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

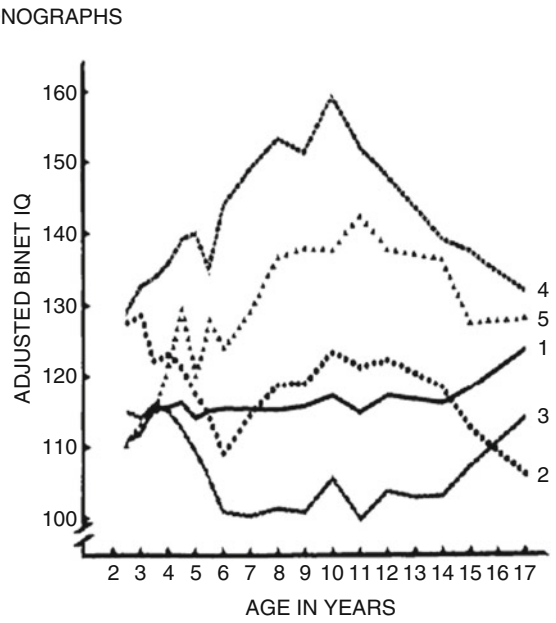

Fig. 1 From McCall, Applebaum & Hogarty (1973, p. 48)

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