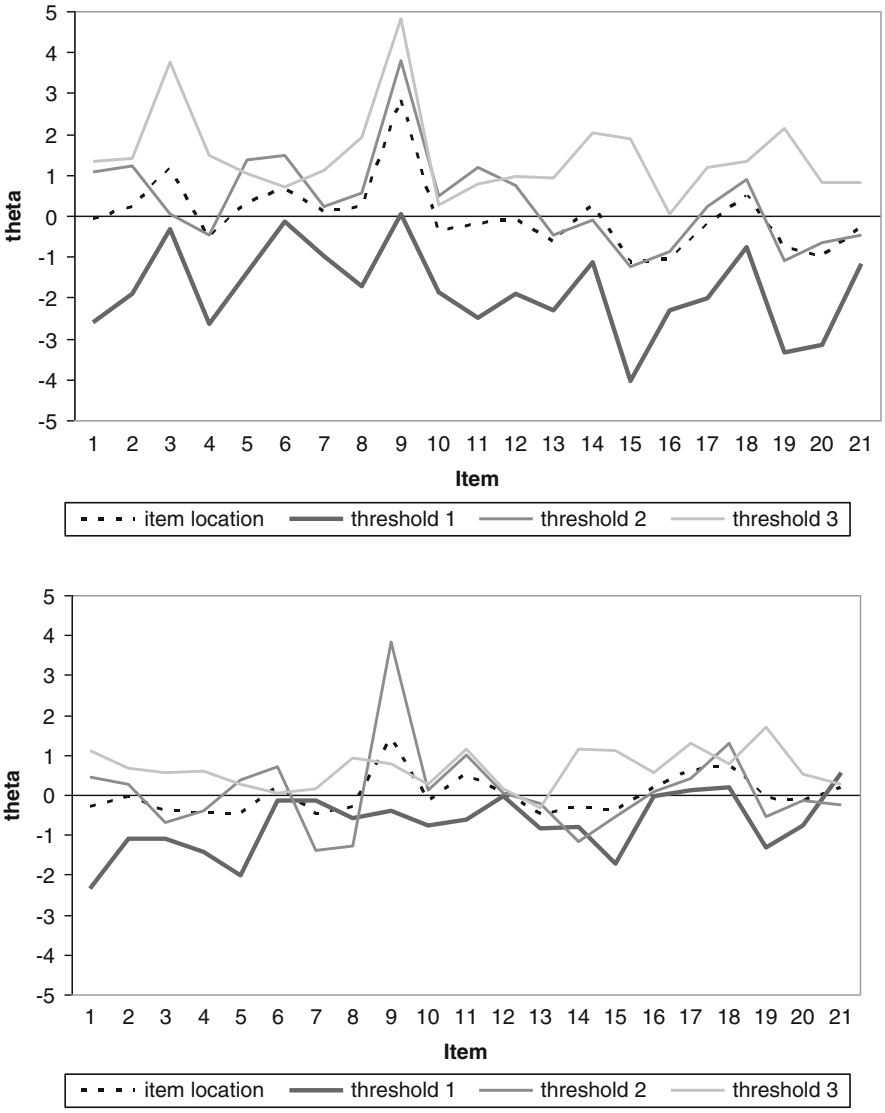


Fig. 2 Threshold parameters and item locations for the unconstrained PCM with two latent classes for the admission data (*upper part*: class 1 (MRS), *lower part*: class 2 (ERS))

for the category 3 are low. The average class membership probabilities indicate good separation in assignment of the individuals to the classes (.935 for class 1 and .907 for class 2).

For the discharge data, the BIC still favours a two-class solution ($\text{Modelfit}_{2\text{Class}}$: empirical $p = .032$), while the CAIC suggests a solution with only one class ($\text{Modelfit}_{1\text{Class}}$: empirical $p = .008$). Since the BIC is usually used as a decision

criterion and also the model-fit statistics favour the 2-RM solution, we selected the PCM with two classes. The class sizes are quite similar (63.2 and 36.7 %) compared to the admission data and also the patterns of the thresholds, indicating a class with tendency to the middle categories (class 1) and a class with tendency to the extreme categories (class 2), although the range of the thresholds increased. Concerning average class membership probabilities, the values are even better than those of the admission data (.944 for class 1 and .924 for class 2).

The constrained PCMs reveal differential results: for the admission data, the fit of the constrained PCM with two latent classes is worse than the unconstrained 2-RM, indicating additional heterogeneity; for the discharge data, the constrained and the unconstrained PCM with two latent classes are similar in model fit, especially in the BIC, indicating no additional heterogeneity.

The mean BDI-II scores at t_1 are different for the two classes ($t = -5.00$, $df = 623.5$, $p < .001$; Cohen's $d = 0.32$), with class 1 (MRS) having a mean score of 20.2 ($SD = 9.0$) and class 2 (ERS) having 23.7 ($SD = 12.8$). At discharge, the difference is larger ($t = -23.7$, $df = 526.6$, $p < .001$; Cohen's $d = 1.57$). Class 1 (MRS) has a low mean value of 5.3 ($SD = 4.2$), whereas the ERS class has a mean value of 16.0 ($SD = 8.7$).

Mixed PCM Estimated for the Virtual Sample

To assess possible qualitative change of response styles across the two time points we applied the MRM on the long format of data. The lower part of Table 1 contains also the indices for the PCM when applied to the virtual sample (long-format, left panel of Fig. 1). Both information criteria, BIC and CAIC, favor a three-class solution ($\text{Modelfit}_{3\text{Class}}$: $p = 0.03$). Inspection of the threshold parameters indicates that the largest class has many unordered thresholds; this class has also a mean raw score of 6.7 ($SD = 5.7$). The other two classes can be interpreted as before: class 2 seems to have a tendency to the middle categories (MRS), and class 3 prefers the extreme values (ERS). The mean class membership probability is sufficient to good with .939, .905, and .892, respectively.

Stability of Class Membership in the Virtual Sample

The members of class 1 show high stability, most of them (93.9 %) stay in the class 1 (see, Table 2). This class, however, is characterized by many unordered thresholds, and inspection of the mean BDI-II scores for this class revealed a low mean value (6.7) suggesting that the higher categories of the BDI-II items are rarely endorsed. Separating the mean BDI-II values for admission and discharge, this class has a mean sum score of 8.7 ($SD = 6.1$) at admission and of 4.8 ($SD = 3.5$) at discharge; that is, this class contains patients with low depression values at admission and even lower ones at discharge.

Table 2 Cross-classification from t_1 to t_2 in the long-format MRM with three latent classes

	Class assignment at discharge (t_2)			
Class assignment at admission (t_1)	Class 1	Class 2	Class 3	Total
Class 1	124 (93.9 %)	7 (5.3 %)	1 (0.8 %)	132 (100 %)
Class 2	510 (72.8 %)	148 (21.1 %)	43 (6.1 %)	701 (100 %)
Class 3	207 (62.5 %)	43 (13.0 %)	81 (24.5 %)	331 (100 %)
Total	841 (72.3 %)	198 (17.0 %)	125 (10.7 %)	1164 (100 %)

Note. Class assignments are given as frequencies and percentages

The majority of patients who are in the response style classes 2 or 3 at admission also move to class 1 at discharge (72.8 or 62.5 %). Obviously, class 1 consists of the much improved patients, but improvement is also remarkable in the other two classes: class 2 has a mean sum score of 21.0 ($SD = 8.6$) at admission and of 8.4 ($SD = 7.3$) at discharge; the values for class 3 are 27.3 ($SD = 11.1$) at admission and 12.4 ($SD = 9.6$) at discharge. Aside from that trend into the low symptom class 1, there is a clear preference to stay in class 2 or in class 3 and not to switch to the respective other response style class. The odds ratio for these four cells ("22," "23," "32," "33") is 6.48 (95 %-CI: 3.92–10.7).

Associations Between Latent Classes and Gender and Age

The cross-classification of gender and the assigned three classes for the long-format gives no significant association, neither at t_1 ($\chi^2 = 0.88$, $df = 2$, n.s.) nor at t_2 ($\chi^2 = 1.93$, $df = 2$, n.s.). The same is true for the separate analysis of t_1 ($\chi^2 = 2.53$, $df = 1$, n.s.); there is an association for t_2 ($\chi^2 = 5.24$, $df = 1$, $p = .022$) with female patients being underrepresented in class 1 (MRS; 62.3 % vs. 69.0 % in class 2 (ERS)), but effect size is low ($\Phi = .067$).

Concerning age, there are significant mean differences between the three classes assigned by the long-format analysis ($F(2,1161) = 14.0$, $p < .001$; $eta^2 = .024$; mean values are 43.9, 46.6 and 43.0 years for the three classes). For the separate analysis of t_1, there is a significant difference in age as well ($t = 5.44$, $df = 1162$, $p < .001$; Cohen's $d = .34$). Class 1 (MRS) is slightly older with a mean value of 46.5 years ($SD = 10.4$) than the ERS class which has a mean value of 42.9 years ($SD = 11.0$).

Associations to Diagnostic Subgroups

The proportion of MRS and ERS at admission is not evenly distributed across diagnostic subgroups (see Table 3). There is preponderance for ERS in individuals with personality disorders, eating disorders, PTSD, and substance-related disorders. Patients with depression are the only group which are overrepresented in the MRS class. The remaining diagnostic subgroups (anxiety, somatoform disorders) are about uniformly distributed.

Table 3 Distribution of response style classes for diagnostic subgroups at admission

Diagnosis (ICD chapter)	Total frequency and percentage	Percentage in class 1 (MRS) and class 2 (ERS)		Odds ratio	95 % CI	
					Lower	Upper
F1 (substance-related disorders)	$N = 254$ (21.8 %)	18.9 %	27.3 %	1.62	1.22	2.15
F3 (depression)	$N = 1001$ (86.0 %)	87.9 %	82.5 %	0.65	0.47	0.91
F4 (anxiety)	$N = 226$ (19.4 %)	18.9 %	20.4 %	1.11	0.82	1.50
F4 (PTSD)	$N = 218$ (18.7 %)	15.7 %	24.4 %	1.73	1.28	2.34
F4 (somatoform disorders)	$N = 177$ (15.2 %)	16.0 %	13.8 %	0.84	0.60	1.19
F5 (eating disorders)	$N = 212$ (18.2 %)	15.0 %	24.1 %	1.80	1.33	2.43
F6 (personality disorders)	$N = 211$ (18.1 %)	12.7 %	28.3 %	2.73	2.01	3.69

Discussion

The current study examined the existence and the stability of the MRS and the ERS response styles with an IRT based approach. For this purpose the mixed PCM was used which combines the Rasch model with latent class analysis. Usually this model is used for the assessment of latent classes in which the Rasch model holds for the data. There are also studies in which the model is used for the assessment of different response styles. There are also applications testing the consistency across several traits and in longitudinal studies, but not for the assessment of response styles across time. Furthermore, our study is more complex than a simple longitudinal study, since we examined response styles in the clinical context in which mentally ill individuals received clinical intervention between the measurement points. For this purpose we used the BDI-II, a questionnaire to assess the severity of depression. For the decision on the number of latent classes, a bootstrap analysis of model fit showed always low fit values and was not very helpful; thus, this decision was based on information criteria.

The application of the mixed PCM shows interesting results for the BDI-II. The main results can be summarized as follows: For the separate analysis of the admission data (t_1), a distinction into two latent classes could be found. The classes could be interpreted as MRS and ERS. Thus, the response styles ERS and MRS that have repeatedly been found in personality and achievement tests could also be replicated with a self-report questionnaire in depression research. The constrained model fitted worse than the unconstrained model; that is, there might be some additional heterogeneity between classes beyond the response style alone (although the differences in mean BDI-II sum score are small).

For the discharge data (t_2), the separation into two latent classes indicating MRS and ERS was questionable. Furthermore, the response style classes seem to be highly confounded with depression severity when comparing the mean sum scores of the two classes. The comparison of the fit of the constrained and the

unconstrained mixed PCM with two classes, however, shows minimal differences; that is, homogeneity can be assumed. In sum, it might be concluded that the model with two classes is probably not necessary and the PCM holds for the discharge data, supporting the finding of Keller (2012) where the PCM showed the best fit in the sample of healthy individuals.

The analysis with the long-format data yields three classes, where one class contains the patients with low depression values and the other two can, again, be described as MRS and ERS. The low symptom class 1 is the largest class at discharge because most of them stay within this class and the major part of the patients in the initial classes 2 and 3 move to the class 1. Within the classes 2 and 3, there is a pronounced stability to stay, i.e., to remain in the same response style. Although additional heterogeneity has to be assumed (the constrained PCM fits worse than the unconstrained PCM with three latent classes), we may take this as a confirmation of the stability of the ERS and MRS response styles over time, as has been found before by Weijters et al. (2010) with a quite different methodological approach (the authors used a second order factor model in which they specified time-invariant and time-specific response style factors based on a coding scheme for weighting the item categories).

There are no significant relations between response style classes and gender except for the separate analysis at discharge, but effect size is low and we may conclude that gender is not related to response style to a relevant degree. However, the small effect would be in line with Weijters et al. (2010) who found that female respondents showed significantly higher levels of ERS. In contrast, Khorramdel and von Davier (2014) found no significant gender differences with regard to ERS and MRS, but their sample of students was relatively homogeneous in age and education.

The difference in age between response style classes was significant, but small in effect size and seems therefore also to be negligible. The uneven distribution in several diagnostic subgroups is an interesting result, but due to the lack of previous findings in the literature, interpretations derived only from clinical impressions may be currently too speculative before replication of these differences.

The emergence of response styles at admission and in the combined sample (long format) has implications for clinical treatment as well as for the evaluation of treatment. For treatment assignment based on the admission BDI-II score, consider a patient with a sum score of 20 which is a commonly used inclusion criterion for depression treatment studies (and may be used also in assigning treatment modules in a psychiatric/psychosomatic clinic). The corresponding person parameter in the PCM would be -0.78; with the additional knowledge of the response style of an individual as provided by the mixed Rasch model, the individual in the ERS class would receive a person parameter of -0.97, while the individual assigned to the MRS class would receive a value of -0.69. For a sum score of 14 (= cutoff for mild depression), the difference would be even larger: -1.69 for the ERS class and -1.13 for the MRS class.

In extension to this cross-sectional differential assignment of patients, one is usually interested in whether a patient has significantly improved during the stay in a clinic/from a treatment approach. One of the most popular approaches is the

Reliable Change Index (RCI—Jacobson & Truax 1991) that is based on classical test theory. Brouwer, Meijer, and Zevalkink (2013b) compare the RCI with an IRT-based change index. For a majority of cases the IRT-based statistic resulted in a similar conclusion as compared to the use of the RCI, but for some patients within the range of lower or higher change scores, IRT provided a more accurate tool (Brouwer et al. 2013b). The addition of response style information may further improve the classification into improved vs. unchanged patients (or deteriorated patients).

Currently, however, our MRM results are explorative and need to be replicated in other samples. Furthermore, other IRT-related methodological possibilities for the assessment of response styles could be examined. Multi-process IRT models have been developed and applied to decompose observed rating data into multiple response processes (Khorramdel & von Davier 2014; Plieninger & Meiser 2014). Wetzel et al. (2013) suggest conceiving response styles as their own dimension in a multidimensional model (e.g., the multidimensional random coefficient multinomial logit model by Adams, Wilson, & Wang 1997). For the purpose of measuring change, e.g., the evaluation of improvement of an individual during therapy, these multidimensional models seem a promising way to answer such research questions in longitudinal designs, and will be assessed in further studies.

Acknowledgement We thank Dr. Robert Mestel, head of Research/Quality Assurance of HELIOS Klinik Bad Grönenbach, for providing us with the dataset.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation, 18*(5). Retrieved from http://pareonline.net/getvn.asp?v=18&n=5.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck depression inventory—Second edition. Manual*. San Antonio, TX: The Psychological Corporation.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytic extensions. *Psychometrika, 52*(3), 345–370.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013a). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment, 25*(1), 136–145.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013b). Measuring individual significant change on the Beck Depression Inventory-II through IRT-based statistics. *Psychotherapy Research, 23*(5), 489–501.

Bühler, J., Keller, F., & Läge, D. (2014). Activation as an overlooked factor in the BDI-II: A factor model based on core symptoms and qualitative aspects of depression. *Psychological Assessment, 26*(3), 970–979.

Eid, M., & Zickar, M. J. (2010). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen

(Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 255–270). New York, NY: Springer.

Embretson, S. E. (2010). Mixed Rasch models for measurement in cognitive psychology. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 235–254). New York, NY: Springer.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitation of four different approaches. *Methods of Psychological Research, 2*(1). Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art6/article.html.

Glück, J., & Spiel, C. (2010). Studying development via item response models: A wide range of potential uses. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 281–292). New York, NY: Springer.

Gollwitzer, M., Eid, M., & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment, 17*(1), 56–69.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *BDI-II. Beck depressions inventar revision—Manual [BDI-II. Revision of the Beck Depression Inventory—Manual]*. Frankfurt, Germany: Harcourt Test Services.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.

Keller, F. (2012). Das Beck-Depressions-Inventar (BDI-II): Psychometrische Analysen mit probabilistischen Testmodellen [The Beck-Depression-Inventory (BDI-II): Psychometric analyses with probabilistic test models]. In W. Baros & J. Rost (Eds.), *Natur- und kulturwissenschaftliche Perspektiven in der Psychologie [Natural science and cultural studies perspectives in psychology]* (pp. 120–132). Berlin, Germany: Verlag irena regener.

Keller, F., & Kempf, W. (1997). Some latent trait and latent class analyses of the Beck-Depression-Inventory (BDI). In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 314–323). Münster, Germany: Waxmann.

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the big five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research, 49*(2), 161–177.

Langeheine, R., van de Pol, F., & Pannekoek, J. (1996). Bootstrapping goodness-of-fit-measures in categorical data analysis. *Sociological Methods & Research, 24*(4), 492–516.

Meiser, T. (2010). Rasch models for longitudinal data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 191–200). New York, NY: Springer.

Meiser, T., Hein-Eggers, M., Rompe, P., & Rudinger, G. (1995). Analyzing homogeneity and heterogeneity of change using Rasch and latent class models: A comparative and integrative approach. *Applied Psychological Measurement, 19*(4), 377–391.

Meiser, T., Stern, E., & Langeheine, R. (1998). Latent change in discrete data: Unidimensional, multidimensional, and mixture distribution Rasch models for the analysis of repeated observations. *Methods of Psychological Research Online, 3*(2), 75–93. Retrieved http://www.dgps.de/fachgruppen/methoden/mpr-online/issue5/art6/meiser.pdf.

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*. doi:10.1177/0013164413514998.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65*(2), 251–262.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271–282.

Rost, J. (1991). A logistic mixture distribution model for polytomous item responses. *The British Journal for Mathematical and Statistical Psychology, 44*(1), 75–92.

Rost, J. (2004). *Lehrbuch Testtheorie—Testkonstruktion [Testtheory—Testconstruction]*. Bern, Germany: Verlag Hans Huber.

Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten [Are the Big Five Rasch scalable? A reanalysis of the NEO-FFI norm data]. *Diagnostica, 45*(3), 119–127.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 257–268). New York, NY: Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–462.

Spiel, C., & Glück, J. (1998). Item response models for assessing change in dichotomous items. *International Journal of Behavioral Development, 22*(3), 517–536.

Spiel, C., & Glück, J. (2008). A model-based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 45–65). Cambridge, MA: Hogrefe.

Von Davier, M. (2001). WINMIRA 2001 user's guide. Kiel: IPN.

Von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). New York, NY: Springer.

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory-II. *Psychological Assessment, 18*(1), 81–88.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*(1), 96–110.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality, 47*(2), 178–189.

Part V
Other Methods for the Analyses of Dependent Data

Studying Behavioral Change: Growth Analysis via Multidimensional Scaling Model

Cody Ding

Abstract In recent years, statistical methods for latent growth modeling have been commonly used in educational and psychological research. The purpose of this chapter is to illustrate growth modeling of change in pattern using multidimensional scaling (MDS) in the context of growth mixture modeling (GMM). We discuss how MDS growth pattern analysis may differ with respect to modeling changes in level, as commonly done with GMM, given that they have similarities in terms of model estimation, latent group identification, classification of individuals, and the interpretation of growth trajectory. We discuss the MDS growth pattern analysis in particular since it is less known. Using two simulated data sets as well as actual data from the Early Childhood Longitudinal Study of the Kindergarten Class of 1998–99 (ECLS-K) study, we demonstrate differences in growth pattern vs. level. It is our goal to provide researchers with a better idea of what MDS growth pattern analysis can accomplish, which may provide them with the knowledge to appropriately utilize this type of analysis in their own research.

Studying change processes has been an area of interest in education and the behavioral sciences for a long time. Researchers and practitioners in the behavioral sciences are concerned with questions about how individuals change over time (Willett & Sayer, 1994; Williamson, Appelbaum, & Epanchin, 1991). In recent years the number of models utilized to address questions of this kind has increased substantially (e.g., Collins & Horn, 1991; Aber & McArdle, 1991). Given the importance of studying change, the purpose of this chapter is to illustrate and discuss how growth patterns are modeled via multidimensional scaling (MDS) growth analysis in the framework of commonly used growth mixture modeling (GMM). We first briefly discuss GMM since this method is well known among researchers. Then MDS analysis of growth is discussed in the same context. The example data are used to illustrate the MDS approach so that developmental researchers can employ the relevant method in their own research.

C. Ding (✉)
University of Missouri-St. Louis, St. Louis, MO, USA
e-mail: dingc@umsl.edu

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_14

Growth Mixture Modeling

One particular kind of latent growth curve modeling[1] receiving more attention in recent years is GMM, the analysis of which can be conducted using a structural equation modeling approach (e.g., Boscardin, Muthén, Francis, & Baker 2008; Hallquist & Lenzenweger 2012; Muthen 1989 2001; Nagin 1999). It is beyond the scope of this chapter to introduce the details of GMM. Those unfamiliar with GMM may wish to consult Jung and Wickrama (2008) or Ram and Grimm (2009) for a good introduction. Suffice it to say here, GMM is an extension of single-population latent growth models, combining latent class analysis and latent growth curve modeling into one coherent modeling system. It is particularly useful when the subpopulation is unobserved or unknown a priori and is designed to identify and describe qualitatively distinct classes of cases with respect to change in level, allowing different growth parameters across the classes. As such, it can be employed to test the hypotheses of (a) whether there are different growth trajectories actually present in the population and (b) if they exist, whether the trajectories are defined by different initial growth status (i.e., initial level) as well as later growth rates in level. Ram and Grimm (2009) specify GMM model as follows:

$$y_{it} = \Sigma \left[\pi_{ic} \left(f_{0ic}\lambda_{0ct} + f_{1ic}\lambda_{1ct} + e_{ict}\right)\right] \quad (1)$$

where y_{it} is an individual's score y at time t. f_{0ic} and f_{1ic} are latent growth factors that represent intercept (i.e., initial score) and slope (i.e., growth shape) of latent class c to which individual i belongs. λ_{0ct} and λ_{1ct} are factor loadings corresponding to the two growth factors. e_{ict} is a time-specific residual. π_{ic} is the probability that individual i belongs to latent class c, with $0 \leq \pi_{ic} \leq 1$, and $\Sigma\pi_{ic} = 1$. Estimated posterior probabilities for each individual's class membership are derived as $\pi_{ic} = p(k_{ic} = 1|y_i)$, with the latent class membership indicators, k_{ic}, being 1 if individual i belongs to class c, and 0 otherwise. The objective of GMM is (1) to represent across-class differences in the initial score and the growth shape, (2) to determine the means of growth factors, and (3) to establish variance and covariance of the growth factors.

As indicated by Jung and Wickrama (2008), there are three main areas of GMM that attract much of the current debates: (1) identification of latent classes, (2) which model fit index to use, and (3) the problem of convergence. The first two issues are not unique to GMM since many other modeling methods encounter the same issues. In this regard, good research should focus on questions that prompt the development of theories and hypotheses. We need to judge the models by whether they conform to our theories. The third issue is more challenging since the computational load of GMM estimation is very heavy and mathematically modeling a sample distribution

[1] As indicated by Ram and Grimm (2009), latent growth modeling is a generic term that include various similar growth modeling approaches, such as latent trajectory analysis, latent curve modeling, mixed effects models of change, and multilevel models of change.

that consists of a mixture of many different kinds of sub-distributions is extremely difficult (Jung & Wickrama 2008). As a result, some models are less stable or difficult to estimate. Therefore, Wang and Bodner (2007) recommend using GMM in a confirmatory manner, although the model may undergo many modifications.

Growth Mixture Modeling via MDS

Different from GMM using a SEM approach, multidimensional scaling growth pattern analysis is an exploratory and data visualization method that focuses on modeling change in pattern only, with level being removed. This is the chief difference between the two approaches. That is, a key distinction between GMM and MDS is that MDS does not accommodate level differences, while GMM can be used with a random intercept factor within-class to account for level differences with differences in shape being accommodated through class level differences. Although MDS analysis has the same objective as GMM, its methodological foundation is a geometric or spatial representation of relationships among repeated measures. Using MDS for identifying growth trajectories in latent pattern has been discussed in a series of papers by Ding, Davison, and colleagues (Davison et al. 1995; Davison, Gasser, & Ding 1996; Davison, Kuang, & Kim 1999; Ding 2005 2007a 2007b; Ding, Davison, & Petersen 2005). Briefly, a dimension from MDS-GM represents changes in pattern when the variable under study is repeated across time. In other words, in MDS models, each dimension k represents a growth curve, or an exemplar of a particular arrangement of scores of different time points, called a prototypical growth pattern or latent growth profile. This growth curve is quantified by a set of scale value estimates x_{kt} from the Euclidean distance model in MDS analysis. In a sense, this set of scale value estimates can be considered a set of polynomial contrast coefficients, which can be used for hypothesis testing in a subsequent analysis.

Technically, the MDS growth pattern model can be represented as

$$y_{it} = c_i + \sum\nolimits_k w_{ik} x_{kt} + \varepsilon_{it} \tag{2}$$

where y_{it} is the observed score for individual i at time t. x_{kt} is a scale value, as described previously. c_i is a level parameter or initial score for person i if the average scale value is centered on the first time of measurement. ε_{it} is an error term for person i at time t. w_{ik} represents the individual's profile match index that quantifies the degree to which an individual manifests the identified growth profile. First, MDS analysis involves estimating scale values from a distance matrix computed for every pair of time points, t_i and t_{i+1}. The distance data for time points t_i and t_{i+1} is the difference between the scores y_i at times t_i and t_{i+1} for person i. The distance measure, $d_{tt'}$, can be computed from observed responses at time pairs across all individuals:

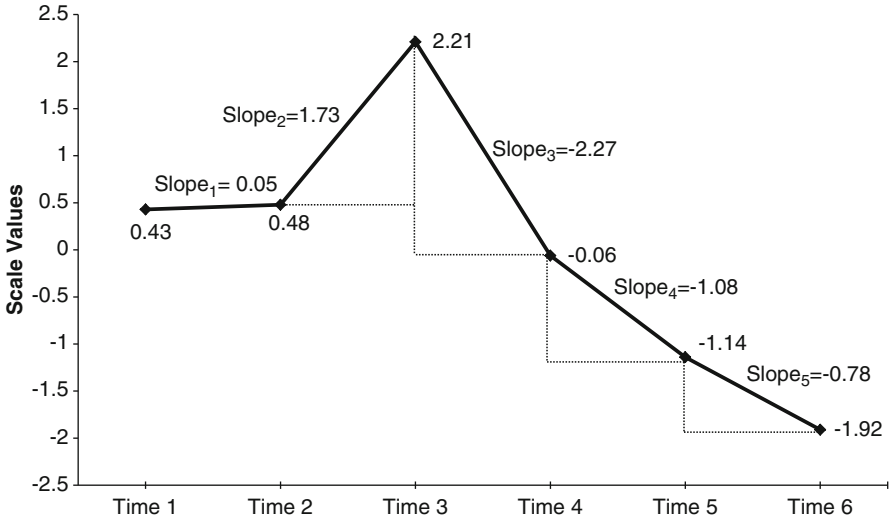


Fig. 1 Distance as representation of growth rates for different time intervals

$$x_{\mathrm{kt}} \rightarrow d_{\mathrm{tt}}, = \sqrt{\sum\nolimits_{\mathrm{i}} \left(\mathrm{y_{it}} - y_{it},\right)^2} \tag{3}$$

For each dimension, a plot of the scale values along the vertical axis against the time points along the horizontal axis shows the pattern of growth. Without loss of generality of the solution, the scale values along the dimension are centered in such a way that the mean scale value is equal to zero or some value that has substantive interpretation. Figure 1 illustrates the concept of distance as a representation of growth rates, as indicated by the scale values. In Fig. 1, a set of six time points is plotted along one dimension.[2] The differences between scale values of adjacent time points indicates the change (i.e., slope) for a given time interval. As can be seen in the figure, little or no growth occurs from time 1 to time 2 ($\mathrm{slope}_1 = 0.05$), but a large change is observed from time 3 to time 4 ($\mathrm{slope}_3 = -2.27$). It should be noted that although the interval must be the same for each individual, time intervals do not need to be equally spaced because growth rate is the slope for each particular interval. If the time unit between time 1 and time 2 is 3 months but the time unit between time 3 and time 4 is 1 year, then slope_1 indicates growth for the 3 months and slope_2 is the growth for 1 year.

[2]In MDS, dimensions are defined as a set of *m* directed axes that are orthogonal to each other in a geometric space. In the applied context, dimensions may be viewed as underlying representations of how the points may form certain groupings, which would meaningfully explain the data. This concept is similar to latent classes or factors in mixture modeling. Distance is defined as distribution of points along *k* dimension among pairs of objects (e.g., time points) in a plane that shows changes.

Based on the information provided by w_{ik} in Eq. (2), individuals are then classified into each growth dimension by using posteriori profile probability (Ding 2007b). For each individual i, the probability of profile membership in profile k can be calculated as follows:

$$p_i(k|w_{ik}) = \frac{p_i(w_{ik}|k)\pi_i}{\sum p_i(w_{ik}|k)\pi_i} \tag{4}$$

where $p_i(k|w_{ik})$ is the estimated probability of observed individual i belonging to profile k, given the individual's profile match index w_{ik} in Eq. (2), and π_i is the estimated proportion of profile variance among the total variance in the observed profiles for a given individual. The quantity $p_i(w_{ik}|k)$ is the probability of observing w_{ik} for a given profile k. In a sense, the probability $p_i(k|w_{ik})$ can be viewed as an approximation of the posterior probability of profile membership. The posteriori profile probability is calculated after estimation of the growth pattern. The resulting profile type can then be used in subsequent analyses. For example, we could investigate the relationships between growth profile type and covariates under inquiry.

To adapt the model for studying growth or change, the origin of the scale values needs to be "centered" appropriately. Because Euclidean distances are invariant with respect to choice of an origin, in MDS analyses based on distance models, the fit of the model to the distance data is invariant with respect to a translation of origin. Therefore, once an initial MDS solution is obtained, the zero point on each dimension can be reset in one of several ways, depending on the desired interpretation of the level parameter. The particular way of "centering" the origin of scale values along each dimension determines whether the model is a growth model or a change model.

If growth along the time dimension is to be studied, the MDS growth profile model can be created by centering the dimension zero point in a way that is meaningful for growth curve analysis. Given the importance of initial level in the literature on growth, the zero point can be set to correspond with the scale value at the first time period (i.e., $x_{k(1)} = 0$ for all k), then scale values will indicate growth rates for different time intervals. The intercept, c_i, becomes the expected score under the model for person i at the initial time t = 1. That is, in Eq. (2), if $x_{k(1)} = 0$ for every k, then the model predicted data point at time 1 for person i, $y_{i(1)}{}' = c_i + \Sigma_k\, w_{ik}\, x_{k(1)}$ reduces to $y_{i(1)}{}' = c_i$, and the intercept corresponds to initial level.[3]

On the other hand, if the data involve change, the abovementioned method of centering would be inappropriate since a change pattern does not follow a

[3]The issue of setting the origin for each dimension in the PAMS model corresponds to the "centering" issue in multiple regression. That is, just as the interpretation of the intercept parameter in multiple regression changes depending on how the predictor variables are centered, the interpretation of the intercept parameter in latent growth curve models changes depending on placement of the zero point along each growth dimension.

monotonic (at least implicitly) trajectory as does a growth curve, especially when change patterns are multidimensional. To adapt the model for this type of data, the zero point of scale values on each dimension is set equal to the mean scale value along that dimension; that is, $0 = (1/T)\Sigma_t x_{k(t)}$ for all k. If the zero point on each dimension is so defined, then scale values will indicate change patterns over time, and the intercept, c_i, becomes the average score of person i over the several time periods; that is, $c_i = (1/T)\Sigma_t y_{i(t)}$. In the example below, adolescent mood change data will be used to illustrate MDS change profile analysis.

MDS growth pattern analysis to modeling growth mixture, as we described above, has three main aspects that differ from commonly discussed GMM. First, the estimation of growth pattern or profile, as indicated by scale values, and the number of growth classes (i.e., growth profile type), as indicated by the number of dimensions, are different. For GMM, latent class analysis model is first used for classification of individuals and a latent class indicator variable is computed. The growth model is then estimated using this information (e.g., Asparouhov & Muthén 2012; Vermunt 2010). MDS growth pattern analysis takes the reverse approach, first identifying the typical growth patterns or profiles and then determining how much each individual resembles a given growth pattern or profile. However both approaches can be subject to classification error.

One practical implication of this difference is that the model building process for MDS growth pattern analysis is easier to implement as one only needs to specify a set of 1 to k dimensional solutions and choose a k dimensional solution that best approximates the data. The measure for model selection is *Stress* value (Kruskal 1964) or R^2, an index of the proportion of variance in observed growth profiles accounted for by the model. There is no need to specify a series of models with respect to growth trajectory and number of classes, as recommended by Muthén and Muthén (2001) and Ram and Grimm (2009) for GMM analysis.

Second, and more importantly, MDS growth pattern analysis typically models change in patterns rather than in levels that mirror the observed trajectories. This is because the growth profiles of MDS solutions have a mean of zero and are represented as deviations about the growth profile's mean of zero. Positive growth profile scale values signify scores above the growth profile's mean; negative scale values signify scores below the growth profile's mean. Thus, the MDS solutions display the patterning of scores in a prototypical growth profile but do not display elevation or level information (Davison et al. 1996). In most GMM analysis, growth trajectory is typically represented by a regression model, either linear or nonlinear, which indicates the change in level. This is a key difference between the two approaches as discussed previously. One implication of this distinction is that MDS and GMM analysis may result in a different number of growth classes/types. Specifically, when change in pattern displays the same information as the level, two approaches reach the similar findings. But if there are fewer changes in pattern than in level, then MDS growth pattern analysis may results in a fewer growth classes/types than those from GMM analysis.

Two examples are used to illustrate the above point. In the first example, we simulate a dataset of two latent classes that have one growth pattern but with

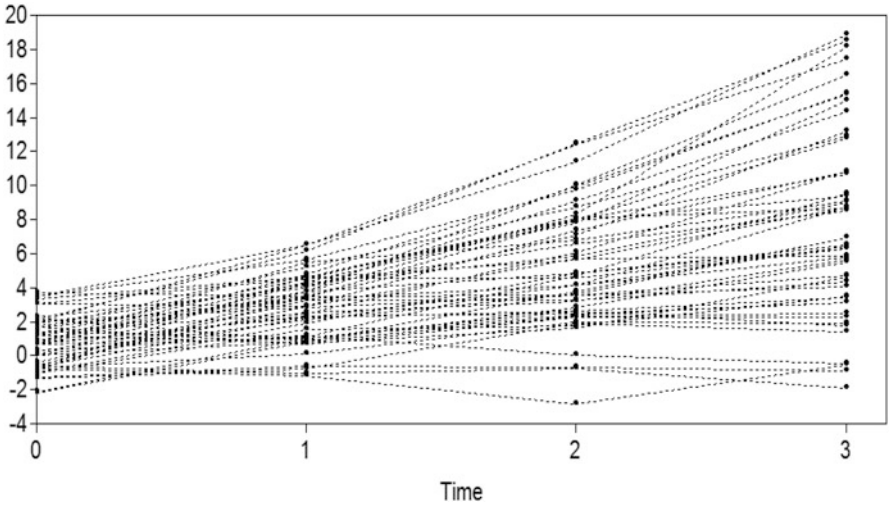


Fig. 2 An example of observed growth trajectories of 50 individuals randomly selected from the first simulated data with two latent classes but one growth pattern

differences in growth levels. That is, with one overall growth pattern (an increased trend), the individuals in each latent class may differ in level with respect to the growth trend, with some individuals having a higher growth level and some having a lower growth level. Figure 2 shows observed growth trajectories of 50 individuals randomly selected from this first simulated dataset. Figure 3 shows the estimated latent trajectory from MDS growth pattern analysis and Fig. 4 shows the estimated latent trajectories from GMM analysis. As can be seen from Fig. 3, the growth trajectory from MDS growth pattern analysis reflects the observed overall patterns with differences in mean or level removed. Thus, one growth profile is estimated to represent the prototypical growth pattern in the observed trends. On the other hand, the growth trajectories in Fig. 4 from GMM analysis indicate two growth classes, which can be expressed as

$$\text{class 1}: \ \widehat{y}_{it} = 1.24 + 1.68time + 0.74time^2$$
$$\text{class 2}: \ \widehat{y}_{it} = 1.10 + 1.67time + 0.03time^2$$

These two classes mainly differ with respect to mean level, although class 1 seems to have a faster growth acceleration.

In the second example, we simulate a dataset that has two latent classes, each with its own growth pattern—one is linear and another is quadratic. Accordingly, the grow trajectory differs not only in terms of level but also in terms of growth pattern. Figure 5 shows observed growth trajectories of 50 individuals randomly selected from this second simulated dataset. Figure 6 shows the estimated latent trajectory from MDS growth pattern analysis. Figure 7 shows the estimated latent trajectories

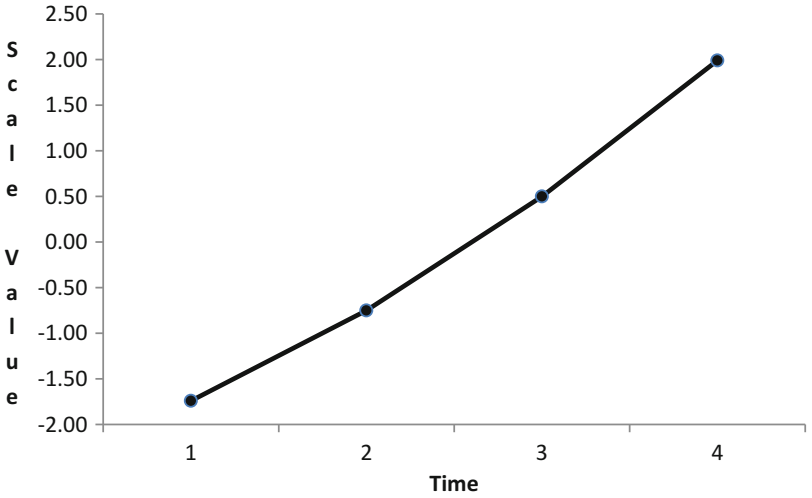


Fig. 3 Estimated growth pattern from MDS growth analysis based on the first simulated data

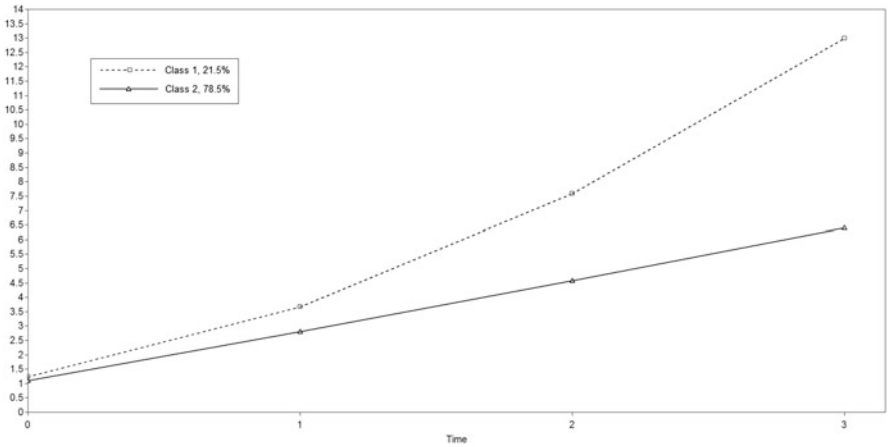


Fig. 4 Estimated mean growth trajectory of GMM analysis based on the first simulated data

from GMM analysis. As can be seen from Fig. 6, two MDS growth patterns reflect the observed growth patterns with differences in mean removed. That is, two growth profiles are estimated to indicate the prototypical growth patterns in the observed trends. Similarly, the growth trajectories in Fig. 7 from GMM analysis also indicate two growth classes, which can be expressed as

$$\text{class } 1: \ \widehat{y}_{it} = 0.96 + 0.56time + 0.48time^2$$
$$\text{class } 2: \ \widehat{y}_{it} = 2.29 + 3.69time - 1.28time^2$$

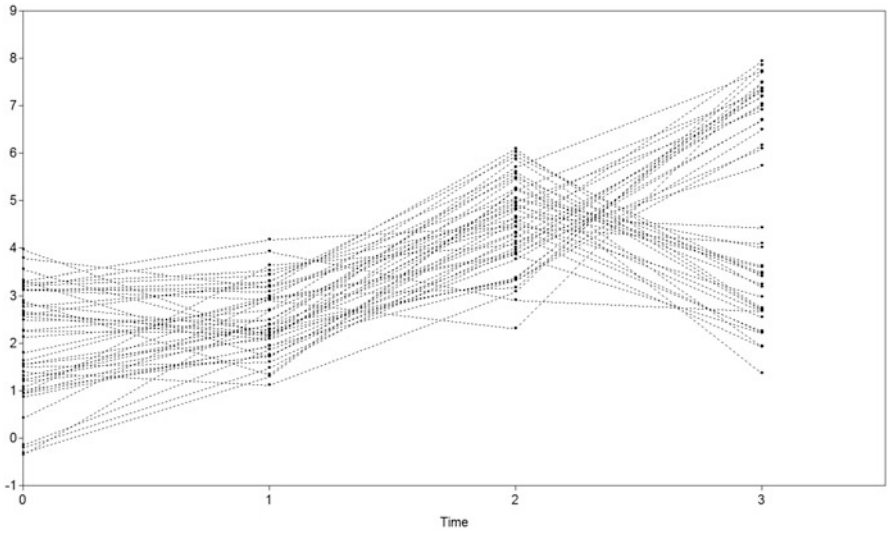


Fig. 5 Observed growth trajectories of 50 individuals randomly selected from the second simulated data with two classes, each with its own growth pattern (one is linear and another is quadratic)

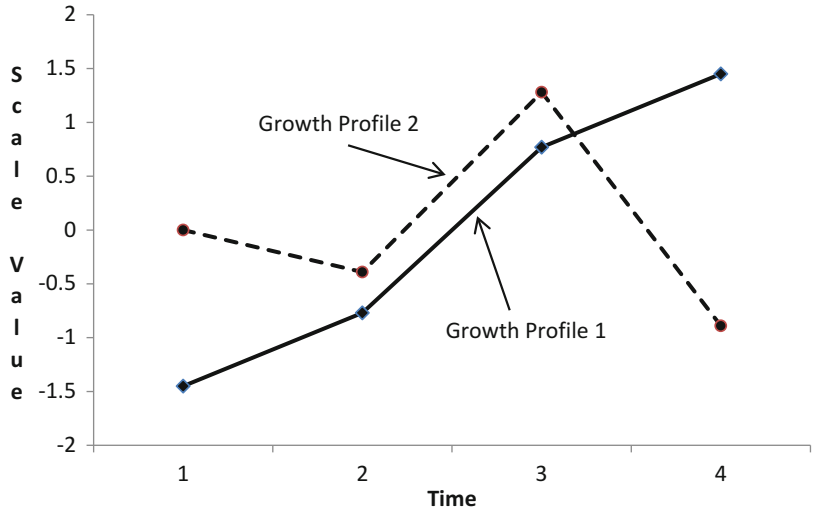


Fig. 6 Estimated growth pattern from MDS growth analysis based on the second simulated data

Thus, the key point is that both analytic approaches are correct in depicting the growth trends, but in a different way. One practical implication of this difference is that the growth trajectory from these two analytic approaches may manifest a different pattern as can be seen here, depending on the degree to which the observed patterns coincide with the level.

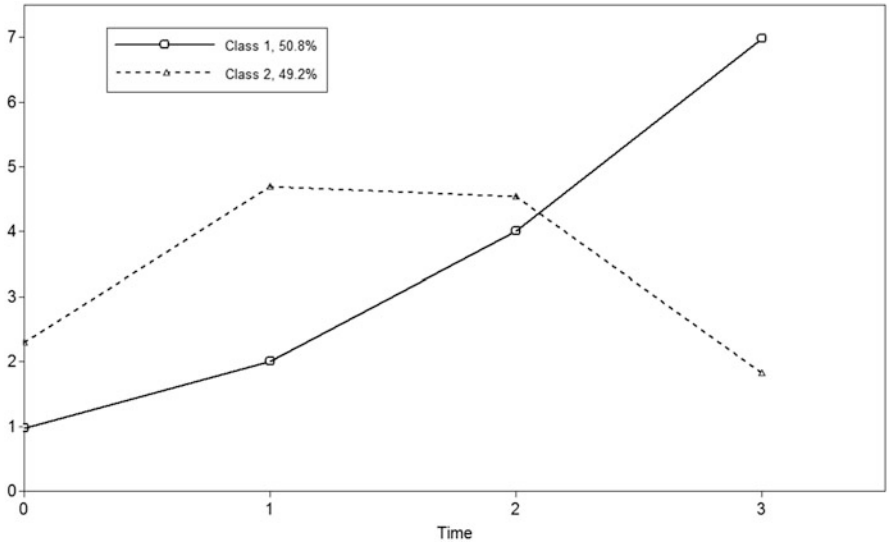


Fig. 7 Estimated mean growth trajectory of GMM analysis based on the second simulated data

Third, GMM and MDS differ with respect to distributional assumption, missing data, and incorporation of sampling weights. GMM requires assumption of normal distribution and can incorporate sampling weights and missing data in the estimation of growth patterns. The MDS approach does not require distributional assumption, nor does it incorporate sampling weights and missing data into the estimation of growth patterns. The practical implication of this difference is that we need to use either list-wise deletion for handling missing values or missing value estimation and imputation in MDS growth pattern analysis. Sampling weights may not have any effects on estimation of scale values.

In the following sections, using empirical data of mathematic achievement for children in the US we demonstrate the growth modeling of growth pattern using MDS analysis.

MDS Analysis of Math Growth

In this section, we examined mathematic achievement among children from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) study, which was following a nationally representative cohort of children from kindergarten and into the later grades (Denton, West, & Walston 2003; Princiotta, Flanagan, & Germino Hausken 2006). Details of the discussion of ECLS-K can be found in the references provided.

The present illustration examined two research issues. First, were there any distinct growth patterns of mathematic achievement across four waves of assessments?

Past work suggests that student reading performance tended to vary as a function of latent groups identified in the data (Ding & Navarro 2004; Friedman 1989). Second, would children with low parental educational level to be more likely than other children to belong to the group(s) that maintained the most negative mathematic achievement across time? This outcome would be consistent with findings that children with a low initial level of mathematic skills remained at the low level of mathematic skills (Ding & Davison 2005).

Data source. For this chapter, we analyzed the mathematic achievement of 9549 children with the complete data across the four waves of the ECLS-K data which were collected when they were in kindergarten (fall-1998 and spring-1999), during first grade (fall-2000 and spring-2001), during third grade (spring-2002), and during fifth grade (spring-2004). Among these children, 49 % were males and 51 % were females; 63 % were White, 12 % were Black, 13 % were Hispanic, 9 % were Asian, and 3 % were multi-race. Only 6 % of children had IEPs. Parents' education level included 21 % high school, 35 % some college, 25 % college, 11 % master, and 7 % doctoral degree.

Measure. The measures used for this illustration included a mathematic assessment, which contains items that assess basic skills such as counting, shapes, addition, fractions, area, and volume. Scale scores derived from item response theory (IRT) were used for the growth analysis. The score ranged from 7.89 to 150.94. Gender and parental educational level of children were used as covariates. For parental educational level, the score was coded from *1* (high school) to *4* (professional).

Analysis. We analyzed the data with MDS growth analysis. In order to make some comparison, we also analyzed the same data with GMM using SEM approach. The MDS growth pattern analysis was performed using SAS (SAS Institute Inc 2011) according to procedures described previously. The GMM analysis was performed using Mplus 7.0 (Muthen & Muthen 1998–2007). In both analyses, we used a two-stage approach to see if a two-stage GMM model could have increased the similarity of its solution to the MDS growth pattern analysis. Thus, we first performed a one- to three-class mixture model without any covariate. As suggested by Muthén and Muthén (2001), deciding on the number of trajectory classes was based on (1) an inspection of the Bayesian Information Criterion (BIC) and (2) the posterior probability of being assigned to a particular class, as well as the utility of the number of classes with respect to substantive considerations such as whether classes have fairly large numbers of assigned cases. We then followed the analysis by adding covariates to the selected growth mixture model.

Results. Table 1 shows the mean and standard deviation of IRT math scores across four data points. Figure 8 shows an increased growth pattern over time based on 150 children randomly selected from the sample. The MDS growth analysis was performed with models from one to three dimensions being fit to the data. An optimal number of dimension was determined using the *Stress* value and R^2. The *Stress* value for one-dimensional solution was 0.001 and R^2 was 1.00, suggesting that the model of one growth pattern fit the data well. The subsequent two- to four-

dimensional solution did not improve the model in any substantive way. Thus one growth pattern seemed to well approximate the data.

Table 2 shows the growth scale value of the MDS growth analysis. Figure 9 depicts one growth profile corresponding to the one-dimensional solution. As shown in Fig. 9, the growth pattern profile revealed an overall increased pattern of achievement over time. Table 3 shows the growth rate in terms of percentage for each time interval. The growth pattern indicated a pretty steady rate, identifying an overall linear trend with a slower growth rate from Grade 3 to Grade 5.

The correlation between initial score and the growth rate was 0.23 ($p < .05$). We performed the multiple regression analyses using gender and parental education level as predictors and the growth profile as dependent variable using the SAS surveyreg procedure, which allowed us to adjust the sample size using the sampling weight. The results of the analyses indicated that the growth profile was statistically significantly related to gender ($b = -1.30$, $p < .001$) and parental education level ($b = 1.63$, $p < .001$). It seemed that female children had a lower growth rate than did male children, and an increase in parental education level was related to the increased growth rate or helped to reduce the negative growth rate in math achievement.

Table 1 Mean and standard deviation of IRT mathematic scores across four waves of assessment

	M (SD)
Kindergarten	23.57(9.01)
Grade 1	58.87(16.66)
Grade 3	93.35(21.21)
Grade 5	114.67(21.06)

Note. Number in parenthesis is standard deviation

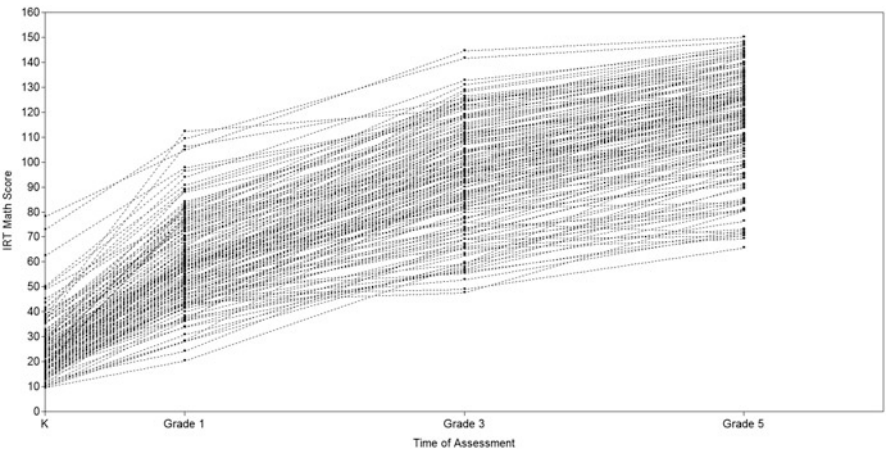


Fig. 8 Observed mathematic growth trajectories of 150 children randomly selected from the sample

Table 2 Growth scale values of one growth pattern corresponding to the one-dimensional solution

	Dim1
Kindergarten	−1.42
Grade 1	0.40
Grade 3	0.61
Grade 5	1.21

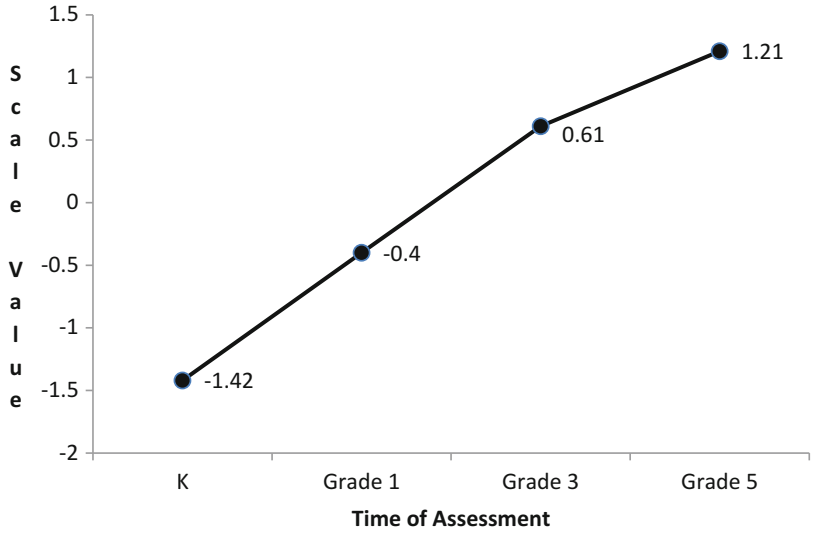


Fig. 9 MDS growth pattern of mathematic achievement

Table 3 Difference in growth scale value and percentage of change over time

Difference in scale values	
T2 − T1=	1.02
T3 − T2=	1.01
T4 − T3=	0.60
% of change	
T2 − T1=	0.39
T3 − T2=	0.38
T4 − T3=	0.23
Overall average	0.33

Note. T1 = Kindergarten; T2 = Grade 1; T3 = Grade 3; T4 = Grade 5

In order to see how the results from MDS growth pattern analysis differ from that of commonly used GMM analysis, we conducted the GMM analysis using the same data. Given the MDS finding of a linear growth pattern, we conducted a linear GMM analysis without any covariates. The sampling weight was incorporated into the analysis. The BIC values for these solutions were 276,671.74 and 273,683.12,

respectively, suggesting the two-class solution. In addition to the BIC, the posterior probabilities of classification based on the two-class solution showed good classification (.44 and .56 for classes 1 and 2, respectively). The Entropy value was .50 and .61 for one-class solution and two-class solution, respectively. Thus, it seemed that the two-class solution approximated the data better than one-class solution and was used for further analysis.

Next we performed a two-class linear growth model analysis, adding gender and parent education level as covariates. The sampling weight was also incorporated into the analysis. Figure 10 shows the estimated growth trajectory, which can be expressed as

$$\text{class 1}: \ \widehat{y}_{it} = 13.84 + 16.10 \text{ time}$$
$$\text{class 2}: \ \widehat{y}_{it} = 23.26 + 18.98 \text{time}$$

Class 1 had a lower initial score and a significant linear growth rate ($p < .01$), indicating an initial low developing group. On the other hand, Class 2 had a higher initial score and a significant linear growth rate ($p < .01$), suggesting an initial high developing group. But both groups had a similar growth rate over time.

For the initial low developing group, the correlation between the initial score and the linear rate was 0.44 ($p < .01$), suggesting that children's initial score was related to the growth rate. In addition, gender and parental education were significantly related to linear growth rate. Females had a slower growth rate than males ($b = -0.08$, $p < .001$). The higher parental education level, the growth rate was faster ($b = 0.12$, $p < .001$). For the initial high developing group, the correlation between the initial score and linear rate was .39 ($p < .001$), suggesting that children

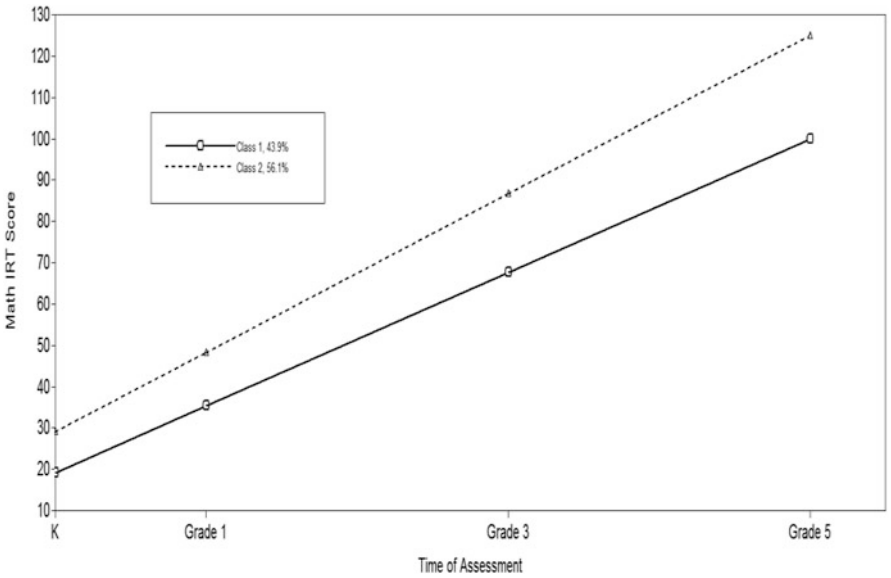


Fig. 10 Estimated mean growth trajectory of mathematic achievement

with a higher initial score had higher growth rate. Females also seemed to have a slower growth rate than males ($b = -0.17$, $p < .001$). The parental education level was also positively related to the growth rate ($b = 0.26$, $p < .001$).

Discussion

The purpose of this chapter was to illustrate MDS modeling of change in pattern, which may differ from change in level using GMM approach. Since the method of MDS growth pattern analysis is less well known, we discussed some major aspects of such an approach as well as some of its differences from GMM. Given that the GMM approach is more common in modeling growth, the significance of the chapter is that it discusses the MDS-based approach in the context of growth mixture modeling, showing that the MDS model can be a viable method for growth analysis that has been exclusively belonging to the realm of SEM technique. Researchers and practitioners should be aware of the utilities of MDS in modeling growth.

Two simulated datasets were used to illustrate MDS growth pattern analysis in the context of GMM. In addition, we conducted MDS growth mixture modeling using IRT scores of mathematic achievement among a group of kindergarteners from kindergarten to fifth grade. GMM analysis was also performed using the same data in order to demonstrate the differences in the results. Based on the results of these analyses, the following points were worth noting. First, the growth pattern from MDS analysis reflected the observed growth trends in the data with differences in level removed. When the observed growth mean trajectories did not significantly differ from patterns, as shown in Figs. 3 and 4, MDS growth pattern analysis seems to capture that pattern as a prototypical pattern in the data regardless of any mean differences in these trajectories. In contrast, GMM analysis captures the mean-level differences in growth trajectory resulting in two growth classes. However, when growth patterns significantly differ from the growth mean trajectories, MDS growth pattern analysis reflects these different growth patterns resulting in two growth profiles. This is similar to those from GMM analysis as shown in Figs. 6 and 7. Thus, the MDS approach is modeling the patterning of scores without level information of growth trajectories. With the same goal of identifying latent growth trajectories, these two approaches focus on different aspects of growth trajectory.

Second, the focus of the MDS analysis on pattern rather than level may account for differences in number of growth classes/types. Since class membership is assigned after the growth pattern is identified, there was only one growth class from MDS analysis. In contrast, GMM analysis takes the reverse approach. The number of latent classes is estimated first and then the growth trajectory is estimated with respect to each class, resulting in two growth classes.

Third, one may naturally ask which approach is better or best reflects the reality? The response can be considered from two angles. First, since GMM approach and MDS analysis are modeling different aspects of trajectory, we should focus on what information is more important or relevant to know. Second, as Cudeck and

Henly (2003) said, a realistic perspective of data modeling is that there are no true models to discover, and searching for the true number of latent classes is "pointless because there is no true number to find" (p. 381). Thus, "the issue of model misspecification is irrelevant in practical terms. The purpose of a mathematical model is to summarize data, to formalize the dynamics of a behavioral process, and to make predictions. All of this is scientifically valuable and can be accomplished with a carefully developed model, even though the model is false" (p. 378). In this regard, MDS analysis provides another perspective in understanding the nature of the change.

Given the previous discussion, one needs to realize that in selecting a growth modeling method, one should consider the desired information to be obtained from such an analysis. We hope that this illustration of MDS latent growth modeling approach can facilitate researchers in better understanding how MDS analysis can shed light on the growth trajectory of individual behaviors in relation to a GMM approach. Besides the pedagogical value of the chapter, we also hope that it can pique the interest of the readers to employ MDS growth analysis in their research.

References

Aber, M. S., & McArdle, J. J. (1991). Latent growth curve approach to modeling the development of competence. In M. Chandler & M. Chapman (Eds.), *Criteria for competence: Controversies in the conceptualization and assessment of children's abilities* (pp. 231–258). Mahwah, NJ: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. (2012). Auxiliary variables in mixture modeling: A 3-step approach using Mplus. statmodel.com.

Boscardin, C., Muthén, B., Francis, D., & Baker, E. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology, 100*, 192–208.

Collins, L. M., & Horn, J. L. (1991). *Best methods for the analysis of change: Recent advance, unanswered questions, future directions*. Washington, DC: American Psychological Association.

Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods, 8*(3), 378–383.

Davison, M. L., Davenport, E., Bielinski, J., Ding, S., Kuang, H., Li, F., et al. (1995). *Utilizing profile analysis via multidimensional scaling to ascertain patterns in course-taking: Mathematics and science course-taking patterns.* Paper presented at the AERA, San Francisco, CA.

Davison, M. L., Gasser, M., & Ding, S. (1996). Identifying major profile patterns in a population: An exploratory study of WAIS and GATB patterns. *Psychological Assessment, 8*, 26–31.

Davison, M. L., Kuang, H., & Kim, S. (1999). The structure of ability profile patterns: A multidimensional scaling perspective on the structure of intellect. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 187–204). Washington, DC: APA Books.

Denton, K., West, J., & Walston, J. (2003). *Reading—Young children's achievement and classroom experiences, NCES 2003–070*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Ding, C. S. (2005). Applications of multidimensional scaling profile analysis in developmental research: An example using adolescent irritability patterns. *International Journal of Behavioral Development, 29*(3), 185–196.

Ding, C. S. (2007a). Modeling growth data using multidimensional scaling profile analysis. *Quality & Quantity, 41*(6), 891–903.

Ding, C. S. (2007b). Studying growth heterogeneity with multidimensional scaling profile analysis. *International Journal of Behavioral Development, 31*(4), 347–356.

Ding, C. S., & Davison, M. L. (2005). A longitudinal study of math achievement gains for initially low achieving students. *Contemporary Educational Psychology, 30*, 81–95.

Ding, C. S., Davison, M. L., & Petersen, A. C. (2005). Multidimensional scaling analysis of growth and change. *Journal of Educational Measurement, 42*, 171–191.

Ding, C. S., & Navarro, V. (2004). An examination of student mathematics learning as assessed by SAT 9: A longitudinal look. *Studies in Education Evaluation, 30*, 237–253.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research, 59*, 185–213.

Hallquist, M. N., & Lenzenweger, M. F. (2012). Identifying latent trajectories of personality disorder symptom change: Growth mixture modeling in the longitudinal study of personality disorders. *Journal of Abnormal Psychology*. doi:10.1037/a0030060.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302–317.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1–27.

Muthen, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–587.

Muthen, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: American Psychological Association.

Muthen, L. K., & Muthen, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus: Statistical analysis with latent variables*. Los Angeles, CA: Muthén & Muthén.

Nagin, D. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods, 4*, 139–177.

Princiotta, D., Flanagan, K. D., & Germino Hausken, E. (2006). *Fifth grade: Findings from the fifth grade follow-up of the early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K)*. Washington, DC: National Center for Education Statistics.

Ram, N., & Grimm, K. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development, 33*(6), 565–576.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's guide*. Cary, NC: SAS Institute Inc.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*, 450–469.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods, 10*(4), 635–656. doi:10.1177/1094428106289397.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363–381.

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analyses of academic achievement. *Journal of Educational Measurement, 28*, 61–76.

A Nonparametric Approach to Modeling Cross-Section Dependence in Panel Data: Smart Regions in Germany

Harry Haupt and Joachim Schnurbus

Abstract In addition to intuitively plausible dependence structures in the time series dimension, in many applications it is reasonable to assume that there are contagion, spill-over, and repercussion effects among cross-sectional units. Modeling those structures in the systematic part of a panel regression requires both information on the underlying sources that drive the dependence and their respective range. The range allows one to define a neighborhood for each unit, a crucial concept for common methods in spatial statistics and econometrics. Furthermore, specification of a parametric regression function requires knowledge of the specific functional form of the spatial associations. However, lacking information on the sources usually leads to accepting misspecification and to including spatial error component or factor structures. As recent research reveals, the consequences of misspecification in both strategies are troubling in many cases. This paper proposes a data-driven nonparametric method for determining neighborhood as a first step. Second step nonparametric panel regressions have several benefits: (i) they allow one to test for misclassification of cross-sectional units to a wrong neighborhood in the first step; (ii) estimation is accomplished using data beyond the respective neighborhood, thus imposing less structure than parametric methods; (iii) neighborhood/location effects can be directly estimated in analogy to spatial statistics; (iv) no assumptions on functional form are required. The proposed method is illustrated with an empirical analysis of spatio-temporal patterns of high-skilled employees across German regions.

Introduction to Cross-Section Dependence

In economics and statistics the concept of cross-section dependence is often closely linked to problems of spatial associations—meaning dependence and/or heterogeneity—structures. This link is established by a plethora of economic applications where it is reasonable to assume that there are contagion, spill-over,

H. Haupt (✉) • J. Schnurbus
Department of Statistics, University of Passau, Passau, Germany
e-mail: harry.haupt@uni-passau.de

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer
Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_15

and repercussion effects among cross-sectional units representing different locations (e.g., sites, cities, regions, countries, etc.). A prominent example is the statistical modeling of prices in urban housing markets, as housing markets are generically spatial, house price functions are generically nonlinear (due to lumpy provision of characteristics at location i) and have non-constant higher moments, both driven by associations between unit i and units i' resident in the neighborhood $\mathcal{N}_i$ (see Haupt & Ng 2014). Further examples and discussions on statistical modeling of spatial association can be found among others in the textbooks of Anselin (1988), Bivand, Pebesma, and Gómez-Rubio (2013), Cressie (1993), Gaetan and Guyon (2010), LeSage and Pace (2009), Pace and LeSage (2010), and Ripley (1981).

Clearly, cross-sectional dependencies may exist on a more general, not exclusively spatial level. A prominent example are peer group effects influencing individual behavior of and social interactions between group members when we study cliques or crowds. Analogous considerations can be applied, for instance, to clusters or segments in entrepreneurial studies. Hence, in the following sections, concepts such as area, neighborhood, sites, position, are to be understood in a wider sense and may or may not refer to a spatial context.

Let us consider a set of neighborhoods $\mathcal{N} \subseteq \mathbb{R}^k$. In a spatial context we usually have $k = 2$ and consider spatial processes Y_n indexed on a spatial set $\mathcal{N}$, i.e., $n = (\text{longitude}, \text{latitude}) \in \mathcal{N}$. Referring to the peer group example, k may be quite large, representing the many characteristics influencing a persons's behavior. The k-dimensional space spanned by those characteristics is the analog to the two-dimensional space spanned by longitude and latitude. Depending on the application $\mathcal{N}$ may be continuous or discrete or mixed continuous-discrete, and the (vector valued) random process $Y = \{Y_n, n \in \mathcal{N}\}$ is observed at m neighborhoods $\{n_1, \ldots, n_m\} \subset \mathcal{N}$, where these neighborhoods may be random (e.g., point data sampled on m random cliques or geographical locations, where $m \leq N$), or fixed (e.g., network data sampled on m fixed jurisdictions of a region or country, where typically $m << N$). In the following we consider a cross-sectional unit i, located in neighborhood n_i.[1] Each neighborhood n_i is defined by a distance between i and every $i' \in n_i$. The distance between the cross-sectional units i and i' characterizes the dependence between the processes Y_{n_i} and $Y_{n_{i'}}$. As emphasized above, such a distance may, for example, refer to geographical, social, or economical dimensions. Intuitively, as the distance increases (beyond a certain point) we expect the dependence to decrease (or independence).

More formally, dependence between cross-sectional units i and i' is present whenever first or higher moments of the process Y_{n_i} depend on $Y_{n_{i'}}$, $i' \in n_i$, where n_i denotes the neighborhood containing cross-sectional unit i. In the following

[1] In the context of panel (or longitudinal data) a cross-sectional unit i at time point t is indexed by it. We consider a (vector valued) random process $Y = \{Y_{n_{it}}\}_{n \in \mathcal{N}}$. For the sake of simplicity, the following considerations refer to a given time period t, as we will not discuss forms of cross-section dependence varying in the time dimension. However, the proposed non-parametric approach allows for such cases, for example spatio-temporal processes by simultaneous smoothing over $n = (\text{longitude}, \text{latitude}, \text{time})$.

discussion we allow for stationary and non-stationary processes in the time dimension. In the cross-section dimension (Conley 1999) discusses weak conditions for stationarity and regularity of such processes:

> Strict stationarity is present, if the joint distribution of Y_n for any collection of neighborhoods $(n_1, \ldots, n_k)$ is invariant to location shifts such as $(n_1 + h, \ldots, n_k + h)$ for all $h \in \mathcal{N}$. If the mean $E(Y_n)$ is constant for all n and the covariance $Cov(Y_{n_i}, Y_{n_{i'}})$ is invariant for all $n_i, n_{i'}$, that is a function of $i - i'$ only (w.r.t. to the respective notion of neighborhood), the process is weakly stationary. If we assume that only the distance $|i-i'|$ matters, irrespective of direction, we consider the case of isotropic dependence. The counterpart is anisotropic dependence.
>
> The most important regularity conditions are assumptions on the nature of cross-sectional dependence. A prominent manifestation of such concepts is m-dependence: For $|i - i'| < m$ we assume dependence, for $|i - i'| \geq m$ independence. Most econometric applications are based on such a concept, where it seems natural to make an a priori assumption about m. Alternatives include concepts of asymptotic independence for $|i - i'| \to \infty$. Examples are mixing concepts, requiring only weak a priori assumptions on the mixing coefficients, while allowing for inference based on well-established results from asymptotic theory.

In general, neighborhoods can be seen as (usually non-disjoint) subsets of an area $\mathcal{A} \subseteq \mathbb{R}^k$. The identification of neighborhoods within an area usually follows one of two approaches:

(i) Subject matter a priori knowledge motivates assumptions on whether a cross-sectional unit belongs to one or more neighborhoods within a given area. For example defining neighbors by sharing borders, by belonging to a specific economic sector or a clique, or by fixing the set of subjects under investigation in an experiment. All examples require that neighborhoods do not form a disjoint partition of an area. Statistical modeling in this case usually either includes all cross-sectional units in one common model (area and neighborhood coincide) or allows for different sets of parameters across neighborhoods containing subsets of cross-sectional units. The latter, less restrictive approach allows to test for common parameters across neighborhoods.

(ii) Data-driven approaches are used, when a priori assumptions as stated above either cannot be motivated or have a history of not fitting the data in the applications at hand. Several methods exist for the clustering of cross-sectional units using suitable indicator variables. The identification of the latter, however, has a considerable impact on number and nature of neighborhoods and again requires subject matter knowledge.

Approach (i) is dominant in the fields of spatial econometrics and in many applications in spatial statistics.[2] Interestingly earlier work such as Conley (1999) call the neighborhood matrix with elements w_{ij}—reflecting whether units i and j share a neighborhood (or not)—"part of the data." Such strong beliefs in being able to pre-specify the complete neighborhood structure for all i are prone to generate data with misclassified cross-sectional units, and, as a consequence, have been replaced by a literature (see, for example, Kelejian & Prucha 2010) dealing with misspecification issues of neighborhood matrices. A problem of clustering methods in approach (ii) is that clusters in general form disjoint neighborhoods $n_1, \ldots, n_m$ and as such a partition of the area $\mathcal{A}$. This can be overcome by allowing some fuzziness in the outcome, for instance by using misclassification probabilities in model-based clustering (see Fraley & Raftery 1998; Handcock, Raftery, & Tantrum 2007).

Avoiding some of the problems of approaches (i) and (ii), an alternative two-step approach is proposed in this paper. The approach does neither rely on any parametric assumptions about the joint distribution of the underlying random processes nor on assumptions about both the neighborhood composition and (within-and-between) association structure. As such both steps are fully non-parametric in nature. In the first step neighborhoods are identified using a non-parametric approach allowing for data-driven quantity and composition of neighborhoods. A second-step non-parametric smoothing method allows one to estimate within- and between-neighborhood effects. A remedy of the problem that the resulting neighborhoods again are partitions can be seen in a typical property of nonparametric estimation method in step two: all cross-sectional units $i = 1, \ldots, N$ are used for estimation at any local position $Y = y_0$, hence adjacent neighborhoods are allowed to affect each other in a data-driven fashion.

The remainder of the paper is organized as follows: section "Regressions Under Cross-Section Dependence" introduces a general panel data regression framework, discusses some recent contributions to cross-section dependence in panels and proposes a simple but flexible non-parametric modeling framework. Section "Smart Regions in Germany" illustrates the proposed method using panel data on German regions.

Regressions Under Cross-Section Dependence

To start, let $(\Omega, \mathcal{F}, P)$ be a complete probability space and let $\{Y_n\}$, $n \in \mathcal{N}$, be an $\mathcal{F}$-measurable scalar random sequence. We consider regression models derived from the identity $Y_n = E(Y_n) + Y_n - E(Y_n)$ where the difference of the last two terms on the right-hand-side defines the centered error process $\{U_n\}$ by $U_n \stackrel{\text{def}}{=} Y_n - E(Y_n)$

[2]The respective goals in those two strands of literature may differ significantly as suggested by the respective discussions of theory and applications in Kauermann, Haupt, and Kaufmann (2012).

and the mean $E(Y_n)$ is modeled as $g(Z_n)$, where Z_n is a vector of explanatory variables and $g(.)$ is a fixed function. As we will not make any parametric assumptions about U_n we only consider semi- and nonparametric regression models, where for the former case $g(.)$ depends on a finite number of parameters collected in a vector θ_0. In particular in the spirit of Andrews (2005), we consider static models but allow for common (economical, sociological, psychological, technological, etc.) shocks across the cross-section dimension (i.e., individuals, cliques, networks, households, firms, industries, regions, etc.): In each time period t the regressors Z_n, the error U_n, and thus the response Y_n may be affected by common shocks λ_t that are captured by sigma-field $\mathcal{C}_t = \sigma(\lambda_t)$, where $\mathcal{C}_t \subset \mathcal{F}$.

In the following sections we employ a classical longitudinal regression framework for response Y_{it} with cross-section index i and time index t, regression function $g(Z_{it})$, and additive error components U_{it}: For $1 \leq i \leq N$, $1 \leq t \leq T$ let

$$Y_{it} = g(Z_{it}) + U_{it}, \tag{1}$$

and (time series as well as) cross-section associations (dependence, heterogeneity) may exist: (first or higher) moments of the response Y_{it} may depend on $Y_{i't}$, $i' \in \mathcal{N}_i$. In a general setting, the covariates may include contemporary and past values of exogenous variables, $X_{it}, X_{it-1}, X_{it-2}, \ldots$, and $Y_{i't}$, for all $i' \in \mathcal{N}_i$. All examples discussed in the following refer to sampling from fixed neighborhoods, such as jurisdictions.

Written in the usual compact notation, the equations in (1) can be written as $Y = g(Z) + U$. In a spatial context, $g(Z)$ may be a simple function of longitude and latitude, for example linear in parameters θ_0, i.e., $g(Z) = Z'\theta_0$. For a sampling scheme based on fixed coordinates as in the example studied in section "Smart Regions in Germany", Z is a non-random vector and $Var(Y) = Var(U) = \Sigma$, where the latter is a non-scalar covariance matrix due to potential spatial associations. Many texts in econometrics then discuss the ordinary least squares (OLS) estimator $\hat{\theta}$ of θ_0 and its alleged property to be unbiased though it neglects the true structure of the covariance matrix. As Spanos (1986) convincingly argues this belief is ill-founded, the respective regression is misspecified as it does not reflect the spatial associations in the systematic part of the regression, and hence the OLS estimator is biased and inconsistent (except for some very specific special cases).

The insights of Spanos (1986) on static versus dynamic regression modeling in the time-series context suggest the following on more general grounds: Whenever cross-sectional (and time-series) dependence is present in model (1), an encompassing model must be in the form of a (nonlinear) stochastic difference equation

$$Y_{it} = h(Y_{i't}, X_{it}, X_{it-1}, X_{it-2}, \ldots, U_{it}, U_{it-1}, U_{it-2}, \ldots, U_{i't}, U_{i't-1}, U_{i't-2}, \ldots),$$

where both the homogeneous and the inhomogeneous parts must allow for potential cross-sectional association structures. As a consequence, any approach neglecting such structures, misspecifying $h(.)$ or the covariance structure is prone to be biased

and inconsistent and suffers from unreliable estimates of precision, and as such from non-interpretability of (economically, psychologically, sociologically, etc.) relevant model parameters or effects.

The next section discusses some existing approaches to identifying the (potentially interrelated) sources of associations between i and i', $i' \in \mathcal{N}_i$. For example,

Quality proximity: (construction, political, etc.) era, (urban, cultural, etc.) development history,
Locational proximity: spatial distribution of (dis)-amenities,
Physical proximity (to jobs, virtual/real friends),
Sociological proximity ("keeping up with the Joneses").

Note that appropriate models should allow for "direct" effects, but also for the possibility that the mentioned proximities induce "indirect" (multiplicator or repercussion) effects. From a statistical point of view this leads to questions beyond economic, psychologic, sociologic, etc. motivation: How to restrict association structures—i.e., define $\mathcal{N}_i$, how to specify the regression function, its arguments and the error component?

Existing Approaches to Modeling Cross-Section Dependence

In the previous section we made the strong point that association structures should enter in the (correctly specified) systematic part of the regression. Besides modern methods of data-driven approaches to model the regression function, however, a prominent assumption in models such as (1) establishes a richer error structure. One example is to consider a two-way-error components structure

$$U_{it} = \mu_t + \eta_i + V_{it}, \tag{2}$$

where μ_t and η_i are unknown, smooth functions depending only on time index t and cross-section index i, respectively, and V_{it} is an idiosyncratic error-component. The spatial econometrics and statistics literature usually deals with modeling spatial associations via the functions g and η.

The literature basically can be grouped into two streams:

Spatial approach: A huge number of works across different disciplines try to implement (predominantly spatial) association structures in the systematic (and error) component under various degrees of structural assumptions. A common theme of this literature is that the true nature of all cross-section associations in (1) and (2) has to be specified a priori. A recent review of Kauermann et al. (2012) contrasts and applies those assumptions, model philosophies, specification strategies, and corresponding modeling goals and interpretations for the strands of spatial econometrics and spatial statistics.
Interestingly, though specific guidelines from theory and previous empirical analyses are lacking or non-existent, assumptions are strong, quite unnaturally,

about "neighborhood" composition of $\mathcal{N}_i$ for all i, as basically one parameter remains to be estimated. As has been stressed by Kauermann et al. (2012), spatial econometrics tries to estimate the "strength" of the relationship between y_{it} and $y_{i't}$, $i' \in \mathcal{N}_i$. For example, estimation of (and subsequent inference on) the parameter ρ_0 in the spatial autoregressive (SAR) model

$$Y = \rho_0 WY + Z'\theta_0 + U, \tag{3}$$

where row (it) of the equation system $Y = \rho_0 WY$ is given by $Y_{it} = \rho_0 \sum_{i' \in \mathcal{N}_i} w_{ii'} Y_{i't}$, a so-called spatial lag structure. In contrast, though relying on the same assumptions about the neighborhood structure given by the matrix $W = (w_{ii'})$, spatial statistical models include spatial random effects S_i in order to be able to visualize and predict the spatial patterns represented. The literature knows many, basically linear variants and extensions of the SAR model: for example, spatial ARMA or Durbin models, nesting the SAR and assuming a spatial lag assumption and parameter λ_0 for the U_{it}. Then, instead of (1) and (2) we consider (3) together with

$$U = \lambda_0 WU + \mu + \eta + V \tag{4}$$

and, as detailed in Elhorst (2010), interest lies in the statistical (and only lately economical) significance of ρ_0 and λ_0, and model selection via restrictions on these parameters.

Latent factor approach: Another strand of literature avoids any assumptions on cross-sectional associations in the systematic part and focusses on modeling (2), assumed to be due to misspecification of the systematic part of the regression. Model assumptions, applications and instructive surveys can be found in Conley (1999), Conley and Topa (2002), and Sarafidis and Wansbeek (2012).

Starting point is the model

$$Y_{it} = Z'_{it}\theta_0 + \eta_i + U_{it}, \tag{5}$$

where, for some t and some $i \neq j$, we may have

$$Cov(U_{it}, U_{jt}) \neq 0, \tag{6}$$

due to model misspecification of Eq. (5). In the so-called factor structure approach it is assumed that

$$U_{it} = \lambda'_i \psi_t + V_{it},$$

with latent factors ψ_t and loadings λ_i, and, again V_{it} is an idiosyncratic error-component. As the number of pairs (ij) with property (6) increases with the number of cross-sectional units N, the recent literature in this field analyzes statistics for $N \to \infty$.

On the positive side the latter approaches seems to be more realistic, as the former suffers from the well-known impossibility of finding the data generating process. On the negative side, extending the error structure of misspecified models seems a bit like medicating the dead. It is fair to say, though, that our excessively brief treatment is little more than a tunnel vision and the interested reader is referred to the beautiful survey of Sarafidis and Wansbeek (2012).

A Relative Similarity Approach to Modeling Cross-Section Dependence

Recently Kuersteiner and Prucha (2013) studied panel data-based approaches with a dynamic factor structure such as Phillips and Sul (2007, 2009), but extended those ideas by allowing for cross-sectional interactions in both systematic and error components of linear (in parameters) panel data regression models. As mentioned above our approach makes use of the idea of Andrews (2005) and Phillips and Sul (2007, 2009) that cross-section dependence is due to common shocks, leading to similar time trajectories of neighborhoods (clubs, cliques, etc.) over time, and hence is denoted as relative similarity approach. It differs from Kuersteiner and Prucha (2013), on the one hand, by not allowing for interactions in the error components, while, on the other hand, avoiding any parametric assumptions in the systematic component.

The relative similarity approach: The approach basically consists of two nonparametric steps. Step one is basically the approach of Phillips and Sul (2007, 2009) for panel data-driven identification of disjunct neighborhoods, step two is a fully nonparametric regression analysis.
The approach does not rely on a priori assumptions on the number (such as in confirmatory clustering) and covariance structure (such as in model-based clustering) of neighborhoods. Each neighborhood contains cross-sectional units with similar trajectories over time. The algorithm of Phillips and Sul (2007, 2009) relates these trajectories to each other. It is based on a sequence of one-sided t-tests of the null hypothesis $\delta_1 \geq 0$ in the auxiliary time series regression

$$Y_t^* = \delta_0 + \delta_1 \log(t) + \epsilon_t,$$

with response $Y_t^* = \log\left(\sum_{i^*=1}^{n^*}(H_{i^*0}-1)^2/\sum_{i^*=1}^{n^*}(H_{i^*t}-1)^2\right) - 2\log(\log(t))$, based on the relative transition path H_{i^*t} (over n^* observations) derived from a selection variable z, i.e., $H_{i^*t} = Z_{i^*t}/\overline{Z_{i^*t}}$. The null $\delta_1 \geq 0$ implies similar relative transition paths (of the considered cross-sectional units) and hence a joint convergence behavior w.r.t. the selection variable z. The algorithm proceeds until every observation is either part of a convergence club (with homogeneous

convergence behavior within the club) or of a remainder group, denoted as divergence group. In essence step one allows to generate a categorical variable η_i, indicating to which neighborhood cross-sectional unit i belongs.
In the second step, we estimate a nonlinear, fully nonparametric regression model

$$Y_{it} = g(Z_{it}^*) + U_{it}, \tag{7}$$

where $Z_{it}^* = (Z_{it}, \eta_i)$, by simultaneous smoothing over all dimensions of the arguments of the regression function $g(.)$, de facto allowing any form of nonlinearities involving the covariates in Z and η. We apply the nonparametric mixed kernel regression approach of Li and Racine (compare Li & Racine 2004, 2007; Racine & Li 2004). For the sake of illustration consider the minimization calculus for a local linear mixed kernel regression at covariate position Z_0^*, for the simple case of a single continuous covariate Z_{it} and discrete covariate η_i, respectively,

$$\min_{\tilde{\alpha}(Z_0^*),\tilde{\beta}(Z_0^*)} \sum_{i=1}^{n}\sum_{t=1}^{T}\Big(Y_{it} - \tilde{\alpha}(Z_0^*) - \tilde{\beta}(Z_0^*)\cdot(Z_{it} - Z_0)\Big)^2 \cdot K(Z_0^*, Z_{it}^*, b). \tag{8}$$

The estimated mean regression effect at covariate position Z_0^* is denoted by $\hat{\alpha}(Z_0^*)$ while the corresponding estimated first partial derivative w.r.t. the continuous covariate Z_{it} is denoted by $\hat{\beta}(Z_0^*)$. All observations are weighted by the generalized product kernel function $K(Z_0^*, Z_{it}^*, b)$, the product of the weight functions (i.e., kernels) of the covariates: Continuous covariates Z are weighted by a second order Gaussian kernel

$$k_Z(Z_0, Z_{it}, b_Z) = \frac{1}{b_Z}\phi\left(\frac{Z_{it} - Z_0}{b_Z}\right), \tag{9}$$

where $\phi(.)$ is the standard normal density and the smoothing parameter $b_Z \in]0, \infty[$. Unordered categorical covariates η are weighted by

$$k_\eta(\eta_0, \eta_{it}, b_\eta) = \begin{cases} 1 & \text{for} \quad \eta_{it} = \eta_0, \\ b_\eta & \text{for} \quad \eta_{it} \neq \eta_0, \end{cases} \tag{10}$$

with smoothing parameter $b_\eta \in [0, 1]$ (see Li & Racine 2004). The interpretation of parameter b is detailed in the empirical analysis in section "Misspecification of Parametric Functional Form". In a mixed covariate context, data-driven estimation of b is required prior to the kernel regression estimation. We estimate b by least-squares cross-validation (see Li & Racine 2007, Chap. 4).
The flexibility of approach (7) has the following merits: First, any form of relevant moderator effect is considered in a data-driven way. Second, it alleviates the problem of misspecification as it enables local approximations of omitted relevant variables. Third, for the local smoothing estimation inside a given

Table 1 Summary statistics of the continuous variables

Variable	Minimum	1. Quartile	2. Quartile	3. Quartile	Maximum	Mean	Std.-Dev.
grschool	−0.208	0.136	0.219	0.289	0.801	0.205	0.149
$\texttt{school}_0$	0.019	0.038	0.054	0.080	0.222	0.063	0.033
$\log(\texttt{school}_0)$	−3.940	−3.258	−2.915	−2.526	−1.504	−2.886	0.482

neighborhood, not only information of the other club members are used, but of all observations. Fourth, the degree of smoothing in each dimension allows to empirically assess the statistical significance of all covariates, including the neighborhood structure.

Smart Regions in Germany

High-skilled employees[3] (HSE) are the basis for developing new technologies and economic growth. The basis of our empirical analyses is the share of HSE in a region, the larger the former, the smarter the latter. From an economic point of view region-specific shares of highly educated employees can be used as a proxy for high-skilled labor. Lumpy provision of high-skilled labor across German regions may slow-down growth and increase already existing gaps in innovation and productivity. It is thus of obvious interest to study the spatial distribution and spatio-temporal diffusion of high-skilled labor and develop statistical methods to study existence and patterns of eventually occurring convergence and divergence processes.

The dependent variable in our model is the growth rate

$$\texttt{grschool}_i \stackrel{\text{def}}{=} \log(\texttt{school}_{i,2005}) - \log(\texttt{school}_{i,1996}),$$

where $\texttt{school}_{i,t}$ represents the share of HSE in region i ($i = 1, \ldots, 439$) (as a place-of-work) and year t ($t = 1996, \ldots, 2005$). Adapting the approach of Barro and Sala-i Martin (1992) our analysis is based on the unconditional β-convergence, where the key explanatory variable is $\texttt{school0}_i \stackrel{\text{def}}{=} \texttt{school}_{i,1996}$, the share of HSE in region i in the year 1996. Descriptive details of these variables are displayed in Table 1.

Figure 1 provides a first impression of the spatial distribution of the key variables grschool and $\texttt{school}_0$. Both maps reveal obvious patterns due to the former separation of Germany into Federal Republic of Germany and German Democratic Republic, hereafter denoted as west and east. To reflect this structural information, the binary variable west—which is equal to one for west regions and zero

[3]Employees liable for social security insurance, who have at least 11 years of schooling and a degree.

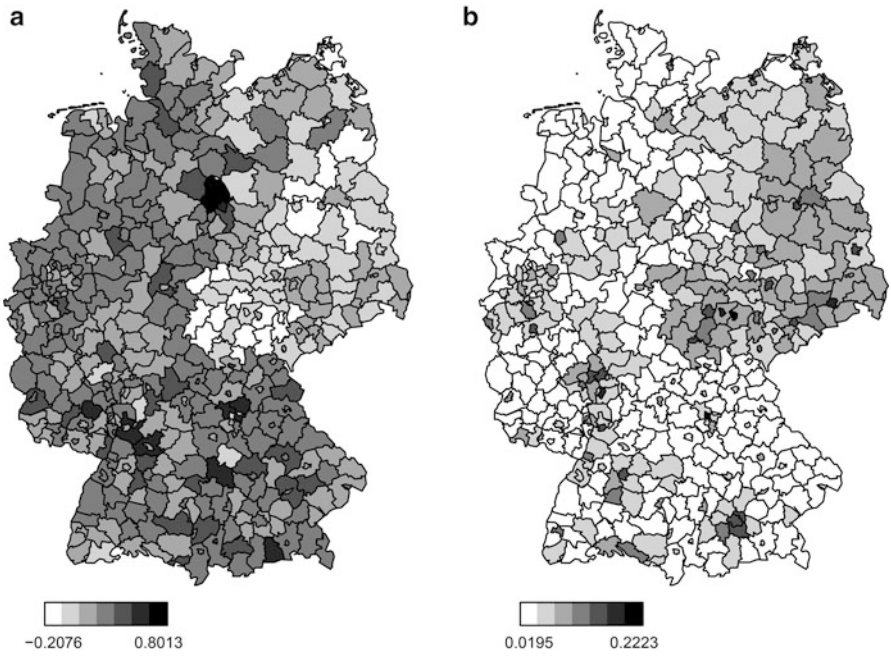


Fig. 1 Spatial maps for grschool and $\texttt{school}_0$. (**a**) grschool. (**b**) $\texttt{school}_0$

for east regions—is included in the subsequent analyses. Note that we primarily consider west as a political variable, although it is of obvious economic and, as a consequence, spatial—due to spill-over effects—relevance, too. Interestingly, the share of HSE in east regions in 1996 seems to be somewhat higher on average compared to the majority of west regions. In sharp contrast the growth-rate (between 1996 and 2005) is higher on average for most of the west regions, whereas some of the east regions even experienced negative growth-rates. This phenomenon, often denoted as the post-reunion brain-drain, is obviously still in progress many years after the official reunion in 1990.

Our baseline model[4] allows for west-east-specific convergence parameters,

$$\begin{aligned}\texttt{grschool}_i = {} & \alpha_1 \texttt{west}_i + \alpha_2(1 - \texttt{west}_i) + \beta_1 \log(\texttt{school0}_i)\texttt{west}_i \\ & + \beta_2 \log(\texttt{school0}_i)(1 - \texttt{west}_i) + \varepsilon_i, \end{aligned} \tag{11}$$

[4]Note that the estimation of Eq. (11) is based on cross-section data, where only information in the initial and final time period is employed. The a priori selection of $t = 0$ and $t = T$, respectively, may have a crucial impact on the outcome. We will not discuss such sources of non-robustness in this study.

Table 2 Regression output for Eq. (11)

	Estimate	Std. error	t value	Pr(> |t|)
`west`	0.08816	0.04038	2.183	0.0296
1-`west`	−0.19617	0.08016	−2.447	0.0148
`west:log(school0)`	−0.05825	0.01318	−4.421	0.0000
(1-`west`):`log(school0)`	−0.09214	0.03233	−2.850	0.0046

$PR^2 = 0.504$, AIC $= -725.31$, SIC $= -704.90$

and β-convergence of west regions is assumed to be present if $\beta_1 < 0$ (an analogous interpretation for east regions applies to β_2).

OLS estimation results for the baseline convergence regression (11) are displayed in Table 2. As $\hat{\beta}_2 < \hat{\beta}_1$ skill concentration differences across the regions seem to decrease as regions with a lower concentration of HSE (usually west) increase their concentration faster than regions with a higher concentration (usually east). Thus, the results may be interpreted as slightly suggestive in favor of converging shares of HSE over all administrative regions. We will not stress these preliminary results further, as the baseline model obviously suffers from lack of economic content and consequently various sources of misspecification (indicated by a battery of tests). For this reason we also do not report adjusted standard errors here. Given this disclaimer, the convergence coefficient is significantly negative for both parts of Germany and the fit, measured as squared correlation of observed and fitted response values (PR^2), is moderate at about 50 %.

Following the main contributions of among others Barro and Sala-i Martin (1992), Barro, Sala-i Martin, Blanchard, and Hall (1991), and Mankiw, Romer, and Weil (1992), a plethora of works appear addressing several strands of criticism confronting the baseline Solow model (see, e.g., Haupt & Petring 2011 for a recent survey). In the following exposition we pick up two main points of criticism.[5]

Equation (11) can be written compactly as $Y_i = X_i'\theta + \varepsilon_i$. However, as motivated in the following sections, it is safe to assume that the true conditional expectation of Y_i given all relevant explanatory variables is equal to $g(X_i, W_i)$, where g is an unknown, smooth function and W contains unobservable explanatory variables. Then the correctly specified model is given by $Y_i = g(X_i, W_i) + \xi_i$, where $\{\xi_i\}$ is an error process. When estimating the misspecified model (11), the error is quite complex as it equals $\varepsilon_i = g(X_i, W_i) - X_i'\theta + \xi_i$. The points of criticism we will consider here reflect two potential sources of the specification error $\Delta_i = g(X_i, W_i) - X_i'\theta$. First, neglected heterogeneity due to incorrectly assuming global convergence, while there may coexist clubs with homogeneous convergence behavior and/or a group of divergent regions. Second, misspecification due to neglected nonlinearities in the regression function. Empirical evidence on both issues is analyzed for the HSE in German regions.

[5]For the sake of brevity we will not discuss issues of neglected heterogeneity induced by spatial association due to spill-over and repercussion effects between German regions here.

Heterogeneity Due to Convergence Clubs and/or Divergence (Group)

One of the main points of criticism confronting the classical convergence regression (11) is that there are several forms of neglected heterogeneity causing invalid estimation results (e.g., Alfò, Trovato, & Waldmann 2008; Canarella & Pollard 2004; Ertur & Koch 2007; Haupt & Petring 2011; Mansanjala & Papageorgiou 2004).

In two seminal contributions Phillips and Sul (2007, 2009) build on the ideas of Durlauf and Quah (1999) and suggest that heterogeneity may occur due to individual effects and different technology levels. Considering these effects they propose a dynamic factor model based on the time trajectory $\{\texttt{school}_{i,t}\}_{t=0,\ldots,T}$ of each region i. Their convergence concept—which we label as "club convergence" hereafter—is based on the idea that convergence is assumed if all regions have the (approximately) same share of HSE in the final period T. Hence club convergence is based on panel data in contrast to β-convergence, the latter only relying on cross-sections for the initial and final periods 0 and T.

If there is no evidence (from a so-called log t regression test) in favor of global convergence,[6] Phillips and Sul (2007, 2009) introduce a clustering algorithm for identifying convergence clubs empirically. The idea of convergence clubs is that there are groups of countries with common convergence behavior. The algorithm proposes a classification of convergence clubs, while it is not possible to analyze convergence behavior on a club-level in the sense of Barro et al. (1991), Barro and Sala-i Martin (1992), and Mankiw et al. (1992). Applying an augmented form (see Haupt & Meier 2011) of the clubbing algorithm of Phillips and Sul (2007, 2009) to German regions yields a discrete covariate $\texttt{club}_i$ with 11 categories, i.e., 10 convergence clubs and a divergence group. Starting point for the clubbing algorithm is the log(school) order (descending) of the last period ($T = 2005$). For the sake of brevity, the algorithm is skipped, but Fig. 2 shows boxplots of the corresponding $\log(\text{school}_T)$-distribution for each club/group, while Fig. 3 shows the relative transition paths. In accordance with the work of Phillips and Sul (2007, 2009), the latter underline the meaning of club convergence.

We augment the classical β-convergence regression of Eq. (11) by m dummy variables $\texttt{club}_{i,j}$ representing the convergence clubs (divergence group), and estimate

$$\begin{aligned}\texttt{grschool}_i = {} & \alpha_0 \texttt{west}_i + \alpha_1 \log(\texttt{school0}_i) \cdot \texttt{west}_i \\ & + \sum_{j=1}^{m} \beta_j \texttt{club}_{i,j} + \sum_{j=1}^{m} \gamma_j \log(\texttt{school0}_i) \cdot \texttt{club}_{i,j} + \varepsilon_i. \quad (12)\end{aligned}$$

[6] In the present case of high-skilled employees in German regions the corresponding log t regression reveals no evidence in favor of global convergence on any reasonable significance level.

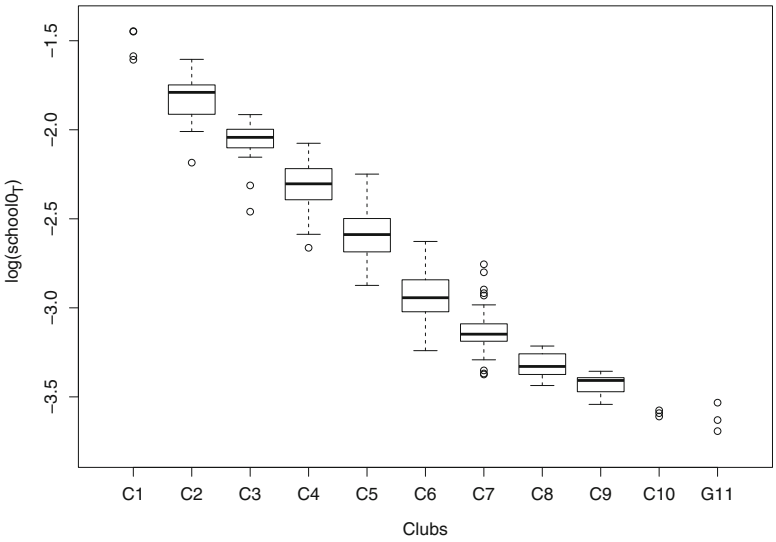


Fig. 2 Boxplots of $\log(\texttt{school}_T)$ for ten convergence clubs (C) and one divergence group (G)

Table 3 Occupation frequency for the category combinations of the discrete covariates

club	1	2	3	4	5	6	7	8	9	10	11	Total
East	1	5	6	18	52	24	6	0	0	0	0	112
West	3	14	18	41	87	86	48	13	11	3	3	327
Total	4	19	24	59	139	110	54	13	11	3	3	439

Table 3 contains the occupation frequencies of all club categories for both German regions. The clubs 1 and 10 as well as the divergence group are poorly occupied each having a total of less than five observations. For clubs 8, 9, and 10, as well as for divergence group 11 there are no observations for the east regions of Germany. In section "Misspecification of Parametric Functional Form" we will address potential issues of sparsely populated cells. The results for OLS estimation of Eq. (12) are displayed in Table 4. The estimated convergence coefficients are significantly negative for club 1–9, indicating β-convergence for each of these clubs. We do not find differences in the convergence behavior between west and east regions, as the coefficient of the interaction between $\texttt{west}_{i,j}$ and $\texttt{school0}_i$ is not significantly different from 0. The PR^2 is approximately 90 % and also the Akaike (AIC) and Schwarz information criteria (SIC) suggest a clear superiority in comparison with the baseline model (11).

The next natural question to ask is whether the latter model is also capable of capturing potential spatial patterns in the data. Table 5 contains the AIC and SIC for the baseline models of Eqs. (11) and (12) as well as the corresponding common spatial competitors. The model of Eq. (11) is clearly outperformed by a spatial error

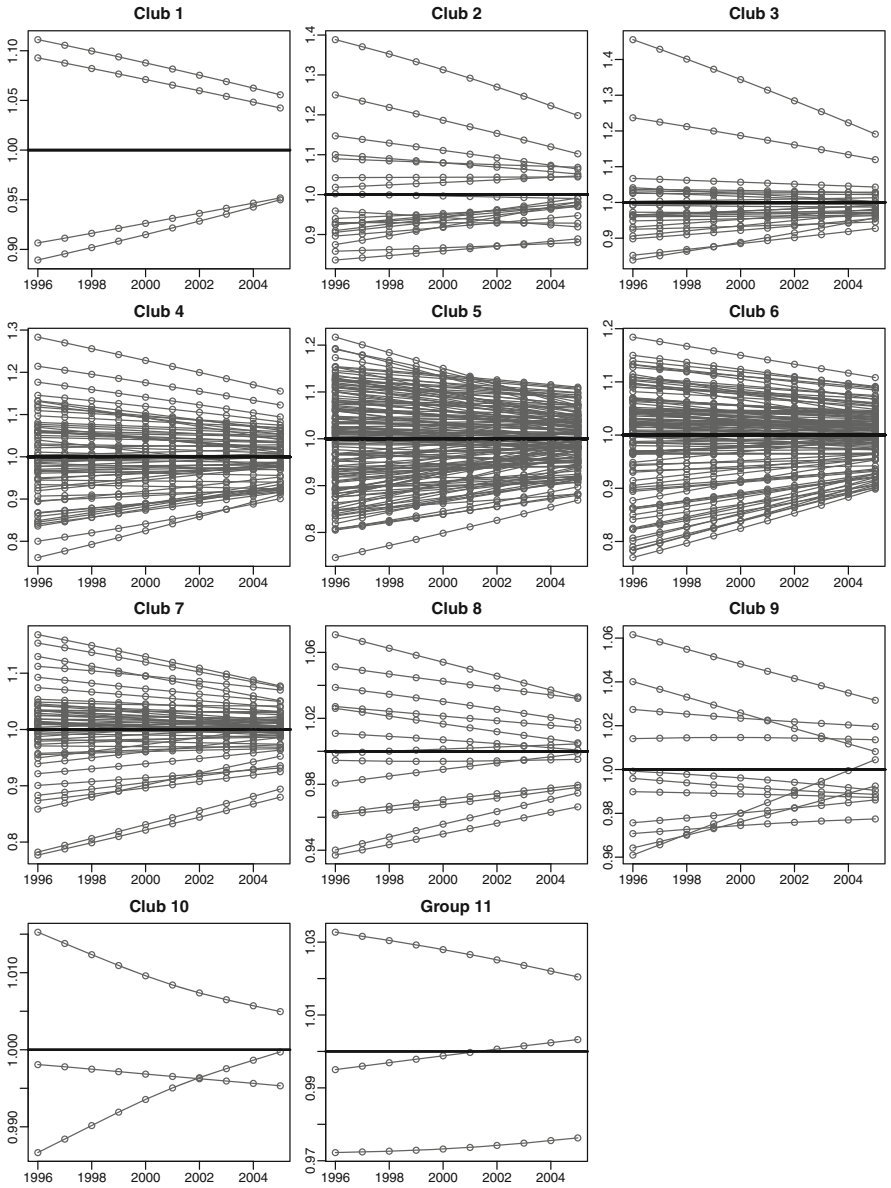


Fig. 3 Relative transition paths from time period 0 to T for convergence clubs and divergence group

model (in terms of AIC by every spatial competitor). Model (12) that is augmented by the club information performs best (i.e., better than every corresponding spatial competitor) in terms of SIC and is only slightly outperformed in terms of AIC by a

Table 4 Regression output for Eq. (12)

	Estimate	Std. error	t value	Pr(> |t|)
Club 1	−0.6634	0.2349	−2.82	0.0050
Club 2	−0.8543	0.0867	−9.86	0.0000
Club 3	−1.0384	0.0876	−11.86	0.0000
Club 4	−0.9782	0.0685	−14.28	0.0000
Club 5	−1.2090	0.0593	−20.39	0.0000
Club 6	−1.2928	0.0760	−17.01	0.0000
Club 7	−1.3670	0.1082	−12.63	0.0000
Club 8	−1.6421	0.3127	−5.25	0.0000
Club 9	−1.7918	0.4597	−3.90	0.0001
Club 10	−0.9817	1.5052	−0.65	0.5146
Group 11	−0.9182	1.3978	−0.66	0.5116
West	0.0525	0.0483	1.09	0.2772
Club 1:log(school0)	−0.4763	0.1379	−3.46	0.0006
Club 2:log(school0)	−0.5105	0.0417	−12.23	0.0000
Club 3:log(school0)	−0.5439	0.0377	−14.43	0.0000
Club 4:log(school0)	−0.4658	0.0279	−16.71	0.0000
Club 5:log(school0)	−0.4973	0.0231	−21.53	0.0000
Club 6:log(school0)	−0.4730	0.0273	−17.32	0.0000
Club 7:log(school0)	−0.4629	0.0354	−13.08	0.0000
Club 8:log(school0)	−0.5194	0.0890	−5.84	0.0000
Club 9:log(school0)	−0.5462	0.1274	−4.29	0.0000
Club 10:log(school0)	−0.2972	0.3975	−0.75	0.4551
Group11:log(school0)	−0.2923	0.3647	−0.80	0.4234
West:log(school0)	0.0071	0.0188	0.38	0.7063

$PR^2 = 0.896$, AIC $= -1373.89$, SIC $= -1270.78$

Table 5 AIC and SIC for baseline, spatial error, spatial lag, and spatial Durbin model with and without convergence clubs

Models without convergence clubs				
	Eq. (11)	Spatial error	Spatial lag	Spatial Durbin
AIC	−725.31	−734.54	−729.10	−735.36
SIC	−704.90	−710.04	−704.55	−698.60
Models including convergence clubs				
	Eq. (12)	Spatial error	Spatial lag	Spatial Durbin
AIC	−1373.89	−1372.01	−1374.70	−1370.73
SIC	−1270.78	−1265.81	−1268.51	−1170.59

spatial lag model. Table 6 reflects these findings as on a 5 % significance level, the null hypothesis of no (necessary) spatial association in model (11) is rejected w.r.t. all alternative hypotheses, while the equivalent tests for the model with additional club information do not yield a rejection of the null. We interpret this in a way that the inclusion of the club information sufficiently captures the spatial associations of the data such that no additional spatial effect has to be included.

Table 6 Results of LM-tests for spatial dependencies in the residuals of Eqs. (11) and (12)

Test results for Eq. (11)			
	Statistic	df	p.value
LM-test for spatial error	10.87	1.00	0.00
LM-test for spatial lag	4.30	1.00	0.04
LM-test for spatial error and spatial lag	17.50	2.00	0.00
Test results for Eq. (12)			
	Statistic	df	p.value
LM-test for spatial error	0.00	1.00	0.98
LM-test for spatial lag	3.20	1.00	0.07
LM-test for spatial error and spatial lag	4.26	2.00	0.12

Misspecification of Parametric Functional Form

Several authors identify neglected nonlinearities as a source of invalidity of classical convergence analysis (e.g., Haupt & Petring 2011; Henderson 2010; Kalaitzidakis, Mamuneas, Savvides, & Stengos 2001; Liu & Stengos 1999; Maasoumi, Li, & Racine 2007; Quah 1993, 1997). Following the proposal of Haupt and Meier (2011) we address this issue by employing a fully nonparametric approach. Keeping the notation of the previous sections, the sample counterpart of the nonparametric convergence regression model (7) is given by

$$\texttt{grschool}_i = f\left(\log(\texttt{school0}_i), \texttt{club}_i, \texttt{west}_i\right) + \varepsilon_i, \tag{13}$$

allowing for nonlinearities and interactions among all covariates within the regression function $f(\cdot)$. In the previous section the club membership is shown to sufficiently reflect the spatial association. Hence we include this information also as unordered discrete covariate `club` in the nonparametric regression. For the present problem we have a mix of continuous and discrete covariates. We apply the nonparametric mixed kernel regression approach of Li and Racine (compare Li & Racine 2004, 2007; Racine & Li 2004). Unsurprisingly Haupt and Petring (2011) find superior in-sample but also out-of-sample performance of this approach (compared to parametric regression function specifications) in the context of growth regressions for the original data of Mankiw et al. (1992).

The corresponding minimization calculus for a local linear mixed kernel regression is

$$\min_{\tilde{\alpha}(z_0^*),\tilde{\beta}(z_0^*)} \sum_{i=1}^{n} \Big(\texttt{grschool}_i - \tilde{\alpha}(z_0^*) - \tilde{\beta}(z_0^*) \cdot \left(\log(\texttt{school0}_i) - \log(\texttt{school0}_0)\right)\Big)^2 \cdot K(z_0^*, z_i^*, b), \tag{14}$$

where the vector $z_i^* = (\log(\texttt{school0}_i), \texttt{club}_i, \texttt{west}_i)$ contains the covariate values of cross-sectional unit i. Analogously, z_0^* refers to the covariate position $(\log(\texttt{school0}_0), \texttt{club}_0, \texttt{west}_0)'$ for local estimation of the regression function. The estimated local mean regression effect at this position is denoted by $\hat{\alpha}(z_0^*)$ while the corresponding estimated local first partial derivative w.r.t. $\log(\texttt{school0})$ is denoted by $\hat{\beta}(z_0^*)$. Observations are weighted by the generalized product kernel function $K(z_0^*, z_i^*, b)$, the product of the weight functions (i.e., kernels) of the three covariates. In a kernel estimation context the smoothing parameters are denoted as bandwidths:

> Small bandwidth values for $b_{\texttt{school0}}$ defined in (9) lead to reasonable weights only for observations i where $|\log(\texttt{school0}_i) - \log(\texttt{school0}_0)|$ is small, i.e., the number of HSE ($\texttt{school0}_i$) is close to $\texttt{school0}_0$. In contrast, large bandwidths yield almost equal weights for all observations, thus indicating an approximately linear relationship between $\log(\texttt{school0})$ and $\texttt{grschool}$.
> The bandwidths for the discrete kernel defined in (10) take values in $[0, 1]$, where a value of 0 means that the regression function is separately estimated for the observations of different covariate categories, i.e., the so-called frequency approach (see Li & Racine 2007, Chap. 3). For a bandwidth of 1 we obtain equal weights for the observations of all categories of the underlying covariate, which is thus irrelevant.

Table 7 displays the estimated bandwidth values for the covariates. The estimated bandwidth of the continuous covariate is about half as large as the standard deviation of $\log(\texttt{school0})$ (which is 0.4816). Thus the model allows for a considerable degree of nonlinearity w.r.t. this covariate while indicating that neglected nonlinearity may indeed be a problem for the present data. The estimated bandwidths of the discrete covariates are low. The bandwidth value of 0.1717 for the covariate $\texttt{west}$ indicates some smoothing of the underlying categorical information, meaning that the observations of East-Germany are also used for estimating the West-German regression relationship and vice versa, where the weight of about six ($\approx 1/0.1717$) observations of the "wrong" category offsets one observation of the corresponding "correct" category. Hence, we see that the nonparametric specification can at least partially deal with poorly occupied category combinations. A bandwidth value for the $\texttt{club}$ variable close to 0 indicates that the convergence clubs are well chosen. An estimated bandwidth of $1.5 \cdot 10^{-15}$ is extremely close to 0. Hence the probability of club-misclassification also seems to be negligible.

The nonparametric mixed kernel regression approach allows for an explicit test for parametric misspecification proposed by Hsiao, Li, and Racine (2007).

Table 7 Estimated bandwidths (using LSCV) for nonparametric mixed-kernel regression

Covariate	Kernel function	$b_k \in$	$\hat{b}_k$
$\log(\texttt{school0})$	of Eq. (9)	$]0, \infty[$	0.2867
$\texttt{club}$	of Eq. (10)	$[0, 1]$	≈ 0
$\texttt{west}$	of Eq. (10)	$[0, 1]$	0.1717

Table 8 *p*-values for test of Hsiao et al. (2007) for misspecification of parametric functional form

Spec.\bootstrap	iid	Wild
Eq. (11)	≈0	≈0
Eq. (12)	0.0476	0.1579

The *p*-values from applying the test are displayed in Table 8. Hsiao et al. (2007) suggest bootstrapping to obtain the null distribution of the test and Haupt, Schnurbus, & Tschernig (2010) show that the test can be quite sensitive w.r.t. the nonparametric configuration (including the bootstrap-type). Hence, we consider iid- as well as wild-bootstrapping to determine the distribution of the test under the null. The correct specification of Eq. (11) is clearly rejected for both bootstrap configurations, while the parametric specification including the club information is only rejected at a 5 %-level by the iid-bootstrap, indicating only minor nonlinearities in the residuals of this parametric specification. The PR^2 of the nonparametric estimation is 0.9015 and thus slightly higher than that of the OLS estimation of Eq. (12).

The estimated partial effects w.r.t. log(`school0`) for the nonparametric mixed kernel approach are obtained as $\hat{\beta}(z_0^*)$, compare Eq. (14). In principle these partial effects can be evaluated for a grid covering the range of log(`school0`)-values for each of the 22 category combinations of the discrete covariates (or more generally for any z_0^*). However, since the data is sampled from a lattice structure, we only compute partial effects for the given 439 observed covariate value combinations, compare Fig. 4. The vertical lines indicate the estimation uncertainty and correspond to pointwise asymptotic confidence intervals. For means of comparison we add the estimated partial effects from OLS estimation of Eq. (12), compare Table 4. A clear difference between parametric and nonparametric estimation is only visible for the clubs 1–6, the partial effects for the other clubs (and divergence group) seem to be reasonably estimated by the parametric specification of Table 4. For the clubs 2–6, the nonparametrically estimated partial effects are not constant, thus hinting at a nonlinear relationship between the underlying variables w.r.t. some of the clubs. Our finding that the estimated partial effects are not constant for some of the clubs corresponds to a moderate amount of nonlinearity in the relationship between log(`school0`) and `grschool` This weak or only local nonlinearity is in line with the results of the test of Hsiao et al. (2007) for Eq. (12), i.e., Eq. (12) seems only to suffer from minor misspecification issues (in terms of neglected additional spatial association).

Discussion

In our empirical exercise we investigate three potential sources of misspecification in convergence regressions: Omitted heterogeneity due to convergence clubs, due to spatial associations between neighboring regions, and due to potential nonlinearities in convergence behavior. As a first step to allow for heterogeneities induced by

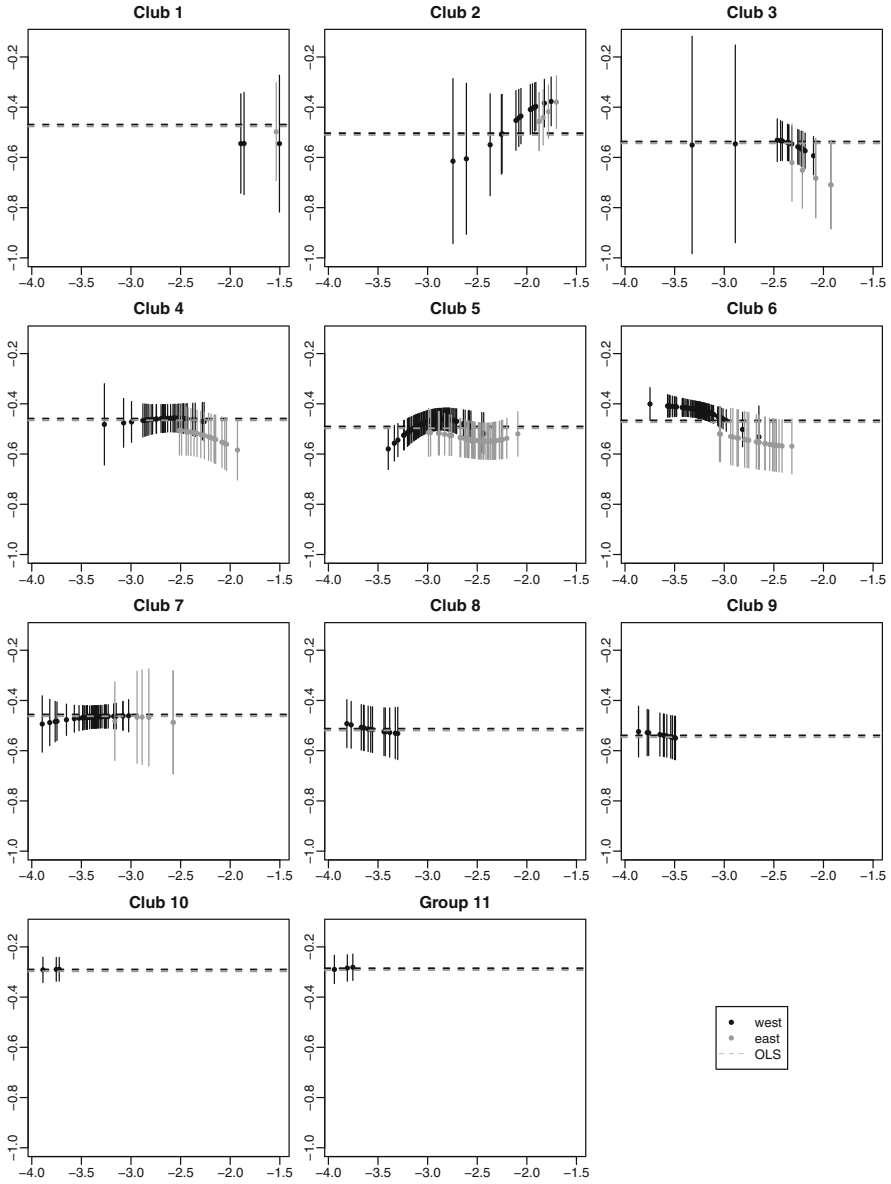


Fig. 4 Plots of $\log(\texttt{school}_0)$ (abscissa) and the estimated partial effects (w.r.t. $\log(\texttt{school}_0)$, ordinate) for the nonparametric regression model (points) of Eq. (13) and the parametric model (*horizontal dashed lines*) of Eq. (12)

non-global convergence processes we identify convergence (and divergence) clubs from a dynamic factor model using panel data. In the second step further potential heterogeneities in the extended model are assumed to be generated by spatial associations between regions in a cross-section model. As an encompassing step we test for parametric misspecification of the extended model and check the validity of the club structure generated from panel data to capture heterogeneity of convergence processes in a cross-section model. The employed nonparametric estimation method allows to investigate potential club-specific nonlinearities.

The proposed modeling framework is applied to analyze the growth convergence of HSE in German regions. Model selection results suggest that there is no clear empirical evidence in favor of including further spatial model components. The residual heterogeneity in classical models can be captured quite well by controlling for the club structure identified in the first step of our analysis. If, however, the club information is neglected, model selection criteria and tests suggest the existence of spatial association in the model. Tests for parametric misspecification and visual inspection of estimated partial effects reveal some but no clear evidence for nonlinearities. We stress that our findings do not suggest that there are no spatial externalities, spill-overs, or repercussion effects. However, it is possible to identify convergence (and divergence) club-level functionals that seem to be capable of controlling for parts of these effects—for the present data set, though equivalent results can be obtained for Penn World Table income data used by Ertur and Koch (2007)—while following a different economic motivation and implication.

On more general grounds the proposed method is applicable to regression problems with cross-section associations in panel data contexts covering one or more cross-sectional dimension. The first step of the method allows for data-driven identification of disjunct neighborhoods without any a priori assumptions on their number, size, or composition, nor on within and between covariance structures. The second step exploits the first-step information on neighborhoods and does not rely on parametric assumptions on a general nesting model, while it allows to estimate, test, and interpret all relevant effects in the usual fashion and beyond. In addition, the undesired concept of disjunct neighborhoods from the first step does not matter because of the fully multiplicative nature of nonparametric smoothing regression: Neighborhood-specific effects are estimated using information from the neighbors, but also from all other cross-sectional units, weighted in a data-driven fashion, respectively. The interplay of these properties renders the method a suitable candidate for addressing the problem that cross-section associations are not a nuisance but a relevant part of both the model and the story the data tries to tell us.

References

Alfò, M., Trovato, G., & Waldmann, R. (2008). Testing for country heterogeneity in growth models using a finite mixture approach. *Journal of Applied Econometrics, 23*, 487–514.

Andrews, D. (2005). Cross-section regression with common shocks. *Econometrica, 73*, 1551–1585.

Anselin, L. (1988). *Spatial econometrics*. Studies in operational regional science. Dordrecht: Kluwer Academic Publishers.

Barro, R. J., & Sala-i Martin, X. (1992). Convergence. *Journal of Political Economy, 100*(2), 223–251.

Barro, R. J., Sala-i Martin, X., Blanchard, O. J., & Hall, R. E. (1991). Convergence across states and regions. *Brookings Papers on Economic Activity, 1*, 107–182.

Bivand, R., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R*. New York: Springer.

Canarella, G., & Pollard, S. (2004). Parameter heterogeneity in the classical growth model: A quantile regression approach. *Journal of Economic Development, 29* 1–31.

Conley, T. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics, 92*, 1–45.

Conley, T., & Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics, 17*, 303–327.

Cressie, N. (1993). *Statistics for spatial data*. New York: Wiley.

Durlauf, S. N., & Quah, D. T. (1999). The new empirics of economic growth. In: J. B. Taylor & M. Woodford (Eds.), *Handbook of Macroeconomics* (Vol.1, Chap. 4, pp. 235–308). Amsterdam: Elsevier.

Elhorst, J. (2010). Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis, 5*(1), 9–28.

Ertur, C., & Koch, W. (2007). Growth, technological interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics, 22*(6), 1033–1062.

Fraley, C., & Raftery, A. (1998). How many clusters? which clustering methods? answers via model-based cluster analysis. *Computer Journal, 41*, 578–588.

Gaetan, C., & Guyon, X. (2010). *Spatial statistics and modeling*. New York: Springer.

Handcock, M., Raftery, A., & Tantrum, J. (2007). Model-based clustering for social networks (with discussion). *Journal of the Royal Statistical Society, Series A, 170*, 301–354.

Haupt, H., & Meier, V. (2011). Dealing with heterogeneity, nonlinearity and club misclassification in growth convergence: A nonparametric two-step approach. Working Papers from Bielefeld University, Institute of Mathematical Economics, No 455.

Haupt, H., & Ng, P. (2014). Smooth quantile smoothing spline estimation of urban house price surfaces under conditional price and spatial heterogeneity. Working Paper.

Haupt, H., & Petring, V. (2011). Assessing parametric misspecification and heterogeneity in growth regression. *Applied Economics Letters, 18*(4), 389–394.

Haupt, H., Schnurbus, J., & Tschernig, R. (2010). On nonparametric estimation of a hedonic price function. *Journal of Applied Econometrics, 5*, 894–901.

Henderson, D. (2010). A test for multimodality of regression derivatives with application to nonparametric growth regression. *Journal of Applied Econometrics, 25*(3), 458–480.

Hsiao, C., Li, Q., & Racine, J. S. (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics, 140*(2), 802–826.

Kalaitzidakis, P., Mamuneas, T., Savvides, A., & Stengos, T. (2001). Measures of human capital and nonlinearities in economic growth. *Journal of Economic Growth, 6*, 229–254.

Kauermann, G., Haupt, H., & Kaufmann, N. (2012). A hitchhiker's view on spatial statistics and spatial econometrics for lattice data. *Statistical Modelling, 12*, 419–440.

Kelejian, H., & Prucha, I. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics, 157*, 53–67.

Kuersteiner, G., & Prucha, I. (2013). Limit theory for panel data models with cross sectional dependence and sequential exogeneity. *Journal of Econometrics, 174*, 107–126.

LeSage, J., & Pace, K. (2009). *Introduction to spatial econometrics*. London: Taylor and Francis.

Li, Q., & Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica, 14*, 485–512.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton: Princeton University Press.

Liu, Z., & Stengos, T. (1999). Non-linearities in cross-country growth regressions: A semiparametric approach. *Journal of Applied Econometrics, 14*(5), 527–538.

Maasoumi, E., Li, Q., & Racine, J. (2007). Growth and convergence: A profile of distribution dynamics and mobility. *Journal of Econometrics, 136*, 483–508.

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics, 107*(2), 407–437.

Mansanjala, W., & Papageorgiou, C. (2004). The solow model with CES technology: Nonlinearities and parameter heterogeneity. *Journal of Applied Econometrics, 19*(2), 171–201.

Pace, R. K., & LeSage, J. (2010). Spatial econometrics. In: A. E. Gelfand, P. J. Diggle, M. Fuentes, & P. Guttorp (Eds.), *Handbook of Spatial Statistics* (pp. 245–262). Boca Raton: Chapman & Hall/CRC.

Phillips, P. C. B., & Sul, D. (2007). Modeling and econometric convergence tests. *Econometrica, 75*(6), 1771–1855.

Phillips, P. C. B., & Sul, D. (2009). Economic transition and growth. *Journal of Applied Econometrics, 24*(7), 1153–1185.

Quah, D. (1993). Empirical cross-section dynamics in economic growth. *European Economic Review, 37*, 426–434.

Quah, D. (1997). Empirics for growth and distribution: Stratification, polarization and convergence clubs. *Journal of Economic Growth, 2*, 27–59.

Racine, J. S., & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics, 119*(1), 99–130.

Ripley, B. D. (1981). *Spatial statistics*. New Jersey: Wiley.

Sarafidis, V., & Wansbeek, T. (2012). Cross-sectional dependence in panel data analysis. *Econometric Reviews, 31*, 483–531.

Spanos, A. (1986). *Statistical foundations of econometric modelling*. Cambridge: Cambridge University Press.

MANOVA Versus Mixed Models: Comparing Approaches to Modeling Within-Subject Dependence

Christof Schuster and Dirk Lubbe

Abstract For inferential purposes such as hypothesis testing or confidence interval calculations, analysis of repeated measures data needs to account for within-subject dependence of observations. Multivariate analysis of variance (MANOVA) is a suitable traditional technique for this purpose. It assumes an unconstrained within-subject covariance matrix and balanced data. However, the so-called mixed-model approach is a viable alternative to analyzing this type of data, because its underlying statistical assumptions are equivalent to the MANOVA model. While MANOVA is the classical approach, the mixed-model methodology, although by now implemented in all major statistical software packages, still is a relatively recent statistical development. The equivalence of both approaches to analyzing repeated measures data has frequently been noted in the literature. Nevertheless, in terms of test-statistics both approaches differ. While in large samples the test-statistics are essentially equivalent, their small sample behavior is not well known. In this article, we investigate by computer simulation the performance of several test-statistics calculated either from the MANOVA or the mixed-model approach for testing the interaction hypothesis with balanced data.

Introduction

An important aspect of experimental design is the control of variation in the dependent variable that is unrelated to the treatment. Among the causes resulting in this type of variation are (1) measurement error and (2) systematic influences. Examples of systematic influences are covariates and individual differences in repeated measures designs.

A well-known technique to eliminate or at least reduce this type of systematic variation is the so-called blocking of observations. A block of observations is made up of observations that are as similar as possible with respect to characteristics that are suspected to influence the outcome. If the observations of each block are

C. Schuster (✉) • D. Lubbe
University of Giessen, Giessen, Germany
e-mail: christof.schuster@psychol.uni-giessen.de

M. Stemmler et al. (eds.), *Dependent Data in Social Sciences Research*, Springer Proceedings in Mathematics & Statistics 145, DOI 10.1007/978-3-319-20585-4_16

randomly assigned to treatment, the treatment comparisons within each block are free of between block characteristics, thereby reducing between block variation from the treatment comparison. As a result, treatment comparisons are typically more powerful.

A standard example of blocking comes from agricultural research, where it is known that fertility of soil may vary considerably within fields. If a field is divided into several plots, then blocks usually comprise plots close to each other. In this way, natural fertility differences in soil can be controlled, inasmuch as they are removed from comparisons of treatments randomly assigned to the plots within each block.

Although the idea of blocking is also popular in the social sciences, its application often requires additional effort. If, for instance, naturally occurring intelligence differences in a sample of individuals are suspected to cause additional variation in the dependent variable of a learning experiment, then blocking of individuals with respect to intelligence would require pre-trial intelligence assessment. Because block size typically is fixed and equal to the number of treatments (randomized complete block design), two appointments per individual have to be arranged: one for pre-trial testing and one for treatment administration.

In an analysis of covariance no such pre-trial assessment of individuals is necessary because information about individual differences, e.g. intelligence, can be accounted for as a covariate at the data analysis stage. Therefore, in social sciences analysis of covariance is more popular than blocking to control for systematic dependent variable variation.

Nevertheless, there is one setting in which blocking in social science research is natural and does not require additional effort. If individuals can be repeatedly observed, then individuals can be considered as blocks. As a result, within-subject treatment comparisons are free of between-subject variation.

If the treatments are randomly assignment to subjects, then the statistical model for a randomized complete block design applies to the repeated measures design. Let Y_{jm} denote the response of the mth individual (block) to the jth treatment, then the model is

$$Y_{jm} = \mu + \beta_j + \pi_m + u_{jm},$$

where μ denotes the grand mean, β_j, $j = 1, \ldots, p$ is the treatment effect, π_m, $m = 1, \ldots, N$ is the random block effect, and u_{jm} is a random residual. Standard assumptions are: (1) $\pi_m \sim \mathrm{N}(0, \sigma_\pi^2)$, (2) $u_{jm} \sim \mathrm{N}(0, \sigma^2)$, and (3) independence of random terms. Because the hypotheses of contrasts pertaining to the β-effects are comparisons within the repeated observations, they are referred to as *within-subject* hypotheses.

Intuition suggests that within-subject observations are more similar than between-subject observations and the above model reflects this. If the repeated measures from the mth individual are collected in a vector $\boldsymbol{y}_m = (Y_{1m}, \ldots, Y_{pm})'$, then the covariance matrix of the within-subject association is $\mathrm{Var}(\boldsymbol{y}_m) = \boldsymbol{\Sigma}$. Assuming $p = 4$, the model for the randomized complete block design implies the covariance pattern

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 + \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma^2 + \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \sigma^2 + \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 & \sigma^2 + \sigma_\pi^2 \end{pmatrix},$$

often referred to as "compound symmetry." Compound symmetry is a simple structure in which only two parameters, σ^2 and σ_π^2, account for the association of the within-subject responses. Specifically, compound symmetry implies equal variances and equal covariances among within-subject observations. In addition, the covariance is necessarily non-negative.

Furthermore, if the individual vectors are collected in one overall $Np \times 1$ vector $\boldsymbol{y}$, where N is the sample size, then the covariance matrix of all observations, $\mathrm{Var}(\boldsymbol{y})$, has block-diagonal form. Thus, if the model for the randomized complete block design applies, the covariance matrix of all observations requires only two parameters. In this case, standard ANOVA procedures can be used to test for treatment differences.

In practice assuming compound symmetry can be justified if treatment assignment is random within individuals. However, without random assignment, such a simple covariance structure is questionable. In particular, if "time" is considered as the within-subject treatment, that is, the dependent variable is simply observed repeatedly after fixed time intervals, observations closer together in time are expected to be more strongly associated than observations further apart. This is often the case in social science studies in which subjects are repeatedly observed at fixed time points. In such research studies, subjects typically belong to one of several groups and comparisons between them are typically referred to as *between-subjects* hypotheses.

In the following sections, we first give an applied example using artificial data of a so-called split-plot design. We then explain three approaches that can be used to analyze this type of data. Specifically, we discuss (1) the univariate analysis, (2) the MANOVA analysis, and (3) the mixed-model approach. We then focus in particular on comparing the test-statistics from MANOVA and mixed-model approaches by computer simulation.

Example

A typical research setting of a repeated measures design is given by a two-way layout in which several groups of individuals are repeatedly observed over time. Thus, there is a between-subject treatment and a within-subject treatment. The between-subject treatment pertains to the groups and the within-subject treatment is time. This design often is referred to as a split-plot experiment, where there are whole-plot experimental units, belonging to one of several groups, and split-plot experimental units, corresponding to repeated observations of the within-subject factor.

Table 1 Artificial data set of repeatedly observed achievement scores of children belonging to one of three groups

Group	Child	Age group			
		1	2	3	4
1	1	48	50	50	47
	2	47	50	53	47
	3	47	48	47	47
	4	47	49	54	60
	5	52	49	49	52
2	6	55	57	60	63
	7	58	56	55	58
	8	48	50	53	58
	9	52	52	52	58
	10	51	52	57	54
3	11	53	64	61	62
	12	50	57	57	68
	13	55	61	61	63
	14	51	63	68	70
	15	48	53	54	61

In this design, the hypotheses refer to the presence of (1) group differences, (2) change over time, and (3) interaction (do the groups change differentially over time). In fact, in such a design, the most important research focus often is the interaction hypothesis.

Assume we have five children in each of three groups whose cognitive performance Y is measured and reported in a T-norm metric (mean of 50 and a standard deviation of 10) at four times. The data set analyzed below is given in Table 1.

The response profiles of the three groups are given in the left panel of Fig. 1. The profile of the first group shows almost no change over time while the second and third groups show considerable development. If the hypothesis of no interaction is true, then the response profiles over time are truly parallel as shown in the right panel. Thus the interaction hypothesis concerns the question of whether the deviations of the actual profiles from parallel profiles can be explained by sampling variation. If so, the interaction hypothesis will not be rejected and although there may be development over time as well as group differences, the different programs appear to be equally effective.

Approaches to Analyzing Repeated Measures

There are three well-known approaches to analyzing repeated measures data of the main-effects and interaction hypotheses. Let μ_{ij} denote the mean outcome of the ith group, $i = 1, \ldots, g$, observed at time j, $j = 1, \ldots, p$. In terms of these means, the three hypotheses are:

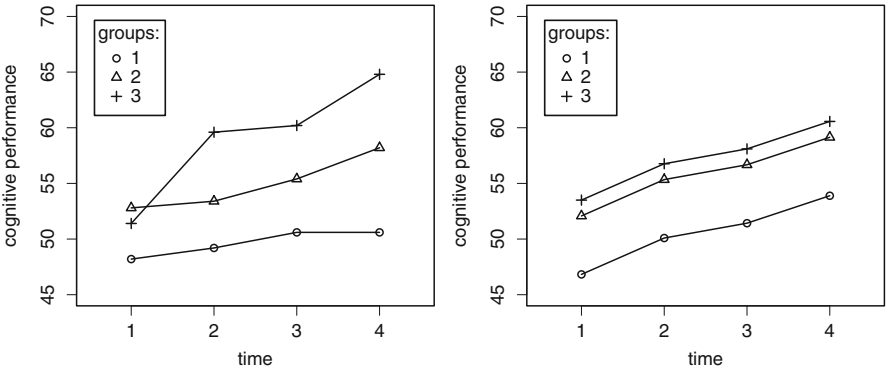


Fig. 1 Mean response profiles of three groups. *Left panel* profiles are produced using cell means. *Right panel* profiles are based on fitted cell means assuming the interaction null hypothesis to be true

$$\begin{aligned} \text{Interaction} &: \mu_{ij} - \mu_{i,j+1} - \mu_{i+1,j} + \mu_{i+1,j+1} = 0, \\ \text{Group main-effect} &: \mu_{i\cdot} = \mu_{i+1\cdot}, \\ \text{Time main-effect} &: \mu_{\cdot j} = \mu_{\cdot j+1}, \end{aligned}$$

where $i = 1, \ldots, g-1$ and $j = 1, \ldots, p-1$.

In the univariate analysis the hypotheses can be tested in the familiar analysis of variance framework, in which a total sum of squares is decomposed into components pertaining to the main-effects and interaction. Although the assumptions of a split-plot analysis are restrictive and are questionable for a repeated measures design, these difficulties can be overcome by a so-called epsilon correction of the numerator and denominator degrees of freedom of the F-tests for testing the time main-effect and the interaction hypotheses.

The second approach to analyzing repeated measures data is to test the three hypotheses in a MANOVA framework. This approach makes no restrictive assumptions about the within-subject covariance structure $\text{Var}(\boldsymbol{y}_m) = \boldsymbol{\Sigma}$. While the computational complexity of both the univariate and the MANOVA approaches is relatively low, MANOVA is conceptually more complex. For instance, the MANOVA approach as implemented in statistical software packages reports four closely related test-statistics. Specifically, these are (1) the Hotelling-Lawley trace, (2) Pillai's trace, (3) Wilks' lambda, and (4) Roy's greatest root. Although these statistics often yield similar conclusions, this is not necessarily the case. The exact distributions of these statistics are known only in special cases.

The third approach to testing the above hypotheses uses mixed models. This approach has become widely available since it has been included in the statistical software package SAS around 1990. This approach is more general and flexible than the other two approaches because it allows for patterned within-subject covariance matrices and does not require balanced data. More specifically, the times at which

observations are made can vary between individuals and the number of observations can differ across individuals as well. While SAS PROC MIXED offers a variety of estimation methods, the likelihood-based approaches (ML and REML) appear to be the preferred methodology. For discussions of the advantages of the mixed-model methodology see, for instance, Wolfinger and Chang (1995).

Univariate Analysis

The univariate analysis of the data example is based on the model

$$Y_{ijm} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \pi_{im} + u_{ijm},$$

where τ_i and β_j are group and time main-effects and $(\tau\beta)_{ij}$ denotes the interaction, $j = 1, \ldots, p$, $i = 1, \ldots, g$, and $m = 1, \ldots, n_i$. In addition, the model assumptions about the between- and within-subject errors are: (1) $\pi_{im} \sim \mathrm{N}(0, \sigma_\pi^2)$, (2) $u_{ijm} \sim \mathrm{N}(0, \sigma^2)$, and (3) independence of random terms (Schuster & von Eye 2001; Winer, Brown, & Michels 1991). In this model the covariance structure of the repeated observations is compound symmetry, which is unrealistic if the within-subject treatment is time as in the example above.

It is well known that univariate analyses, justified under compound symmetry, are also valid under more general covariance structures, specifically, the so-called H-pattern. Huynh and Feldt (1970) have shown that this condition is necessary and sufficient for the hypothesis tests of the univariate analysis to yield valid results. Morrison (1976, p. 215) gives the H-pattern for the design considered in the data example. However, if the covariance structure is not of the type H-pattern, then the actual type I error rate of within-subject tests may change. Usually, it becomes too large resulting in too many true null hypothesis rejections. If the type H-pattern is violated, Geisser and Greenhouse (1958) and Huynh and Feldt (1976) give ϵ-corrections with which the numerator and denominator degrees of freedom of the univariate F-Tests are adjusted downward. As a result, the size of the F-test will not exceed its nominal level.

Although we focus in this article on the comparison between MANOVA and the mixed-model approach, the analysis of the interaction hypothesis of the data example will be reported briefly.

Testing the Interaction Hypothesis

The interaction hypothesis test of the data example yields $F = 3.819244$ based on 6 numerator and 36 denominator degrees of freedom. However, the test for sphericity rejects the hypothesis that the within-subject covariance pattern is of the H-type. Thus, the degrees of freedom should be corrected using either the epsilon proposed

by Greenhouse-Geisser or Huynh-Feldt, which for the data example are 0.6606 and 0.9226, respectively.

Correction of the numerator and denominator degrees of freedom yields the p-values 0.0157 for the Greenhouse-Geisser correction and 0.0062 for the Huynh-Feldt correction. If the nominal significance level had been set to $\alpha = 0.01$, the two corrections would yield different conclusions. Specifically, whereas the Greenhouse-Geisser correction would retain the null hypothesis of no interaction, the Huynh-Feldt correction would reject it. The different conclusions accord with the general observation that the Greenhouse-Geisser correction tends to be conservative, i.e. produces p-values that tend to be too large.

MANOVA Analysis

The multivariate analysis considers the observations of one individual as a $(p \times 1)$ observation vector. Because each individual belongs to one group, the data in Table 1 show five observation vectors in each group. Thus, the model for the data is a one-factor MANOVA model. If the model is parameterized in terms of the cell means, the equation for the mth observation in the ith group is:

$$\begin{array}{ccccc} \boldsymbol{y}'_{im} & = & \boldsymbol{\mu}'_i & + & \boldsymbol{u}'_{im}. \\ (y_{i1m} \ \ldots \ y_{ipm}) & = & (\mu_{i1} \ \ldots \ \mu_{ip}) & + & (u_{i1m} \ \ldots \ u_{ipm}) \end{array} \tag{1}$$

In terms of the distributional assumptions, the model requires that $\boldsymbol{u}_{im}$ follows a multivariate normal distribution with expected value of zero and covariance matrix $\boldsymbol{\Sigma}$, which is required to be symmetric and positive definite but is otherwise unspecified. Since the residual vectors $\boldsymbol{u}_{im}$ are assumed to be independent across individuals, the covariance matrix of all observations collected in a vector $\boldsymbol{y}$ is block diagonal. Based on the model equation of the one-way MANOVA the null hypothesis is:

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_g. \tag{2}$$

This hypothesis claims that the mean vectors for all groups are identical. In this case, the four means of $\boldsymbol{\mu}_i$ account for the total of twelve means (four means in each of three groups). In terms of the cell means, this hypothesis can be expressed as the following collection of contrast statements

$$\mu_{ij} - \mu_{i+1,j} = 0$$

for $i = 1, \ldots, g-1$ and $j = 1, \ldots, p$. If these statements are all true, then the interaction hypothesis and the group main-effect hypothesis are true jointly.

Thus, hypothesis (2) is not particularly interesting because it confounds the interaction hypothesis (differential change over time between the groups) with the between-subjects hypothesis.

The hypothesis of primary interest concerns the interaction between group and time. In other words, it addresses the question of whether the groups change differently across time. In terms of the cell means, this hypothesis has already been given above. To find test statistics for the interaction hypothesis in the MANOVA framework, the corresponding collection of contrast statements is expressed in terms of the so-called general linear hypothesis, which is

$$H_0 : \boldsymbol{CBM} = \mathbf{0},$$

where $\boldsymbol{C}$ and $\boldsymbol{M}$ are known $(g-1) \times g$ and $p \times (p-1)$ matrices with ranks $(g-1)$ and $(p-1)$, respectively, and $\boldsymbol{B}$ is the $g \times p$ matrix of cell means. The columns of $\boldsymbol{M}$ correspond to within-group mean comparisons at different observation times while the rows of $\boldsymbol{C}$ correspond to between-group mean comparisons. To simultaneously test all $(g-1)(p-1) = 2(3) = 6$ interaction contrasts, as is the case in the data example, one uses (e.g., Khattree & Naik 1999, p. 180 or Mardia, Kent, & Bibby 1979, p. 348)

$$\underset{\boldsymbol{C}}{\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}} \underset{\boldsymbol{B}}{\begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} \\ \mu_{31} & \mu_{32} & \mu_{33} & \mu_{34} \end{pmatrix}} \underset{\boldsymbol{M}}{\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}}.$$

To see how these $\boldsymbol{C}$ and $\boldsymbol{M}$ matrices achieve testing the interaction contrasts, we calculate the first element of the matrix $\boldsymbol{CBM}$, denoted as $[\boldsymbol{CBM}]_{11}$. Using $\boldsymbol{c}_1'$, the first row of $\boldsymbol{C}$, and $\boldsymbol{m}_1$, the first column of $\boldsymbol{M}$, yields

$$[\boldsymbol{CBM}]_{11} = \boldsymbol{c}_1' \boldsymbol{B} \boldsymbol{m}_1 = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}.$$

The resulting collection of these interaction contrasts, when tested simultaneously, is equivalent to the interaction hypothesis given above.

To obtain test statistics, the MANOVA model is written as a multivariate regression model. Having parameterized the models in terms of the g group means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$, Eq. (1) can be generalized in the following way:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U},$$

where $\boldsymbol{Y}$ and $\boldsymbol{U}$ are $N \times p$ matrices of the dependent variables and residuals, respectively, $\boldsymbol{X}$ is the $N \times g$ design matrix, and $N = \sum n_i$ denotes the total number of subjects. If we define the $n_i \times 1$ vector $\mathbf{1} = (1, \ldots, 1)'$, then the design matrix of the example is simply

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{1}\,\mathbf{0}\,\mathbf{0} \\ \mathbf{0}\,\mathbf{1}\,\mathbf{0} \\ \mathbf{0}\,\mathbf{0}\,\mathbf{1} \end{pmatrix}.$$

One then calculates the so-called hypothesis and error matrices:

$$\boldsymbol{H} = (\boldsymbol{C}\hat{\boldsymbol{B}}\boldsymbol{M})'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{B}}\boldsymbol{M})$$

$$\boldsymbol{E} = (\boldsymbol{Y}\boldsymbol{M})'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'](\boldsymbol{Y}\boldsymbol{M})$$

Although there are several multivariate statistics for testing the general linear hypothesis, we report Wilks' lambda only because it corresponds to the likelihood-ratio test of the fixed effects. Lambda is defined in terms of the above matrices as

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|},$$

and follows under the null hypothesis a Wilks' lambda distribution with parameters $p - 1$, $N - g$, and $g - 1$, where $N = \sum n_i$ (Mardia et al., 1979, p. 163).

In some cases a transformation of Λ follows an exact F-distribution. Otherwise a chi-square approximation due to Bartlett and an F-approximation due to Rao is available (see Rao 1973, p. 555). Statistical packages typically report the F-statistic only, which follows either exactly or approximately an F-distribution.

Testing the Interaction Hypothesis

From the error and hypothesis matrices for testing the interaction hypothesis one obtains Wilks' Lambda as $\Lambda = 0.14718$. This value can be converted to the likelihood-ratio, Bartlett's chi-square statistic, or Rao's F-statistic. The formulas for the likelihood-ratio and Bartlett chi-square statistics, based on df $= 6$, are

$$-2 \log \Lambda^{N/2} = -2 \log 0.14718^{15/2} = 28.74,$$

and (for $p = 3$, $q = 2$, and $t = 14$ in Rao's notation)

$$-(t - (p + q + 1)/2) \log \Lambda = -(14 - (3 + 2 + 1)/2) \log 0.14718 = 21.08$$

respectively. It is well known that the Bartlett chi-square statistic provides a better approximation to the likelihood-ratio statistic. Thus, the likelihood-ratio chi-square is seen to be too large. Usually statistical software packages report Rao's F-statistic, which is generally considered an even better approximation when compared to the Bartlett chi-square statistic. Rao's statistic yields $F = 5.356$. For the data example this F-value has 6 numerator and 20 denominator degrees of freedom and follows

Table 2 Test-statistics for evaluating the interaction hypothesis

	LR-χ^2	Barlett-χ^2	Rao-F
Value	28.74	21.08	5.356
df	6	6	6 and 20
p-Value	0.000068	0.0018	0.0019

The distribution of the Rao-F-statistic is exact for the data example

an exact F distribution. From the p-values of the Bartlett and Rao statistics given in Table 2, it can be seen that both statistics yield similar conclusions.

This example illustrates that in the MANOVA framework based on ML estimation, there exist test-statistics (Rao's F or Bartlett's chi-square) that closely approximate the nominal sampling distributions. The uncorrected likelihood-ratio test, however, yields values that tend to be too large resulting in too many true null hypothesis rejections.

Linear Mixed-Model Analysis

Linear mixed models is a well-established powerful and flexible methodology that is available today in virtually all general purpose statistical packages (e.g., SAS, SPSS, R). A linear mixed model specifies a linear function in fixed parameters and random effects for the mean of the dependent variable. Specifically,

$$\boldsymbol{y}_m = \boldsymbol{X}_m\boldsymbol{\beta} + \boldsymbol{Z}_m\boldsymbol{\gamma}_m + \boldsymbol{u}_m,$$

where $\boldsymbol{X}_m$ and $\boldsymbol{Z}_m$ are known matrices and $\boldsymbol{\gamma}_m$ is a vector containing the random effects of the mth subject having zero mean and covariance matrix $\boldsymbol{G}$, and $\boldsymbol{u}_m$ is a vector of random residuals having zero mean and covariance matrix $\boldsymbol{R}$. In addition, $\boldsymbol{\gamma}_m$ and $\boldsymbol{u}_m$ are assumed to be independent. The covariance matrix $\text{Var}(\boldsymbol{y}_m) = \boldsymbol{\Sigma}$ implied by this model is

$$\boldsymbol{\Sigma} = \boldsymbol{Z}_m\boldsymbol{G}\boldsymbol{Z}_m' + \boldsymbol{R}.$$

Three cases can be distinguished depending on how the covariance of the within-subject observations $\boldsymbol{\Sigma}$ is modelled:

1. The covariation can be modelled in terms of the matrix $\boldsymbol{R}$ only. One possibility is to let this matrix be arbitrary, as is assumed throughout this article. In this case, linear mixed models can also be used to analyze multivariate models, which typically assume an unconstrained within-subject covariance matrix. If $\boldsymbol{R}$ is unconstrained, a $\boldsymbol{G}$ matrix is not needed. Alternatively, a particular pre-specified pattern for $\boldsymbol{R}$ can be used. Software packages (e.g., SAS, SPSS)

typically offer a wide selection of such patterned matrices. Because, for arbitrary $\boldsymbol{R}$, the number of covariance parameters increases quickly with the number of repeated observations, an arbitrary $\boldsymbol{R}$ will be most useful in cases where the number of repeated observations is relatively small.

2. In social sciences linear mixed models have become very popular because random effects and random coefficient models, also known as hierarchical linear models (Bryk & Raudenbush 1992), are special cases of this methodology. In these models, uncorrelated and homoscedastic residuals are typically assumed. This implies $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ and the within-subject covariation results from the random-effects covariance matrix $\boldsymbol{G}$. In random-coefficient models the within-subject observations are modelled as a linear function of random intercept and slope parameters. This leads to a parsimonious description of the within-subject observations because $\boldsymbol{G}$ contains only two variance and one covariance parameter.
3. If a patterned matrix for $\boldsymbol{R}$ is selected, it is also possible to allow random effects so that both $\boldsymbol{G}$ and $\boldsymbol{R}$ contribute to the covariation between repeated observations. Singer (1998) gives examples of this approach.

Likelihood Estimation

If the within-subject observations of a balanced design are assumed to follow a multivariate normal distribution with arbitrary $\boldsymbol{\Sigma} = \boldsymbol{R}$, then the likelihood-function on which estimation of model parameters is based is

$$L = (2\pi)^{-Np/2}\,|\boldsymbol{\Sigma}|^{-N/2}\exp\left[-\frac{1}{2}\sum_{m=1}^{N}(\boldsymbol{y}_m - \boldsymbol{X}_m\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_m - \boldsymbol{X}_m\boldsymbol{\beta})\right], \qquad (3)$$

where $\boldsymbol{y}_m$ is the $p \times 1$ within-subject observations vector, $\boldsymbol{X}_m$ is a $p \times q$ design matrix of rank q, $\boldsymbol{\beta}$ is a $q \times 1$ fixed-effects parameter vector, and $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix of the within-subject observations.

If the covariance matrix $\boldsymbol{\Sigma}$ is completely known or known only up to a scalar multiple, e.g. $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, then the distribution of the maximum-likelihood estimate of $\boldsymbol{\beta}$ is known exactly (McCulloch, Searle, & Neuhaus, 2008, Sect. 6.3) and the maximum-likelihood estimates of the fixed-effects parameters are

$$\hat{\boldsymbol{\beta}} = \left(\sum_{m=1}^{N}\boldsymbol{X}'_m\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_m\right)^{-1}\left(\sum_{m=1}^{N}\boldsymbol{X}'_m\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_m\right). \qquad (4)$$

As a result, hypothesis testing with respect to the fixed-effects parameters does not depend on a large sample size. However, if $\boldsymbol{\Sigma}$ is not known, it needs to be estimated from the data.

Two popular approaches to estimating the covariance parameters are maximum-likelihood (ML) and restricted maximum-likelihood (REML) estimation. In maximum-likelihood estimation (3) is maximized with respect to all unknown parameters, both fixed-effects parameters and covariance parameters. In REML estimation, the $N \times 1$ vector $\boldsymbol{y}$ is transformed to a set of q new variables $\boldsymbol{Z} = \boldsymbol{K}\boldsymbol{y}$, where $\boldsymbol{K}$ is any $(N - q) \times N$ matrix of full-row rank satisfying $\boldsymbol{K}\boldsymbol{X} = 0$. The likelihood-function of the transformed variables $\boldsymbol{Z}$ is commonly referred to as restricted likelihood or marginal likelihood. For expressions of this likelihood function, see Kenward and Roger (1997) or Schluchter and Elashoff (1990). For estimating the covariance parameters, REML estimation is often preferred (McCulloch et al. 2008, Sect. 6.10).

Regardless of whether the covariance parameters are estimated with ML or REML, the fixed-effects parameters are estimated from (4), where $\boldsymbol{\Sigma}$ is replaced by its estimate $\hat{\boldsymbol{\Sigma}}$. The large sample covariance matrix of the fixed-effects parameters is then given by

$$\boldsymbol{A} = \left(\sum_{m=1}^{N} \boldsymbol{X}_m' \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_m \right)^{-1}.$$

In small samples, however, $\boldsymbol{A}$ is a biased estimate of the fixed-effects variability. One reason for this bias is that the sampling variance of the covariance parameters contained in $\hat{\boldsymbol{\Sigma}}$ is not accounted for. As a result, $\boldsymbol{A}$ underestimates (in a matrix sense) the variability of the fixed-effects parameter estimates (Kenward & Roger, 1997).

Testing hypothesis about the fixed effects parameters can be expressed as

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = 0,$$

where $\boldsymbol{C}$ is an $\ell \times q$ matrix of rank ℓ. This hypothesis can be evaluated using either the likelihood-ratio statistic or the Wald test statistic, which is

$$W = \hat{\boldsymbol{\beta}}' \boldsymbol{C}' (\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}')^{-1} \boldsymbol{C}\hat{\boldsymbol{\beta}}. \tag{5}$$

In large samples, both the likelihood-ratio and the Wald statistics follow a chi-square distribution with ℓ degrees of freedom. In small samples, however, their exact distributions are unknown. Exceptions occur for balanced data (Schluchter & Elashoff, 1990). For instance, when the covariance parameters are estimated by ML, the likelihood-ratio test is equivalent to the Wilks' Lambda statistic for which an approximation to the F-distribution due to Rao is exact in certain special cases (Rao 1973, p. 555). According to Roger and Kenward (1993) the multivariate F-test based on Wilks' Lambda as implemented in PROC GLM has a better approximation to the actual distribution of the test statistic than that given for the F-tests as implemented in PROC MIXED. However, Kenward and Roger (1997) have developed a Wald-type test statistic that yields an improved approximation to an F distribution in small samples, when the covariance parameters are estimated by REML.

This statistic is developed from the expression $F = W/\ell$, which in large samples follows an F-distribution with ℓ numerator degrees of freedom and infinite denominator degrees of freedom. In small samples, Kenward and Roger suggest two modifications of F that yield a test-statistic following approximately an F-distribution. First, F is scaled by a factor λ. Thus, the test-statistic is $F^* = \lambda F$, where λ is typically smaller than 1.0. Second, the denominator degrees of freedom m have to be estimated. The approach is similar to Satterthwaite (1946), where λ and m are obtained by equating first and second moments of F^* to the first two moments of the F distribution with ℓ numerator and m denominator degrees of freedom (McCulloch et al. 2008, p. 168). The resulting formulas for λ and m are given in Kenward and Roger's equations (7) and (8). However, these formulas are not suitable for hand calculation. The Kenward-Roger test-statistic is available in SAS PROC MIXED but not in SPSS MIXED. PROC MIXED reports the denominator degrees of freedom m, but not the scale factor λ.

Testing the Interaction Hypothesis

First, we analyze the interaction hypothesis of the data example using the likelihood-ratio test when parameter estimation is based on ML. Note that the likelihood-ratio statistic is not available with REML estimation. Specifically, the model is fit with and without the interaction. This yields the following log-likelihood values: $-2 \log L = 277.62751129$ with interaction and $-2 \log L = 306.36936631$ without interaction. Thus, the likelihood-ratio statistic based on df $= 6$ is $\text{LR} - \chi^2 = 306.36936631 - 277.62751129 = 28.74$. This is exactly the value obtained above from Wilks' lambda given in Table 2 (Wright 1998).

While the likelihood-ratio statistic is not available when parameter estimation is based on REML, the Wald chi-square statistic can be calculated with both ML and REML estimation. Testing the interaction hypothesis yields the results given in Table 3.

Comparing the Wald chi-square values of Table 3 with the chi-square values of Table 2, it can be seen that the Wald chi-square values are even larger than the likelihood-ratio statistic, which was seen to be too large, when Bartlett's chi-square statistic is used as the criterion for comparison. Note that the F values in Table 3 should not be directly compared to Rao's F in Table 2 because they are based on different denominator degrees of freedom.

This example illustrates two points: First, the likelihood-ratio statistic is often preferable to Wald chi-square when the sample size is small. Second, while the

Table 3 Wald-type chi-square statistics of the data example using ML and REML estimation

Estimation	ℓ	m	Wald	F	p
ML	6	12	78.72	13.12	0.0001
REML	6	12	62.98	10.50	0.0004

The F values are obtained as W/ℓ

REML test is somewhat better (closer to the Bartlett chi-square) than the ML test, there is definitely a need for a small-sample correction of the REML test-statistic, such as the one suggested by Kenward and Roger (1997).

The Kenward and Roger small sample correction, which is based on REML estimation, yields $F = 8.39$ based on $\ell = 6$ and $m = 12.018510$ with a p-value of 0.0010. Note that the correction has scaled the REML F value of 10.50, see Table 3, down to 8.39 and adjusted the denominator degrees of freedom slightly. As a result, the p-value is increased and, thus, approaches the p-values of the Bartlett chi-square and the Rao F-statistic.

Simulation Study

As the data example has illustrated, the MANOVA approach has excellent small-sample approximations of the Wilks' Lambda test-statistics. However, the MANOVA approach is generally only applicable if the data come from a balanced design. If the design is unbalanced, mixed models can be used to analyze the data without the need to eliminate subjects with missing values. Nevertheless, the balanced case provides a basis on which the mixed-model test statistics can be evaluated and compared to the MANOVA approach. Because not all statistical software packages have implemented small-sample corrections in their mixed-model procedures, it is of interest to examine how the uncorrected Wald statistic compares to the corrected Wald statistic and the MANOVA statistics.

To investigate the small-sample behavior of the various test statistics we consider the test for interaction only and compare the following three test-statistics: (1) the Rao F-approximation to Wilks' Lambda, (2) the scaled Wald-type F-statistic of Kenward and Roger (1997) assuming REML estimation, and (3) the default Wald-type F-statistic as implemented in SAS PROC MIXED or SPSS MIXED using REML estimation. All simulations are based on a multivariate covariance structure, that is, the within-subject covariance matrix is unconstrained.

The default Wald-type test, case (3) above, is particularly important because the options available for selecting a particular statistic are limited across statistical packages. For instance, SPSS does not offer options for small sample corrections. The default test for the interaction in SAS PROC MIXED and SPSS MIXED divides the Wald chi-square in (5) by the interaction degrees of freedom, which for the balanced design are $\ell = (g-1)(p-1)$ and then treats this value as approximately F-distributed with ℓ numerator and $m = N-g$ denominator degrees of freedom. The denominator degrees of freedom correspond to the denominator degrees of freedom for the group main-effect in a balanced design, which is commonly referred to as the between-subjects effect. Therefore, we use the label BE in Table 4 below to label the default Wald-type F-statistic.

Multivariate normally distributed data were simulated using the following three compound symmetric covariance structures:

Table 4 Size of interactions hypothesis F-tests across 10,000 replications having nominal size of $\alpha = 0.05$

Number of groups	Group size	Test statistic	ℓ	m	Within-subject correlation		
					0.2	0.5	0.8
g = 3	n = 5	KR	6	12.018510	0.0469	0.0506	0.0520
		BE	6	12	0.0879	0.0886	0.0944
		WI	6	20	0.0463	0.0497	0.0524
g = 3	n = 10	KR	6	31.943571	0.0477	0.0509	0.0524
		BE	6	27	0.0613	0.0655	0.0670
		WI	6	50	0.0468	0.0507	0.0531
g = 5	n = 5	KR	12	26.746232	0.0477	0.0471	0.0472
		BE	12	20	0.0666	0.0672	0.0667
		WI	12	47.915	0.0493	0.0499	0.0503
g = 5	n = 10	KR	12	70.334982	0.0473	0.0510	0.0473
		BE	12	45	0.0531	0.0572	0.0521
		WI	12	114.06	0.0470	0.0507	0.0479

KR denotes Wald-type F-Test using the Kenward-Roger test-statistic; BE denotes Wald-type F-Test using denominator degrees of freedom of between-subjects test; WI denotes MANOVA F-Test of Wilks' Lambda

$$\Sigma_1 = \begin{pmatrix} 5.0 & 1.0 & 1.0 & 1.0 \\ 1.0 & 5.0 & 1.0 & 1.0 \\ 1.0 & 1.0 & 5.0 & 1.0 \\ 1.0 & 1.0 & 1.0 & 5.0 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 5.0 & 2.5 & 2.5 & 2.5 \\ 2.5 & 5.0 & 2.5 & 2.5 \\ 2.5 & 2.5 & 5.0 & 2.5 \\ 2.5 & 2.5 & 2.5 & 5.0 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 5.0 & 4.0 & 4.0 & 4.0 \\ 4.0 & 5.0 & 4.0 & 4.0 \\ 4.0 & 4.0 & 5.0 & 4.0 \\ 4.0 & 4.0 & 4.0 & 5.0 \end{pmatrix}.$$

The within-subject correlations are 0.2, 0.5, and 0.8. The means in all simulations were set to zero, so that the null-hypotheses for all effects were true.

Similar to the above data example we assumed a balanced design having three between-subject groups with five observations in each group and four within-subject observations. Thus the total sample size N of each replication was 15. In each replication we calculate the actual size of the tests for interaction for which the significance level was set to $\alpha = 0.05$. To broaden the scope of the simulations, we factorially combined an increased sample size per group, 10 instead of 5 observations with five instead of three groups. The results of all four simulation studies are contained in Table 4.

The results of the simulation confirm that Rao's F-approximation to the Wilks' Lambda statistics available from a MANOVA analysis is very close to the nominal level of $\alpha = 0.05$ under all simulation conditions. The same is true for the Kenward-Roger small sample correction of the Wald test statistic. Therefore, a mixed-model analysis with small-sample correction or a MANOVA analysis yields virtually identical results in the balanced case. Of course, if the data are unbalanced, the mixed-model analysis with small-sample correction should be used.

If a statistical software package does not provide a small-sample correction when analyzing a mixed model assuming a multivariate covariance structure, which corresponds to the BE-test statistic in Table 4, then the test-statistic tends to reject true interaction null hypotheses too frequently if the denominator degrees of freedom, m, are small. More specifically, with $m = 12$ the actual type I error rate is close to 0.10, roughly double the nominal level of 0.05. As m increases, the actual type I error rate is closer to the nominal level, see BE-tests for $m = 20$, $m = 27$, and $m = 45$.

Discussion

We have examined various test statistics available from the MANOVA and mixed-model approaches for testing the hypothesis of no interaction in a repeated measures design, with one between-subject and one within-subject factor assuming an unconstrained within-subject covariance matrix. We restricted attention to the interaction hypothesis to reduce the complexity of the presentation. In addition, the interaction hypothesis, do groups change differentially over time, is often the most interesting research question. With unbalanced designs, a mixed-model analysis is definitely preferable, because likelihood-based mixed-model estimation does not rely on balanced data. However, if balanced data are available, the distribution of Wilks' Lambda obtained from a MANOVA analysis can be closely approximated even in small samples using either Bartlett's chi-square or Rao's F statistic. While mixed-model theory is based on large samples, it is well known that both the likelihood-ratio test statistic and the Wald test statistics of fixed effects are typically too large, leading to too many true null hypotheses rejections. While SAS PROC MIXED has implemented a small-sample correction developed by Kenward and Roger (1997), it is not clear whether the MANOVA approach or the small-sample corrected Wald-type statistic of the mixed-model approach is preferable with balanced data.

Our limited simulation study suggests that for a multivariate covariance structure, the REML estimation together with the Kenward-Roger small-sample correction yields a test statistic which performs very well and gives virtually identical results as well-known MANOVA approximations to the Wilks' lambda statistic. However, if such a small-sample corrections are not offered by the software used to fit the mixed model assuming an unstructured covariance matrix (e.g., SPSS MIXED), then the MANOVA approach is preferable with small samples.

References

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the *F*-distribution in multivariate analysis. *Annals of Mathematical Statistics, 29*, 885–891.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*, 1582–1589.

Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*, 69–82.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983–997.

Khattree, R., & Naik, D. N. (1999). *Applied multivariate statistics with SAS software* (2nd ed.). Cary, NC: SAS Institute Inc.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

McCulloch, D. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed.). Hoboken, NJ: Wiley.

Morrison, D. F. (1976). *Multivariate statistical methods* (2nd ed.). New York: McGraw-Hill.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Roger, J. H., & Kenward, M. (1993). Repeated measures using PROC MIXED instead of PROC GLM. In *Proceedings of the First Annual South-East SAS Users Group Conference* (pp. 199–208). Cary, NC: SAS Institute Inc.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

Schluchter, M. D., & Elashoff, J. D. (1990). Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computation and Simulation, 37*, 69–87.

Schuster, C., & von Eye, A. (2001). The relationship of ANOVA models with random effects and repeated measurement designs. *Journal for Adolescence Research, 16*(2), 205–220.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical linear models, and individual growth models. *Journal of Educational Statistics, 24*(4), 323–355.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Wolfinger, R. D., & Chang, M. (1995). Comparing the SAS GLM and MIXED procedures for repeated measures. In *Proceedings of the Twentieth Annual SAS Users Group Conference.* Cary, NC: SAS Institute Inc.

Wright, S. P. (1998). *Multivariate analysis using the MIXED procedure.* Paper presented at the 23 Annual Meeting of the SAS Users Group International Conference.