Structural Equation Modeling for Description, Prediction, and Causation

Structureel Vergelijkingsmodeleren voor Beschrijven, Voorspellen en Oorzakelijkheid Vaststellen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 7 juni 2024 des ochtends te 10.15 uur

door

Jeroen Dick Mulder

geboren op 2 oktober 1994
te IJsselstein

Promotor:

Prof. dr. E. L. Hamaker

Copromotor:

Dr. S. Usami

Beoordelingscommissie:

Prof. dr. S. van Buuren
Prof. dr. L.G.M.T. Keijsers
Prof. dr. D.L. Oberski
Prof. dr. J. Vermunt
Prof. dr. M. Voelkle

Jeroen D. Mulder

Structural Equation Modeling for Description, Prediction, and Causation

Structural Equation Modeling for Description, Prediction, and Causation
Dissertation Utrecht University, Utrecht, the Netherlands
Met een samenvatting in het Nederlands

Cover idea: Jeroen D. Mulder
Cover photo: Niels Struis

Print: Ridderprint, Alblasserdam
ISBN: 978-90-393-7667-6
DOI: https://doi.org/10.33540/2253

The studies in this dissertation were done as part of Consortium on Individual Development, and were financially supported through the Gravitation Program of the Dutch Ministry of Education, Culture and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003).

Contents

CHAPTER 1

Introduction

Since the early work a century ago by Sewall Wright and Charles Spearman, and through important contributions by the likes of Karl G. Jöreskog, Otis Dudley Duncan, and many, many others, structural equation modeling (SEM) has emerged as a powerful and versatile statistical modeling framework (Matsueda, 2023; Tomarken & Waller, 2005). One of the defining features of this framework is the ability to define latent variables herein, and connect these to other latent and observed variables through structural equations. Latent variables can be used to model unobserved constructs (e.g., self-esteem or intelligence through confirmatory factor analysis), but their uses extend far beyond this: Latent variables can also capture statistical concepts such as measurement errors, clusters, random effects, and variance components (L. K. Muthén & Muthén, 2009). Furthermore, the SEM framework itself has been extended to include mixture modeling, multilevel modeling, missing data modeling, and Bayesian estimation, further broadening its statistical capabilities and areas of application. Combined with implementation of these techniques in user-friendly software such as Mplus (L. K. Muthén & Muthén, 2017) or the R package lavaan (Rosseel, 2012), and with powerful algorithms for estimation, researchers have developed a wide range of SEM models for a diverse set of research questions. By now, SEM has become established as one of the main statistical modeling frameworks in the social and behavioural sciences.

My first encounter with SEM was in 2017, during the first year of my research master. I was immediately drawn to it. Its main appeal was the visual aspect: Using a combination of circles, squares, one-headed, and two-headed arrows, a SEM model, and the set of regression equations that are implied by it, can be visualized in a *path diagram*. These diagrams offer an intuitive connection between the theoretical phenomenon that is the object of investigation and the statistical analysis, mapping hypothesized relationships (based on theory) to parameters in regression equations.

Perhaps because it is so simple to draw a set of boxes and circles with arrows connecting them, path diagrams can be extended beyond situations of mere multiple regression with ostensible ease. For example, path diagrams can be used to represent moderation, mediation, multiple outcomes, longitudinal processes, with causal relationships going into multiple directions, and combinations thereof. As human behavior, cognition, emotions, abilities, and other psychological features are complex systems—interacting with each other and evolving over time—the possibility to visualize such systems and map them to a set of equations in a (seemingly) straightforward manner is a huge appeal.

1.1 The PhD project

My PhD project started as part of the Consortium on Individual Development (CID), an interuniversity research consortium in the Netherlands that investigated how child characteristics and environmental factors impact a child's development of social competence and behavioural control (Consortium on Individual Development, 2023). The goal of the PhD project was to evaluate and develop methods for studying the causes and effects in psychological and behavioural developmental processes. Given the expertise of my promotor, Dr. Ellen Hamaker, the project quickly centered on longitudinal SEM models, and how (applications of) this broad class of models could be improved by learning from causal inference approaches in other disciplines (e.g., through the use of instrumental variables, or the potential outcomes framework).

The first subproject concerned developing and describing extensions of the random intercept cross-lagged panel model (RI-CLPM). It is a longitudinal SEM model for investigating lag-1 relationships between constructs over time (i.e., cross-lagged effects), and it is part of the larger class of cross-lagged panel models in psychology (Usami et al., 2019; Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020). After the RI-CLPM was introduced by Hamaker et al. (2015), the model rapidly increased in popularity. It addressed some long-standing concerns that psychological researchers have had about the analysis of panel data, such as unobserved heterogeneity, and the conflation of trait-like and state-like variance. Hence, it is not surprising that applied researchers were interested in how the RI-CLPM could be adapted to accommodate specific data and research interests. Frequently asked questions concerned extensions of the RI-CLPM to the use of multiple indicators, multiple groups, the inclusion time-invariant predictors, and sample size recommendations. These questions ended up as the focus of two separate papers: The first paper described three extensions of the RI-CLPM, and included an online website with elaborately annotated Mplus syntax and R code for fitting these particular models; The second paper outlined a strategy for power analysis that is tailored to the particularities of the RI-CLPM (this strategy was also implemented in the R package powRICLPM as part of this subproject). These papers are included as Chapters 4 and 5 in this dissertation, respectively.

By the second year of the PhD project, I had gotten involved in a considerable amount of statistical consultation about longitudinal SEM models. Some of the questions I received were in response to the RI-CLPM-related papers that had been published, some were asked during consultation duties at the Methodology and Statistics department at Utrecht University, and others resulted from collaborations with colleagues within CID. Ultimately, two applied papers evolved from these consultations. The first is a collaborative project with clinical psychologists from Utrecht University and the Altrecht Academic Anxiety Center, and concerned the development of post traumatic stress disorder (PTSD) symptoms of patients throughout a newly developed two-week clinical treatment program. The second is a collaborative project with developmental neuroscientists from the Erasmus University Rotterdam and Leiden University, in which we investigated the developmental trajectories of children's neural and behavioural (aggressive) responses to social rejection. In both papers we used growth curve models (GCMs), which is another broad class of longitudinal models that is well-suited for describing development in a construct over time, and individual differences herein. However, these applied subprojects were not run-of-the-mill applications of GCMs: Each subproject had its own set of fundamental and/or statistical challenges that made these projects interesting from a methodological perspective. A personal challenge that I had set for myself, was ensuring that I described our use of statistics in an accessible manner for an applied audience, while not compromising on statistical rigor. These values sometimes clashed in practice: The complexity of empirical data often dictated going beyond the use of standard, off-the-shelf statistical models, which made the description of these methods in the papers more complex as well. Here, I found the use of Rmarkdown- and Quarto-websites as online supplementary materials for these papers quite useful. Websites offer an essentially unlimited amount of space for additional explanation of, and rational for the statistical methods that were used, and can include interactive plots and annotated code to support additional explanation. Both applied projects are included as Chapters 2 and 3 in this dissertation.

The latter halve of my PhD project is characterized by an increasing focus on formal causal inference, specifically the potential outcomes framework that is widely used in disciplines like epidemiology and biostatistics. This focus emerged from a fundamental interest in causality within CID, and psychology more generally. In Hamaker et al. (2020), we evaluated a hundred randomly sampled research papers published by researchers that were part of CID. In each paper, we identified sentences regarding its research question, hypothesis, discussion, and conclusions, and categorized these sentences as being descriptive, predictive, or causal in nature. We found that CID studies were mostly driven by descriptive and causal interests. This is in line with Grosz et al. (2020), who argues that much of the psychological research interest is essentially causal in nature, but often implicitly so (Hernán, 2018). At the same time, there is critique in the formal causal inference literature on the use of longitudinal SEM models for causal inference from nonexperimental data. For example, Van der Laan and Rose (2011) and VanderWeele (2012) argue that the traditional use of SEM models relies on a large number of parametric assumptions that are likely to be wrong, resulting in biased estimates of model parameters. Instead, they promote the use of a class of methods called generalized methods (g-methods), which have been developed specifically in the potential outcomes framework (Rubin, 1987; Vansteelandt & Sjolander, 2016). These methods have been developed to minimize reliance on parametric assumptions and therefore, in principle, should lead to more robust causal inference.

However, the uptake of g-methods, and the potential outcomes framework more generally, is still limited in the psychological literature. One of the main obstacles that psychological researchers face, is that the causal inference literature on these methods is not easily accessible due to the sometimes technical descriptions of the methods, and examples that do not connect to the modeling practices that psychological researchers are familiar with. To clarify the critique on the use of SEM models for causal inference, and to enable applied researchers to make better informed decisions about which particular model and modeling approach is most useful for their research project, my promoter and I started two subprojects comparing the use of longitudinal SEM models, specifically cross-lagged panel modeling approaches, to g-methods in a psychological context. The first is a collaboration with an epidemiologist and a biostatistician from the University Medical Center Utrecht, focusing on cross-lagged panel modeling and inverse probability weighting estimation of marginal structural models (one of the g-methods). The second is a collaboration with my co-promotor, associate professor Satoshi Usami at the University of Tokyo, centering around cross-lagged panel modeling and structural nested mean modeling (another one of the g-methods). Both subprojects are included in this dissertation as Chapters 6 and 7, respectively.

1.2 Dissertation outline

The current dissertation is a reflection of the work done in my PhD project. With the exception of Chapter 2, it centers around applications and evaluations of longitudinal SEM models for nonexperimental data. At the same time, the dissertation is diverse in focus, ranging from applied, to methodological studies, and describing statistical methods across disciplines, from clinical psychology to neuroscience and epidemiology.

The organizing principle for the chapters in this dissertation is the distinction of research questions into those that are descriptive, predictive, or causal . This distinction is absolutely critical for making well-informed decisions about the study design and data analysis in research projects, as the fundamental issues underlying each type of research question are different (Hamaker et al., 2020). Although this distinction may appear evident at first, there is evidence that in practice a study's research goals are not always clear (Grosz et al., 2020; Haber et al., 2022; Hamaker et al., 2020). The problem with ambiguity in research questions is that it encumbers critical assessment of the methodological approach that was taken, and whether or not the conclusions drawn are valid in relation to the research question (Hernán, 2018). By organizing the chapters in this dissertation by the type of research question that each chapter addresses, I hope to clarify what I believe these models are used for in practice.

1.2.1 Description

Chapter 2 is a collaboration with Dr. Michelle Achterberg, assistant professor at Erasmus University Rotterdam, and Dr. Simone Dobbelaar, a neuroscience researcher from Leiden University. This applied project consisted of two parts. First, we investigated the developmental trajectories of children's neural and behavioural (aggressive) responses to social rejection (from childhood to emerging adolescence), as well as individual differences herein. Numerous statistical challenges were present here, such as (a) individually-varying times of observation, (b) censoring of behavioural measurements, (c) nonnormality of the data, (d) missing data, (e) nonindependence of measurements due to twinning, and (f) expected nonlinear development. Bayesian multilevel growth curve models were used to model the development in neural and behavioural responses, and to take the statistical challenges into account. Second, we explored if individual differences in development of neural and behavioural responses were related to social well-being in early adolescence. For this, individual-level growth components were extracted from the Bayesian multilevel model, and used as predictors in a SEM model and with social well-being items as outcomes.

1.2.2 Prediction

Chapter 3 is an applied project with Valentijn Alting van Geusau, a PhD candidate in clinical psychology at Utrecht University, and Dr. Suzy Matthijsen at the Altrecht Academic Anxiety Center. At Altrecht, a new two-week clinical PTSD treatment program had been developed. Due to the high costs of the treatment program and the high rate of dropout, medical practitioners were interested in predicting early on in the treatment program (in the first week), who would benefit from continuing treatment into the second week. Therefore, the goal of the study was to predict PTSD reduction four weeks after the treatment program from daily PTSD symptom measurements during the program. Five different latent growth curve models (LGCMs) were used to capture PTSD symptom reduction throughout the treatment program, and to predict PTSD reduction at four-week follow-up. Through the use of k-fold cross-validation we compared the out-of-sample prediction performance of the different LGCMs models.

1.2.3 Causation

I categorize the majority of the chapters in this dissertation as pertaining to causation, including the chapters about the RI-CLPM. While some might question this decision—pointing out that the RI-CLPM is merely a statistical model, and that applications thereof in psychological research rarely evaluate the causal identification assumptions and parametric assumptions needed to interpret estimates causally—estimates of the RI-CLPM are commonly interpreted as causal. Therefore, we should be clear about what this model is used for by researchers, such that we can have an honest discussion about the advantages and shortcomings of this modeling approach (Hernán, 2018).

Chapter 4 is a statistical paper in collaboration with Ellen Hamaker. We describe three extensions of the RI-CLPM, including (a) the inclusion of stable, person-level characteristics as predictors and/or outcomes; (b) specifying a multiple-group version; and (c) including multiple indicators. A core element of this paper is its online supplementary material: It is a website on which we provide elaborately annotated Mplus syntax and R code for fitting the RI-CLPM and the extensions we described, as well as an example dataset for practice. The website also includes a section with answers to frequently asked questions that reached us since the publication of this paper.

Chapter 5 proposes a strategy for performing a power analysis for the RI-CLPM. The strategy was designed to be user-friendly, and is implemented in the R package powRICLPM. Various extensions to the basic power analysis analysis strategy are described, including the use of bounded estimation, imposing various constraints on parameters over time, and inclusion of measurement errors in the data generating model and estimation model (leading to the stable trait autoregressive trait state model).

Chapter 6 is a methodological project in collaboration with Dr. Kim Luijken, an epidemiologist at the University Medical Center Utrecht, Dr. Bas Penning de Vries, a biostatistician at the University Medical Center Utrecht, and Ellen Hamaker. We aimed to clarify some of the critique in the causal inference framework on the use of SEM models for causal inference. First, we described how the use of SEM models fits within the potential outcome framework for causal inference. Second, we compared SEM methods to potential outcome methods (specifically, path analysis versus inverse probability weighted estimation of marginal structural models) using an empirical example on smoking cessation and weight gain. Third, we zoomed in on the critique that path analysis relies too heavily on parametric assumptions. As such, we performed a simulation study to investigate the finite sample performance of path analyses and inverse probability weighted estimation under violations of parametric assumptions.

Chapter 7 is a methodological project with Dr. Satoshi Usami and Ellen Hamaker. In this chapter, we bridged the disciplinary disconnect between the SEM literature and causal inference literature, by comparing a cross-lagged panel modeling approach with structural nested mean modeling for investigating effects of time-varying exposures. We introduced and compared the causal effects that are targeted by both methods (i.e., cross-lagged effects versus joint effects), as well as the causal and parametric assumptions that both methods rely on. Their use was illustrated using an empirical psychological example regarding the joint effect of self-esteem on depression. To facilitate further integration of the SEM literature and causal inference literature, we linked our comparison to other methodological and statistical discussions that are taking place in the SEM literature, such as the decomposition of observed variance into within- and between-unit variance, and the inclusion of contemporaneous effects.

CHAPTER 2

Individual differences in developmental trajectories of social emotion regulation from childhood to emerging adolescence

Revised version published as:
Mulder, J. D., Dobbelaar, S., & Achterberg, M. (2024). Behavioral and neural responses to social rejection: Individual differences in developmental trajectories across childhood and adolescence. *Developmental Cognitive Neuroscience*, *66*, 101365. https://doi.org/10.1016/j.dcn.2024.101365

Abstract

Dealing with social rejection is challenging, especially during childhood when behavioral and neural social emotion regulation is still developing. Prior research has focused largely on group-based averages of this development, obscuring meaningful individual variation. In the current longitudinal study, we used a Bayesian multilevel growth curve model to describe individual differences in the development of behavioral and neural responses to negative social feedback in a large sample ($N > 500$). We found a slight peak in aggression following negative feedback (compared to neutral feedback) during late childhood, as well as individual differences during this developmental phase, possibly suggesting a sensitive window for social emotion regulation development across late childhood. Moreover, we found evidence for individual differences in the linear development of neural responses to social rejection in our three brain regions of interest: The anterior insula, the medial prefrontal cortex, and the dorsolateral prefrontal cortex. In addition to providing insights in the individual trajectories of social emotion regulation during childhood, this study also makes a meaningful methodological contribution: Our statistical analysis strategy (and online supplementary information) can be used as an example on how to take into account the many complexities of developmental neuroimaging datasets, while still enabling researchers to answer interesting questions about individual-level relationships.

2.1 Introduction

During the transition from childhood to emerging adolescence (approximately between the ages of 7- to 14-years-old) peer relations and long-lasting friendships become more salient. Social emotion regulation, that is, regulating one's emotions in social situations, for example after receiving negative peer feedback, is an important prerequisite for developing and maintaining such relationships. A broad range of literature has shown that receiving negative social feedback can result in reactive aggressive behavior (Dodge et al., 2003; Leary et al., 2006; Nesdale & Lambert, 2007), and that the regulation thereof is related to neural activation (Achterberg et al., 2016; Chester et al., 2014; Riva et al., 2015).

These behavioral and neural responses to social emotion regulation develop across childhood and adolescence (Achterberg et al., 2020; Dobbelaar et al., 2023). However, existing research into development of behavioral aggression has focused largely on group-based averages, obscuring meaningful individual variation across children in development (Chester, 2019). To move towards a more nuanced understanding of behavioral aggression and neurocognitive changes, developmental neuroimaging studies need to characterize individual differences as a variable of interest, as argued by Foulkes and Blakemore (2018) and Telzer et al. (2018), amongst others. By addressing individual variability in adolescent development, researchers acknowledge the fact that adolescents, and their brains, develop in meaningfully different ways. This is particularly important when studying behavioral and neural responses to social interactions, as adolescents substantially vary in the quantity and quality of friendships they have, affecting both their behavioral and neural responses to social interactions (Lamblin et al., 2017; Van Harmelen et al., 2017). Some researchers have even proposed that adolescent development is shaped by brain-based individual differences in sensitivity to social contexts, and that individual differences in neurobiology might determine how sensitive an adolescent is to the social context (Schriber & Guyer, 2016). Additionally, a focus on individual differences in behavioral and neural development allows for investigating whether such differences are useful predictors for future mental health and well-being (Copeland et al., 2013; Foulkes & Blakemore, 2018; Van Harmelen et al., 2017).

Therefore, the current preregistered study investigates individual differences in developmental trajectories of social emotion regulation (the preregistration is published as Achterberg et al., 2022). Our focus is on behavioral (aggressive) responses, and neural responses to negative social feedback, specifically in three brain regions that have previously been related to the processing of social feedback, namely the anterior insula (AI), the medial prefrontal cortex (MPFC), and the dorsolateral prefrontal cortex (DLPFC). To understand the underlying brain mechanisms, we additionally examine how developmental trajectories of aggression regulation following negative social feedback relate to each other, and to social well-being in early adolescence. To address individual variability in developmental trajectories, we analyze longitudinal behavioral and fMRI data (three waves, measured during childhood and emerging adolescence) in a multilevel modeling framework.

2.1.1 Behavioral and neural correlates of social emotion regulation

Social emotion regulation, defined here as aggression regulation following negative social feedback, is an essential quality for children to develop in order to establish and maintain relationships with peers. A recently introduced experimental method for measuring this, which is also used in this study, is the Social Network Aggression Task (SNAT; Achterberg et al., 2016; Achterberg et al., 2018). Using this method, it has been demonstrated that negative social feedback, compared to neutral or positive feedback, can lead to aggression in 7-9-year-old children (Achterberg et al., 2018; Achterberg et al., 2017; Dobbelaar et al., 2022), in 9-11-year-old children (Achterberg et al., 2020), in typically developing young adults (Achterberg et al., 2016; Van de Groep et al., 2021), and in young adults with a history of antisocial behavior (Van de Groep et al., 2022).

By extending the SNAT with fMRI measurements, researchers have investigated relations between social emotion regulation and neural (brain) responses, particularly in the AI, MPFC, and DLPFC brain regions. It has been shown that both positive and negative social feedback (comparedto neutral feedback) result in increased neural activation in the Anterior Cingulate Cortex (ACC) gyrus and bilateral AI (Achterberg et al., 2016; Achterberg et al., 2018; Achterberg et al., 2020; Dobbelaar et al., 2022; Van de Groep et al., 2021). These findings fit with the literature suggesting that the ACC and AI signal for social salience in general (Cheng et al., 2019; Dalgleish et al., 2017; Somerville et al., 2006). Moreover, the social salience networks reported in adults (Achterberg et al., 2016; Van de Groep et al., 2021), middle childhood (Achterberg et al., 2018; Dobbelaar et al., 2022) and late childhood (Achterberg et al., 2020) show remarkable resemblances, indicating that *on average* this mechanism is already developed in middle childhood. More importantly, variation in AI activation following negative social feedback has been related to variation in aggression regulation. That is, Achterberg et al. (2020) previously found that children with increased activation in the AI showed more aggression after negative social feedback. Interestingly, Chester et al. (2014) found a similar association, but only in adults with low executive control (and not in adults with high executive control). Possibly, the association between AI activation and behavioral aggression is stronger in childhood than adolescence, as executive control functions increase across development. The current study includes longitudinal measures across childhood and emerging adolescence, such that we can test developmental changes in brain-behavior associations.

Second, the MPFC has been shown to play an important role in social cognition and behavior (Adolphs, 2009; Blakemore, 2008), and is specifically implicated when thinking about others (Apps et al., 2016; D. Lee & Seo, 2016). Receiving negative social feedback may leave the children wondering what the other might have thought about them (Gallagher & Frith, 2003). Interestingly, when conducting whole brain analyses, previous studies often failed to find significant neural activation after negative social feedback (Gunther Moor et al., 2010; Guyer et al., 2012). However, studies with a larger sample, and increased statistical power, reported strong activation in the MPFC after social rejection in childhood (Achterberg et al., 2018; Achterberg et al., 2020). As social cognition and behavior are increasingly important during adolescence, activation in this region might show strong development—and strong individual differences in development—during the transition from childhood to adolescence. Previous studies did not reveal associations between aggression regulation following negative social feedback and MPFC activation. However these studies were often underpowered, examined group differences, and/or used aggregated scores (Chester, 2019).

Third, a brain-behavior association that has been consistently found using the SNAT is the negative association between DLPFC activation after negative social feedback and reactive aggression. That is, consistent with prior experimental studies, increased activation in the DLPFC after social rejection was followed by decreased aggression in adults, suggesting that these individuals were more successful at regulating their behavioral aggression (Achterberg et al., 2016; Riva et al., 2015). Region of interest analyses of the DLPFC in 7- to 9-year-olds provided some indications of an aggression regulation network, but this was not strong enough to be depicted using whole brain-behavior analyses (Achterberg et al., 2018). When examining these same children two years later—now during late childhood—there was a significant association between brain and behavior. Similar to adults, increased neural activation in the DLPFC was related to decreased behavioral aggression after negative social feedback (Achterberg et al., 2020). Importantly, the children who displayed the largest developmental increases in DLPFC activity across childhood also displayed the largest changes in behavioral aggression. These results suggest that, in addition to being an important region for cool (nonemotional) cognitive control (Crone & Steinbeis, 2017; Luna et al., 2004; Luna et al., 2010) the DLPFC is also important in controlling hot emotional control (Welsh & Peterson, 2014; Zelazo & Carlson, 2012). The current study expands this knowledge by examining functional DLPFC development across a broader age range, including emerging adolescence, and by including both linear and nonlinear development.

2.1.2 Study aims

The aim of this study is threefold. First, we describe developmental trajectories of neural and behavioral (aggressive) responses to social emotion regulation, *allowing for individual differences herein.* We focus specifically on the AI, MPFC, and DLPFC brain regions, as these have previously been related to the processing of social feedback. Second, we examine associations between the individual developmental trajectories of the behavioral and neural responses to social emotion regulation. Third, we test whether individual differences in developmental trajectories of brain and behavior across childhood (7- to 14-year-olds) are predictive for social well-being in (early) adolescence (12- to 15-year-olds). For readability, we discuss our usage of the Bayesian multilevel framework for the analyses only in general terms, and provide (technical) details, elaborate explanations, and R code in our online supplementary materials at https://jeroendmulder.github.io/social-emotion-regulation.

2.2 Methods

2.2.1 Participants and procedure

Participants in this study took part in the longitudinal twin study of the Leiden Consortium on Individual Development (L-CID; Crone et al., 2020). The procedures were approved by the Dutch Central Committee for Human Research and written informed consent was obtained from both parents. Invitations to participate were sent to families with same-sex twins born between 2006 and 2009, within a two-hour radius around the city of Leiden, the Netherlands. Participants were fluent in Dutch and were excluded when they had visual or physical impairments that could disable them from performing the behavioral tasks. The data were collected during annual visits between 2016 and 2021. Annual visits were either a home visit, in which families performed behavioral tasks at home, or a lab visit, in which families were invited to participate in an fMRI session. The sixth visit consisted of digital questionnaires that participants filled in at home. For the current study, data from the Middle Childhood Cohort collected at the lab visits during waves 1, 3, and 5, and the social well-being questionnaire at wave 6 were used. For details regarding the L-CID study and procedure, see Crone et al. (2020).

At wave 1 (first fMRI visit, September 2015 to August 2016), 512 children were included (7.02–9.68 years old, M = 7.94), with 55% being monozygotic. The majority of the sample (91%) was Caucasian and had normal IQ (M = 103.58, SD = 11.76), as measured using two subsets of the WISC (for details, see Achterberg et al., 2018). Socioeconomic status (based on parental education) was high for 45% of the sample, middle for 46%, and low for 9% of the sample (Crone et al., 2020). 489 children completed the fMRI scan at wave 1. At wave 3 (second fMRI visit, September 2017 to August 2018, 8.98–11.67 years old, $M = 9.98$), 456 participants were included, of whom 406 completed the fMRI scan. Wave 5 (third fMRI visit, September 2019 to April 2021, 11.15–14.11 years old, $M = 12.38$) included 336 participants, of whom 236 completed the fMRI scan. At wave 6 (June 2021 to October 2021, 11.98–15.10 years old, $M = 13.34$), 294 children filled in the digital social well-being questionnaires. Further details about the sample characteristics can be found in Table 1 of this study's preregistration (Achterberg et al., 2022), and in Dobbelaar et al. (2023).

2.2.2 Measurements

There are three outcomes of interest that were measured for this study: (a) Behavioral aggression following social feedback, measured simultaneously with the fMRI sessions at waves 1, 3, and 5; (b) neural responses in the AI, MPFC, and DLPFC following social feedback, measured at waves 1, 3, and 5; and (c) social well-being measured at wave 6.

2.2.2.1 Behavioral aggression following social feedback

Behavioral aggression after social feedback was measured using the Social Network Aggression Task (SNAT), which was programmed in Eprime, version 2.0.10.356 (see also Achterberg et al., 2016; Achterberg et al., 2018; Achterberg et al., 2020; Achterberg et al., 2017). One to four weeks prior to the fMRI session, participants filled in a personal profile at home, which was handed in at least one week before the actual fMRI session. The profile page consisted of questions such as: "What is your favorite color?", "What is your favorite food?", and "What is your biggest wish?". Participants were informed that their profiles were reviewed by other, unfamiliar, peers. During the SNAT the participants were presented with pictures and feedback to their personal profile from those unfamiliar peers. Unbeknownst to the participants, others did not judge the profile, and the photos were created by morphing two peers of an existing data base (matching the participants' age range) into a new, nonexistent peer. Every trial consisted of feedback from a new unfamiliar peer. This feedback could either be positive (visualized by a green thumb up), negative (red thumb down), or neutral (grey circle) as visualized in Figure 2.1. Peer pictures were randomly coupled to feedback, ensuring equal gender proportions for each type of feedback.

Following each peer feedback, the participants were instructed to send a loud noise blast to this peer. The longer they pressed the button, the more intense the noise would be, which was visually represented by a volume bar (Figure 2.1). To keep task demands as similar as possible between the conditions, participants were instructed to always press the button, but they could determine the intensity and duration of

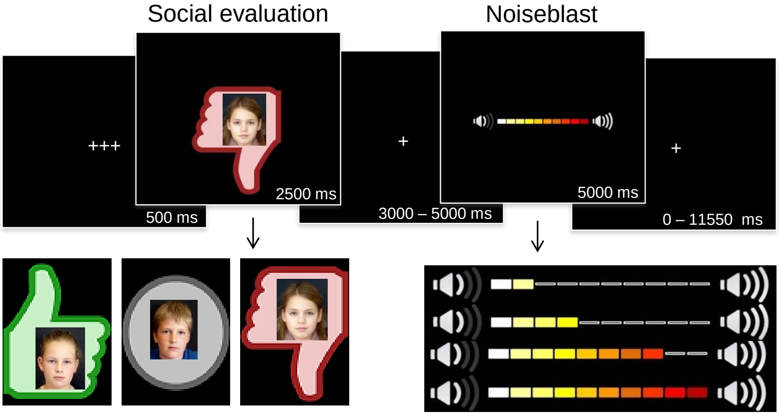


Figure 2.1: Visualization of the Social Network Aggression Task. After the participants viewed positive, neutral or negative social feedback on their personal profile, participants got the opportunity to blast a loud noise towards the peer, which was taken as a proxy for behavioral aggression following social evaluation.

the noise blast. Participants were instructed to deliver the noise blast by pressing one of the buttons on the button box attached to their legs, with their right index finger. As soon as the participant started the button press, the volume bar started to fill up with a newly colored block appearing every 350 ms. After releasing the button, or at maximum intensity (after 3500 ms), the volume bar stopped increasing and stayed on the screen for the remainder of the 5000 ms. The duration of the button press (in ms) to each negative, neutral, or positive trial was recorded and used as measurement of behavioral aggression in the statistical analyses (see Section 2.2.3). Participants were aware that the peers were not actually receiving the noise blast, but were instructed to respond as if the other peer would receive the noise blast.

The SNAT consisted of sixty trials (twenty per condition). An overview of trial order of the SNAT including jitter times is available at https://osf.io/ycgqe/. Each trial started with a fixation screen (500 ms), followed by the social feedback (2500 ms). After another jittered fixation screen (3000-5000 ms), the noise screen with the volume bar appeared, which was presented for a total of 5000 ms. Before the start of the next trial, another jittered fixation cross was presented (0-11550 ms; Figure 2.1). The order of trials was semirandomized to ensure that no condition was presented more than three times in a row. The optimal jitter timing and order of events were calculated with Optseq 2 (Dale, 1999). For each wave, the same version of the task was used. In the third fMRI wave we selected different photos of peers, such that they matched the age range of participants. For the current study, we specifically focused on noise blast duration after negative social feedback, compared to neutral social feedback.

2.2.2.2 Neural responses following social feedback

MRI scans were acquired with a Philips Ingenia 3.0 Tesla MR scanner. A standard whole-head coil was used, with foam inserts added to minimize head motion. A screen was placed behind the MRI scanner, such that participants could view the screen displaying the stimuli through a mirror on the head coil. T2*-weighted echo planar imaging (EPI) was used to collect the fMRI scans. The first two volumes were discarded to allow for equilibration of T1 saturation effects (field of view = 220 × 220 × 111.65 mm, TR = 2.2 s, TE = 30 ms, FA = 80°, sequential acquisition, 37 slices, voxel size = 2.75 × 2.75 × 2.75 mm). A high-resolution 3D T1 scan was collected as anatomical reference (field of view = 224 × 177 × 168 mm, TR = 9.72 ms, TE = 4.95 ms, FA = 8°, 140 slices, voxel size = 0.875 × 0.875 × 0.875 mm).

fMRI data were analyzed in SPM12 (Wellcome Department of Cognitive Neurology, London). Preprocessing included slice timing correction and correction for rigid body motion. Images were normalized to T1 templates (based on MNI-305 stereotaxic space; Cocosco et al., 1997) using 12-parameter affine transform mapping and nonlinear transformation with cosine basis functions. Volumes of each participant were resampled to 3 × 3 × 3 mm voxels and were spatially smoothed using a 6 mm full-width-at-half-maximum isotropic Gaussian kernel. Data of participants with at least two blocks of fMRI data with less than 3 mm movement in every direction were included in the analyses. Individual participants' data at each wave were analyzed using a general linear model in SPM12. The onset of feedback delivery was modeled as a zero duration event with positive, neutral and negative feedback added as separate regressors. To model the start of noise blast, the hemodynamic response function (HRF) was modeled for the length of the noise blast duration. Noise blasts following positive, neutral, and negative feedback were modeled as separate regressors (Achterberg et al., 2018). This study focuses specifically on the feedback event. Longitudinal trajectories of the noise blast event are described in Dobbelaar et al. (2023). Trials on which participants did not respond in time were marked invalid and excluded from further analyses. Six motion regressors were added as covariates of no interest. Least-squares parameter estimates of height of the best fitting canonical HRF for each condition were used in pairwise contrasts. The focus of this study was on the contrast negative versus neutral feedback.

Based on previous findings in an adult sample (N = 30, 18–30 years old) by Achterberg et al. (2016), the AI, MPFC, and right DLPFC were selected as regions of interest (ROI, see Figure 2.2). Parameter estimates were extracted using the MarsBar toolbox (Brett et al., 2002) for the contrast "negative feedback > neutral feedback", which was used as a measure of neural activity of social emotion regulation. These fMRI brain data analyses resulted in individual- and wave-specific contrast scores per ROI, representing the mean difference in brain activity between the negative and

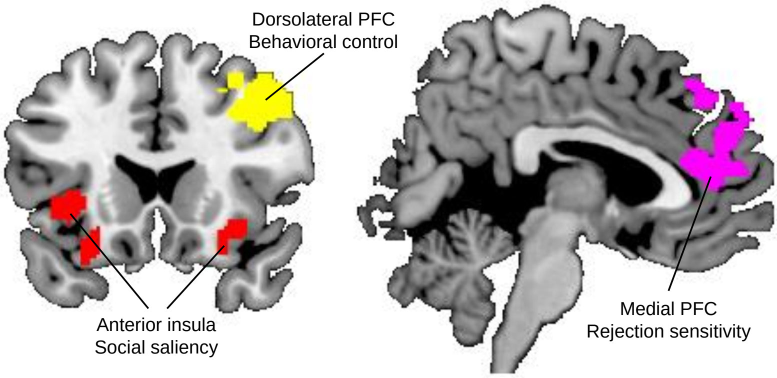


Figure 2.2: Regions of interest (ROIs) for the anterior insula (AI), the medial prefrontal cortex (MPFC), and the right dorsolateral prefrontal cortex (DLPFC). ROIs are available as .png, .nii, and .mat files at https://osf.io/byn7r/files/.

neutral social feedback conditions.

2.2.2.3 Social well-being questionnaire

The social well-being questionnaire was filled in by participants at wave 6 and consisted of 35 items. A complete overview of the questionnaire including all items and response categories is available at https://osf.io/fseq8/. It was constructed from five subscales: Ten items from the Adolescent Wellbeing Paradigm (AWP; Green et al., 2023), ten items from the World Health Organization Quality of Life Scale (WHOQoL; Vahedi, 2010), and three subscales (each five items) from the Harter's Self-Perception Profile for Adolescents (SPPA; Harter, 1988; Wichstraum, 1995), specifically the subscales Social Competence (SC), Close Friendships (CF) and Global Self-worth (GS). All items were answered on a four-point Likert scale, with low scores indicating low social well-being and high scores indicating high social well-being. Instructions in each of the subscale manuals were followed for the handling of missing data and scoring of subscale scores, resulting in simple mean scores per subscale.

2.2.3 Statistical analyses

In this section, the statistical analyses are described in general terms. For Aim 1—describing development in behavioral and neural responses to social emotion regulation, and individual differences herein—brain and behavioral data were analyzed with growth curve models in a Bayesian multilevel model framework. For Aims 2 and 3—investigating the relationships between individual development in behavioral responses, individual development in neural responses, and later social well-being—a structural equation modeling (SEM) approach was used. Technical details on these analyses (e.g., model equations, the fitting procedure, assessment of convergence and model fit), R code, and a rationale for the modeling decisions that were made, can be found in this study's online supplementary materials.

2.2.3.1 Bayesian multilevel growth curve models (Aim 1)

To describe developmental trajectories of (aggressive) behavioral and brain responses, growth curve models were fitted for each outcome in a Bayesian multilevel framework. The multilevel framework was used to allow for individual differences in the development of brain and behavioral responses, and to more easily accommodate the individual variation in age at each measurement wave (i.e., there is substantial variability in participants' age at each measurement occasion, see Section 2.2.1). The Bayesian framework was used because it is more flexible in accommodating some characteristics of the data, such as dropout of participants across time, censoring of the behavioral response data at 3500 (ms), and potential nonnormality. The models were fitted using the package brms (version 2.18.0; Bürkner, 2017) in R (version 4.2.2 R Core Team, 2022).

Analysis of the behavioral response data is discussed first. The data have a four-level structure, with the sixty repeated trials nested within three measurement waves, nested within individuals, nested within families. Using a multilevel model, we can estimate individual behavioral responses to social rejection at the trial level (level 1), model the development in these responses across a participant's age at the wave level (level 2), describe individual differences within families in this development at the individual level (level 3), and account for twin-dependence in the measurements at the family level (level 4). It is important to note that because our data is twin data, individual differences here are a combination of differences between individuals *within a given family/twin-pair* (level 3) and differences *between such families/twin-pairs* (level 4). For the current study, this differentiation is not of substantive interest, and is only made to control for the nonindependence of observations in our statistical analyses.

From a multilevel model we can extract various components relevant for Aim 1. The model's *fixed effect* (FE) parameters capture *average* change, that is, averaging across individuals within families (level 3) and across families (level 4), does an individual's behavioral response to social rejection change as the individual's age increases? Because behavioral change across time is hardly ever linear, we include FE parameters for both linear and quadratic changes across time. In total, three FE parameters from the model are of interest: An intercept, which captures the expected behavioral response to negative feedback (compared to the neutral condition) at the mean age (approximately nine years and nine months)[1], an expected linear slope

[1] Because the variable *time* was grand mean centered before use in the multilevel model, the growth

in behavioral response at mean age, and an expected quadratic slope in behavioral response at mean age. From hereon we jointly refer to the intercept and slopes as *growth components*.

The multilevel model contains random effect (RE) terms for the growth components at the individual level and the family level. The inclusion of these terms in the model implies that the estimated development (as captured by the growth components) can vary from individual to individual within a family (i.e., through the RE terms at level 3), and between families (i.e., through the RE terms at level 4). Standard deviations of the RE terms are then measures of across adolescent (but within family) and between-family variability in the development of behavioral and neural responses, respectively. By extracting RE terms for each individual and family, we can create individual-specific growth components. These components serve as input for the second-part of the data analysis (see Section 2.2.3.2).

The analysis procedure of the neural responses was largely similar to the analysis procedure as just-described for the behavioral responses: For each ROI, growth curve models were fitted in a Bayesian multilevel framework, and FEs (averaged across individuals and families) and REs (both individual- and family-specific) of the growth components were extracted herefrom. There was one notable exception. As described in Section 2.2.2, preprocessing of the fMRI data resulted in contrast score averages across trials rather than trial-specific scores. Hence, for the fMRI data, the neural responses to social rejection (compared to the neutral condition) do not have to be estimated anymore as part of the multilevel model. Therefore, for the fMRI data, a three-level multilevel model was used in which the trial level was omitted.

The Bayesian framework was used to handle multiple complicating factors of the data. First, it accommodated censoring in the behavioral data (at 3500 ms) by integrating censored values out. Second, to prevent unnecessary loss of data, missing data for the outcomes were imputed as part of the model fitting procedure under the assumption of missing at random. Third, because the data showed increased kurtosis, a Student t distribution was used for the outcome to increase model fit (compared to assuming a Gausian-distributed outcome). Ultimately, a Bayesian fitting procedure does not result in a single point estimate of the model parameters, but rather in a distribution of likely values for each parameter (i.e., the posterior distribution). We specified the Bayesian fitting procedure such that it resulted in a thousand sets of plausible values for individual-specific growth components for each outcome. These data sets were used as input for the structural equation model for investigating Aims 2 and 3.

Finally, the above-described analysis procedure deviates slightly from the preregistered analyses. Initially, we described the use of intercept-only models to compute

components represent development at the mean age of participants. This was done to prevent issues with multicollinearity of the linear and quadratic growth components in the model.

the intraclass correlations (ICC) for each level. Based on the ICC, we could decide to remove the family level if no substantial amount of variance was captured at this level, making the multilevel models considerably simpler. However, these analyses were not performed because a three- or four-level structure, albeit complex, is theoretically fitting for the brain and behavioral data, respectively. Moreover, by fitting multilevel models with the same number of levels the same analysis procedure (i.e., extracting the REs, combining them, etc.) could be used for all brain areas.

2.2.3.2 Structural equation model (Aims 2 and 3)

Structural equation modeling was used to estimate the associations between the individual-specific growth components of the behavioral and neural responses (Aim 2), and predict later social well-being (Aim 3). First, a one-factor confirmatory factor analysis was performed on the five social well-being subscale means. If the one-factor model for social well-being showed good model fit, we would use the growth components to predict a common social well-being factor. If the one-factor model showed bad model fit, the growth components would predict each of the social well-being subscales separately.

Second, a multivariate regression model was specified with a latent social well-being factor (or the subscales separately) as the outcome(s), and the estimated growth components from the Bayesian multilevel model as predictors. In this model, the predictors were allowed to covary freely with each other such that associations between development in behavioral responses and development in neural activation could be estimated (Aim 2). The regression coefficients represent the relationship between development in social emotion regulation and social well-being (subscales) in adolescence (Aim 3). The models were fitted using the R package lavaan (version 0.6.16; Rosseel, 2012).[2]

As explained in Section 2.2.3.1, a thousand sets of plausible values for the growth components were extracted from the Bayesian multilevel model. Hence, the multivariate regression model was fitted a thousand times, once for each set of plausible values. This was done using the R package semTools (version 0.5.6; Jorgensen et al., 2022). Parameter estimates of the thousand fitted SEM models were averaged to create a single point estimate of the associations amongst the growth components, and their relation with later social well-being (either a single common social well-being factor, or its five separate subscales). Standard errors for these parameters were pooled following the rules by Rubin (1987).

[2] In contrast to the preregistration, we did not use the software package Mplus. This decision was made for convenience as R packages are freely and openly available.

2.3 Results

In this section we present model results that are directly related to this study's Aims. The full set of numerical results can be found in the online supplementary materials.

2.3.1 Individual differences in development of neural and behavioral responses (Aim 1)

For our first aim, Bayesian multilevel growth curve models were fitted to the behavioral an brain data. Table 2.1 contains the 95% credible intervals for the FEs of the growth components, and standard deviations of the REs (at both the individual- and family-level) of the growth components. Results for the FEs are also visualized in Figure 2.3, which shows the model-predicted development across adolescence for behavioral aggression and the neural responses in the AI, MPFC, and DLPFC.

For behavioral aggression, results show that there is 95% certainty that the expected behavioral response at the mean age (approximately nine years and nine months) lies between 1.31 and 1.49 seconds. The REs imply that there is evidence of differences between individuals (within families) herein—with the standard deviation of the RE at the individual level estimated to be between 0.03 and 0.42—as well as differences between families—with the standard deviations of the REs at the family level estimated to be between 0.15 and 0.43. Linear development of behavioral aggression at the mean age is estimated to lie between -0.05 and 0.02, with the standard deviation of the RE estimated to lie between 0.00 and 0.11 for the individual level, and between 0.01 and 0.16 at the family level. This implies that there is no, to little evidence of differences between individuals and families in linear development, respectively. Quadratic development at the mean age is estimated to be slightly negative, lying between -0.06 and -0.03. This implies that expected development herein follows an inverted-U shape, with behavioral aggression following negative feedback peaking in late childhood, and decreasing thereafter. There is no evidence of between-individual or between-family differences herein.

Table 2.1: 95% credible intervals for the fixed effects (FEs) and the standard deviations of the random effects (REs) of the growth components. REs exist at both the individual-level (i.e., within families) and the family-level (i.e., between families). Results are shown for development in behavioral aggression, and neural responses in the anterior insula, media prefrontal cortex, and dorsolateral prefrontal cortex. The asterisk * denotes credible intervals not containing zero.

	FE	*SD*(RE) individual-level	*SD*(RE) family-level
Aggression (noise)			
Intercept	[1.31, 1.49]*	[0.03, 0.42]*	[0.15, 0.43]*
Linear slope	[-0.05, 0.02]	[0.00, 0.11]	[0.01, 0.16]*
Quadratic slope	[-0.06, -0.03]*	[0.00, 0.03]	[0.00, 0.05]
AI			
Intercept	[0.42, 1.03]*	[0.02, 1.01]*	[0.01, 0.81]*
Linear slope	[-0.29, -0.02]*	[0.03, 0.66]*	[0.01, 0.35]*
Quadratic slope	[-0.01, 0.13]	[0.00, 0.17]	[0.00, 0.14]
MPFC			
Intercept	[0.52, 1.20]*	[0.03, 1.37]*	[0.13, 1.62]*
Linear slope	[-0.17, 0.10]	[0.01, 0.56]*	[0.00, 0.31]
Quadratic slope	[-0.01, 0.13]	[0.00, 0.23]	[0.00, 0.27]
DLPFC			
Intercept	[-0.65, -0.09]*	[0.02, 0.97]*	[0.01, 0.84]*
Linear slope	[-0.14, 0.11]	[0.02, 0.53]*	[0.01, 0.43]*
Quadratic slope	[-0.02, 0.11]	[0.00, 0.17]	[0.00, 0.16]

For neural responses in the AI, the REs show that there is 95% certainty that expected response at the mean age lies between 0.42 and 1.03. The REs imply that there is evidence of between-individual (within-families), and between-family differences herein. Linear development of AI response at the mean age is estimated to lie between -0.29 and -0.02, with results indicating some evidence of differences between individuals and between families herein. Quadratic development at the mean age is estimated to lie between -0.01 and 0.13, implying that there is no evidence of a quadratic trend in AI development across adolescence. Additionally, there is no evidence of differences between individuals or between families herein. Thus, in general we found evidence for increased AI activity following social rejection, and a linear decrease herein (but no quadratic development). Furthermore, results also show individual differences in linear development.

For neural responses in the MPFC, the results for the REs show that there is 95% certainty that expected response at mean age lies between 0.52 and 1.20. The REs imply that there are significant differences between individuals (within families) herein, as well as significant between-family differences. Linear development of MPFC at mean age is estimated to lie between -0.17 and 0.10, with only marginal evidence of

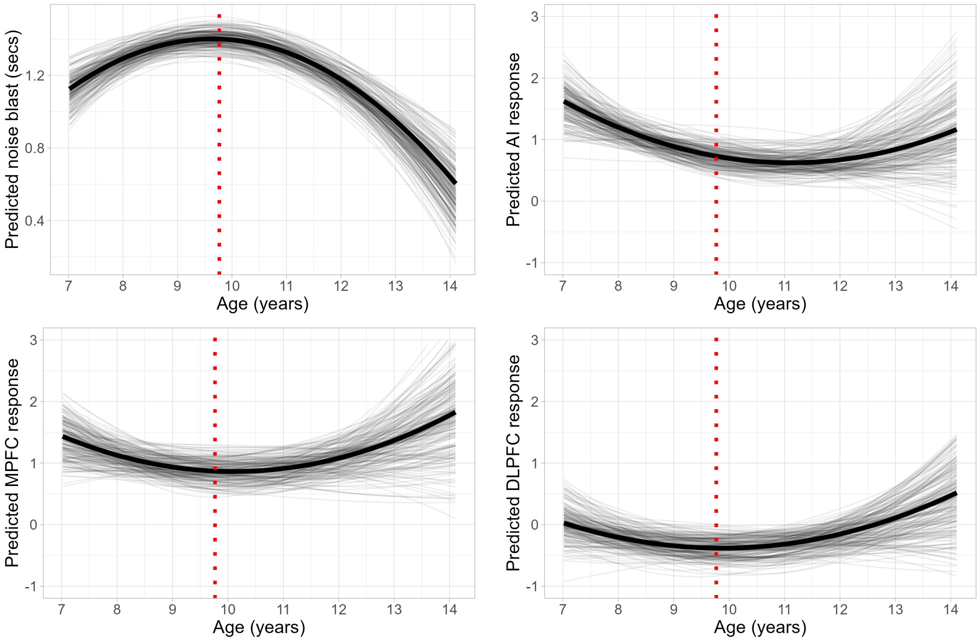


Figure 2.3: Predicted development in behavioral and neural responses to social emotion regulation (i.e., negative feedback versus neutral feedback). The bold black line represents the predicted (based on the REs) average development across adolescence. The gray lines represent uncertainty around this prediction, based on draws from the posterior distribution for the REs. The vertical dotted (red) line represents the mean age of approximately 9 years and 9 months.

differences between individuals in linear development, and no evidence of differences between families. Quadratic development at mean age is estimated to lie between -0.01 and 0.13. Results show no evidence of differences between individuals herein. Thus, in general we found evidence for increased MPFC activity following social rejection, but no overall (linear or quadratic) development herein. However, results do show individual differences in linear development (i.e., for some individuals there is a positive linear development, and for some a negative).

Finally, for neural responses in the DLPFC, results for the REs show that there is 95% certainty that expected response at mean age lies between -0.65 and -0.09. The REs provide only marginal evidence that are between individual-, and between family differences families differences herein. Linear development at mean age is estimated to lie between -0.14 and 0.11, with again only marginal evidence of between individual and between family differences. The results show no evidence for a significant quadratic trend on average for the development of DLPFC responses, and do not suggest differences between individuals or between families. Thus, in general we found evidence for decreased DLPFC activity following social rejection, but no overall (linear or quadratic) development herein. However, results do show individual differences in linear development.

2.3.2 Associations between growth components of behavioral and neural responses (Aim 2)

For Aims 2 and 3, a single multivariate regression model was fitted in which the growth components predicted later social well-being (of interest for Aim 3), and the predictors freely covaried with each other (of interest for Aim 2). In total, 66 covariances between the individual-level growth components of neural and behavioral responses were estimated. Of these, two covariances were significant at the $\alpha < .05$ significance level. The covariance between the expected AI response at mean age (intercept) and the linear development in AI response at mean age (linear slope) was estimated to be -0.123, $SE = 0.054$, $t(978.689) = -2.280$, $p = .023$. This implies that individuals with a higher AI response at mean age tend to have a steeper linear decrease in AI response. Furthermore, the covariance between the expected MPFC reactivity at mean age (intercept) and the quadratic slope of MPFC at mean age was estimated to be -0.111, $SE = 0.051$, $t(1012.368) = -2.194$, $p = .028$. This implies that individuals with a higher expected MPFC at mean age also show a less curvilinear (i.e., more linear) development.

Parameter	Est.	SE	95% CI
Factor loadings:			
λ_{AWP}	1	-	-
λ_{WHO}	0.996	0.057	[0.884, 1.108]
λ_{SC}	0.766	0.096	[0.578, 0.954]
λ_{CF}	0.628	0.092	[0.448, 0.808]
λ_{GS}	1.161	0.099	[0.967, 1.355]
Unique variances:			
θ_{AWP}	0.051	0.007	[0.037, 0.065]
θ_{WHO}	0.019	0.006	[0.007, 0.031]
θ_{SC}	0.299	0.026	[0.248, 0.350]
θ_{CF}	0.281	0.024	[0.233, 0.328]
θ_{GS}	0.275	0.025	[0.226, 0.324]
Common variance:			
ψ	0.141	0.017	[0.108, 0.174]

Table 2.2: Parameter estimates of the measurement model of social well-being. The factor loading of the first indicator was set to one for scaling. AWP = ten items from the Adolescent Wellbeing Paradigm; WHO = ten items from the World Health Organization Quality of Life Scale; SC = Social Competence subscale from Harter's Self-Perception Profile for Adolescents; CF = Close Friendships subscale of Harter's Self-Perception Profile for Adolescents; GS = Global Self-Worth subscale of Harter's Self-Perception Profile for Adolescents.

2.3.3 Prediction of social well-being subscales (Aim 3)

The exact specification of social well-being in the multivariate regression model was determined based on how well the social well-being subscales could be represented as a unidimensional construct. To this end, a one-factor confirmatory factor analysis model was fitted to the five subscale measures. Estimates of the factor loadings λ, unique subscale variances θ, and the common social well-being factor variance ψ are reported in Table 2.2. This unidimensional structure for the social well-being subscales showed substantial misfit to the data, $\chi^2(5) = 49.269$, $p < .001$, $CFI = .922$, $TLI = .844$, $RMSEA = 0.178$. Therefore, subscales were included as separate outcomes in the multivariate regression model rather than a common social well-being factor. Such a model, in which the exogenous predictors freely covary amongst each other, and in which all outcome residuals also freely covary amongst each other, is saturated, implying perfect fit.

None of the growth components significantly predicted any of the social well-being subscales.

2.4 Discussion

The regulation of negative emotions during social interaction is an essential quality for developing and maintaining social relations, and there are many individual differences in how children deal with social rejection. Although prior literature has linked the development of social emotion regulation to changes in behavioral (aggressive) responses and neural activation, previous literature has mostly focused on group-based aggregates, limiting our knowledge on individual differences in development (Chester, 2019). Complementing existing research, this preregistered study focuses on the development of behavioral aggression and neural responses in the AI, MPFC, and DLPFC following social rejection, and places individual differences in such development front and center. The renewed focus on individual variability endorses the fact that adolescents' behavioral and neural responses to social interaction develops in meaningfully different ways (Foulkes & Blakemore, 2018; Telzer et al., 2018), and allows for investigating if such individual differences are predictive of, for example, future health outcomes (Copeland et al., 2013; Foulkes & Blakemore, 2018; Van Harmelen et al., 2017). In this study, we made use of data of the Leiden Consortium on Individual Development (Crone et al., 2020), which is a longitudinal (experimental) data set containing neural (fMRI) and behavioral measurements following social interaction (for more information, see https://www.developmentmatters.nl/). To describe linear and quadratic development of behavioral and neural responses, as well as individual differences herein (Aim 1), we fitted Bayesian multilevel growth curve models. Results from the multilevel models served as input for a structural equation model, in which we simultaneously investigated intercept-slope associations among brain and behavioral development (Aim 2), and whether or not individual behavioral and neural development could predict social well-being (Aim 3). All research aims and analyses were preregistered in Achterberg et al. (2022). Technical details and R code for the analyses can be found in the online supplementary materials at https://jeroendmulder.github.io/social-emotion-regulation.

The main findings of this study are threefold: First, average behavioral development was found to be nonlinear (quadratic), with a peak in behavioral response during late childhood. Individual differences were found primarily in the intercept (expected behavioral response at mean age) and to a lesser degree in the linear slope. Secondly, in line with our expectations, we found individual differences in the linear development of neural responses to social rejection. Third, we did not find associations between the estimated individual trajectories of brain and behavioral response, nor were these estimated individual trajectories predictive for future self-reported social well-being. Below, we discuss the theoretical and methodological implications of these main findings further.

2.4.1 Late childhood as a sensitive window for social emotion regulation

Behaviorally, we found that social emotion regulation (as measured by aggression following negative versus neutral feedback) peaks during late childhood. The REs in the multilevel models described general linear and quadratic development at the mean age of approximately nine years and nine months. Based on the estimated standard deviations of the REs, we found evidence for individual differences in the intercept (i.e., expected behavioral response at mean age). Note that here, individual differences are a combination of both differences within- families at the individual level, and between-families at the family level. Furthermore, there was some evidence for individual differences in linear slope between families, but these effects were less pronounced. This suggests that children may differ in their response to rejection in late childhood, but that the developmental trajectories (i.e., a peak in aggression in late childhood) are relatively similar between children. Although most prior developmental studies have focused on adolescent specific peaks in social behavior (cf. Brechwald & Prinstein, 2011; Casey et al., 2010; Somerville & Casey, 2010; Steinberg, 2008; Steinberg & Morris, 2001), our results suggest that late childhood is also an important period for social development, specifically for dealing with social rejection. Prior work on reactive aggression also reported a peak in late childhood (Cui et al., 2016), with decreases in aggression towards adolescence (Fite et al., 2008). This peak in aggression in late childhood may be explained by delayed development of inhibition of aggression following negative feedback, compared to inhibition of aggression following neutral feedback (Dobbelaar et al., 2023). However, although social rejection is a challenging experience for all children, there are pronounced differences in how children deal with such rejection. While some socially rejected children suffer from widespread and persistent impairments in mental health (i.e., internalizing and externalizing problems; Ladd, 2006; Prinstein & Aikins, 2004; Prinstein & La Greca, 2004), other children seem more resilient in dealing with social rejection (Ioannidis et al., 2020; Van Harmelen et al., 2021). Until now there was little insight on where in the developmental process these individual differences emerge. Our findings add to the existing literature by providing evidence for individual differences during late childhood. Possibly, the peak in aggression following negative feedback during late childhood, and individual differences herein, suggests an undiscovered sensitive period in development. This sensitive window might provide a window of opportunity for interventions that foster social development in youth.

2.4.2 Individual differences in the linear development of neural responses to social rejection

With regards to overall development of neural responses, the results provide evidence of a negative linear development in the AI. This implies that, in general, the AI response to social rejection is expected to decrease between ages nine and ten which levels off again in emerging adolescence. Additionally, in line with earlier empirical and theoretical studies, we report evidence for individual differences in linear development of all ROIs (Bottenhorn et al., 2023; Foulkes & Blakemore, 2018). Furthermore, neural sensitivity to social feedback may be shaped by social experiences (Rudolph et al., 2021), that can substantially differ between individuals. However, very few studies have investigated brain development across childhood. The main reason for this is that scanning children is more challenging than scanning adolescents or adults (Achterberg & Van der Meulen, 2019; O'Shaughnessy et al., 2008). Nevertheless, our findings indicate that there is evidence of individual differences in brain development during childhood, and highlight that future studies should also include participants below the age of twelve. Notably, we did not find evidence of quadratic trends in the developmental trajectories, nor in general, nor at an individual level. Prior studies have suggested nonlinear development across puberty and adolescence and our results add to this literature by showing that functional brain development across childhood seems mostly linear (Gracia-Tabuenca et al., 2021; Vijayakumar et al., 2019).

2.4.3 Testing brain-behavior associations: Methodological considerations

We did not find evidence for associations between the estimated growth components of behavioral and neural responses themselves (Aim 2), nor were we able to predict future social well-being from the individual growth components (Aim 3). That is, our data analysis did not provide any evidence that these individual differences in development are meaningfully related to each other, or to future social well-being. This stands in contrast to what is described in the literature as previous studies based on (parts of) the same data and/or experiment report significant brain-behavior associations (cf. Achterberg et al., 2016; Achterberg et al., 2020; Dobbelaar et al., 2022; Dobbelaar et al., 2023; Van de Groep et al., 2021). For example, it was found that behavioral aggression regulation across time was associated with DLPFC activation across time (Achterberg et al., 2020).

There are a couple of potential explanations for this seeming discrepancy. First, this research project is ambitious in its scope, and utilized a complex study design (e.g., longitudinal twin data, in which individuals inevitably drop out, and in the presence of censoring). Our specific setup therefore requires a large number of individuals and repeated measures in order to achieve adequate statistical power. While this study is amongst the first in the literature to attempt to collect repeated MRI data in children at this scale, the sample size might still be too small to detect the many, and arguably small neural relationships that are targeted here (Marek et al., 2022). Second, the statistical analyses in this study deviate in some important ways from previous studies into this topic. The deviations concern the handling of missing data, censoring in the data, and individually-varying times of observations of participants. Such methodological and statistical differences between studies can lead to differences in results, and consequently differences in conclusions that are drawn. This underlines the importance of making informed decisions about the methodological and statistical choices that researchers have apriori, and recording these in a preregistration, or even better, a registered report, and with the inclusion of extensive peer reviewing. It is also important to engage in team science, with interdisciplinary collaborations on research projects to get different perspectives on the subject-matter and analysis strategy (Fair et al., 2021).

2.5 Conclusion

Dealing with social rejection, or negative peer feedback, can be challenging, specifically for children as their social emotion regulation is still developing. Prior research has focused largely on group-based averages of this development, obscuring meaningful individual variation in development. Here, we employed a Bayesian multilevel modeling framework to describe individual differences in the development of behavioral and neural responses to negative social feedback. We found a slight peak in behavioral social emotion regulation development across late childhood, as well as individual differences during this developmental phase. Moreover, we report evidence for individual differences in the linear development of neural responses to social rejection in our three brain regions of interest: the AI, MPFC and DLPFC. Our follow-up analyses did not provide evidence for associations between individual trajectories of brain and behavior, or later social well-being. In addition to providing insights in the individual trajectories of social emotion regulation during childhood, this study also makes a meaningful methodological contribution. That is, our statistical analysis strategy can be used as an example of how to take into account the many complexities of developmental neuroimaging datasets, while still enabling researchers to answer interesting questions about individual-level relationships.

Acknowledgements: Prof. dr. Marinus van IJzendoorn, prof. dr. Eveline Crone, and prof. dr. Marian Bakermans-Kranenburg designed the Leiden Consortium on Individual Development (L-CID) experimental cohort-sequential twin study "Samen

Uniek" as part of the Consortium on Individual Development. We thank the participating families for their enthusiastic involvement in L-CID. We are grateful to the data-collection and data-processing team, including all current and former students, research assistants, PhD students and postdoctoral researchers for their dedicated and invaluable contributions. We also thank Eveline Crone and dr. Stephan Heunis for their help in the preregistration of this study.

Author contributions: Conceptualization: JM and MA - Data curation: JM and MA - Formal analysis: JM – Investigation: MA and SD - Methodology: JM, SD, and MA – Project administration: JM and MA - Resources: MA, JM, and SD - Software: JM - Visualization: JM and MA - Writing (original draft preparation): JM and MA – Writing (review and editing): JM, SD, and MA

Funding: This study is part of the Consortium on Individual Development, which was supported by a Gravitation grant awarded by the Dutch Ministry of Education, Culture, & Science, and the Netherlands Organization for Scientific Research [NWO grant number 024.001.003; 2013].

Supplementary materials: This study's online supplementary materials can be found at https://jeroendmulder.github.io/social-emotion-regulation.

Preregistration: This study was preregistered as Achterberg et al. (2022).

CHAPTER 3

Predicting outcome of an intensive outpatient PTSD treatment program using daily measures

Adapted from:

Alting van Geusau, V. V. P., Mulder, J. D, & Matthijssen, S. J. M. A. (2021). Predicting outcome in an intensive outpatient PTSD treatment program using daily measures. *Journal of Clinical Medicine*, *10*(18), 4152. https://doi.org/10.3390/jcm10184152

Abstract

The Altrecht Academic Anxiety Center has developed a new intensive six-day treatment program for patients with posttraumatic stress disorder. Due to the high dropout rates in trauma treatments generally, and high costs of the newly-developed treatment program, clinicians were interested in predicting treatment outcomes after completion of the program from development of patients during the program. The current study investigates daily treatment progress as a predictor for treatment success at four-week follow-up. Data from 109 PTSD-patients (87.2% female, mean age = 36.9, *SD* = 11.5) were used. PTSD symptoms were measured with the CAPS-5 and the self-reported PTSD checklist for DSM-5 (PCL-5). Daily PTSD symptoms were measured with an abbreviated version of the PCL-5 (8-item PCL). Multiple latent growth curve models were used to describe changes in daily PTSD symptoms and predict treatment outcome. Cross-validation was used to compare the prediction performance (in terms of mean square error) these models. Overall, results showed that a greater decline in daily PTSD symptoms measured by the 8-item PCL predicts better treatment outcome (CAPS-5 and PCL-5), but that a patient's PTSD symptoms on the first day of treatment has no predictive effect. A decline in PTSD symptoms only during the first half of treatment was also found to predict treatment outcomes.

3.1 Introduction

Posttraumatic stress disorder (PTSD) is a stress-related disorder that one can develop after being exposed to one or more traumatic events (American Psychiatric Association, 2013). The lifetime prevalence of PTSD is around 7.4–8% (De Vries & Olff, 2009; Kessler et al., 2012). According to multidisciplinary guidelines, there are several evidence-based treatments for PTSD (American Psychological Association, 2017; International Society of Traumatic Stress Studies, 2018). Among these are eye movement desensitization and reprocessing (EMDR), trauma-focused cognitive behavioral therapy (TF-CBT), prolonged exposure (PE), and cognitive processing therapy (CPT) which all show good effect sizes in reducing PTSD symptoms (Lewis et al., 2020). EMDR therapy seems to be the most cost-effective treatment (Mavranezouli et al., 2020).

Although there are effective trauma treatments, dropout rates are often high. In a meta-analysis by Imel et al. (2013), an average dropout rate of 18% was found among active treatments in clinical trials for PTSD, but dropout rates as high as 54% are reported in some studies (Schottenbauer et al., 2008). Furthermore, of all the patients who complete treatment, 30–50% still show symptoms (Bradley et al., 2005). Therefore, there is much room for improvement. A first step would be to find out who is likely to benefit from treatment and who is not, and to see if treatment success can already be predicted in an early phase. If so, practitioners may decide to scale up or alter treatment during early stages of treatment, which would prevent patients from undergoing treatment that is predicted to have little effect in the long term.

In some studies, factors related to treatment outcome for psychotherapeutic interventions for PTSD were identified, including comorbidity, cognitive dimensions, suicide risk, and characteristics of the patient such as gender (Ehlers et al., 1998; Forbes et al., 2003; Tarrier et al., 2000). Results of a study investigating predictors of treatment outcome and dropout in two samples of PTSD patients who were treated with PE, showed that higher PTSD symptom scores at pretreatment were correlated with more PTSD symptoms at posttreatment and at follow-up (Van Minnen & Hagenaars, 2002). Another study found that lower pretreatment *clinician-rated* PTSD symptoms were associated with better treatment outcomes, whereas higher baseline *self-rated* PTSD symptoms were associated with better treatment outcomes (Karatzias et al., 2007). In one study, indications were found that benzodiazepine use was related to worse treatment outcomes, and alcohol use was related to increased dropout rates. However, demographic variables; depression; general anxiety; personality pathology; trauma characteristics; feelings of anger, guilt, and shame; and nonspecific variables regarding therapy were not related to either treatment outcomes or dropping out (Van Minnen & Hagenaars, 2002). The result that the use of benzodiazepines was related to worse PTSD psychotherapy outcomes has also been found in a meta-analysis (Guina et al., 2015). Although in some studies factors were identified that were related to treatment outcomes, in other studies contradictory results were found. Hence, there are not many clear, convincing and reliable pretreatment predictors for treatment outcomes, a result that has also been found in other studies (Ehlers et al., 2013; Karatzias et al., 2007). However, how about predictors during treatment? Is it possible to predict treatment outcomes during early stages of treatment?

Several factors during treatment have been shown to predict treatment outcome. A strong therapeutic alliance has been linked to better treatment outcome in psychotherapeutic interventions (Horvath & Symonds, 1991; Martin et al., 2000). Between-session habituation has been identified as a predictor for treatment outcomes in PE treatment programs, with patients who showed more between-session habituation being more likely to show better treatment outcomes (Cooper et al., 2017; Hendriks et al., 2018). It has also been shown that trauma-related belief change predicted subsequent PTSD-symptom change in PE (Cooper et al., 2017). Higher fear activation during the first session of PE, as measured with subjective units of distress (SUDs) and facial expression, was found to be correlated with better treatment outcome (Foa et al., 1995). Higher emotional engagement during PE in the first session, as measured with SUDs, predicted better treatment outcomes (Jaycox et al., 1998). For EMDR, it was found that lower SUD scores at the end of the first session predicted better treatment results (D. Kim et al., 2008).

In identifying predictors of treatment outcome, one could argue that it is clinically relevant to identify treatment response in an early stage in order to be able to adjust treatment strategies when deemed necessary. In a study examining PE effects for PTSD symptoms of veterans of the war in Iraq, the greatest reduction in symptoms was found in the first five sessions (Tuerk et al., 2011). In another study, comparing EMDR to brief eclectic psychotherapy, it was also found that the largest reduction in PTSD symptoms was achieved in the first five sessions in the EMDR condition (Nijdam et al., 2012). However, only a few researchers have studied whether the early response progress predicts posttreatment outcomes. For example, one study found that PTSD patients receiving PE or CPT who did not improve much after the first eight sessions were not likely to improve much subsequently (Sripada et al., 2020). Another study found that the probability of achieving meaningful symptom amelioration decreased after every session for patients receiving CPT, indicating that patients who show little PTSD-symptom change during early stages of treatment are likely to show worse overall treatment outcomes (Byllesby et al., 2019).

The present study aims to respond to the limited evidence for early treatment response as a predictor for treatment success. Treatment for PTSD is commonly delivered once or twice a week over the course of several months. Since PTSD interferes with social and occupational functioning (Ehlers et al., 2014), it is desirable for patients to make rapid progress. Several intensive treatment programs have been set up, with good results and significantly lower dropout rates of below 10% (Ragsdale et al., 2020). The current study aims to determine the predictive value of treatment response on treatment outcome in such an intensive treatment program, which consists of two weeks of treatment for three consecutive days each. Patients receive three hours of trauma therapy (PE and EMDR), one hour of physical activity and one hour of psychoeducation every day. The results of a meta-analysis showed that adding physical activity to usual care improved the health of PTSD patients, and was effective in decreasing PTSD symptoms (S. Rosenbaum et al., 2015). A combination of PE, EMDR, physical activity and psychoeducation in an inpatient intensive treatment program was found to be effective in reducing PTSD symptoms (Van Woudenberg et al., 2018).

The goal of the present study is to investigate whether change in PTSD symptomatology *during* the current intensive treatment program predicts PTSD reduction four weeks *after* completions of the program. It is expected that a greater decline in PTSD symptoms—based on development during the entire two-week treatment program—predicts greater PTSD symptom reduction four weeks after treatment completion. Moreover, it is expected that a greater decline in PTSD symptoms—based on development during only the first half of the treatment program—also predicts greater PTSD symptom reduction four weeks after treatment completion.

3.2 Materials and methods

3.2.1 Participants

The current study used a self-select sample comprised of 109 PTSD patients who participated in the program between April 2018 and November 2019 (from hereon referred to as participants). The mean age was 36.9 years old ($SD = 11.5$), ranging from 20 to 64 years; 14 identified as male (12.8%), 95 as female (87.2%). The trauma types of participants varied (e.g., sexual abuse, physical abuse, and accidents). Inclusion criteria were (a) having a PTSD diagnosis according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5); (b) having experienced multiple traumatization; (c) no alcohol or drug use during treatment; (d) no acute suicidality risk; (e) sufficient proficiency in the Dutch language; (f) the absence of comorbid psychiatric disorders that would seriously interfere with treatment; and (g) no, or in exceptional cases, minimal, use of sedating medication during treatment (for example, participants prone to mania who would be deprived of sleep without sleeping medication were allowed to continue their medication).

3.2.2 Procedure

The intensive trauma treatment program was provided at the Altrecht Academic Anxiety Center, a center specialized in the treatment of severe anxiety disorders, obsessive-compulsive disorder, and trauma-related disorders in Utrecht, the Netherlands. Prior to starting treatment, participants were screened on diagnoses and inclusion criteria, and an individual treatment plan was made. Participants were asked to select the six subjectively most disturbing traumatic memories for treatment. Each day, one memory would be treated.

3.2.3 Treatment

The treatment program consisted of two consecutive weeks, with treatment provided on Tuesday, Wednesday and Thursday in an outpatient setting (i.e., totalling six days of treatment). Each treatment day had the same outline (except for days 1 and 6, when participants filled out additional measurements, see Section 3.2.4 and Table 3.1). Treatment consisted of two evidence-based treatments for PTSD: PE and EMDR. For the PE sessions, a slightly modified version of the PE protocol was used, where participants did not make audio recordings to listen to as homework in between sessions (Foa et al., 1995). EMDR was delivered according to standard protocol (Jongh & Broeke, 2020; Shapiro, 2018). The combination of these treatments was used because PE and EMDR supposedly differ in underlying working mechanism, and for this reason could possibly complement each other in treatment effect. It has also been found that these treatments can be successfully combined (Van Minnen et al., 2020). Participants received PE in the morning, and EMDR in the afternoon. It has been shown that this sequence resulted in better treatment outcomes than the reversed sequence (Van Minnen et al., 2020). Treatment was delivered with therapist rotation to ensure that participants were treated by many different therapists, and that the therapists had daily multidisciplinary meetings in between sessions to ensure treatment was given according to protocol. It has been suggested that therapists' shared responsibility for treatment leads to better implementation thereof due to decreased therapist drift, and reduced negative concerns by therapists (Van Minnen et al., 2018). In between PE and EMDR sessions, participants conducted physical activity: Either trauma-sensitive yoga (Emerson et al., 2009; Nolan, 2016), walking or jogging, or physical exercises.

3.2.4 Measurements

Multiple instruments were used to assess PTSD symptomatology before, during, and after the treatment program:

Table 3.1: Daily treatment program.

Activity	Duration (minutes)
Pretreatment measurement[a]	45
PE session	90
Short break	15
Physical activity (yoga, exercise, running)	60
Lunch break	30-45
EMDR session	90
Short break	90
Psycho-education	60
Measurements[b]	45

[a] Only at day 1.
[b] Only at day 6.

- The Dutch version of the Clinician Administered PTSD Scale for DSM-5 (CAPS-5) assesses the frequency and intensity of the twenty DSM-5 PTSD symptoms (Boeschoten, Bakker, Jongedijk, et al., 2014). It was administered to participants for evaluating the existence of a PTSD diagnosis at screening, and at one-week, four-week, and six-week follow-up. Severity scores were computed as a sum score of the 20 symptom-specific severity scores, ranging from 0-80. Boeschoten et al. (2018) found the CAPS-5 to have adequate validity and reliability.

- The Dutch version of the PTSD checklist for DSM-5 (PCL-5) is a twenty-item self-report questionnaire intended to measure PTSD symptomatology, with scores ranging from 0–80 (Boeschoten, Bakker, & Jongedijk, 2014; Weathers et al., 2013). It was administered to participants at screening, at the start of day 1 of treatment, at the end of day 6 of treatment, and at one-week, four-week, and six-week follow-up. It has been found to show strong validity and reliability (Blevins et al., 2015).

- An abbreviated 8-item version of the Dutch PCL-5 (from hereon referred to as the 8-item PCL) was used to monitor PTSD symptoms on each day during treatment. This self-report instrument consists of eight of the original twenty questions from the PCL-5, with scores ranging from 0–32. The 8-item PCL strongly correlated with the complete PCL-5 and has been recommended for use to monitor treatment progress (Price et al., 2016). For interpretive data on the 8-item PCL, readers are referred to Price et al. (2016).

PTSD reduction four-weeks after completion of the program was operationalized as the difference in CAPS-5 score between screening and four-week follow-up, ΔC; and as the difference in PCL-5 score between the start of day 1 and four-week follow-up, ΔP.

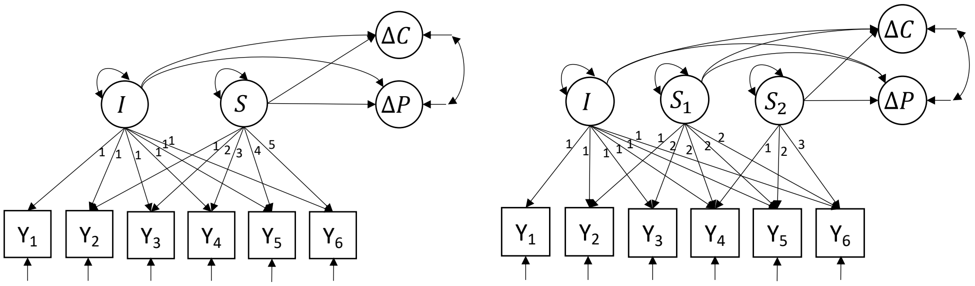
3.2.5 Data analysis

Latent growth curve models (LGCMs) were fitted to the daily PTSD measurements (measured with the 8-item PCL) to capture the change in PTSD symptomatology of participants during the treatment program (Meredith & Tisak, 1990). In LGCMs, this change is captured by latent factors, which we refer to as growth factors. To investigate to what degree progress during treatment could predict PTSD reduction four weeks after treatment, PTSD reduction four weeks after completion of the program (i.e., ΔP and ΔC) were regressed on the growth factors.

As LGCMs are flexible models that can differ in the number and type of growth factors used to capture progress during treatment, multiple candidate LGCMs (i.e., linear, quadratic, and piece-wise versions) were compared with respect to how well ΔP and ΔC could be predicted. Full details on the different candidate models considered, assessment of their out-of-sample prediction performance using k-fold cross-validation, and model fit can be found in the online supplementary materials at https://jeroendmulder.github.io/predicting-PTSD-using-LGCM.

Here, we focus only on the linear LGCM (L-LGCM, Figure 3.1a), and the piece-wise linear LGCM (P-LGCM, Figure 3.1b). The L-LGCM uses two growth factors: An intercept factor I to capture initial PTSD-symptomatology of participants at start of treatment, and a linear slope factor S to capture linear change in symptomatology across the six daily measures. These factors were then used to predict treatment reduction at four-week follow-up ΔC and ΔP. The L-LGCM is interesting as it is the most parsimonious model, while retaining relatively good out-of-sample prediction performance (see cross-validation results in the online supplementary materials). The P-LGCM extends the L-LGCM by including a second linear slope factor S_2: The first linear slope factor S_1 captures linear change in symptomatology across the first three daily measurements (in week 1), whereas the second linear slope factor captures linear change in symptomatology across the last three daily measurements (in week 2). All three growth factors then predict PTSD reduction at four-week follow-up. This particular variant of LGCMs is interesting from a clinical perspective, as it allows for investigating how well ΔP and ΔC can be predicted from change in PTSD symptomatology across the first three daily measurements alone.

Analyses were performed in R (version 3.6.1; R Core Team, 2022). The LGCMs were fitted to data using the R package lavaan (version 0.6–7; Rosseel, 2012). Missing data were handled using full information maximum likelihood, such that all available data points were used in the analyses (i.e., no participants were listwise-deleted).



(a) Linear latent growth curve model

(b) Piece-wise linear latent growth curve model

Figure 3.1: Y_t = daily measurement of PTSD-symptomatology on day t using the 8-item PCL; I = intercept growth factor; S = linear slope growth factor; S_1 = linear slope growth factor week 1; S_2 = linear slope growth factor week 2; ΔC = difference in CAPS-5 score between screening and four-week follow-up; ΔP = difference in PCL-5 score between score at start of day 1 and at four-week follow-up.

3.3 Results

In this section, we present and visualize the data, and discuss results from the L-LGCM and P-LGCM. Descriptive statistics for the 8-item PCL daily measurements and PTSD reduction at four-week follow-up can be found in Table 3.2. At screening, participants had an average CAPS-5 score of M = 43.10 (SD = 9.93). On day 1 before treatment, patients had an average PCL-5 score of M = 55.08 (SD = 11.04). Figure 3.2 depicts development of daily PTSD symptomatology throughout treatment for a subsample of ten participants (to avoid overplotting). These participants were selected to illustrate the variability in PTSD symptoms across participants and throughout treatment. The solid black line represents the average change in PTSD symptoms over the entire sample. An interactive plot containing the daily PTSD measurements from the entire sample can be found in the online supplementary materials. Figure 3.2 shows that there are large differences in PTSD symptoms of participants at the start of treatment (as can also be inferred from the standard deviation of the daily measurements in Table 3.2), as well as in how much participants change, and the form of this change during the treatment program. Four weeks after completion of the treatment program, 45 participants (41.3%) still met the criteria for PTSD according to the CAPS-5, 48 participants (44.0%) did not meet the criteria for PTSD anymore, and 16 (14.7%) cases were missing (i.e., the participants did not show up for the four-week follow-up measurement, or they showed up too late). Participants had an average CAPS-5 score of M = 25.82 (SD = 17.00) and an average PCL-5 score of M = 34.02 (SD = 20.08) at four-week follow-up. The average PTSD reduction at four-week follow-up was $\Delta P = -22.05$ (SD = 19.61) in terms of PCL-5 scores, and $\Delta C = -17.29$ (SD = 15.47) on the CAPS-5 scale.

Table 3.2: Descriptive statistics for daily 8-item PCL scores, and PTSD reduction at four-week follow-up.

Variable	N	M	SD	Min.	Max.
ΔP	89	-22.05[a]	19.61	-69[b]	+12[b]
ΔC	99	-17.29	15.47	-53[b]	+14[b]
8-item PCL - Day 1	107	22.49	35.54	5	32
8-item PCL - Day 2	105	20.60	40.64	2	32
8-item PCL - Day 3	102	19.52	48.45	1	32
8-item PCL - Day 4	106	19.29	45.09	0	32
8-item PCL - Day 5	102	18.08	48.35	1	32
8-item PCL - Day 6	96	17.37	55.25	1	31

[a] The sum of ΔP and mean PCL-5 score at four-week follow-up does not exactly equal the mean PCL-5 score on day 1. This is due to 14 participants that are included in the PCL-5 measurements on day 1, but are missing at four-week follow-up, and hence not included when computing ΔP.

[b] ΔP and ΔC represent a change in PTSD-symptomatology. Therefore the column "Min." actually represents the *greatest* observed decrease in symptomatology, and the column "Max." the *smallest* observed decrease (in fact, an increase) in symptomatology.

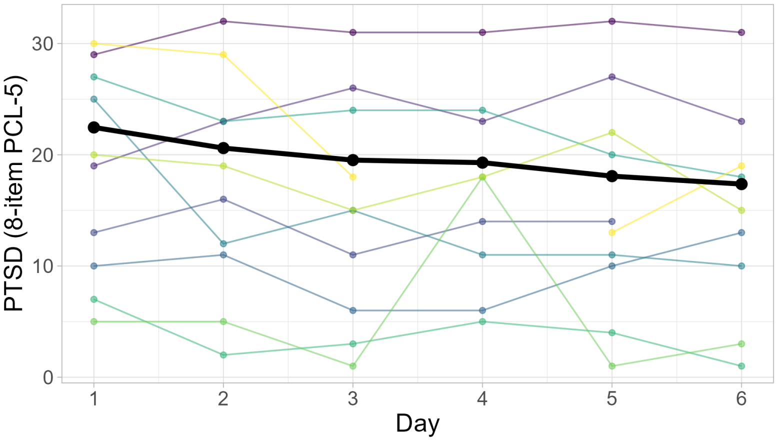


Figure 3.2: PTSD symptomatology during treatment (measured with the 8-item PCL) of a subsample of ten participants. The y-axis represents total scores on the 8-item PCL. The x-axis represents treatment day. The thick black line represents the observed mean PTSD symptomatology over time, whereas the colored lines represent individual trajectories. See the online supplementary materials for an interactive spaghetti plot of the complete sample.

3.3.1 The linear LGCM (L-LGCM)

The L-LGCM showed good model fit: $\chi^2(25) = 35.98$, $p = .072$; root-mean-square error of approximation (RMSEA) = 0.06; comparative fit index (CFI) = 0.99; Tucker-Lewis index (TLI) = 0.98. The mean of the intercept factor was estimated to be 21.94 ($SE = 0.56$, $p < .001$) with a variance of 29.45 ($SE = 4.56$, $p < .001$); the mean of the linear slope factor was estimated to be -0.96 ($SE = 0.12$, $p < .001$) with a variance of 1.15 ($SE = 0.23$, $p < .001$). This implies that, on average, participants started treatment with a PTSD score of 21.94 on the 8-item PCL scale, which decreased linearly each day by 0.96 points. However, the variances of the growth factors imply that there are large differences between participants in the starting point and change during the treatment program.

When predicting the follow-up treatment outcome ΔP from the growth components, we found a nonsignificant regression coefficient (unstandardized) for the intercept ($b = 0.20$, $SE = 0.30$, $p = .492$), but a significant coefficient for the slope ($b = 15.44$, $SE = 1.81$, $p < .001$). Combined, the intercept and slope result in an R^2 of .69. This implies that the linear slope in the L-LGCM is a significant predictor of PTSD reduction at four-week follow-up, explaining approximately 69% of the variance. Participants who show a one-point-per-day greater decrease in their PTSD symptomatology during treatment are predicted to have a 15.44 point greater PTSD reduction on the PCL-5 scale at four-week follow-up. The symptomatology at the start of the treatment program does not provide any predictive information about ΔP. For PTSD reduction as measured using the CAPS-5, ΔC, we again found a nonsignificant regression coefficient (unstandardized) for the intercept ($b = 0.25$, $SE = 0.24$, $p = .298$), and a significant coefficient for the slope ($b = 11.76$, $SE = 1.47$, $p < .001$). Combined, the intercept and slope produce an R^2 of 0.66. This suggests that the intercept is not predictive of PTSD reduction at four-week follow-up as measured with the CAPS-5 scale, but participants with one-point-per-day greater decrease in PTSD symptomatology are predicted to have in a 11.76 greater decrease in ΔC. Both the intercept and slope explain approximately 66% of the variance.

3.3.2 The piecewise LGCM (P-LGCM)

The P-LGCM showed adequate model fit: $\chi^2(21) = 33.99$, $p = .036$; $RMSEA = 0.08$; $CFI = 0.98$; $TLI = 0.98$. The mean of the intercept was estimated to be 22.18 ($SE = 0.57$, $p < .001$) with a variance of 28.04 ($SE = 4.45$, $p < .001$), the mean of the slope across the first three daily measurements was -1.20 ($SE = 0.26$, $p < .001$) with a variance of 3.51 ($SE = 0.95$, $p < .001$), and the slope across the last three daily measurements was -0.85 ($SE = 0.16$, $p < .001$) with a variance of 1.36 ($SE = 0.37$, $p < .001$). This implies that on average, participants started treatment with a score of 22.18 on the 8-item PCL. Over the first two full treatment days, PTSD symptomatology decreased on average with approximately 1.20 points per day, whereas in the second week the symptomatology decreased with approximately 0.85 points per day.

ΔP and ΔC were regressed on the growth components. For ΔP we found an unstandardized regression coefficient of $b = 0.12$ for the intercept ($SE = 0.32$, $p = .693$), $b = 6.47$ for the first slope ($SE = 1.41$, $p < .001$) and $b = 8.83$ for the second slope ($SE = 2.03$, $p < .001$). These results indicate that both changes in PTSD symptomatology in the first week, and changes in the second week, are significant predictors for PTSD reduction at four-week follow-up on the PCL-5 scale. We found an R^2 of .67, meaning that the three growth components were able to account for 67% of the variance in the outcome. Next, we inspected the standardized regression coefficients to compare which predictor (either the change in PTSD symptomatology the first week or the second week) had greater predictive power. For the first slope we found $\beta = 0.62$ ($SE = 0.11$, $p < .001$), and for the second slope $\beta = 0.53$ ($SE = 0.11$, $p < .001$). We therefore concluded that both the first and the second slope are useful in predicting follow-up treatment outcomes (PCL-5), but relatively speaking, the change in PTSD symptomatology during the first week holds more predictive power than the change in symptoms during the second week. When using only the change in PTSD symptomatology in the first week to predict PTSD reduction at four-week follow-up, 33% of the variance in ΔP was explained.

For PTSD reduction on the CAPS-5 scale, we again found a nonsignificant regression coefficient (unstandardized) for the intercept ($b = 0.28$, $SE = 0.26$, $p = .270$), but a significant coefficient for the first slope ($b = 4.17$, $SE = 1.09$, $p < .001$), and the second slope ($b = 7.73$, $SE = 1.76$, $p < .001$). The growth components combined explain approximately 63% of variance in the outcome. Looking at the standardized effects, we found $\beta = 0.52$ ($SE = 0.12$, $p < .001$) for the first slope and $\beta = 0.59$ ($SE = 0.11$, $p < .001$) for the second slope. Therefore, we concluded that, again, both slopes are useful for predicting PTSD reduction using the CAPS-5 scale. However, relatively speaking, it is the change in PTSD symptomatology during the second week that is more informative for predicting PTSD reducation at four-week follow-up (measured using the CAPS-5) compared to the change in the first week. When using only the change in PTSD symptomatology in the first week to predict PTSD reduction, the explained variance in ΔC was reduced to 28.3%, implying that it is the combination of all three growth components that is most useful for predicting the follow-up treatment outcomes.

3.4 Discussion

This paper studied whether change in PTSD symptomatology during treatment could be used to predict PTSD reduction four weeks after treatment was completed. This was assessed in an outpatient intensive treatment program for PTSD, in which PE sessions, EMDR sessions, physical activity, and psychoeducation were combined. Progress during the six treatment days (i.e., three consecutive days per week, for two consecutive weeks) was monitored by assessing symptoms of PTSD with an abbreviated PTSD self-report measure (8-item PCL). It was expected that early improvement in PTSD symptoms during treatment would be a predictor for PTSD reduction at four-week follow-up. Consistent with expectations, the results indicate that a greater decline in self-reported PTSD symptoms during the complete treatment program is a predictor for a greater decline in PTSD symptoms at four-week follow-up, as measured with both self-rated, and clinician-rated instruments. Additionally, it was expected that a greater decline in PTSD symptoms during the first half of treatment would predict PTSD reducation as well. The results show that a greater decline in self-reported PTSD symptoms during the first week of treatment, after completing two of the total six treatment days, was indeed a predictor for PTSD reduction at four-week follow-up, as measured with self-rated, and clinician-rated instruments. These findings are consistent with earlier findings which showed that early PTSD-symptom change was related to overall treatment outcome (Byllesby et al., 2019; Sripada et al., 2020). Interestingly, discrepant results have been found concerning the predictive value of the first and second treatment week. When using self-report PCL-5 as the overall treatment outcome measure, development in the first treatment week had more predictive value, whereas when using the clinician-rated CAPS-5, development in the second week had more predictive value. Moreover, the results show that pretreatment PTSD symptoms, as measured with the self-report 8-item PCL, do not predict treatment outcomes. This is an interesting finding because the findings of previous studies showed that pretreatment PTSD symptoms were in fact related to treatment outcome (Karatzias et al., 2007; Van Minnen et al., 2002).

The inconclusive results of studies on the value of pretreatment factors (cf. Ehlers et al., 1998; Ehlers et al., 2013; Forbes et al., 2003; Karatzias et al., 2007; Tarrier et al., 2000; Van Minnen et al., 2002) stress the need for more approaches to predict treatment outcome. Although the current study awaits replication, the results imply that using a measure to monitor treatment progress, and evaluate early progress, could be valuable in decision-making (e.g., adjusting treatment based on treatment response). When a patient does not show a large improvement during the first few days, practitioners may decide to scale up treatment or change the type of evidence-based treatment provided, which is what guidelines recommend (Akwa GGZ, 2020).

Doing so might prevent participants from continuing a treatment with little expected benefit. Although in this case, scaling up treatment by intensifying seems difficult considering the treatment program for the current sample is already an intensive one. Other remaining guideline suggestions include switching to other evidence-based treatments (e.g., cognitive processing therapy), pharmacotherapy, or experimental treatments.

The current study has a number of limitations. First, there were numerous missing cases, which were mainly participants missing, or showing up too late for follow-up measurements. This might compromise the generalizability of this study's results to the population. One might argue that a reason why participants did not show up, or showed up too late to be included in follow-up measurements, is because they were unsatisfied with treatment and have seen little improvement in their symptoms. Another reason might be that participants who did not show PTSD symptoms after treatment may have refused to spend time and effort on measurement, as they already finished treatment, and did not see any personal added value to inclusion in follow-up measurements. A second limitation is that the patient sample predominantly consisted of women, which again negatively affects generalizability of the results. It is unclear why the sample had such a notable gender imbalance, however the results should be interpreted with this in mind. Third, the size of the sample can be seen as a limitation, since it made it impossible to investigate if the presence of comorbid disorders could influence treatment progress.

A strength of the current study is that PTSD reducation at four-week follow-up was measured using two different measurements: The clinician-rated CAPS-5, and the self-report PCL-5. Although the CAPS-5 and PCL-5 correlate strongly (Blevins et al., 2015), one study found contradictory results in prediction models using the CAPS (Blake et al., 1995) and PCL (Davidson et al., 1989) as treatment outcome measures (Karatzias et al., 2007). They found that lower baseline CAPS scores were associated with better treatment outcomes, whereas higher baseline PCL scores were associated with better treatment outcomes. Using the clinician-rated CAPS-5 as well as the self-report PCL-5 to measure treatment outcome controls for these possibly ambiguous results.

Future research should be aimed at improving the reliability and generalizability of the results by replicating the current study in different cultures, and in samples that vary in gender, age, and other demographic variables. One could study if there is a difference in the results for trauma type, and different types and intensities of treatment as well. Since the results were found in an intensive treatment program including PE and EMDR, the study should be replicated in nonintensive treatment programs and other types of treatments (Byllesby et al., 2019; Sripada et al., 2020). Another recommendation for future research is to explore possible adaptations in intensive treatment programs for patients who show little response, since it is yet unclear which adaptations can be made. Therefore, future research should first be focused on investigating relevant processes during treatment that could influence outcomes in the current treatment program. The current treatment consists of several components, and one should also identify which treatment components are responsible for treatment progress, and also in whom. It is useful to know that the absence of response during treatment indicates less beneficial overall treatment outcome. The next step should be investigating what is causing patients to show little or no response, or even deterioration. As mentioned before, factors like therapeutic alliance, trauma-related belief change, between-session habituation, and SUD scores during the first PE or EMDR session have been shown to be related to treatment outcome (Cooper et al., 2017; Foa et al., 1995; Hendriks et al., 2018; Horvath & Symonds, 1991; Jaycox et al., 1998; D. Kim et al., 2008; Martin et al., 2000). These are possible relevant factors during treatment that could explain why some patients show little reduction in PTSD symptoms. Another reason why it is useful to identify which processes are responsible for less beneficial treatment response is that the predictive value of the current study is based on means, and the results are not necessarily applicable to every patient. This poses an ethical dilemma for clinical use because practitioners would have to make a decision for the individual to change treatment (intensity) based on those means.

In conclusion, the present study indicates that a greater decline in PTSD symptoms during the course of an intensive treatment program is a predictor for greater PTSD reduction at four-week follow-up. This prediction can also be made using the progress measured only during the first treatment week, after completing two of the six treatment days. Pretreatment PTSD symptoms had no predictive value for treatment outcome at the four-week follow-up. Being able to predict treatment outcomes using the progress measured during treatment shows a large potential for clinical use. Future research should mainly be focused on replicating the current results and improving the reliability and generalizability of the results. A next step would be investigating the factors responsible for poorer treatment responses.

Acknowledgments: We thank Sophie Menses and Mathilde Huisman for managing the data. We thank students Pepijn van der Post and Sanne Hofman for their aid in collecting the data.

Author contributions: Conceptualization: VA and SM - Data curation: VA and JM - Formal analysis: JM – Investigation: VA - Methodology: VA, JM, and SM – Project administration: VA and SM - Resources: SM - Software: JM - Supervision: SM - Visualization: VA and JM - Writing (original draft preparation): VA – Writing (review and editing): JM and SM.

Data availability statement: The data presented in this study are available on request from the corresponding author.

Informed consent statement: Informed consent was obtained from all subjects involved in the study.

Institutional Review Board statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Scientific Board of Altrecht GGz.

Supplementary materials: This study's online supplementary materials can be found at https://jeroendmulder.github.io/predicting-PTSD-using-LGCM.

CHAPTER 4

Three extensions of the random intercept cross-lagged panel model

Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 638-648. https://doi.org/10.1080/10705511.2020.1784738

Abstract

The random intercept cross-lagged panel model (RI-CLPM) is rapidly gaining popularity in psychology and related fields as a structural equation modeling (SEM) approach to longitudinal data. It decomposes observed scores into within-unit dynamics and stable, between-unit differences. This paper discusses three extensions of the RI-CLPM that researchers may be interested in, but are unsure of how to accomplish: (a) including stable, person-level characteristics as predictors and/or outcomes; (b) specifying a multiple-group version; and (c) including multiple indicators. For each extension, we discuss which models need to be run in order to investigate underlying assumptions, and we demonstrate the various modeling options using a motivating example. We provide fully annotated code for the R package lavaan, and Mplus on an accompanying website.

The random intercept cross-lagged panel model (RI-CLPM) proposed by Hamaker et al. (2015) is an extension of the traditional cross-lagged panel model (CLPM). It was introduced to account for stable, trait-like differences between units (e.g., individuals, dyads, families, etc.), such that the lagged relations pertain exclusively to within-unit fluctuations.[1] The idea that we should decompose longitudinal data into stable, between-unit differences versus temporal, within-unit dynamics is closely linked to the multilevel literature on cluster-mean centering (Bolger & Laurenceau, 2013; Enders & Tofighi, 2007; Kievit et al., 2013; Kreft et al., 1995; Mundlak, 1978; Neuhaus & Kalbfleisch, 1998; Nezlek, 2001; Raudenbush & Bryk, 2022; Snijders & Bosker, 2012). Alternatively, it can also be linked to the discussion in panel research on the need to account for *unobserved heterogeneity* in longitudinal data (Allison et al., 2017; Bianconcini & Bollen, 2018; Bollen & Brand, 2010; Bou & Satorra, 2018; Finkel, 1995; Hamaker & Muthén, 2020; Liker et al., 1985; Ousey et al., 2011; Wooldridge, 2002, 2013). A detailed discussion of how other common panel models account for unobserved heterogeneity (as well as for *measurement error* and *developmental trajectories*) is provided by Usami et al. (2019), Zyphur, Allison, et al. (2020), and Zyphur, Voelkle, et al. (2020).

The appeal of the RI-CLPM can be attributed to three factors. First, the basic idea that one needs to decompose the observed variance into two sources resonates with a concern many researchers have had about the traditional CLPM (Keijsers, 2016). In fact, there have been numerous other proposals aiming to do exactly this (e.g., Allison et al., 2017; Bianconcini & Bollen, 2018; Kenny & Zautra, 1995; Ormel et al., 2002; Ormel & Schaufeli, 1991; Ousey et al., 2011). Second, the model can be applied if one has three occasions of data or more, using any structural equation modeling (SEM) software package, which makes the approach broadly applicable and easy to implement. Third, the RI-CLPM tends to fit empirical data (much) better than the traditional CLPM, as is corroborated by empirical work of for instance Borghuis et al. (2020), Burns et al. (2020), and Keijsers (2016). The second-order lagged relations that are often needed to get a CLPM to have an acceptable fit, are typically not needed in the RI-CLPM, because the long-run, trait-like stability is now captured by the random intercepts instead of by the second-order lagged relations.

Given the growing popularity of the RI-CLPM, it is not surprising that researchers are interested in how they can adapt the basic model to accommodate their particular data and research interests. Examples of this can be found in the Mplus Discussion Board thread on the RI-CLPM,[2] the lavaan forum,[3] and RI-CLPM-related posts on

[1] While the original paper by Hamaker et al. (2015) uses the terms within-person and between-person, we use within-unit and between-unit here to emphasize that the cases are not necessarily individuals, but can also be dyads, families, companies, or individuals and their context, peers, etc.

[2] Accessible via http://www.statmodel.com/discussion/messages/11/25297.html?1579816772

[3] Accessible via https://groups.google.com/forum/#!forum/lavaan.

SEMNET.[4] Some of the most frequently asked questions are how to extend the model by (a) including person-level characteristics (e.g., social economic status, personality factors, age, health) as a predictor or outcome variable, (b) performing a multiple-group version of the model to investigate whether lagged relationships are different across groups, and (c) using multiple indicators for latent variables in the model. The purpose of the current paper is to elaborate on these extensions and help researchers navigate the different modeling options and assumptions.

This paper is organized as follows. In the first section we begin with presenting the RI-CLPM, and discuss how it is related to the traditional CLPM. In the following three sections we discuss the three different extensions described above and we will focus on the modeling options available. To facilitate the explanation of the model and its results we will use a motivating example about the reciprocal effects of *sleep problems* and *anxiety* in young adolescents based on Narmandakh et al. (2020). Furthermore, to allow the reader to obtain hands-on experience with this modeling approach, we provide a simulated data set of our motivating example, as well as annotated lavaan code and Mplus syntax in the online supplementary materials at https://jeroendmulder.github.io/RI-CLPM.

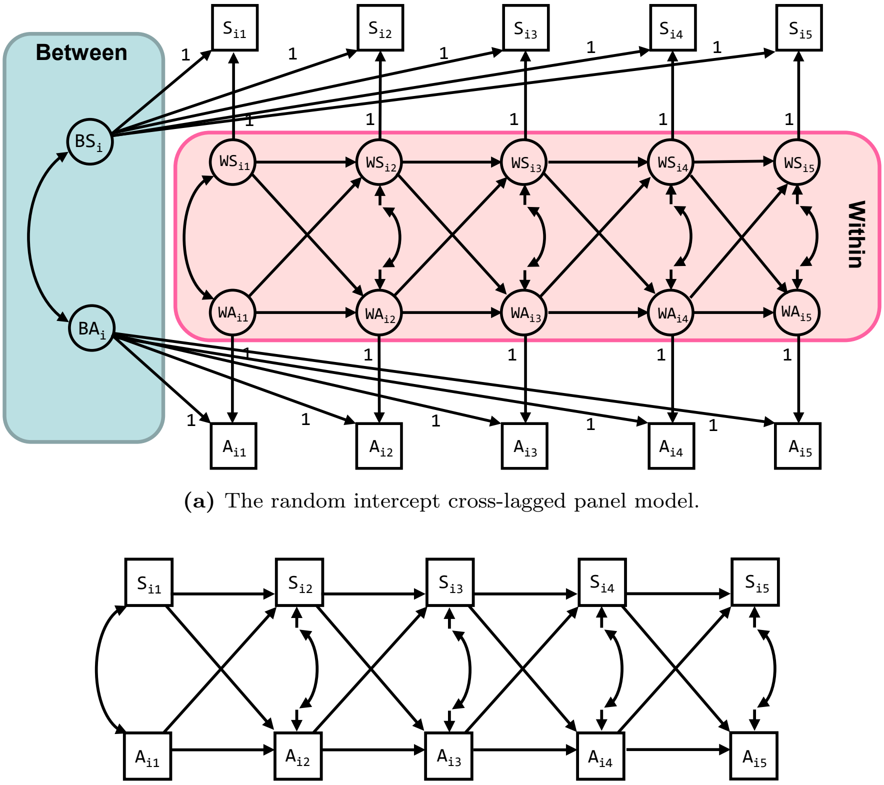
4.1 The RI-CLPM and the traditional CLPM

Below, we begin with discussing how the RI-CLPM is build up. Subsequently, we discuss diverse constraints over time that can be imposed or relaxed. We end by briefly discussing how this model is related to the traditional CLPM. While the terminology used here is clearly inspired by the multilevel literature (where there is a between-cluster level and a within-cluster level), the RI-CLPM is estimated in wide-format using structural equation modeling (SEM), rather than in long-format with multilevel modeling. Throughout we make use of a simulated data set that was motivated by Narmandakh et al. (2020). In their study, five waves of data were obtained from 1189 adolescents on their sleep problems and anxiety during the past 15 years.

4.1.1 Building up the basic RI-CLPM

To fit an RI-CLPM, we need to decompose the observed scores into three components: Grand means, stable *between* components, and fluctuating *within* components. This decomposition is illustrated in Figure 4.1a. Let S_{it} and A_{it} represent the *observed* scores on sleep problems and anxiety for person i at occasion t, respectively. The first components are the grand means, which are the means over all units per occasion t, and represented by μ_t for sleep problems and π_t for anxiety. These grand means may

[4] Accessible via https://listserv.ua.edu/cgi-bin/wa?A0=SEMNET.


Between
Within
BS_i
BA_i
S_{i1} S_{i2} S_{i3} S_{i4} S_{i5}
WS_{i1} WS_{i2} WS_{i3} WS_{i4} WS_{i5}
WA_{i1} WA_{i2} WA_{i3} WA_{i4} WA_{i5}
A_{i1} A_{i2} A_{i3} A_{i4} A_{i5}
1

(a) The random intercept cross-lagged panel model.

(b) The traditional cross-lagged panel model.

Figure 4.1: S_{it} = observed sleep problems of unit i at occasion t; A_{it} = observed anxiety of unit i at occasion t.

be time-varying, or may be fixed to be invariant over time. Second, the *between* components, indicated by the letter B, are the random intercepts: BS_i for sleep problems and BA_i for anxiety. They capture a unit's time invariant deviation from the grand means and thus represent the stable differences between units. The random intercepts are specified in SEM software by creating a latent variable with the repeated measures as its indicators, and fixing all the factor loadings to 1. Third, the *within* components, indicated by the letter W, are the differences between a units observed measurements and the unit's expected score based on the grand means and its random intercepts. WS_{it} and WA_{it} thus represent the within components of sleep problems and anxiety, respectively. We create these components in SEM software by specifying a latent variable for each measurement and constraining its measurement error variances to 0. As a result, we have $S_{it} = \mu_t + BS_i + WS_{it}$ and $A_{it} = \pi_t + BA_i + WA_{it}$.

Next, we specify the structural relations between the within components. The autoregressive effects (i.e., α_t from $WS_{i\,t-1}$ to WS_{it} and δ_t from $WA_{i\,t-1}$ to WA_{it}) represent the within-person carry-over effects. If α_t is positive, this implies that an individual who experiences elevated sleep problems relative to his/her own expected score, is likely to experience elevated sleep problems relative to his/her own expected score at the next occasion as well. The same logic applies to the interpretation of δ_t. For this reason, the within-person autoregressive effects are sometimes referred to as inertia (i.e., the tendency to not move; see Suls et al., 1998). The cross-lagged effects in the model represent the spill-over of the state in one domain into the state of another domain. Here, β_t represents the effect of $WS_{i\,t-1}$ to WA_{it} and γ_t the effect of $WA_{i\,t-1}$ to WS_{it}. A positive β_t implies that a positive (negative) deviation from an individual's expected level of sleep problems will likely be followed by a positive (negative) deviation in the individual's expected level of anxiety at the next occasion in the same direction. The same logic applies to γ_t.

Finally, we need to include covariances for both the within, and between components of the model. For the within part we specify that the components at occasion 1 and the within-person residuals at all subsequent occasions are correlated within each occasion. For the between part we specify that the random intercepts are correlated. We are *not* including covariances between the within-person components at the first occasion and the random intercepts because typically the observations have started at an arbitrary time point in an ongoing process, and there is no reason to assume that the within components at the first occasion are correlated to the random intercepts.[5]

Applying this model to our simulated example data, we find that both random intercepts have significant variance, which implies that there are stable, trait-like

[5] This is in contrast to other SEM approaches that combine lagged relations with stable components, such as the one presented by Allison et al. (2017) and Bianconcini and Bollen (2018). The defining difference between these approaches and the RI-CLPM discussed here is whether or not the lagged relations are modeled between the observed variables, or between the within-person components. For more details, see Usami et al. (2019).

differences between persons on sleep problems and anxiety. Moreover, we find a significant positive covariance between the random intercepts of 0.01 with $SE = 0.001$ (the correlation is .59, $SE = 0.050$), suggesting that individuals who have more sleep problems in general are also more anxious in general.

If, in contrast to our findings here, the variance of a random intercept does not significantly differ from 0, this means that there are little to no stable between-unit differences, and that each unit fluctuates around the same grand means over time. Including a random intercept in the model can then be regarded as "redundant"; such a model would be too complex for the data. In that case one can choose to either fix the nonsignificant variance (and all the covariances between this random intercept and the other intercepts) to 0, or simply remove the random intercept from the model and include lagged-relations between the observed variables instead of between the within-unit components. These two solutions are statistically equivalent and will lead to the same lagged-parameter estimates and model fit. Note that it is possible to have a model in which one variable needs to be decomposed into a between-unit and a within-unit part, while the other variable does not require such a decomposition.

Looking at the within part of the model we find the following standardized autoregressive effects for sleep problems, $\alpha_2 = 0.29$ ($SE = 0.034$), $\alpha_3 = 0.24$ ($SE = 0.036$), $\alpha_4 = 0.27$ ($SE = 0.036$), $\alpha_5 = 0.29$ ($SE = 0.035$), and for anxiety, $\delta_2 = 0.004$ ($SE = 0.045$), $\delta_3 = 0.25$ ($SE = 0.036$), $\delta_4 = 0.29$ ($SE = 0.033$), $\delta_5 = 0.40$ ($SE = 0.030$). There are also significant cross-lagged effects of sleep problems to anxiety, $\beta_2 = 0.15$ ($SE = 0.039$), $\beta_3 = 0.10$ ($SE = 0.035$), $\beta_4 = 0.11$ ($SE = 0.034$), $\beta_5 = 0.08$ ($SE = 0.031$), which means that individuals with relatively little sleep problems (relative to an individual's own mean) will likely experience relatively little anxiety at the next occasion. However, none of the cross-lagged effects from anxiety to sleep problems are significant, which means that an individual's temporary elevated or damped amount of sleep problems does not depend on that individual's temporary level of anxiety at the previous occasion.

4.1.2 Imposing constraints over time

To test specific hypotheses, researchers can decide to impose constraints on the model and test the tenability of these constraints. This can be done by comparing the fit of a (nested) model with constraints to the fit of the more general model using a chi-square difference test ($\Delta\chi^2$); if the constrained model fits the data significantly worse, the imposed constraints are untenable. Alternatively, one can use the AIC or BIC as measures of model fit to compare both non-nested and nested models, where the model with the lower AIC or BIC should be preferred.

The use of the chi-square difference test is wide-spread in the SEM community, but a few cautionary notes are in order. First, parameters should only be constrained if the constraints make theoretical sense, and not solely because it leads to a more parsimonious model. Second, failing to detect a significantly worse fitting model in a sequence of chi-square difference tests does not imply that the constrained model represents the population well. It is possible that the unconstrained base model was misspecified in the first place and this misspecification will carry on into the constrained model. In that case, the chi-square difference test is unable to control for Type I error rates and retain adequate power (Yuan & Bentler, 2004). Careful consideration should always be given to the fit of the models themselves by looking at a variety of model fit indices.

In the RI-CLPM, there are several constraints over time that can be added. We discuss two common ones here. First, we may consider testing if the lagged regression coefficients are time-invariant. This can be done by comparing the fit of a model with constrained regression coefficients (over time), with the fit of a model where these parameters are freely estimated (i.e., the "unconstrained" model). If this chi-square difference test is nonsignificant, this implies the constraints are tenable and the dynamics of the process are time-invariant. If the constraints are not tenable, this could be indicative of some kind of developmental process taking place during the time span covered by the study.

In this context it is important to realize that the lagged regression coefficients depend critically on the time interval between the repeated measures. Hence, constraining the lagged parameters to be invariant across consecutive waves only makes sense when the time interval between the occasions is (approximately) equal (Gollob & Reichardt, 1987; Kuiper & Ryan, 2018; Voelkle et al., 2012). If the time intervals between subsequent occasions vary, we are estimating different autoregressive and cross-lagged effects between each pair of adjacent measurements. In such a situation, constraining the lagged regression coefficients leads to an uninterpretable blend of different lagged relationships. Furthermore, even when the lagged parameters are invariant over time, this will typically not be true for the standardized lagged parameters, because these are a function of the within-unit variance of the predictor and the within-unit variance of the outcome. As these variances are typically not (constrained to be) equal across the occasions (which is complicated due to the recursiveness in the model), the standardized lagged parameters can differ even if the unstandardized lagged parameters are constrained to be the same (Hamaker et al., 2015).

To test if the lagged relations in our sleep problems and anxiety example are invariant over time, we fit a model with constrained lagged regression coefficients and find $\chi^2 = 90.97$ with 33 degrees of freedom. The unconstrained model (the basic RI-CLPM fitted before) has $\chi^2 = 25.81$ with 21 degrees of freedom. The chi-square difference test of these two nested models is thus $\Delta\chi^2(12) = 65.16$, with $p < .001$. Hence, constraining the lagged effects to be the same over time results in a significantly worse model fit. We therefore conclude that the constraints are untenable and that there appears to be a change in within-person dynamics over time. Upon closer inspection of the autoregressive effects of anxiety δ_t in the unconstrained model, this makes sense: These estimates increase with each subsequent occasion, from .004 to .40.

Second, we may investigate whether the grand means, μ_t and π_t, are invariant over time. This can be done by constraining the means to be the same across occasions and performing the chi-square difference test to determine whether this constraint can be imposed. If this is the case, this implies we are dealing with a construct that is stable at the population level for the duration of the study. In contrast, if the grand means cannot be constrained to be invariant over time, this implies that on average there is some change in this variable over time, which may reflect some occasion-specific effect, or a developmental trend. By allowing the means to freely vary over time, we account for such average changes over time. In our example a comparison of the constrained and the unconstrained models yields a chi-square difference test of $\Delta\chi^2(8) = 434.20$, $p < .001$, which implies that the constraints are untenable and that the grand means vary over time.

Alternatively, one can choose to relax, instead of impose, constraints over time to allow for a more flexible and better fitting model. The RI-CLPM is based on the assumption that the random intercepts have the exact same influence on the observed variables at each occasion, which is reflected by the factor loadings that are all constrained to be 1 over time. However, researchers may want to test this, which can be done by comparing the model with these constrained factor loadings to a model in which the factor loadings are estimated freely; the latter model implies that there are stable, trait-like differences between individuals, but the *size* of these differences can change over time. The between components are then no longer random intercepts, but can be interpreted as traits. To fit a model with freely estimated factor loadings, at least four occasions of data are needed; in contrast, with the fixed factor loadings, the model is already identified with only three waves of data.

4.1.3 Relatedness to the traditional CLPM

If we constrain the variances of all random intercepts (and their covariance) in the RI-CLPM to zero, we obtain a model that is nested under the RI-CLPM, and no longer accounts for stable between-unit differences. This model is actually *statistically equivalent* to the traditional CLPM (represented in Figure 4.1b), which implies that we can compare these two models using a chi-square difference test.[6]

[6] Actually, it requires a chi-bar-square test, as it is based on constraining two of the parameters on the bound of the parameter space, see Stoel et al. (2006). The regular chi-square test is too strict, which means that if it is significant, the chi-bar-square test would also be significant, while the reverse is not true.

In comparison to the traditional CLPM, the RI-CLPM often leads to autoregressive parameters that are closer to zero with larger standard errors. As a result, the autoregressive parameters that are significantly different from zero in the CLPM, may not be significant in the RI-CLPM. This has led some to speculate that the reliability of the within-unit components in the RI-CLPM is low. However, it is important to realize that the autoregressive parameters represent quite different phenomena in these two models. In the traditional CLPM, the autoregressive parameter captures the stability of the rank-order of individuals from one occasion to the next. It is closely related to the idea of test-retest reliability, which uses the autocorrelation as a measure of reliability of a time-invariant, trait-like construct. In the RI-CLPM however, the trait-like features are captured by the random intercepts, such that the autoregressive parameters are not there to capture rank-order stability due to a trait, but to account for additional moment-to-moment stability (i.e., inertia or carry-over) of the within-unit fluctuations over time. Hence, in the RI-CLPM, the autoregressive parameters should not be considered as measures of reliability, because reliability and stability do not coincide for state-like concepts (Hertzog & Nesselroade, 1987).

With respect to the cross-lagged parameters, there can be a number of differences between the two models. As discussed in the original paper by Hamaker et al. (2015), we may find cross-lagged paths in the CLPM that seize to exist in the RI-CLPM or vice versa, the standardized absolute values of the cross-lagged parameters may lead to a different ordering, and even the sign of a cross-lagged path may change. The latter result has been corroborated in empirical research by Dietvorst et al. (2018). The extent to which results change depends on various factors, including the relative contributions of the within-unit and the between-unit components to the total variance. For instance, when the relative contribution of the between-unit components is small, the lagged parameters of the two models will be quite similar.

Furthermore, Dormann and Griffin (2015) have recently argued that many of our conventional panel studies are probably based on intervals that are too large to capture the underlying within-unit dynamic relationships. Instead, the lagged effects that are found with the CLPM might result from stable between-unit differences rather than dynamic within-unit relations. This would imply that many of the significant results that are obtained with the CLPM, will not be replicated when using an RI-CLPM because the stable between-unit differences, captured by first and second-order lagged effects in the CLPM, are now captured by the random intercept in the RI-CLPM (Keijsers, 2016). Yet, the extent to which the results from the traditional CLPM and the RI-CLPM will differ, cannot be predicted; the discrepancy or similarity will have to be established empirically through fitting both models to the data and comparing the results.

4.1.4 Conclusion

We have provided a brief introduction to the modeling and reasoning behind the RI-CLPM, and illustrated the basic steps researchers should consider when using this modeling approach. For more details on how this model is related to other longitudinal SEM approaches, the reader is referred to Hamaker et al. (2015) and Usami et al. (2019). In the remainder of this paper, we discuss several extensions of the basic RI-CLPM.

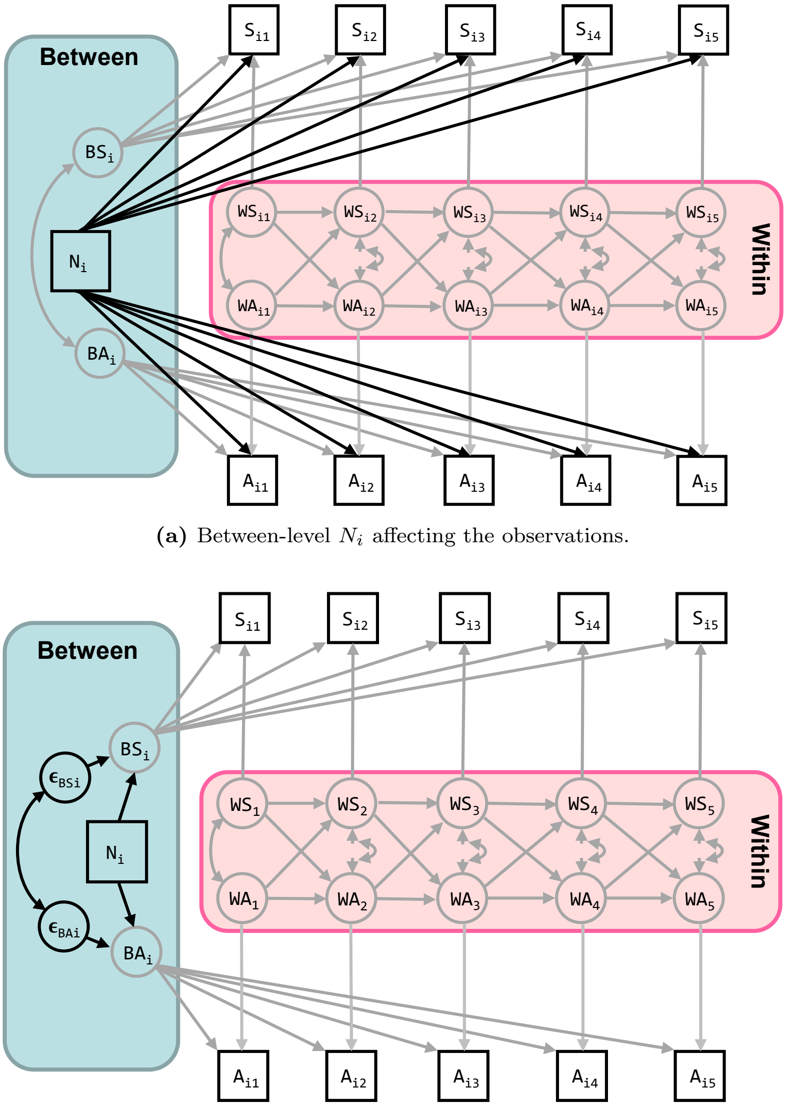
4.2 Extension 1: Including time-invariant predictors and outcomes

If we have obtained certain time-invariant person characteristics prior to the repeated measures—such as social economic status, personality, age, or gender—we may want to include these as predictors in the RI-CLPM. A question that arises in this context is whether these variables should be used to predict the observed variables or the random intercepts. These two options for an observed predictor variable are represented in Figure 4.2. In this section, we discuss both options in more detail and show how they are related. Additionally, we discuss how one may include time-invariant distal outcomes—such as later educational level, life satisfaction, or depression—in the RI-CLPM.

It is important to realize that adding variables to our model changes the covariance structure that is being analyzed, and in SEM we can only compare models that are based on the same set of variables. As a result, a model with a time-invariant predictor is not comparable to a model that excludes it. Likewise it is possible to have a well-fitting model, which is then extended with a predictor that proves significant, while this extended model no longer fits. The reason for this is that the two models are based on different covariance and mean structures.

4.2.1 Including a time-invariant predictor

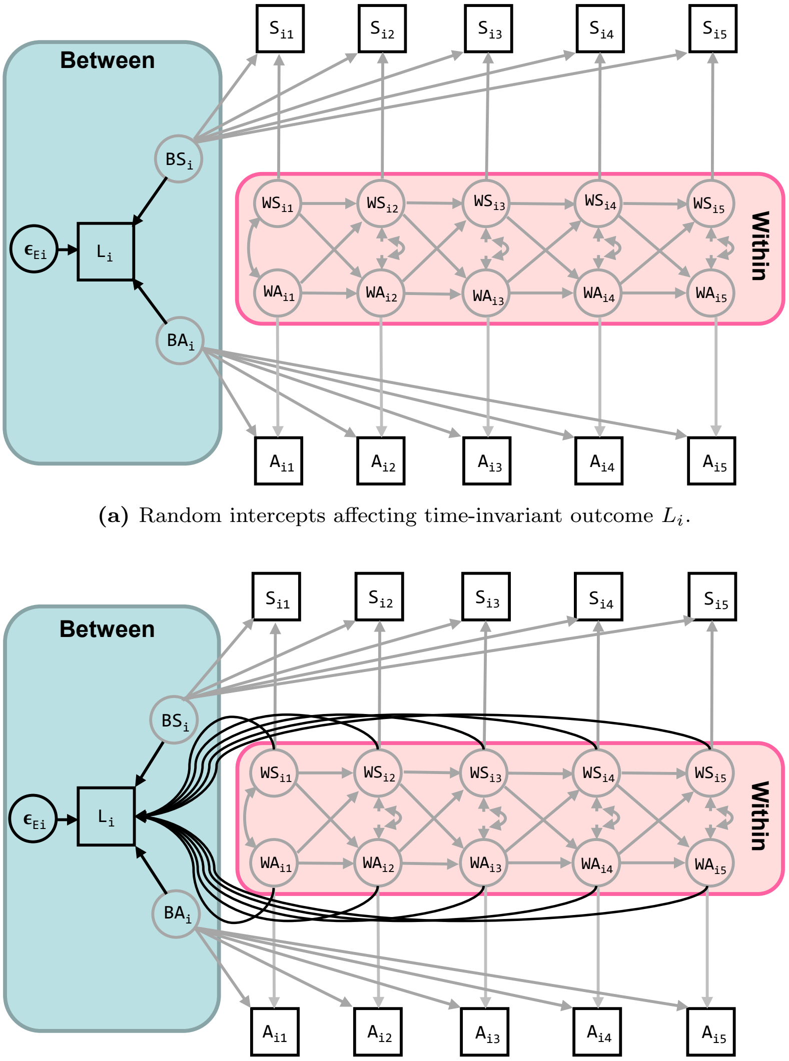
Let N_i be a measure of an individual's neuroticism, which we want to include as predictor of the observed variables S_{it} and A_{it}, as represented in the top left panel of Figure 4.2a. This allows the effect of neuroticism on sleep problems and the effect of neuroticism on anxiety to be different at each occasion t. In the particular case that N_i is a dummy variable (as in our example here) the regression coefficients can be interpreted as mean differences between the group represented by the dummy variable, and the reference group (represented by zero scores on all dummy variables). We include a dummy for individuals who are high on neuroticism, which results in significant positive effects of neuroticism on both sleep problems and anxiety. This



(a) Between-level N_i affecting the observations.

(b) Between-level N_i affecting the random intercepts.

Figure 4.2: Two options for including a between-level predictor N: In Figure 4.2a N_i influences the observed variables directly; in Figure 4.2b this occurs indirectly through the random intercepts. The model in Figure 4.2b is nested under the model in Figure 4.2a (fixing the regression coefficients to be identical over time results in a version that is equivalent to the model on the right).



(a) Random intercepts affecting time-invariant outcome L_i.

(b) Observations affecting time-invariant outcome L_i.

Figure 4.3: Two options for including a between-level outcome: In Figure 4.3a L_i is explained by the random intercepts which includes only between variance; in Figure 4.3b the distal outcome is regressed on both the random intercepts and the within components such that we use both between- and within-level variance to predict L_i. These two models are not nested.

suggests that highly neurotic adolescents experience more sleep problems and have more anxiety symptoms than adolescents in the low-neuroticism group, and this result holds for all occasions. As a restricted version of this model, we can constrain the effects of neuroticism on sleep problems and anxiety to be the same at each occasion t. Because these models are nested, we can perform a chi-square difference test to determine whether these constraints can be imposed.

The latter constrained model is statistically equivalent to a model in which the random intercepts, rather than the observed variables, are regressed on N_i (represented in Figure 4.2b). This is only the case however if the factor loadings of the random intercept are all fixed at 1 like in the basic RI-CLPM discussed before. Imposing the constraints leads to a chi-square difference test of $\Delta\chi^2(8) = 8.91$ with $p = .350$, which implies that the effects of neuroticism on the random intercepts of sleep problems and anxiety are time-invariant: The estimated standardized effects are 0.27 ($SE = 0.040$) and 0.24 ($SE = 0.035$), respectively. Therefore, we conclude that high-neuroticism adolescents experience more sleep problems and anxiety in general than low-neuroticism individuals.

4.2.2 Including a time-invariant outcome

Suppose we have measured later life satisfaction L_i after the repeated measures, and we want to predict this using sleep problems and anxiety. We can do this by regressing L_i either on the random intercepts BS_i and BA_i, the within-person fluctuations WS_{it} and WA_{it}, or on the observed variables S_{it} and A_{it}. The first two options are represented in Figure 4.3. From a substantive point of view, regressing life satisfaction on the random intercepts implies that temporal within-person fluctuations in sleep problems and anxiety, WS_{it} and WA_{it}, are not informative for predicting later life satisfaction as the random intercepts only contain stable between person information. This assumption is defensible as later educational level is a time-invariant outcome and therefore belongs to the between part of the model.

Alternatively, one can decide to regress the outcome on both the random intercepts and the temporal deviations. The regression on the random intercepts then represents the predictive value of between components *net* the predictive value of the within part, and the regression on the temporal deviations represents the predictive value of the within components *net* the predictive value of the between part. As such, we separate the total predictive power of our variables into a uniquely between and uniquely within component. The decision to use only between-unit variance, or both within- and between-unit variance to predict the outcome, should ideally be based on theoretical grounds. However, if this is something that the researcher explicitly wants to test one can fit the above two models and compare them using a chi-square difference test where the model with the outcome regressed on the random intercepts is nested under the current model.

A third option is regressing L_i on the observed variables, which implies that one assumes that both between-person variance that comes from the random intercepts, and temporary, within-person variance that comes from the within-person components, are informative about later depression. However, we find this modeling option less defensible as it again blends stable between-effects and fluctuating within-effects, an issue that the RI-CLPM aims to address in the first place. By regressing the outcome on both the within-components and between-components separately, researchers can check if within variance provides additional predictive value over the between variance.

4.2.3 Including both a predictor and outcome

We can also consider including both neuroticism as a predictor and later life satisfaction as an outcome at the between level. If this is all specified at the between level, this implies neuroticism has an indirect effect on life satisfaction through the random intercepts and this can be considered as a case of mediation at the between level. We can also include the direct effect of neuroticism on life satisfaction to allow for partial mediation.

4.3 Extension 2: The multiple-group RI-CLPM

In the previous section we used a dummy variable for neuroticism as a predictor in our model, which allowed us to investigate whether there are mean differences between the group high on neuroticism, and the group low on neuroticism. Alternatively, one can use such a categorical variable as a grouping variable in multiple group analysis (e.g., Van Lissa et al., 2019; Vangeel et al., 2018). This approach implies that not only the means can differ across the groups (as is the case when including dummy variables as predictors of the random intercepts or the observed variables, as described in the previous section), but also the lagged regression coefficients, the (residual) variances, and the (residual) covariances.

Group differences in lagged regression coefficients can be thought of as *moderation* or *interaction effects*, and may therefore be of specific interest to researchers. This can be investigated by comparing a multiple group version of the RI-CLPM in which there are no constraints across the groups, with a model in which the lagged regression coefficients are constrained to be identical across the groups. If the chi-square difference test indicates that this constraint cannot be imposed, this implies that (some of) the lagged coefficients differ across the groups: The lagged effects of the variables on each other depend on the level of the grouping variable. In contrast, when the equality constraints on the lagged parameters across the groups hold, this implies there is no moderation effect. However, note again that the constraints only imply that the raw coefficients are invariant across groups; the standardized lagged effects may still differ across the groups in case the variances differ across groups.

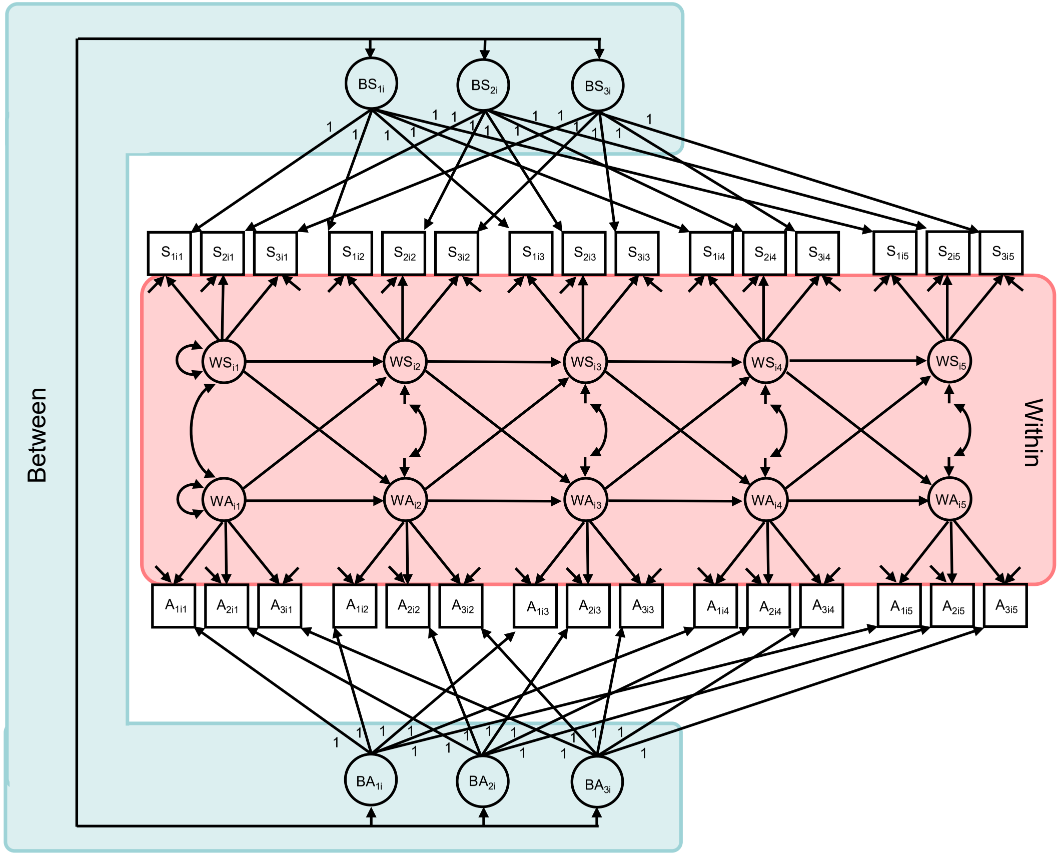
To test if the reciprocal effects between sleep problems and anxiety are the same for those high in neuroticism versus those low in neuroticism, we perform a multiple group analysis. First, we fit a multiple group RI-CLPM without constraints across the groups and find $\chi^2(42) = 45.64$. Subsequently we fit a model in which lagged parameters are invariant across groups and find $\chi^2(58) = 54.80$. The chi-square difference test of these two nested models yields $\Delta\chi^2(16) = 9.16$, $p = .907$, which implies that imposing the constraints is tenable: The lagged effects for individuals with different levels of neuroticism appear to be the same.

4.4 Extension 3: The multiple-indicator RI-CLPM

Another way in which researchers may wish to extend the RI-CLPM is by including multiple indicators for each of the constructs, while formulating the dynamics over time between the latent variables. There are two ways in which this can be done. First, a random intercept can be included for each indicator, as shown in Figure 4.4a, and these random intercepts are allowed to be correlated with each other. In addition, a common factor of the multiple indicators is included per occasion to capture the common within-unit variability over time. Second, the random intercepts can be included at the latent level as shown in Figure 4.4b (e.g., Seddig, 2020). There is a common factor for each construct at each occasion, which is then being further decomposed into a time-invariant part captured by the random intercept, and a time-varying part that is used to model the within-unit dynamics. These two approaches are nested with the second being a special case of the first.

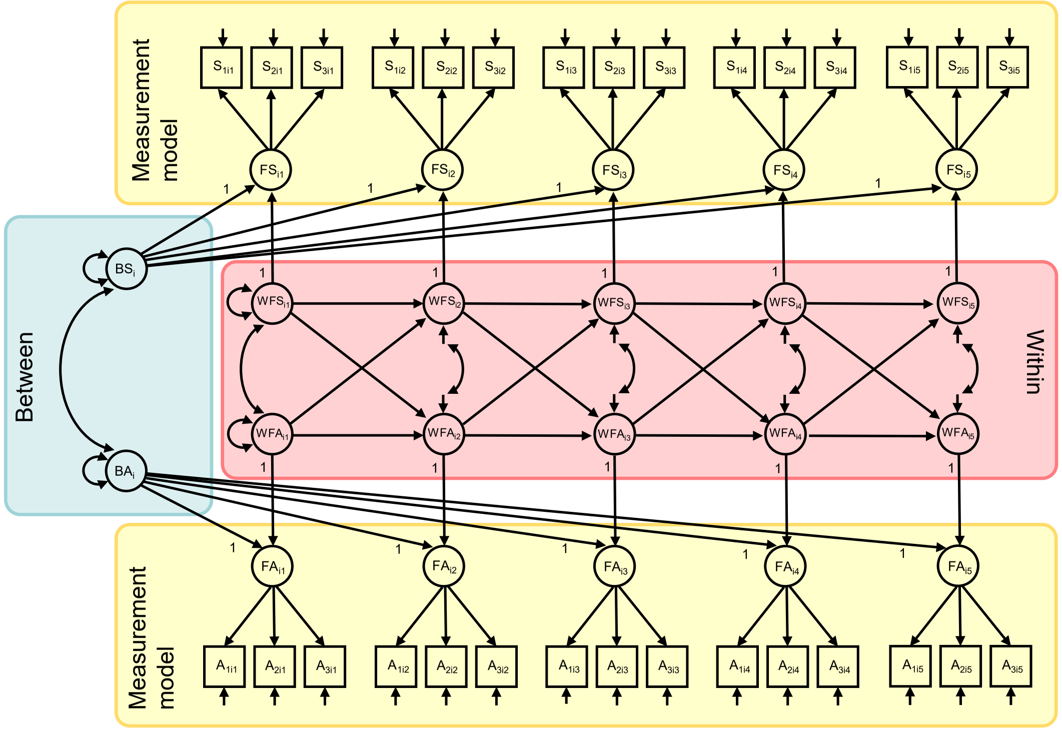
To allow for a meaningful comparison of factors over time, the factor loadings should be time-invariant, such that there is (at least) weak factorial invariance over time (Meredith, 1993; Millsap, 2011). If we are unable to establish this invariance, it implies that the constructs that we try to measure are interpreted differently over time, and it is difficult to make meaningful comparisons between the constructs measured at different occasions. Below we discuss the sequence of models that needs to be considered to establish longitudinal measurement invariance, and detail how the decomposition into within-unit and between-unit variance can be obtained in the context of multiple indicators.

In the first model, we decompose each observed variable into two parts: A stable, between-unit part, and a time-varying, within-unit part that indicators have in common (see Figure 4.4a). Thus, if we use three indicators to measure sleep problems, S_{1it}, S_{2it}, and S_{3it}, and three indicators to measure anxiety, A_{1it}, A_{2it}, and A_{3it},



(a) Multiple-indicator RI-CLPM with indicator-specific random intercepts that capture trait-like differences between units, and occasion-specific factors that capture the within-unit dynamics.

Figure 4.4: Two options for incorporating multiple indicators in a RI-CLPM.


Measurement model
S_{1i1} S_{2i1} S_{3i1} S_{1i2} S_{2i2} S_{3i2} S_{1i3} S_{2i3} S_{3i3} S_{1i4} S_{2i4} S_{3i4} S_{1i5} S_{2i5} S_{3i5}
FS_{i1} FS_{i2} FS_{i3} FS_{i4} FS_{i5}
1
Between
BS_i
BA_i
Within
WFS_{i1} WFS_{i2} WFS_{i3} WFS_{i4} WFS_{i5}
WFA_{i1} WFA_{i2} WFA_{i3} WFA_{i4} WFA_{i5}
FA_{i1} FA_{i2} FA_{i3} FA_{i4} FA_{i5}
Measurement model
A_{1i1} A_{2i1} A_{3i1} A_{1i2} A_{2i2} A_{3i2} A_{1i3} A_{2i3} A_{3i3} A_{1i4} A_{2i4} A_{3i4} A_{1i5} A_{2i5} A_{3i5}

(b) Multiple-indicator RI-CLPM in which there is a latent variable per occasion, which contains a trait-like part that is captured by the higher-order random intercepts, and a state-like part that is used to capture the dynamics over time.

Figure 4.4: Two options for incorporating multiple indicators in a RI-CLPM. (continued)

we specify six random intercepts to capture the trait-like part of each indicator. In addition, since we have five measurement occasions, we need to specify five within-unit components for sleep problems, WS_{it}, and five for anxiety, WA_{it}, that capture the common state-like part at each occasion. Moreover, we allow there to be an occasion- and indicator-specific residual, that captures what each observed variable does not share with itself at other occasions or with the other variables within the same occasion, thus capturing measurement error. At the latent within-unit level, we specify the dynamic model. Furthermore, we allow the within-person factors at the first occasion, and their residuals at subsequent occasions to be correlated within each occasion. The six random intercepts are allowed to be freely correlated with each other. In this model there are no constraints on the factor loadings over time for the within-unit factors; hence, this can be considered a model for *configural invariance*.

In the second model, we constrain the factor loadings to be invariant over time. This model is nested under the previous model, such that we can do a chi-square difference test. Fitting both models to our example data and comparing them yields $\Delta\chi^2(16) = 10.12$, $p = .861$ and we conclude that the model with invariant factor loadings over time does not fit significantly worse. Therefore, we can assume *weak factorial invariance* holds. In contrast, a significant test implies that the factor loadings cannot be constrained over time, making further comparisons between the latent variables problematic or even impossible. There are however two ways of dealing with this problem (Lek et al., 2018; Seddig & Leitgöb, 2018). First, by checking the modification indices, we can determine whether there is a specific factor loading at a particular measurement occasion that wildly deviates from the other factor loadings that it is constrained to be equal to. In such a case, researchers can choose to freely estimate this particular factor loading, resulting in a model that is based on *partial measurement invariance*. The model then accounts for a large measurement difference associated with a particular indicator while retaining weak measurement invariance for the rest of the indicators. Second, recently researchers have argued that the traditional concepts and tests of measurement invariance are too strict for small measurement differences. They advocate the use of approximate measurement invariance which allows for these minor differences through the use of priors in Bayesian estimation procedures. An introduction to the concept of approximate measurement invariance can found in Van de Schoot et al. (2013).

Assuming that weak factorial invariance holds, we can proceed with the third model and test whether *strong factorial invariance* holds. To this end, we specify a model in which we constrain the intercepts of the observed variables over time to be invariant, and estimate the latent means from the second occasion onward.[7] Again,

[7] Note that if we would not freely estimate the latent means, we would not only specify strong factorial invariance, but also specify a model in which there cannot be mean changes over time. Such a model may be of interest, for instance if you want to test for developmental trends, but that should

this model is nested under the previous model, such that a chi-square difference test can be performed to see whether the constraints hold. Applying this test to our example data, we find $\Delta\chi^2(16) = 21.64$, $p = .155$, which means we can assume that strong factorial invariance holds over time. In contrast, a significant chi-square difference test would mean strong factorial invariance does not hold, implying that the actual scores cannot be compared over time, but individual differences in scores can still be meaningfully compared since weak factorial invariance holds. As the focus in cross-lagged panel modeling is primarily on comparing individual differences (by decomposing the observed scores into between-unit and within-unit components) rather than mean scores over time, weak factorial invariance may be enough. However, from a measurement point of view, having strong factorial invariance would be considered more ideal.

Instead of including a random intercept at the observed level for each indicator separately, as shown in Figure 4.4a, we can also choose to specify the entire RI-CLPM at the latent level; this is illustrated in Figure 4.4b. This can be done in either a model with weak or strong factorial invariance over time. To this end, we specify the common factors that capture both trait-like and state-like common variance, and thereby make the assumption that the trait- and state-structures coincide. We then decompose these latent variables into a stable, between-unit part and the within-unit components. Although not immediately apparent, this model is nested under the model specified before. Instead of having free correlations between the six random intercepts as in the first model, we can model the connections between them by including two second-order factors: One for BS_{1i}, BS_{2i}, and BS_{3i}, and one for BA_{1i}, BA_{2i}, and BA_{3i}. We set the factor loadings of these second-order factors to be identical to the corresponding factor loadings of the within-unit factors. Additionally, we constrain the residual variances for the first-order factors to zero. This model is nested under the model presented in Figure 4.4a, and is statistically equivalent to the model presented in Figure 4.4b. This implies that we can use a chi-square difference test to compare the current model, as presented in Figure 4.4b, to the previous model, represented Figure 4.4b.

Comparing the current and previous model on our example data yields $\Delta\chi^2(18) = 17.23$, $p = .508$. This nonsignificant result implies that the current model does not have to be rejected, and we can say that there is measurement invariance across the stable between structure and fluctuating within-structure. If however the chi-square test is significant, then we need to conclude that these structures do not coincide, and temporal fluctuations within individuals take place on a different underlying dimension than the stable differences between units (see Hamaker et al., 2017, for further discussion on this).

be tested separately.

Finally, there are two important considerations that we want to emphasize in the context of having multiple indicators for the constructs on which one wants to perform the RI-CLPM. First, researchers commonly use a two-step procedure, in which they first compute factor scores, sum scores, or mean scores, which are then submitted to the RI-CLPM as if they were observed variables (e.g., Burns et al., 2020; Hesser et al., 2018; Keijsers, 2016). The disadvantage of using sum and mean scores however is that one assumes an absence of measurement error, which often is an unrealistic assumption, especially within the social sciences (Griliches & Hausman, 1986). Failing to properly account for measurement error can bias lagged-parameter estimates downwards, leading to a loss of power. Also, the estimation of factor scores is difficult due to the problem of factor indeterminancy (i.e., there are multiple ways to obtain factor scores, each with their own set of advantages and disadvantages), and it is unclear how this affects the results of the RI-CLPM.

Second, the procedure described above for establishing measurement invariance relies heavily on chi-square difference testing which, as mentioned before, can have serious disadvantages such as an increased Type I and Type II error rate when the base model is misspecified (Yuan & Bentler, 2004). Alternatively, researchers can use equivalence testing (Yuan & Chan, 2016), which allows researchers to explicitly specify an acceptable level of model misfit in their null-hypotheses when comparing the above sequence of models, and thereby retain acceptable Type I an Type II error rates.

4.5 Conclusion

The extensions discussed in this paper adhere to requests from researchers who want to use the decomposition into *time-varying within-unit dynamics* and *stable between-unit differences* in their panel research. While these extensions are mostly straightforward from a modeling point of view, they involve important assumptions, and researchers have to make important decisions with regards to this. The current paper therefore elaborated on diverse extensions, what choices can be made, how these are related, and provides hands-on experience with this modeling approach through our online supplementary materials. We hope that this enables researchers to tailor the RI-CLPM to their own research projects.

Author contributions: Conceptualization: EH - Data curation: EH and JM - Formal analysis: EH and JM – Investigation: EH and JM - Methodology: EH and JM – Project administration: EH and JM - Software: JM - Supervision: EH - Visualization: EH and JM - Writing (original draft preparation): EH – Writing (review and editing): JM and EH.

Funding: This study is part of the Consortium Individual Development, which was supported by a Gravitation grant awarded by the Dutch Ministry of Education, Culture, & Science, and the Netherlands Organization for Scientific Research [NWO grant number 024.001.003; 2013].

Supplementary materials: This study's online supplementary materials can be found at https://jeroendmulder.github.io/RI-CLPM.

CHAPTER 5

Power analysis for the random intercept cross-lagged panel model using the powRICLPM R-package

Mulder, J. D. (2023). Power analysis for the random intercept cross-lagged panel model using the powRICLPM R-package. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(4), 645-658. https://doi.org/10.1080/10705511.2022.2122467

Abstract

The random intercept cross-lagged panel model (RI-CLPM) is a popular model among psychologists for studying reciprocal effects in longitudinal panel data. Although various texts and software packages have been published concerning power analyses for structural equation models (SEM) generally, none have proposed a power analysis strategy that is tailored to the particularities of the RI-CLPM. This can be problematic because mismatches between the power analysis design, the model, and reality, can negatively impact the validity of the recommended sample size and number of repeated measures. As power analyses play an increasingly important role in the preparation phase of research projects, an RI-CLPM-specific strategy for the design of a power analysis is detailed, and implemented in the R package powRICLPM. This paper focuses on the (basic) bivariate RI-CLPM, and extensions to include constraints over time, measurement error (leading to the stable trait autoregressive trait state model), non-normal data, and bounded estimation.

A popular model among psychologists for the analysis of panel data is the random intercept cross-lagged panel model (RI-CLPM). It was first formally introduced by Hamaker et al. (2015) as an extension of the traditional cross-lagged panel model (CLPM; Rogosa, 1980) to account for stable, between-unit differences in the data. Unlike the CLPM, the RI-CLPM separates stable, between-unit variance from fluctuating, within-unit variance: The autoregressive effects can then be interpreted as purely within-unit effects and carry-over (rather than estimates of stability of the rank-order of units, as is the case in the CLPM), and cross-lagged effects can be interpreted as the within-unit effect or "spillover" of one domain into another (Mulder & Hamaker, 2021). This feature addresses some long-standing concerns that researchers have had about panel data analysis, such as the conflation of within- and between-unit variance for studying within-unit processes, unobserved heterogeneity, and bias in the cross-lagged effects due to omitted variables (Andersen, 2022; Hamaker et al., 2015; Heise, 1970; Kenny & Zautra, 1995, 2001). The reader is referred to Usami et al. (2019), Zyphur, Allison, et al. (2020), and Zyphur, Voelkle, et al. (2020) for an overview of how related SEM models address these concerns.

A frequently asked question by substantive researchers in relation to the RI-CLPM, is about the required sample size for detecting hypothesized effects. Such questions of statistical power are especially relevant in the design phase of a study: Underpowered study designs are more likely to result in Type II errors (i.e., incorrectly failing to reject the null-hypothesis of no effect), whereas overpowered studies (i.e., study designs with a sample size larger than necessary to find hypothesized effects) can place an unreasonable burden on the research resources (Zhang & Liu, 2018). While there are some general rules of thumb in the structural equation modeling (SEM) literature for what is regarded an adequate sample size (cf., Barrett, 2007; Jackson, 2003; Little, 2013; MacCallum et al., 1996), in practice, statistical power depends on many factors and assumptions, making it difficult to come up with a generally applicable sample size recommendation. When planning a longitudinal study, it is therefore advized to perform a power analysis that is tailored to a particular research context and research question, to find the optimal study design (Oertzen et al., 2010; Wang & Rhemtulla, 2021; Wolf et al., 2013). However, it can be challenging to design and perform such a study for researchers who are inexperienced with simulation-based power analyses, the particularities of a model, and the software required to automate the process.

This paper proposes a strategy for setting up and executing a power analysis for the RI-CLPM based on Monte Carlo simulations, and implements it in the R package powRICLPM. Although treatments on the design and implementation of Monte Carlo studies have appeared before (cf., S. Lee, 2015; L. K. Muthén & Muthén, 2002; Paxton et al., 2001; Wang & Rhemtulla, 2021; Zhang & Liu, 2018), these texts are not particular to the characteristics of the RI-CLPM, or target model (mis)fit rather than specific parameters within the model. Performing a power analysis for the RI-CLPM involves numerous model-specific and complex decisions which have not been described in the literature yet. The focus of this paper is on à priori power analysis (e.g., during the planning phase of a study, or as part of a grant proposal), but the procedure can similarly be used for post hoc power analysis (e.g., at a reviewer's request, or because dropout or non-response resulted in a lower-than-expected sample size; Hancock & French, 2013).

The paper is organized as follows: First, an illustrative example concerning academic amotivation is introduced that is used throughout. Second, the RI-CLPM itself is presented, as well as the factors influencing its power. Third, a six-step power analysis strategy is laid out. Fourth, this strategy is demonstrated using the powRICLPM package and the illustrative example. Fifth, extensions of the power analysis strategy are described, including the addition of constraints on parameters over time, measurement error (thereby leading to the bivariate stable trait autoregressive trait state model by Kenny & Zautra, 2001), non-normal data, and bounded estimation. This paper concludes with limitations of the proposed procedure, a comparison with other software packages for power analysis, and directions for future development.

5.1 Illustrating example: Self-alienation and academic amotivation

Suppose we are interested in the prevention of loss of academic motivation in students, and have reason to believe (based on previous research and expert opinion) that self-alienation is a driving factor herein. More specifically, we want to investigate the reciprocal effects of self-alienation X (the feeling that one does not know oneself) and academic amotivation Y (a lack of intrinsic or extrinsic motivation for pursuing academic goals) in college students over time. Unfortunately, due to time and money constraints, we are unable to design a randomized experiment in which self-alienation X is intervened on, and are therefore bound to observational data. As such, we want to use the RI-CLPM to estimate cross-lagged effects while controlling for stable, between-person differences in self-alienation and academic amotivation. Suppose further that we deem the assumptions underlying the RI-CLPM plausible, namely that the reciprocal effects between self-alienation and academic amotivation are (a) linear; (b) constant across units, that is, homogeneity; (c) constant across the values of our observed variables and error terms, that is, no effect modification; (d) not affected by unobserved time-varying confounding; and (e) the error terms (approximately) follow a multivariate normal distribution (Gische & Voelkle, 2022).

Prior to the start of the data collection, we want to perform a power analysis

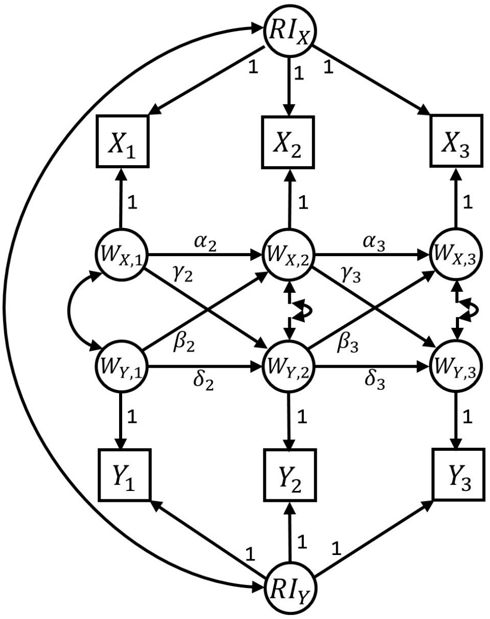


Figure 5.1: A bivariate random intercept cross-lagged panel model with three waves of data. α_t and δ_t are autoregressive effects of W_X and W_Y, respectively. γ_t and β_t are the cross-lagged effects of $W_{X,t-1}$ on $W_{Y,t}$ and $W_{Y,t-1}$ on $W_{X,t}$, respectively. The grand means $\mu_{X,t}$ and $\mu_{Y,t}$ are not included here.

to determine the required sample size N and number of repeated measures T for detecting potential reciprocal effects with a power level of 0.8. Planning for N and T is a matter of balance: It can be beneficial in terms of time and costs to collect an additional wave of data rather than additional participants, or vice versa, while maintaining the desired level of statistical power. Some researchers, like Winkens et al. (2006), explicitly include a "costs function" in their power analysis to determine an optimal trade-off in terms of sample size and number of repeated measures (as well as other factors).

5.2 The model

Figure 5.1 presents an RI-CLPM for a study design with three repeated measures. Let X_{it} and Y_{it} be the observed values for self-alienation and academic amotivation for individual i at time point t, respectively. By fitting an RI-CLPM, these observed variables are decomposed into three independent components: Grand means $\mu_{X,t}$ and $\mu_{Y,t}$ for each time-point, time-invariant random intercept factors $RI_{X,i}$ and $RI_{Y,i}$, and time-varying within-components $W_{X,it}$ and $W_{Y,it}$, for self-alienation and academic amotivation respectively.These decompositions are represented by

$$X_{it} = \mu_{X,t} + RI_{X,i} + W_{X,it}, \tag{5.1}$$

$$Y_{it} = \mu_{Y,t} + RI_{Y,i} + W_{Y,it}. \tag{5.2}$$

The grand means are time-specific means across all individuals, and they can be freely estimated in the model or constrained over time. The random intercepts RI_X and RI_Y are latent factors, with the observed measures for self-alienation as indicators for RI_X, and the observed measures for academic amotivation for RI_Y. They capture individuals' stable, time-invariant (i.e., for the duration of the study) deviations from the grand means $\mu_{X,t}$ and $\mu_{Y,t}$, such that the random intercept factors exclusively represent between-person variance. In the standard RI-CLPM presented here, the factor loadings of the random intercept factors are fixed at 1, implying that the size of the stable, between-person differences is invariant over time. However this may be an assumption that researchers want to check by freely estimating these factor loadings and comparing model fit (Mulder & Hamaker, 2021). Finally, the within-components $W_{X,it}$ and $W_{Y,it}$ represent the deviation of an individual at a specific time-point from the individual's expected score based on the grand mean and the random intercept.

Next, autoregressive and cross-lagged effects are added between the within-components at subsequent waves, such that

$$W_{X,it} = \alpha_t W_{X,i,t-1} + \beta_t W_{Y,i,t-1} + u_{it}, \tag{5.3}$$

$$W_{Y,it} = \delta_t W_{X,i,t-1} + \gamma_t W_{Y,i,t-1} + v_{it}, \tag{5.4}$$

where α_t represents the autoregressive effect of self-alienation from wave $t-1$ to wave t, δ_t represents the autoregressive effect of academic amotivation from wave $t-1$ to wave t, β_t is the cross-lagged effect from academic amotivation at wave $t-1$ to self-alienation at wave t, and γ_t is the cross-lagged effect from self-alienation at wave $t-1$ to academic amotivation at wave t. u_{it} and v_{it} are zero-mean normally distributed residuals with variances $\sigma^2_{u,t}$ and $\sigma^2_{v,t}$, respectively, and they are allowed to covary with each other within each wave. Because the within-person variance is separated from the stable, between-person variance, the lagged effects pertain exclusively to within-person fluctuations. The autoregressive parameters α_t and δ_t can then be interpreted as carry-over or inertia (Kuppens et al., 2010; Suls et al., 1998), whereas the cross-lagged parameters β_t and γ_t represent the within-person spill-over of the one construct into the other (i.e., controlled for stable, between-person differences), and vice versa. Finally, including a covariance between the between-components RI_X and RI_Y completes the basic setup of the RI-CLPM. The model is flexible, and can be extended to include constraints over time, time-invariant and time-varying predictors and outcomes, multiple groups, multiple indicators (Mulder & Hamaker, 2021), and interactions to test for moderation (Ozkok et al., 2022; Speyer et al., 2023).

5.2.1 Factors influencing power

Besides sample size and the number of repeated measures, there are many other factors that influence the RI-CLPM's power to detect individual non-zero parameters.[1] It is important to carefully consider and include these in the setup of a power analysis as they can impact the validity of the power analysis results. Here, these factors are divided into two groups: Characteristics of the study design, and characteristics of the data.

Characteristics of the study design that influence statistical power are interesting because they are under control of the researcher, and can be tweaked to achieve the desired amount of power. Sample size and number of repeated measures are the two most obvious examples of such factors. Others include (a) the significance criterion, where a larger criterion leads to a higher probability of rejecting the null-hypothesis of no effect, but also increases the probability of Type I errors (Zhang & Liu, 2018); (b) model complexity, where it has been suggested that models with fewer freely estimated parameters, for example due to imposed parameter constraints over time, have more power to detect non-null effects (Wang & Rhemtulla, 2021); and (c) in the case of a multiple-indicator RI-CLPM (MI-RICLPM), the number of indicators, as the inclusion of multiple indicators allows for controlling for measurement error, thereby increasing power (Oertzen et al., 2010; Wang & Rhemtulla, 2021).

Characteristics of the data that impact power are important to consider as well. Even though these cannot be controlled by the researcher, failing to adequately represent these data characteristics in the simulated data in the power analysis can negatively affect the validity of the power analysis results. These factors include (a) the effect size, where larger effects in the data result in larger test statistics, and thus greater power to reject the null-hypothesis of no effect (Wang & Rhemtulla, 2021); (b) non-normality, because many SEM models actually assume multivariate normal data, and non-normality then negatively impacts power (Zhang, 2014); (c) missing data, although some missing data patterns have a larger impact than others (Zhang & Liu, 2018); (d) the reliability of indicators, where smaller measurement error vari-

[1] To get some intuition of why increasing the number of repeated measures increases power in the RI-CLPM, it is useful to consider the concept of *cluster-mean centering* from the multilevel literature (Kreft et al., 1995). Rewriting Equation 5.1 shows that the within-components in the model are obtained by subtracting the between-components from the observed variables. The parameters governing RI_X and RI_Y are unknown, however, and must be estimated from the data. This introduces measurement error in the between-components, and by Equation 1, also in the within-components (Asparouhov & Muthén, 2019). A larger number of repeated measures reduces the measurement error in RI_X and RI_Y, resulting in less error in the within-components, and thereby increasing the power to detect lagged effects (Zhang & Liu, 2018).

ances leads to larger power (this is also related to the impact of the use of multiple indicators on power; Oertzen et al., 2010; Wang & Rhemtulla, 2021); and (e) the proportion of between-unit variance in the observed data. This last factor warrants additional explanation as the decomposition of observed variance into independent between-unit and within-unit variance is particular to the RI-CLPM. If a large portion of the observed variance is captured by the random intercepts, this implies that relatively little variance remains in the within-components. Consequently, point estimates of parameters at the within-unit level of the model, including the lagged effects of interest, are less certain, leading to higher standard errors and lower power. This point will be illustrated in Section 5.4 using the illustrative example.

5.3 The power analysis strategy

With the illustrative example, the RI-CLPM, and the factors influencing its statistical power introduced, a strategy for RI-CLPM power analysis is presented next. Because analytical solutions to power-related questions often become intractable in realistic situations, with small sample sizes, complex models, and when the underlying assumptions of a model are not met (Bandalos & Leite, 2013; S. Lee, 2015), this power analysis strategy relies on Monte Carlo simulations instead (Paxton et al., 2001). In general, a Monte Carlo study is based on generating R samples from a model that is thought to represent the population of interest (referred to as the *population model*), and then estimating the parameters in each sample $r = 1, ..., R$. The parameter estimates from each sample are then collected, forming an (artificial) empirical sampling distribution for each parameter. The performance of the estimated parameters is then based on properties of this sampling distribution. In the case of a power analysis, the population model is the same as the estimated model: Here, the RI-CLPM. Sampling distribution properties of interest include the proportion of times the confidence interval for the parameter(s) of interest does not include zero (i.e., the power), the width of the associated confidence interval(s) (i.e., the accuracy), and the mean square error (MSE). Other properties exist, like the percentage and relative bias, the standard deviation around the mean parameter estimate, and the coverage rate of the confidence interval, but these are typically not the primary focus of a power analysis.

The power analysis strategy presented here contains 6 steps:

1. Determine experimental conditions of interest (e.g., with varying sample sizes, numbers of repeated measures, or proportions of between-unit variance, amongst other things).

2. Choose and compute population parameter values.

3. Generate data from an RI-CLPM using the population parameter values from step 2.
4. Estimate an RI-CLPM on the data generated in step 3.
5. Repeat steps 3 and 4 R times for each experimental condition.
6. Summarize the R results and compare across experimental conditions.

5.3.1 Step 1: Define experimental conditions

The first step entails determining the experimental conditions that you are interested in simulating the power for. In this context, an experimental condition is a combination of values for each factor that influences the RI-CLPM's power, for instance, the experimental condition with a sample size of 500, three repeated measures, a significance criterion of 0.05, a 50:50 proportion of within- and between-unit variance, data with a skewness of 0, etc. In an à priori power analysis, a range of experimental conditions is included, where sample sizes and numbers of repeated measures are typically varying across conditions. If none of the included experimental conditions leads to the desired amount of power, the range of experimental conditions can be extended.

The key issue here is determining what are realistic values for the factors (other than the sample size and the number of repeated measures) that make up an experimental condition. If data are generated under conditions that are not representative of empirical data, the validity of the power analysis results can be severely limited (S. Lee, 2015; Paxton et al., 2001). This can happen, for example, when researchers wrongly assume a 90:10 proportion of within-:between-unit variance, whereas in reality it is approximately 50:50. Therefore, it is recommended to define values for these factors using theory, such that any decisions can be explained and defended. Previous studies on the same topic and expert knowledge can be important sources of information for deciding what are realistic values (Bandalos & Leite, 2013; L. K. Muthén & Muthén, 2002). When the appropriateness of certain choices are ambiguous, it might be recommendable to limit the values to conservative options: For example, a higher proportion of between-unit variance, or increased levels of non-normality. Alternatively, these factors can be allowed to vary across simulation conditions as well, rather than relying on a single (ambiguous) decision. This allows the researcher to determine what conditions are tolerable without loss of the desired power level (Bandalos & Leite, 2013).

For the illustrative example, let the sample size range from 200 to 2000 using steps of 200, and the number of repeated measures range from three to five. Regarding appropriate values for the proportion of between-unit variance, J. Kim et al. (2018) found 56% and 59% stable, between-unit variance for self-alienation and academic amotivation, respectively, for biweekly measurements, for a total of eight weeks. However, there is likely to be some uncertainty in basing the proportion of between-unit variance on previous research because research designs and contexts are never perfectly equivalent. For example, intervals of one month between repeated measures might be thought to better reflect the time it takes for the causal effect under study to take place (Heise, 1970; Mitchell & James, 2001), and therefore plan to take monthly measurements rather than biweekly measurements. This difference in the timing of measurements is likely to affect the proportion of between-unit variance in the collected data. Therefore, to take this uncertainty into account, a range of proportions of between-unit variance is included in the power analysis, namely 0.3, 0.5, and 0.7. Furthermore, for the sake of interpretability of this illustrative example, the traditional significance criterion of 0.05 is used to denote significance, and we assume no deviations from normality, no missing data, and no measurement error (i.e., perfect reliability of the indicators).

5.3.2 Step 2: Choose and compute population parameter values

To generate data in step 4, a population model needs to be specified that acts as a data generating mechanism. To this end, population values need to be specified for each parameter in the RI-CLPM first. In this strategy, populations values need to be set for (a) the standardized autoregressive and cross-lagged effects; (b) the correlations between within-components; and (c) the correlation between the random intercepts. Similar to defining the experimental conditions in step 1, the key issue is to choose population parameter values that are realistic. Again, it is recommended to base these decisions on previous literature, or expert opinion. In case of uncertainty, a good strategy might be to be conservative: Pick values on the smaller side of the plausible range. Other parameters in the RI-CLPM are either computed from these population parameter values (like the residual variances and covariances of the within-components at wave 2 and further), determined based on the experimental conditions as defined in step 1 (the random intercept variances), or set to 0 because they are not of primary interest here (the mean structure).

5.3.2.1 Step 2.1: Within-unit parameters

Within-unit parameters include the autoregressive effects α_t and δ_t, cross-lagged effects β_t and γ_t, variances and covariance for the within-components at wave 1, and the *residual* variances and covariances for the within-components at wave 2 and further. For the lagged effects we rely on the specification of *standardized* effects in the population model such that the power analysis does not depend on any particular metric. J. Kim et al. (2018) report standardized autoregressive effects of self-alienation and academic amotivation between 0.206 and 0.266, and between 0.294 and 0.529, respectively. The standardized cross-lagged effect from self-alienation to academic amotivation was estimated to be between 0.08 and 0.104, while the reverse effect was estimated to be between 0.154 and 0.301. Our strategy is to be conservative, and as such we specify a small effect of 0.20 for the autoregressive effect of self-alienation, and a small to medium effect of 0.30 for the autoregressive effect of academic amotivation (following guidelines by Cohen, 1988). This conservative approach would imply that the population parameter values for the cross-lagged effect of self-alienation to academic amotivation are set to be extremely small (i.e., 0.08). However, such small effects are arguably not interesting in practice and it is recommended to use a cutoff value—for example 0.10 as recommended by Paxton et al. (2001)—for population parameter values. We set the cross-lagged effects to be 0.10 for the effect from self-alienation to academic amotivation, and to be 0.15 for academic amotivation to self-alienation.

For these population values to be interpreted as standardized effects, the variances of the within-components need to be 1. At the first wave, the variances can be set to one directly as these variables are exogenous, and their covariance (which is now also the correlation) is set to 0.26, as reported by J. Kim et al. (2018). However, setting the variance of the within-components at wave 2 and further is more involved because these variables are endogenous. Hence, only the variances and covariance of the *residuals* can be set directly, rather than the variances of the within-components *themselves*. Taking the within-components of academic amotivation and self-alienation at wave 2 as an example, it can be shown using the path-tracing rules that their variances are a function of the variance they "inherit" from their predictors (the within-components of academic amotivation and self-alienation at the first wave), as well as the variance from their residuals (see Appendix 5.A for details). Therefore, we must compute how much variance the residuals should have such that, together with the variance from their predictors, the variance of the within-components add up to one.

The residual variances and covariance can be expressed as a function of the population values for the lagged effects and the correlations between the within-components (C. J. Kim & Nelson, 1999), as derived in Appendix 5.B. Using this relationship, the residual variances and residual covariance for a wave t, given the lagged effects and the correlation between the within-components at the previous wave $t-1$, can be computed such that it results in a variance of one for the within-components at wave t. For our example, this results in a residual variance of the within-component of self-alienation of 0.9219, a residual variance of the within-component of academic amotivation of 0.8844, and a residual covariance of 0.1755 at wave 2. Under the assumption of stationarity, whereby the lagged effects and the variances of and correlations between the within-components do not change over time, the residual variances and the residual covariance similarly apply to the within-components at wave 3 and further, ensuring that *all* within-components have a variance of one, and that the population lagged effects can be interpreted as standardized effects at each time point.

5.3.2.2 Step 2.2: Between-unit parameters

Next, the population parameter values for the between-unit parameters need to be set, including the variances of, and covariance between the random intercepts. As the variances of the within-components are designed to be one, the ratio of between- and within-unit variance is determined by the variance of the random intercepts: Setting the random intercept variances to one implies a 50% between- and 50% within-unit variance, while setting the random intercept variance to three leads to 75% between- and 25% within-unit variance in the observed variables. The proportion of between-unit variance is already chosen in step 1, so the exact population value of the random intercept variances can simply be computed from this. Finally, along the lines of J. Kim et al. (2018) we set the correlation between the random intercepts to be 0.35.

Optionally, the means can be set in the population model. While this can be of interest in the case of a multiple indicator RI-CLPM, the mean structure is typically not of (primary) interest in any basic RI-CLPM. In the former case, it is recommended to test if strong measurement invariance over time holds and therefore researchers should constrain the factor loadings and intercepts/means over time. As this is not of interest in the illustrative example, the mean structure is ignored here and the grand means are set to zero, $\mu_{X,t} = \mu_{Y,t} = 0$.

5.3.3 Steps 3-5: Generate data, estimate RI-CLPM, repeat

Once the design of the power analysis has been decided upon (i.e., the experimental conditions and population parameter values are defined), it can be implemented. The process of running a Monte Carlo power analysis—repeatedly generating a sample of data from the population model and estimating parameter values—creates a lot of data and requires adequate computing power. It is therefore important to automate the process, and the R package powRICLPM has been created specifically for this purpose, implementing the power analysis strategy outlined here. The package is demonstrated in Section 5.4.

There are two more factors to consider in these steps: The number of replications R, and the seed. First, the number of replications needs to be large enough to ensure that the results have converged to a stable solution (L. K. Muthén & Muthén, 2002): Too few replications will lead to a large uncertainty around the results, whereas too many replications can take a long time to run, especially for complex models, and large numbers of experimental conditions. An alternative strategy for dealing with many experimental conditions is to first run a power analysis with a reduced number of replications (e.g., 50 or 100 replications) to get preliminary results, and then validate these results, using a larger number of replications only for those experimental conditions that are close to the desired power levels. Second, use a seed to determine the starting point for the random simulation of data. This ensures that the results can be replicated.

5.3.4 Step 6: Summarize results

Before interpreting the results, it is important to check if certain experimental conditions resulted in a high number of convergence issues or inadmissible results (e.g., negative variances). This indicates that the results of the power analysis for these conditions might be unreliable, and that estimation of the model in these conditions is unstable. The powRICLPM package keeps track of convergence issues, inadmissible parameter estimates, and fatal errors terminating an estimation procedure, and can report to the user the number of times each occurs per experimental condition.

Next, multiple metrics can be computed that summarize the simulated sampling distributions for the parameter of interest, per experimental condition. The powRICLPM package reports (a) the mean estimate over all R replications; (b) the standard deviation of the estimates; (c) the mean standard error of the estimates; (d) the mean square error; (e) the accuracy, computed as the mean length of the confidence interval; (f) the coverage rate, computed as the proportion of times the confidence interval included the true population value; and (g) the power (L. K. Muthén & Muthén, 2002, 2017). In addition, the powRICLPM can visualize how these metrics change across experimental conditions, for example as a function of sample size, number of repeated measures, and the proportion of between-unit variance. Using these metrics and visualizations, the sample sizes and number of repeated measures that lead to the desired amount of power can be determined.

5.4 The powRICLPM package

The powRICLPM package provides functions to automate steps 2 to 5, as well as methods for summarizing the results of the analysis as described in step 6. For steps 3 and 4, it uses the R-package `lavaan` on the back-end. Up-to-date information about the functionality of the package, as well as instructions on how to install it and fully annotated R-code for the illustrating example in this article, can be found in the package's documentation at https://jeroendmulder.github.io/powRICLPM.

The main function powRICLPM() implements steps 2 to 5, and follows the procedure as outlined in Figure 5.2. First, users must specify (a) which experimental conditions they want to explore using the sample_size, search_*, time_points, and ICC arguments; (b) the population parameter values of the lagged effects, and correlations between the within-components and between the random intercepts in the Phi, within_cor, and RI_cor arguments, respectively; and (c) their desired power level in the target_power argument. Optionally, users can specify skewness and kurtosis values to generate non-normal data, impose constraints on the estimation model, include measurement error in the simulated data, estimate measurement error variances (leading to the stable trait autoregressive trait state model), and use bounded estimation (De Jonckere & Rosseel, 2022), among other things. Second, this input is used to compute the residual variances and covariances, and the random intercept variances for the population model. Third, for each experimental condition lavaan model syntax is generated to simulate data and estimate the RI-CLPM. Fourth, this syntax is used to repeatedly simulate data and estimate the RI-CLPM. Parameter estimates and standard errors are collected, and summaries are saved in a powRICLPM object. To quantify the uncertainty around the simulated power, powRICLPM implements a non-parametric bootstrapping procedure of the results: It involves taking B bootstrap samples (by default: 1000) of the significance of the parameter estimates to create a bootstrap distribution of the power for each experimental condition (Constantin et al., 2023). The 95% confidence intervals of these bootstrap distributions represent the uncertainty around the simulated power. Finally, the user can use various methods such as summary(), give(), and plot() to explore the results.

5.4.0.1 Illustrating example

The illustrating example concerned determining the required sample size and number of repeated measures for detecting cross-lagged effects between self-alienation and academic amotivation with a power of 0.80. In step 2.1, it was argued that small cross-lagged effects of 0.10 and 0.15 were reasonable effect sizes to include in the power analysis (based on J. Kim et al., 2018), yet large enough to be practically interesting. Furthermore, in step 1 it was determined to include a range of proportions of between-unit variance in the experimental conditions, namely 0.3, 0.5, and 0.7. Continuing with this example, the experimental conditions to be included in the power analysis are further defined by selecting sample size candidates ranging from $N = 200$ to $N = 2000$, increasing with steps of 100, and numbers of repeated measures from $T = 3$ to $T = 5$. In total, this results in 19 sample sizes × 3 numbers of repeated measures × 3 proportions of between-unit variance, totalling 171 experimental conditions.

Following the suggestion in Section 5.3.3, the analysis is partitioned into a preliminary phase with reduced number of replications ($R = 100$), and a validation phase (R

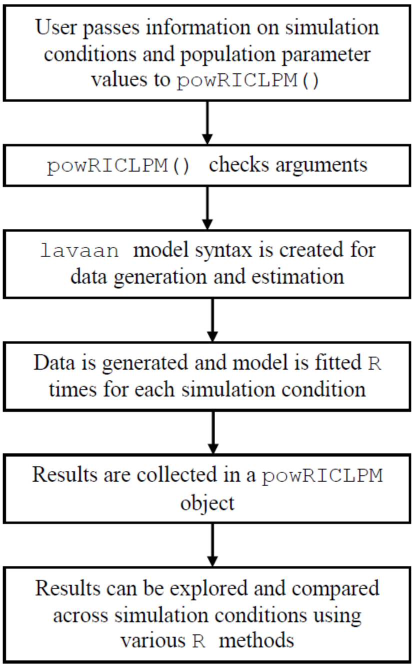

User passes information on simulation conditions and population parameter values to powRICLPM()
powRICLPM() checks arguments
lavaan model syntax is created for data generation and estimation
Data is generated and model is fitted R times for each simulation condition
Results are collected in a powRICLPM object
Results can be explored and compared across simulation conditions using various R methods

Figure 5.2: Overview of power analysis procedure used by powRICLPM.

= 2000) with only those experimental conditions that are close to the desired results. The preliminary power analysis can be run using

```
out_preliminary <- powRICLPM(
    target_power = 0.8,
    search_lower = 200,
    search_upper = 2000,
    search_step = 100,
    time_points = c(3, 4, 5),
    ICC = c(0.3, 0.5, 0.7),
    RI_cor = 0.35,
    Phi = Phi,
    within_cor = 0.26,
    reps = 100,
    seed = 123456
)
```

where target_power denotes the desired power level, the search_ arguments define the lower bound, upper bound, and step size of the range of sample sizes to include, respectively, time_points denotes the numbers of time points, ICC denotes the proportions of between-person variance, RI_cor denotes the correlation between the random intercept factors, Phi refers to a matrix of lagged effects (see Appendix 5.B), within_cor defines the correlation between the within-components, reps sets the number of Monte Carlo replications, and seed sets a seed for replicability. A visualization of the preliminary results across all 171 experimental conditions, specifically the power to detect a cross-lagged effect of 0.10 (standardized), can be obtained using

```
plot(out_preliminary, parameter = "wB2~wA1")
```

and is displayed in Figure 5.3. Details about the naming conventions of parameters, ways to speed up the analysis using multicore processing, and tracking the analysis progress can be found in the online documentation of the powRICLPM, and in the powRICLPM() function documentation (accessible via ?powRICLPM()).

Inspecting the results using

```
summary(out_preliminary)
```

shows that there were no fatal errors or convergence issues across any conditions in the preliminary power analysis. However, for the condition with three time points, 30% between person variance, and sample sizes from 200 to 500, there were eight, five, four, and two replications with inadmissible results, respectively. Investigating this further for the case with a sample size of 200, using

```
summary(
    out_preliminary,
    sample_size = 200,
    time_points = 3,
    ICC = 0.3
)
```

shows that the problematic parameter is likely the variance of the random intercepts: The minimum estimate (across all replications) is negative, which leads to the inadmissible value warning. These inadmissible results might lead to bias in other parameters as well, and hence it is advisable to err on the side of caution while interpreting results for these experimental conditions (De Jonckere & Rosseel, 2022). A solution might be the use of bounded estimation, which is introduced in Section 5.5.

While these are only the preliminary results, the influence of sample size, number of repeated measures, and proportion of between-unit variance on power are already clearly visible in Figure 5.3: Experimental conditions with a higher number of repeated measures have more power to detect the cross-lagged effect of 0.10, and similarly for conditions with a relatively small proportion of stable, between-unit variance in the observed data. Focusing on the relation between number of time points and power, the preliminary results suggest that with the current range of sample sizes and

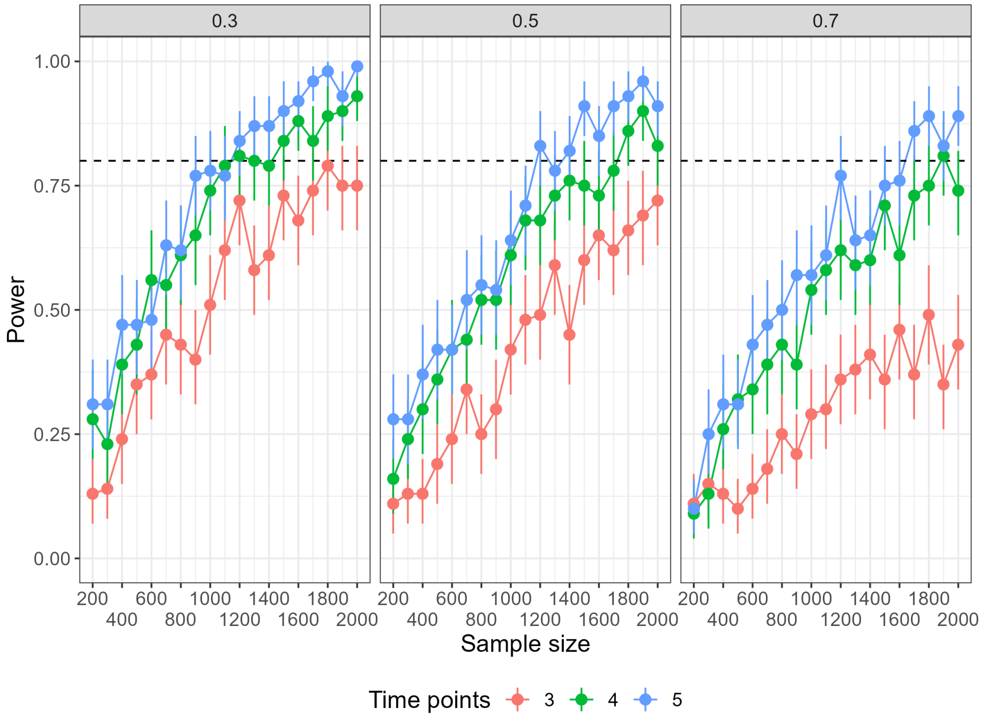


Figure 5.3: Results of preliminary power analysis for the RI-CLPM, based on 100 replications, for a cross-lagged effect of 0.10 (standardized). The different panels display results for conditions with a 0.3, 0.5 and 0.7 proportion of between-unit variance, respectively. The vertical error bars represent the uncertainty around the simulated power.

proportions of between-unit variance, we cannot achieve desirable power to detect a small cross-lagged effect with three time points. Furthermore, the results suggest that at least four time-points and a sample size upwards of a 1000 are required in the condition with the most advantageous proportion of between-unit variance (where proportion of between-unit variance is 0.3). For conditions with a 0.7 proportion of between-unit variance, sample sizes of approximately 1500 are needed with five repeated measures, whereas sample sizes upwards of 1700 are needed for four repeated measures. Based on these results, experimental conditions for validation are selected: The range of sample sizes is reduced to 900 to 1800, and experimental conditions with three repeated measures are omitted, resulting in 10 sample sizes × two numbers of repeated measures × three proportions of between-unit variance, totalling 60 experimental conditions for validation.

The validation results (obtained by increasing the `reps` argument of the R code above to `reps = 2000`) are displayed in Figure 5.4. The error bars representing the uncertainty surrounding the simulated power have shrunk considerably due to the

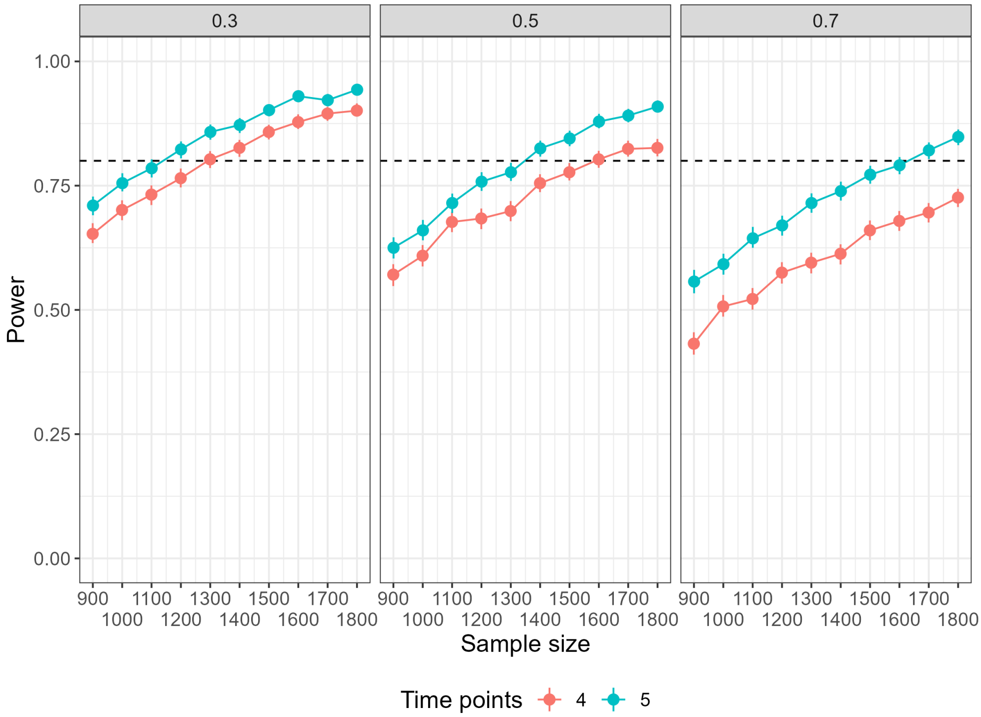


Figure 5.4: Results of the validation phase of the power analysis for the RI-CLPM, based on 2000 replications. The different panels display results for conditions with a 0.3, 0.5 and 0.7 proportion of between-unit variance, respectively. The vertical error bars represent the uncertainty around the simulated power.

higher number of replications, leading to more stable results. With equal proportions between- and within-unit variance (the middle panel of Figure 5.4, a sample size of approximately 1400 is needed for an RI-CLPM with five time points, and a sample size of 1600 is needed for RI-CLPM's with four time points, for detecting small cross-lagged effects. Compare this to data containing higher proportions of between-unit variance, where sample sizes of 1600 or more and more than 2000 are required for RI-CLPM's with five and four time points, respectively. In conditions with lower proportions of between-unit variance, a sample size of approximately 1100 is adequate for detecting small cross-lagged effects with five repeated measures, whereas a sample size of 1300 is needed for the case with four repeated measures.

5.5 Extending the power analysis

So far, the primary focus has been on a basic bivariate RI-CLPM with experimental conditions varying over sample size, number of repeated measures, and proportion of between-unit variance. However, researchers might want to include additional factors in their experimental conditions to better align the power analysis to their research question or empirical context. Below various extensions that have been build into the powRICLPM package are discussed briefly, specifically: (a) imposing various constraints over time on the estimation model; (b) including measurement error in the simulated data and in the estimation model; (c) simulating non-normal data (i.e., skewness and kurtosis; Blanca et al., 2013); and (d) the use of bounded estimation (De Jonckere & Rosseel, 2022). Again, technical details on the implementation of these extensions, as well as example code, can be found in the online documentation powRICLPM. Extensions such as the multiple-group RI-CLPM or the multiple-indicator RI-CLPM are not supported by the package (yet), but are briefly discussed in the Discussion section. Moreover, note that the RI-CLPM is a flexible model, and that further extensions of the model are likely to be developed (e.g., Ozkok et al., 2022), each with its own idiosyncrasies when it comes to power analysis.

5.5.1 Constraints over time

Means, autoregressive and cross-lagged effects, and residual variances and covariances can vary freely over time in the RI-CLPM. This is useful if the process under study is characterized by some kind of development, or if there are unequal intervals between repeated measures. For example, changes in the cross-lagged effects over time can be representative of a maturation process in individuals: The influence of one variable becomes more or less important in driving the other variable (and vice versa) as one gets older. However, imposing constraints on some of the parameters over time can be useful as well. It leads to more parsimonious results (e.g., a single set of lagged effects rather than different lagged effects between each pair of adjacent within-components), can reduce convergence issues, leads to interesting statistical equivalences with other popular panel models (for example, see Andersen, 2022; Hamaker, 2005), and increases power for the constrained parameters. In the proposed power analysis strategy above, the population model used to simulate data implicitly imposes constraints over time on the grand means μ (fixed to 0), lagged effects, and the residual variances and covariances. Essentially, the population model implies a stationary process such that for a person i the expected values $\mathbb{E}(X_{it})$ and $\mathbb{E}(Y_{it})$, variances $Var(X_{it})$ and $Var(Y_{it})$, and autocovariances $Cov(X_{it},\ X_{i,t+1})$ and $Cov(Y_{it},\ Y_{i,t+1})$ are independent of the time point t (Hamaker & Dolan, 2009). This was done for didactic purposes and ease of use of the power analysis strategy, as now only a *a single set* of population values for the lagged effects and within-component variance-covariance matrices has to be found, which can already be challenging. In contrast, the estimation model does not impose any of these constraints and freely estimates these parameters at each time point by default.

To accommodate researchers who choose to impose constraints over time on the estimation model, the powRICLPM-package includes various constraint specifications via the `constraints` argument. It allows users to simulate the power for their specific RI-CLPM specification of interest, including (a) constraints on the lagged effects over time with `constraints = "lagged"`, (b) constraints on the residual variances and covariances over time with `constraints = "residuals"`, or (c) constraints on both the lagged effects and residual variance and covariances over time with `constraints = "within"`. Note that constraining the lagged effects to be time-invariant is only advized when the time interval between repeated measures is (approximately) equal (Gollob & Reichardt, 1987; Kuiper & Ryan, 2018). Furthermore, constraints on the lagged effects pertain only to the unstandardized effects, while the standardized effects are likely to be time-varying still. This is because standardization uses the variances of the within-component predictor and outcome, and these are typically not constrained to be the same over time, even with constraints on the *residual* variances and covariances. To obtain both time-invariant unstandardized and standardized effects, *full stationarity* constraints need to be imposed, which can be done with `constraints = "stationarity"`. These constraints are a function of the estimated autoregressive and cross-lagged effects, residual covariances, and the covariance between the within-components at the first wave. The derivations for these constraints can be found in the online supplementary materials of Mulder and Hamaker (2021).

Finally, full stationarity constraints have a long tradition in the econometric literature on dynamic panel models (for example, see Hamilton, 1994). Therefore, it is understandable that some researchers are interested in this specific specification of the RI-CLPM. However, researchers should feel comfortable with the assumptions one makes à priori when incorporating these constraints in the power analysis, as there can be various reasons why such constraints are not justified (e.g., varying time-intervals, maturation processes, development, etc.). An alternative approach is to not assume time-invariant lagged effects and residual variances beforehand (as one does if these constraints are included in the power analysis), but instead test the tenability of them using the collected data (Mulder & Hamaker, 2021).

5.5.2 Measurement error

It is generally advisable to control for measurement error when analysing psychological data, as it is widely accepted that measurement error is likely to be present in psychological measurements (Steyer et al., 1992). While the RI-CLPM actually does not include measurement error, it can in theory be added if four or more waves of data are available. This would make the model equivalent to the bivariate trait state error (TSE) model by Kenny and Zautra (1995)—later referred to as the bivariate stable trait autoregressive state trait (STARTS) model (Kenny & Zautra, 2001)—without constraints over time and with the stable trait factor loadings fixed to one. However, the STARTS model is notorious for being empirically underidentified, commonly resulting in inadmissible solutions when sample sizes are small, leading Cole et al. (2005) to recommend a minimum of 8 waves of data (given a sample size of 500) or more. Therefore, the inclusion of measurement errors in the RI-CLPM can greatly impact the recommended sample size and number of repeated measures, not for reasons related to the power, but for reasons of empirical identification.

Within the powRICLPM package, users can include measurement error for the simulation of data (step 3) via the `reliability` argument, and in the estimation model (step 4) via the `estimate_ME` argument. The population values for the measurement error variances are determined by the package itself given the specified reliability of the indicators, the specified proportion of between-unit variance, and within-unit component variances of one.

5.5.3 Non-normally distributed data

The RI-CLPM is typically fitted in SEM software using maximum likelihood estimation, thereby assuming multivariate normally distributed data. However, Micceri (1989) concludes that asymmetry in empirical distributions appears to be the rule for psychometric measurements rather than the exception. This is problematic as non-normality of the data can negatively impact the power of SEM models (Yuan & Chan, 2016). Therefore, if researchers have reason to believe that multivariate normality might not be a reasonable assumption for the data they plan on collecting, the power analysis should incorporate non-normal data as well (Yuan & Chan, 2016). powRICLPM allows for incorporating data with various degrees of skewness and kurtosis via the `skewness` and `kurtosis` arguments of the `powRICLPM()` function.

5.5.4 Bounded estimation

Nonconvergence of the estimation model is disadvantageous for Monte Carlo power analyses because it reduces the effective number of replications the power analysis results are based on, and can slow down the analysis considerably as the optimization algorithm takes a long time searching for a solution that it ultimately does not find. It typically occurs when sample sizes are small (e.g., smaller than 100) and when the model is complex (e.g., when measurement error is included). Therefore, De Jonckere and Rosseel (2022) have implemented so-called bounded estimation in the R-package `lavaan`, placing bounds on the parameter space of the model. This prevents the optimization algorithm from searching in the completely wrong direction for parameters, for example, for negative solutions to the (residual) variances of latent variables.

Users of powRICLPM can make use of bounded estimation via the `bounded = TRUE` argument. Automatic wide bounds are then used as recommended by De Jonckere and Rosseel (2022), implying that (residual) variances (e.g., the random intercept variance, and residual variances of the within-components) have a small negative value as a lower bound, and the variances of the observed variables they load on as an upper bound. In the context of the RI-CLPM, the factor loadings are (usually) fixed, and no bounds are included for these parameters. The lagged effects are theoretically infinite, and hence there are no sensible bounds that can be placed à priori on these parameters.

5.6 Discussion

It is easy to underestimate the time and effort it can take to set up and execute a valid power analysis (e.g., see Footnote 2 of Paxton et al., 2001). While the increased focus within the scientific community on à prior power analysis is helpful for the progress of cumulative science, the design and execution of a valid power study is far from trivial for many applied researchers (De Jonckere & Rosseel, 2022; Maxwell et al., 2008). Nevertheless, investing time in a proper power analysis that is tailored to the particularities of one's study is well-worth the effort, as it helps in the prevention of under-powered studies and can reduce unnecessary demand on study resources.

In this article, a six-step Monte Carlo power analysis strategy that is tailored to the random intercept cross-lagged panel model was proposed and demonstrated. It was created with usability for applied researchers in mind and has been implemented in the R-package powRICLPM. For a basic power analysis, four sets of population parameter values are required as input, namely autoregressive and cross-lagged effects, variances and covariances for the within-unit components, the proportion of between-unit variance, and the correlation between the random intercepts. Choices for these population parameter values should be based on expert opinion or literature, or be grounded in theory. The powRICLPM package then computes the remaining population parameter values (e.g., the residual variances and covariances) and automates the process of repeatedly simulating data and estimating the model. Users can use the `summary()`, `give()`, and `plot()` functions to inspect the results, including convergence rates, mean square error, coverage rate, and power, among other metrics, across experimental conditions. Currently, the basic power analysis can be extended to include constraints over time on the estimation model, measurement error (i.e., the STARTS models), non-normal data, and bounded estimation.

5.6.1 Limitations

Step 2 of the strategy involves choosing population parameter values for the lagged effects and correlations between the within-unit components. While it is recommended to base these on theory and literature, this does not imply that any set of population parameters goes. There is a mathematical restriction on the population model-implied variance-covariance matrix that adds a degree of difficulty to the determination of these population parameter values, and introduces an element of trial-and-error to this step. Specifically, there are two restrictions that impact the population parameter values that users can specify. First, population values for the lagged effects should be chosen such that the data that are generated from these form a stable stationary system.[2] Second, the correlation matrix of the within-components is required to be positive definite in order to generate data from it.[3]. The powRICLPM package automatically checks if these restrictions are met, and throws an error otherwise. In that case, researchers should adjust the population parameter values for the lagged effects and correlations between the within-unit components accordingly, which often implies that these should be made smaller.

Furthermore, within a power analysis context one would expect the population model and the estimation model to be the same (i.e., it is assumed that the estimation model is actually the data generating model). However, as discussed in Section 5.5.1, there is a discrepancy in the proposed power analysis strategy *by design* between the population model used to simulate the data, and the model that is estimated. The population model is based on full stationarity constraints, affecting the lagged effects, residual variances and covariances and grand means, while the estimated model allows

[2] In technical terms, the eigenvalues of the matrix of lagged effects Φ should be within unit-circle.

[3] In technical terms, the eigenvalues of the variance-covariance matrix of the within-unit residuals should be positive.

all parameters to be freely estimated over time. This setup was chosen for reasons of usability, without compromising the validity of the power analysis results. It ensures that users need to specify only a single set of lagged parameters and a single correlation for the within-components, which can be quite challenging already. It also implies that the power analysis results are conservative for situations where these constraints are valid, in the sense that a higher power would be achieved if constraints over time had been imposed in the estimation model. Further note that this difference between the data generating mechanism and the estimation model can be overruled using the `constraints` argument.

Moreover, it is possible that small sample sizes (< 100) not only result in low statistical power, but also in bias in the parameter estimates. This is a phenomenon related to the large sample properties of maximum likelihood estimation, something that has been repeatedly reported on in the SEM literature (cf. De Jonckere & Rosseel, 2022; Rosseel, 2020; Wolf et al., 2013). The effect of small samples and a limited number of repeated measures on bias in RI-CLPM parameter estimates was not investigated here. However, it is advisable to check that bias is not a limiting factor (rather than power) for sample size recommendations when performing a power analysis using such limited sample sizes. For this, the bias as reported by powRICLPM package can be used.

A final limitation to take into account is that the sample size recommendations following the illustrating example assume a complete dataset, multivariate normally-distributed data, and no measurement error. However, missing data often do pose a problem in empirical datasets (especially in social sciences, it is nearly inevitable; van Buuren, 2018, p. 7), observed data can show considerable deviations from normality (Blanca et al., 2013; Micceri, 1989), and many (indirect) psychological and behavioural measures are likely to include measurement error. Therefore, the conclusions from this illustrating example should be considered as lower bounds, and in practice greater sample sizes might be required to counter the negative effects of these suboptimal conditions on the power.

5.6.2 Comparison to other packages

Many different software programs have been developed for doing power analyses for SEM. They can be roughly categorized based on whether the power analysis is analytical or simulation-based, the price (free or paid option), and their generality (focusing on SEM models in general, or specific to a particular model). Below the focus is on some software packages that can be useful alternatives for RI-CLPM power analyses. For a more extensive overview of software packages available for Monte Carlo simulation studies for SEM, the reader is referred to S. Lee (2015).

The software package Mplus by L. K. Muthén and Muthén (2017) is a latent variable modeling program with a wide range of analysis options including Monte Carlo simulation analyses, and can be used for power analysis for the RI-CLPM. The main advantage compared to powRICLPM is that it is much faster: Although no formal comparison of computation time was performed, from personal experiences a Monte Carlo power analysis for the RI-CLPM with a single experimental condition can take up to 10 minutes using powRICLPM, whereas it takes less than a minute using Mplus.[4] Disadvantages of Mplus are that it is a paid option, does not run multiple experimental conditions simultaneously, and is not tailored to the RI-CLPM. As such, more steps need to be taken by the user to specify the power analysis for the RI-CLPM, including, for example, computing the residual variances and covariance of the within-unit components. To accommodate users of Mplus, the powRICLPM includes the `powRICLPM_Mplus()` function to generate Mplus syntax for RI-CLPM power analysis (for multiple experimental conditions simultaneously), which can be run subsequently in Mplus itself.

Various analytical power analysis options for SEM are available as well, including WebPower by Zhang and Liu (2018) or functions within the semTools R-package by Jorgensen et al. (2022). These options are useful, especially for the multiple group extension of the RI-CLPM (Mulder & Hamaker, 2021). The multiple group RI-CLPM is based on fitting a multiple group version of the RI-CLPM both with and without constraints across groups (e.g., the constraint of equal lagged effects), and comparing the model fit to determine whether the imposed constraints are tenable. Power thus refers to the probability of rejecting a bad-fitting model due to untenable across-group constraints in this context, rather than rejecting the null-hypothesis for a specific parameter (Wang & Rhemtulla, 2021). The effect size then refers to how much worse the constrained model fits the data compared to the more general model (with less, or no across-group constraints). Analytic solutions, like the likelihood ratio test by Satorra and Saris (1985) or power analyses based on the RMSEA by MacCallum et al. (1996), are more efficient to use for these types of power analyses than computationally intensive Monte Carlo simulation studies. For example, Jak et al. (2021) describes how the `SSpower()` function from the R package semTools can be used for a multi-group SEM power analysis. It requires users to provide a SEM model without, and a model with (a single, or multiple) equality constraints across groups. The `SSpower()` function then performs a chi-square-based power analysis across a range of sample sizes to assess the tenability of the constraints (Jorgensen et al., 2022; Satorra & Saris, 1985).

[4] Computation time depends on many factors, including the speed of the CPU, the number of cores you are using, and the complexity of the model, etc.

5.6.3 Conclusion

In conclusion, this paper proposes a strategy for performing a power analysis specifically tailored to the particularities of the RI-CLPM. It is implemented in the R package powRICLPM, which is designed to be as user-friendly as possible for applied researchers, and accommodates various extensions. Together, this paper and the R package provide researchers with the resources to design a power analysis that produces valid recommendations for planning future research involving the RI-CLPM.

Acknowledgments: I would like to thank Ellen Hamaker for the discussions that have influenced this work, and her useful feedback.

Funding: This study is part of the Consortium Individual Development, which was supported by a Gravitation grant awarded by the Dutch Ministry of Education, Culture, & Science, and the Netherlands Organization for Scientific Research [NWO grant number 024.001.003; 2013].

Online supplementary materials: Annotated R code for the analyses performed in this article can be found in the online documentation of the powRICLPM R package at https://jeroendmulder.github.io/powRICLPM/.

Appendices

5.A Variance of within-components

The variance for the within-component of X at wave 2, $Var[W_{X,2}]$, can be expressed as

$$\begin{aligned}
Var[W_{X,2}] &= Var[\alpha_1 W_{X,1} + \beta_1 W_{Y,1} + u_1], && \text{(5A.1)}\\
&= Var[\alpha_1 W_{X,1}] + Var[\beta_1 W_{Y,1}] + 2\alpha_1\beta_1 Cov[W_{X,1}, W_{Y,1}] + Var[u_1], && \text{(5A.2)}\\
&= \alpha_1^2 Var[W_{X,1}] + \beta_1^2 Var[W_{Y,1}] + 2\alpha_1\beta_1 Cov[W_{X,1}, W_{Y,1}] + Var[u_1], && \text{(5A.3)}\\
&= \alpha_1^2 + \beta_1^2 + 2\alpha_1\beta_1 Cov[W_{X,1}, W_{Y,1}] + \sigma_u^2, && \text{(5A.4)}
\end{aligned}$$

which shows that it is a function of the lagged effects, $\alpha_1^2 + \beta_1^2$, the covariance between the predictors at the previous wave, $2\alpha_1\beta_1 Cov[W_{X,1}, W_{Y,1}]$, and the residual variance, σ_u^2. This logic similarly applies to the variance of the within-component of academic amotivation.

5.B Residual variances and co-variances at wave 2 and further

Let Φ be a square matrix of lagged effects with the diagonal elements representing autoregressive effects, and off-diagonal elements cross-lagged effects. Collecting these population parameter values for the illustrating example gives

$$\Phi = \begin{bmatrix} 0.20 & 0.15 \\ 0.10 & 0.30 \end{bmatrix}.$$

Furthermore, let Σ be a variance-covariance matrix for the within-components at each time point. For the illustrating example, this results in

$$\Sigma = \begin{bmatrix} 1 & 0.26 \\ 0.26 & 1 \end{bmatrix}$$

with the diagonal elements representing the variances of the within-components, and the off-diagonal elements representing the correlation between the within-components.

C. J. Kim and Nelson (1999, p. 27) present an expression for the unconditional covariance matrix of a stationary process as a function of the lagged effects and the residual variance covariance matrix. Rewriting this equation, the residual variances and covariances can be expressed as

$$vec(\Psi) = (I - \Phi \otimes \Phi)vec(\Sigma) \tag{5B.5}$$

with Ψ the residual variance-covariance matrix, I the identify matrix, and $vec(\cdot)$ denoting the operation of putting the elements of a matrix into a column. Applying Equation 5B.5 to the population parameter values of the illustrating example result in

$$\Psi = \begin{bmatrix} 0.9219 & 0.1755 \\ 0.1755 & 0.8844 \end{bmatrix}$$

where the diagonal elements represent the residual variances, and the off-diagonal represent the residual covariances needed to get within-components with a variance of 1.

CHAPTER 6

Estimating causal effects of time-varying exposures: The overlap and differences between structural equation modeling and marginal structural models

Revised version published as:
Mulder[a], J. D., Luijken[a], K., Penning de Vries, B. B. L., & Hamaker, E. L. (2024). Causal effects of time-varying exposures: A comparison of structural equation modeling and marginal structural models in cross-lagged panel research. *Structural Equation Modeling: A Multidisciplinary Journal.* https://doi.org/10.1080/10705511.2024.2316586

Abstract

Structural equation modeling (SEM) has become established as one of the main statistical modeling frameworks in psychology and related fields for investigating prospective effects of variables on each other. However, the use of SEM models for causal inference from panel data is critiqued in the causal inference literature for unnecessarily relying on a large number of parametric assumptions, and alternative methods originating from the potential outcomes framework have been recommended, such as inverse probability weighting (IPW) estimation of marginal structural models (MSMs). To help SEM users understand this criticism we describe three phases of causal research. We explain (differences in) the assumptions that are made throughout these phases for SEM and IPW-MSM approaches using an empirical example. Second, using simulations we compare the finite sample performance of path analysis (a SEM approach) and IPW-MSM for the estimation of time-varying exposure effects on an end-of-study outcome under various violations of parametric assumptions. We conclude that although increased reliance on parametric assumptions does not always translate to increased bias (even under model misspecifcation), psychological researchers are still well-advised to acquaint themselves with causal methods from the potential outcomes framework to investigate time-varying exposure effects.

[a] Mulder and Luijken contributed equally to the work.

A common question shared across research disciplines is how one variable has a prospective effect on another. In psychology and related fields, this question is often tackled using panel data, in which the same people are measured multiple times on the same variables. A particularly popular modeling approach to such data is cross-lagged panel modeling, which falls within the broader context of the structural equation modeling (SEM) framework (Usami et al., 2019; Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020). The cross-lagged effects that are obtained with them are often interpreted as causal effects, sometimes quite explicitly (Asendorpf, 2021; Orth et al., 2021), but oftentimes in a more implicit way through the use of specific language (e.g., when one variable is described to "react to", "respond to", "impact", or "spill over into" another variable; Hamaker et al., 2020; Hernán, 2018). While the SEM framework has been commended by researchers like Bollen and Pearl (2013) for the purpose of causal inference, there is also criticism of this practice. In particular, Van der Laan and Rose (2011) and VanderWeele (2012) point out that SEM models depend heavily on parametric assumptions; since these are likely to be violated—at least to some degree—in practice, SEM is prone to bias when used for causal inference, according to these researchers.

Obviously, this claim should raise concerns among SEM users. Yet, disciplinary differences can hinder SEM users, for instance from the field of psychology, to appreciate the arguments, concerns, and solutions put forward by SEM critics who come from fields like epidemiology and biostatistics. To fully comprehend whether, when, and to what extent the critique of SEM is relevant, one first needs to be well-versed in the principled approach to causal inference (based on the potential outcomes framework) that is currently used in these disciplines. Additionally, one needs to be aware of typical presumptions in these disciplines: Oftentimes, the focus is on a binary causal variable, which is typically referred to as the treatment or exposure; furthermore, when the state of this variable can vary over time, the focus is often on contrasting treatment regimes—that is, specific sequential patterns of being (not) exposed at particular time points—rather than the effect of the exposure at one specific time point only; in that case, the focus is often on an end-of-study outcome, rather than multiple repeated outcomes. Finally, to understand what is meant with the unrealistic parametric assumptions made in the SEM framework, and how these can be avoided using an alternative estimation framework, one needs to be able to compare the SEM approach with a possible alternative that is proposed, such as inverse probability weighting (IPW) estimation of marginal structural models (MSMs; Robins et al., 2000; Vansteelandt & Sjolander, 2016). Hence, bridging this disciplinary gap is quite challenging, and it is therefore likely that the criticism of SEM does not end up with SEM users.

The goal of this paper is therefore twofold. First, we want to provide SEM users from disciplines like psychology with the necessary knowledge to understand the voiced criticism of SEM for causal inference. To this end we introduce the reader to the principled approach to causal inference that has been developed within the potential outcomes framework, and discuss to what extent SEM can be considered compatible with this approach. Moreover, we will explain the main idea and purpose of IPW estimation of an MSM as an alternative that is based on fewer parametric assumptions, making it less susceptible to violations of these assumptions. Second, we will perform a simulation study to assess the finite sample performance of path analysis (a SEM method) versus IPW estimation of MSMs under various violations of the parametric assumptions that path analysis relies on. Throughout, our focus will be specifically on panel data where we want to make inferences about the effect of a time-varying exposure on an end-of-study outcome, in the presence of both baseline and time-varying confounding.

This paper is organized as follows. Section 6.1 introduces the potential outcome framework and the phases of causal research. These phases are illustrated for both SEM and IPW estimation in Section 6.2 using an empirical example concerning the effect of smoking cessation on body weight. Section 6.3 describes the set-up of our simulation study for comparing the bias and mean squared error (MSE) of path analysis and IPW estimation in estimating the effect of a time-varying binary exposure on a continuous end-of-study outcome under different violations of parametric assumptions. Section 6.4 describes the results of the simulations. We end with a discussion and conclusion.

To facilitate understanding of terminology more common in the potential outcomes framework, we provide a glossary in Table 6.1 that explains important causal inference related terms that our discussion relies on. Boldfaced words in this paper are included in the glossary. Annotated R code used for the analyses in this paper can be found in the online supplementary materials at https://jeroendmulder.github.io/SEM-and-MSM.

6.1 Causal inference in the potential outcomes framework

In this paper, we make use of the Neyman-Rubin potential outcomes framework for causal inference (Rubin, 1974; Splawa-Neyman et al., 1990). This framework is centered around randomization as a principle of causality, which is used in experimental trials (Fisher, 1935). These trials are considered the gold standard to causal inference, and are closely mimicked in nonexperimental studies using the potential outcomes framework (Hernán & Robins, 2016). The phases of causal inference in the potential outcomes framework can be regarded as a principled approach for making explicit un-

Table 6.1: Glossary of causal inference-related terms used in this article.

Term	Description and related terms
Exposure regime	A predetermined rule that determines the value of a time-varying exposure *for each time point* (Hernán & Robins, 2020). Here, we discuss *static* regimes, implying that exposure values are all predetermined. *Related term:* exposure sequence.
Always-exposed	An exposure regime in which a binary exposure is set to “exposed” for all predefined number of time points.
Never-exposed	An exposure regime in which a binary exposure is set to “not exposed” for all predefined number of time points.
Causal estimand	A precise description of an effect, reflecting the research question of a research project. It summarizes at a population-level what the outcomes would be in the same individuals under different exposure conditions (European Medicines Agency, 2020). Causal estimands are often defined as functions (e.g., contrasts) of potential outcomes. *Related term*: target causal quantity (Petersen & Van der Laan, 2014).
Causal identification	The process of translating a causal estimand to a statistical estimand, which is defined as a function of observed data. It involves evaluation of the causal identification assumptions of exchangeability, consistency, and positivity.
Exchangeability	A causal identification assumption restricting the exposure to be independent from the potential outcomes (Angrist & Pischke, 2009; Hernán & Robins, 2020; Imbens & Rubin, 2015). It is violated in setting with confounding/selection bias. *Related terms*: unconfounded assignment, unconfoundedness, no unmeasured confounding, (conditional) independence of treatment and potential outcomes, exogeneity (P. R. Rosenbaum & Rubin, 1983).
Consistency	A causal identification assumption linking potential outcomes to observed outcomes. It is violated when exposure is not well-defined and/or there exist multiple versions of intervention/treatment (Hernán, 2016).
Causal directed acyclical graphs (DAGs)	A diagram, consisting of nodes and edges connecting them, visualizing a data generating process. Nodes represent variables in the phenomenon under study, and edges the causal relationships between them. All variables thought to play a role in the causal process should be included. *Related terms*: causal diagrams, non-parametric structural equation model (Pearl, 2009)

der what assumptions statistical effects in nonexperimental research can interpreted as causal effects (Goetghebeur et al., 2020; Petersen & Van der Laan, 2014). In this section, we first give a brief overview of the core phases of causal inference. Then, we discuss in which phases differences between SEM and MSM approaches manifest themselves. Note, that Phases 1 and 2 tend to be left more implicit in empirical research with the SEM framework. As such, researchers working predominantly with SEM may be less familiar with them. For more elaborate introductions, we refer to Goetghebeur et al. (2020) and Hernán and Robins (2020).

6.1.1 Phases of the causal inference process

Generally, the process of causal inference contains three phases, namely (1) the *formulation* of a causal research question using potential outcomes, resulting in a causal estimand; (2) the *identification* of the causal estimand in terms of observed data, translating the causal estimand into a statistical estimand; and (3) *estimation* of the statistical estimand from a finite sample using a statistical model (Goetghebeur et al., 2020; Petersen & Van der Laan, 2014). Below, each phase is discussed in more detail.

Phase 1 concerns the formulation of a causal research question in terms of a contrast between possible scenarios. In the case of a time-varying exposure, the question can be of the form "What would happen to an outcome variable if a time-varying exposure had been fixed to a certain **regime** versus another regime?"[1] These questions are thus expressed as contrasts of potential outcomes, that is, values of an outcome that would have been observed if the exposure had been set to a particular regime (Rubin, 1974; Splawa-Neyman et al., 1990). Phase 1 involves, amongst other things, specifying a population, an exposure contrast, and an outcome. The population indicates which specific group of individuals the study aims to make inferences about (which is referred to as the target population in the causal inference literature), that is, who is eligible for inclusion in the study? This includes a specification of the moment at which individuals become eligible for the study (Brookhart, 2015; Edwards et al., 2016; Hernán et al., 2016; Suissa, 2008). The exposure contrast reflects which specific exposure regimes will be compared. The outcome is specified by defining the measure that is a relevant outcome, including when this is measured. Thinking about these questions and using the potential outcomes language helps researchers to formalize their research question into an explicit *causal estimand* that describes in great detail what causal effect is of interest.

Phase 2 concerns the translation of the causal estimand (which is a hypothetical, potential outcomes concept), into a statistical estimand that can be estimated

[1] Readers more familiar with the potential outcomes literature might recognise this research question as pertaining to *static exposure regimes* rather than *dynamic exposure regimes*. For accessibility of the paper, we focus exclusively on the simpler case of static exposure regimes.

using observed data. The process of equating a causal estimand to a function of the population distribution of observed variables is also know as *identification.* This is done by evaluating a set of **causal identification assumptions**, typically including **consistency**, **exchangeability**, and positivity (Hernán & Robins, 2020). The assumption of consistency relates the potential outcomes that form the basis of the causal estimand to observed outcomes. It requires interventions on the exposures to be sufficiently well-defined, implying that researchers need to clearly define an intervention on exposures, even if the intervention is purely hypothetical (e.g., carrying out the intervention would be unethical, impractical, or impossible; Hernán & Robins, 2020). The assumption of conditional exchangeability states that, conditional on covariates, the potential outcomes are independent from the observed exposures of individuals. One often-discussed scenario in which this assumption is violated, is when there exist unmeasured covariates that confound the relationship between an exposure and an outcome. Hence, this assumption is closely associated to the assumption of no unmeasured confounding that psychological researchers might be more familiar with, but note that the assumption of conditional exchangeability is more general (i.e, there exist situations other than the presence of unobserved confounding in which conditional exchangeability is violated; Bollen, 1989). The positivity assumption indicates that there is a non-zero probability for individuals to be in either exposure condition. This assumption would be violated when, in practice, there is perhaps a policy or condition due to which an individual has a zero probability of either one exposure values.

Phase 3 concerns the translation of the statistical estimand, which still refers to the *entire population*, to an estimator, which is a method to estimate the statistical estimand from a *finite random sample.* We compare two methods here: Path analysis (a SEM approach), and IPW regression of an MSM (a potential outcomes approach). These different methods make different parametric assumptions—such as linearity for certain relations, and whether or not interactions are present—which imply a particular probability distribution. Whenever an estimator relies on parametric assumptions, it comes with the risk of model misspecification, and violation of parametric assumptions can result in a *biased* estimator.[2] Parametric assumptions and violations thereof can also influence other properties of estimators, such as statistical convergence, or sampling variability. It is therefore important to decide a priori which properties of estimators are most desirable for a particular research problem, and then to find an estimator that has these properties.

[2] We discuss parametric assumptions in Phase 3, but it is possible that parametric assumptions are already incorporated in the statistical estimand, and are thus part of Phase 2. A more estimation-specific (i.e., Phase 3-specific) matter is how the parameters of the statistical models are estimated using finite samples. Different estimators (e.g., maximum likelihood with or without penalization, or random forests) need not have the same statistical properties (e.g., finite sample bias, statistical convergence, sampling variability).

Sometimes a fourth phase is described, in which researchers evaluate how particular assumptions made throughout the first three phases impact their results. Through a sensitivity analysis, it can be determined how large the violations of an assumption need to be before this changes the conclusions that were drawn (based on the results in Phase 3). In the current study, we do not further discuss this, but the interested reader is referred to Imbens and Rubin (2015), Lash et al. (2009), and VanderWeele and Ding (2017).

6.1.2 Differences between SEM and MSM approaches

The phases for empirical causal research are equally applicable to both SEM and potential outcome approaches. However, notions about causality are explicit in the latter; for instance, an MSM is defined in terms of potential outcomes, rather than in terms of the observed outcome variable, and thus invites explicit examination of causal identification assumptions. SEM can also be used within the potential outcome framework (e.g., De Stavola et al., 2015; Moerkerke et al., 2015; B. O. Muthén et al., 2016), but common applications of SEM focus mainly on estimation of (complex) statistical models (Phase 3) with little or no attention paid to the formulation of a causal research question (Phase 1), and identifying it (Phase 2). Without careful formulation and evaluation of the causal identification assumptions, it remains unclear if the estimates that result from a statistical analysis actually provide an answer to the causal question of interest.

Another difference between SEM and potential outcome approaches concerns their modeling "focus". While typically only one (or a limited number of) causal effect(s) is targeted in a research question, SEM approaches usually attempt to model the entire causal process under study. That is, SEM models make parametric assumptions about the causal dependencies of the outcome, the exposure, and all time-varying covariates that are thought to play a role. By estimating each and every individual path-specific effect, SEM models rely on a large number parametric assumptions in total. This is a valid approach assuming all of these assumptions hold (e.g., if, in fact, all effects are linear and there are no interactions). However, one of the points made by critics of the use of SEM models for causal inference, is that these parametric assumptions are unlikely to start with, and the potential for violations thereof only increases as the size of SEM models grows (VanderWeele, 2012). Instead, IPW estimation of MSMs does not require a model for the distribution of covariates, and their relation to previous variables. Compared to SEM, this reduced reliance on parametric assumptions therefore should, in principle, lead to more robust causal inference.

A third difference is how both modeling approaches handle the problem of *exposure-confounding feedback*. This issue occurs whenever an exposure affects subsequent time-varying confounding variables and is itself influenced by the confounding variable (Robins et al., 2000). This type of confounding cannot be adjusted for using standard regression techniques which attempt to estimate effects of a time-varying exposure simultaneously, for example by a single linear regression of the outcome on previous time-varying exposures and time-varying covariates. G-methods such as IPW estimation for MSMs have been developed to resolve these issues and estimate time-varying exposure effects (Daniel et al., 2013; Naimi et al., 2016; Robins et al., 2000). However, exposure-confounding feedback is not a topic in the SEM literature, as modeling the entire assumed data generating mechanism forgoes this issue. Hence, the issues introduced by exposure-confounding feedback are likely unfamiliar to researchers predominantly working with SEM.

6.2 Investigating time-varying exposure effects: An example using smoking cessation and body weight

To illustrate the three phases of causal inference described above, we make use of an empirical example about the causal effect of smoking cessation on body weight. These data come from the health survey of the Longitudinal Internet studies for the Social Sciences panel, administered by Centerdata (Tilburg University, The Netherlands). The LISS panel consists of a random sample of Dutch households representative of the Dutch-speaking population in the Netherlands aged 16 years or older (more information about the LISS panel can be found at https://www.lissdata.nl; Scherpenzeel, 2018). The empirical example focuses on self-reported measurements of smoking cessation, body weight, and a set of covariates in the period 2007 to 2020. Some simplifying decisions were made throughout the three phases. This was done for illustrative purposes, and to keep the focus of this comparison on the (parametric) assumptions underlying both approaches (rather than on differences in, for example, techniques for missing data handling).

6.2.1 Phase 1: Formulation of the research question and causal estimand

Formulating a research question and causal estimand is similar for SEM and potential outcome approaches. Suppose we are interested in the impact of smoking cessation on body weight. Rather than focusing on the effect of smoking cessation at one particular wave on body weight at the next wave (e.g., as done in cross-lagged panel modeling), we may decide to focus on the effect of smoking cessation at multiple waves on an end-of-study measure of body weight. Our research question about the average causal effect (ACE) of a change in exposure (i.e., smoking) regimes can then be: “What would be the difference in average body weight after two years if all currently smoking Dutch adults quit smoking, and refrained from smoking for two years, *compared to* if they continued smoking for two years?”. This research question describes a *joint effect*, as it refers to a change in smoking status at multiple exposure times, that is smoking cessation in year 1 and year 2, and its combined (joint) effect on end-of-study body weight (Daniel et al., 2013).

The target population in this example are adults who smoke in the general Dutch population. The moment that individuals become eligible is the moment they enroll in the LISS cohort. Note that this is an eligibility criterion that is difficult to translate into a meaningful event in everyday life (i.e., outside the context of the LISS study; Suissa, 2008). We explored whether we could define a meaningful moment of eligibility, such as “the first time that their physician indicated they are at cardiovascular risk (i.e., suffer from high blood pressure, high serum cholesterol, or diabetes)”. However, this left us with fewer than 80 individuals in the LISS data set, which would inhibit us from fitting the models of interest in this illustration. We make this remark for future empirical studies. The exposure contrast is “quitting and refraining from smoking for two years” versus “continuing smoking for two years”. The outcome is defined as body weight in kilograms measured by a scale two years after the moment of becoming eligible.

To formalize this research question as a causal estimand, we introduce some notation. In terms of timing, we denote $t = -1$ as the time at which eligibility is assessed. From time point $t = 0$ onward, the exposure can vary for everyone. Let Y_2 represent the end-of-study outcome, observed body weight in kilos at time point $t = 2$. Let A_t denote the exposure variable of interest at time point t, in this case quitting smoking ($A_t = 1$) or not ($A_t = 0$). Let L_t denote a set of covariate values at time point t (including body weight at $t = 0, 1$), and with baseline covariates measured at $t = -1$, L_{-1}. We abbreviate the history of the exposure and covariates up to t, that is, $(A_0, ..., A_t)$ and $(L_0, ..., L_t)$, by $\bar{A}_t$ and $\bar{L}_t$, respectively. Finally, let $Y_2^{\bar{a}_1}$ be the potential outcome weight under smoking regime $\bar{A}_1 = (A_0, A_1) = (a_0, a_1)$. The potential outcome of a smoker who continues smoking for two years is then $Y_2^{\bar{0}_1}$, and $Y_2^{\bar{1}_1}$ if the smoker quits and refrains from smoking for two years.

The causal estimand for our research question can be formalized as a contrast of two potential outcomes:

$$\text{ACE} = \mathbb{E}[Y_2^{\bar{1}_1}] - \mathbb{E}[Y_2^{\bar{0}_1}]. \tag{6.1}$$

The causal estimand in Equation 6.1 can be referred to as an “**always-exposed** versus **never-exposed** effect”.

In Phase 1, we can also specify an MSM to formalize the research question. An MSM is a model for the marginal distribution (i.e., summarizing across all possible subpopulations) of potential outcomes. For our research question, in which $\bar{a}_1$ can only be $\bar{0}_1$ or $\bar{1}_1$, it can specified as

$$\mathbb{E}[Y_2^{\bar{0}_1}] = \beta_0, \tag{6.2}$$

$$\mathbb{E}[Y_2^{\bar{1}_1}] = \beta_0 + \beta_1 \tag{6.3}$$

where β_0 represents the expected end-of-study body weight if all individuals continue smoking for two years, and β_1 represents the difference between the expected end-of-study body weight if all individuals quit smoking for two years (i.e., $\bar{a}_1 = \bar{1}_1$) versus if they continue smoking for two years (i.e., $\bar{a}_1 = \bar{0}_1$).

6.2.2 Phase 2: Assess identifiability of causal estimands

Evaluation of the causal identification assumptions, particularly exchangeability, can be done with help of a visual diagram such as a **causal directed acyclic graph** (DAG; Hernán, 2016; Pearl, 2009, 2010; VanderWeele, 2019). A causal DAG encodes causal assumptions about the data-generating mechanism based on domain knowledge (for an introduction on DAGs in the context of psychological science, we refer to Rohrer, 2018). While causal DAGs appear similar to path diagrams commonly used in SEM, there are some crucial differences (Kunicki et al., 2023; Moerkerke et al., 2015). Importantly, the causal relations between variables in a causal DAG do not encode parametric assumptions about those relations, such as linearity assumptions or normality of residuals, which are typically assumed in a path diagram. Furthermore, causal DAGs only depict direct causal relationships represented by one-headed arrows, whereas path diagrams can also include covariances represented by two-headed arrows to account for unexplained relationships between variables. Yet, both types of diagrams can help a researcher to assess whether the causal identification assumptions can be plausibly invoked in theory. To illustrate this, we examine the identifiability of the causal estimand in our empirical example.

First, we visually represent existing knowledge about the causal system (as well as uncertainty). Such knowledge can be obtained by a review of the literature and expert consultations. We pragmatically drew information from a systematic review into smoking cessation and body weight gain by Tian et al. (2015). Potential confounding variables in the causal system of interest are time-invariant covariates age, sex, and ethnicity. Time-varying covariates are body weight, alcohol consumption, physical activity, socioeconomic factors, energy intake, and comorbidities. Knowledge of the involvement of these covariates in the causal system is represented in the causal DAG in Figure 6.1. For readability, we simplified the DAG by omitting relations between covariates themselves, and denoted the set of three time-fixed covariates at baseline simply as “Baseline covariates”, and the set of five time-varying covariates as “Time-

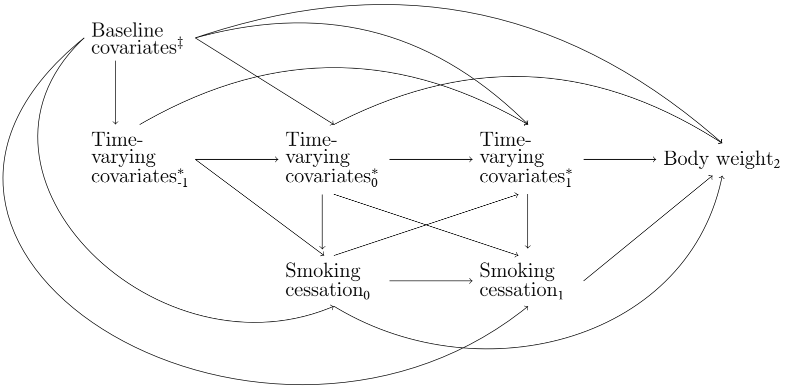

Baseline covariates‡
Time-varying covariates$^*_{-1}$
Time-varying covariates$^*_0$
Time-varying covariates$^*_1$
Body weight$_2$
Smoking cessation$_0$
Smoking cessation$_1$

Figure 6.1: A simplified representation of the causal DAG relating smoking cessation and body weight. It includes the variables smoking cessation, body weight, baseline covariates, and time-varying covariates. The arrows represent the nonparameteric links between them.
‡ Age, sex, ethnicity.
* Body weight, socioeconomic factors, alcohol consumption, physical activity, energy intake, and comorbidities.

varying covariates$_t$" for $t = 0, 1, 2$.

It is of paramount importance that the causal system is drawn based on background knowledge, and is not based on data availability. In this process, the omission of variables or arrows from the causal DAG is a stronger assumption than including them, as omissions of arrows amounts to constraining causal effects to exactly zero (Bollen & Pearl, 2013). In longitudinal studies, this might imply that not only lag-1 effects are included in the causal DAG, but also lag-2 and longer relations (VanderWeele, 2021). As encoded in the simplified causal DAG in Figure 6.1, we do not assume only lag-1 effects, but additionally allows for lag-2 effects and longer. This causal DAG does not assume any particular probability distribution for the causal system, nor does it specify the functional form of the causal relationships in the graph. This means that there may be linear but also non-linear relations, and that there may be interactions in addition to main effects.

We can now determine whether the causal estimand that was specified in Phase 1, can be expressed as a function of the observed data (i.e., the statistical estimand), given the background knowledge encoded in the causal DAG in Figure 6.1 and the available data. The causal identification assumption of consistency entails that the *observed outcome* of an individual who quits smoking for two years is equal to their *potential outcome* if quitting smoking for two years, that is, $Y_2^{\bar{1}_1} = Y$ for individuals with observed $\bar{a}_1 = \bar{1}_1$. Similarly, the observed outcome of an individual who continues smoking for two years should be equal to their potential outcome if continuing smoking for two year, that is, $Y_2^{\bar{0}_1} = Y$ for individuals with observed $\bar{a}_1 = \bar{0}_1$ (Hernán & Robins, 2020). This seemingly obvious assumption implies that the exposure itself, as well as (hypothetical) interventions on it, must be sufficiently well defined such that it is clear what specific exposure the causal effect refers to (Hernán, 2016; VanderWeele, 2018). For example, smoking cessation can be achieved with the help of nicotine pills, therapy, a supporting friend, or a combination of these; setting the exposure to "quit smoking" leaves it open which of these exposures an individual undergoes. Because the different strategies might have different causal effects on body weight, the observed outcome need not necessarily equal the potential outcome. Information about the distribution of strategies to quit smoking might help to link the potential outcomes to observed data (Hernán & Robins, 2020), but this information is not collected in the LISS study, meaning that consistency is compromised in our example.

The conditional exchangeability assumption states that, given a set of covariates, the potential outcomes are independent of the observed exposures. In longitudinal settings, with multiple exposure-times, conditional exchangeability must hold at each time point. This can be denoted as $A_t \perp\!\!\!\perp Y^{\bar{a}_t} | \bar{L}_t, \bar{A}_{t-1} = \bar{a}_{t-1}$, with $\bar{L}_t$ denoting the set of baseline and time-varying covariates up to and including time t and $\bar{A}_{t-1} = \bar{a}_{t-1}$ representing the sequence of exposures a person received up to $t-1$ (Hernán & Robins, 2020; Naimi et al., 2016). To be able to achieve this in practice, we must have collected data (without measurement error) on all relevant covariates that, based on the causal DAG, could confound the relationship between exposure and outcome. Based on Tian et al. (2015), ethnicity, socioeconomic factors, and energy intake were identified as relevant confounders. However, energy intake, for example, is not measured in the LISS data set, and therefore cannot be adjusted for in the analyses. As such, conditional exchangeability is compromised for our example. In practice, this conclusion would imply that additional data needs to be collected or identified to be able to provide a valid answer to the research question. Additionally, a sensitivity analysis can give insight into how strong the confounding by energy intake must be to substantively affect the conclusions derived from the primary analysis.

The sequential positivity assumption indicates that, at each time point and across all values of the covariates in the data, there is a non-zero probability for individuals to be in either exposure condition. This seems to be the case in this example on smoking cessation, because it is hard to conceive a policy or condition due to which an individual has a zero probability to quit or continue smoking.

Based on our evaluation of the causal identification assumptions for the empirical example, we conclude that additional data needs to be collected or identified to provide a valid answer to the causal research question: Such a finding is in itself is a useful contribution for the design of future studies (Petersen & Van der Laan, 2014). This example also underscores the importance of carefully considering Phases 1 and 2 in causal research *before* data is collected to ensure that the causal identification assumptions are as plausible as possible. If no additional data can be collected and the assumptions are compromised, then sensitivity analyses can be performed to determine, for example, how strong the relations of a confounding covariate must be to substantively change the conclusions of the primary analysis. For illustrative purposes we continue with the current example, but emphasize that causal interpretation of findings would be incorrect.

Using the causal identification assumptions, the causal estimand can be reexpressed as a statistical estimand. These steps, which are provided in detail in Appendix 6.B, are a formalisation of the causal identification process as described above. It yields:

$$
\begin{aligned}
\text{causal estimand} :&= \mathbb{E}[Y_2^{\overline{1}_1}] - \mathbb{E}[Y_2^{\overline{0}_1}], \\
&\;\;\vdots \\
&= \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2 \mid \overline{A}_1 = \overline{1}_1, \overline{L}_1] \;\Big|\; A_0 = 1, L_0\big)\Big\} \\
&\quad - \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2 \mid \overline{A}_1 = \overline{0}_1, \overline{L}_1] \;\Big|\; A_0 = 0, L_0\big)\Big\} \qquad (6.4) \\
&=: \text{statistical estimand.}
\end{aligned}
$$

Notice how the identification process starts with the causal estimand in terms of potential outcomes (hypothetical quantities), and ends in a statistical estimand with only observed variables. However the statistical estimand in Equation 6.4 is just one "form", known in the causal inference literature as the "g-formula representation", but can be further rewritten such that it takes a different form. This is illustrated in Appendix 6.B where we further rewrite the g-formula representation of the statistical estimand to the "IPW representation". Different representations of a statistical estimand invite different modeling approaches for Phase 3, and this can have advantages (or disadvantages) for particular research designs.

6.2.3 Phase 3: Estimation using finite sample data

The terms in the statistical estimand can be estimated from finite random samples (taken from the population distribution) under a statistical model. The statistical estimand in Equation 6.4 suggests that we impose a statistical model on the distribution of the outcome given the exposure and covariate history, $\mathbb{E}[Y_2 \mid \overline{A}_1 = \overline{1}_1, \overline{L}_1]$ and $\mathbb{E}[Y_2 \mid \overline{A}_1 = \overline{0}_1, \overline{L}_1]$; and for the conditional distribution of (post-baseline) time-varying covariates at $t = 1$, $\mathbb{E}_{L_1}\big[(...) \;\big|\; A_0 = 1, L_0\big]$ and $\mathbb{E}_{L_1}\big[(...) \;\big|\; A_0 = 0, L_0\big]$. In situations with many time-varying covariates, working with the g-formula representation might be problematic as a statistical model must be specified for all covariates in L_1, thereby increasing the risk of model misspecification. In contrast, the statistical estimand can be reexpressed (see Appendix 6.B) to the IPW representation, such

Table 6.2: Overview of covariates included in the LISS panel study. All variables are self-reported measures taken from a questionnaire.

Covariate	Measurement level	Measurement time	Time-span
age	continuous	baseline	right now
sex	nominal	baseline	right now
body weight	continuous	time-varying	right now
alcohol consumption	ordinal	time-varying	average last year
hours physical activity	continuous	time-varying	average last week
number of comorbidities[a]	ordinal	time-varying	last year

[a] Self reported information on diagnosis by a physician.

that it does not suggest that the conditional distribution of the outcome be modelled; instead, the reexpressed statistical estimand suggests that the time-varying exposures are modelled. When adjustment for many covariates is required, working with the IPW representation of the statistical estimand might thus be advantageous.

Here, we compare path analysis to IPW estimation—in which path analysis is more in line with the g-formula representation of the statistical estimand, and IPW estimation (obviously) with the IPW representation (Naimi et al., 2016)—and attempt to answer our research question using the LISS data.

6.2.3.1 Establishing the study sample from the LISS data

The LISS panel study is based on a rolling enrollment, meaning that each year, a new group of individuals is added to the existing participant pool. Table 6.2 contains an overview of covariates that were included in the LISS data. We established the study sample for the target population "currently smoking Dutch adults" from the LISS data as follows. From each participant, the first four yearly measures were selected (regardless of the year in which participants enrolled) corresponding to time anchors $t = -1$ to $t = 2$ in our study. If participants indicated affirmatively on the question "Do you smoke now?" at their first measurement wave, they were included in the sample of this study starting from the wave after (i.e., their second measurement wave is at $t = 0$). The sample included 2,736 participants. Participants with implausible or impossible values on variables were deleted (i.e., weight higher than 200 kg or lower than 20 kg, yearly weight increase followed by weight decrease of more than 50 kg, more than 150 hours of physical activity per week). To keep the focus of this analysis on the parametric assumptions underlying both approaches, we filled in missing values in this sample by single imputation using the mice package (version 3.16.0; van Buuren & Groothuis-Oudshoorn, 2011) in R (version 4.2.2; R Core Team, 2022).

6.2.3.2 Path analysis with an additional joint effect parameter

In a typical SEM approach, the entire causal system as illustrated in the simplified DAG of Figure 6.1 would be interpreted as a path diagram. This implies that each dependency is specified in a SEM model as a linear effect, with independent residuals that are multivariate normally distributed. All exogenous variables (i.e., the baseline covariates) are allowed to freely covary with each other. This path diagram represents a set of linear equations, the parameters of which are estimated from the data. If the entire causal system is correctly specified (i.e., all dependencies in Figure 6.1 are indeed linear, there is no measurement error or effect modification, error terms follow a multivariate normal distribution, etc.; Gische & Voelkle, 2022), then this approach results in unbiased estimates of each path.

Estimates of our joint effect of interest can then be obtained as linear combinations of path-specific estimates. In particular, the joint effect is a linear combination of all regression coefficients on paths from exposures (both at time points 0 and 1) to the outcome, not going through later exposures. For the empirical example, this includes (a) the lag-2 path "Smoking cessation$_0$"→ "Body weight$_2$"; (b) the set of indirect paths "Smoking cessation$_0$" → "Time-varying covariates$_1$" → "Body weight$_2$"; and (c) the path "Smoking cessation$_1$"→ "Body weight$_2$". The combinations of these paths can be specified as additional parameters in a SEM model, such that point estimates for the targeted effects can be obtained directly. Confidence intervals can be obtained by nonparametric bootstrap.

One major issue of this approach is that it is not obvious how to combine the effect estimates on these paths when the paths include both categorical and continuous covariates. In those situations, one would have to combine linear regression coefficients with odds ratios, and there is no simple way to do this. In simple situations, with one, or a limited number of categorical time-varying covariates, one can rely on g-computation in order to get controlled direct effects from SEM models (B. O. Muthén et al., 2016; Nguyen et al., 2016). However, to make the causal identification assumption of conditional exchangeability plausible in complex nonexperimental settings, researchers would likely want to include a large number of covariates, and there are likely to be numerous categorical covariates as well. For our example, some of the time-varying covariates were categorical in nature, or have been measured in a categorical manner in the LISS data (e.g. alcohol consumption, and number of comorbidities). For this reason, a path analysis using the LISS data that incorporates all variables mentioned in Table 6.2 is not a viable option.

For purely illustrative purposes, we discard the categorical covariates in this example such that we can continue our comparison of SEM and potential outcome approaches, and (the impact of) the parametric assumptions underlying path analysis and IPW estimation. We stress that this decision is far from satisfying from a causal inference point-of-view because the exchangeability assumption would consequently be violated. The decision is a necessary consequence of choosing path analysis as an estimation strategy in this phase. A path model based on Figure 6.1 was fitted to the LISS data in Mplus version 8.9 (L. K. Muthén & Muthén, 2017). Only body weight and hours of physical activity were included as time-varying covariates. The probit-link was used for regressing the time-varying exposures on covariates.

6.2.3.3 IPW linear regression

In brief, the aim of IPW is to create a pseudo-population in which the exchangeability assumption holds conditional on the measured covariates (Robins et al., 2000). This is achieved in three steps. First, probability of exposure is estimated using a propensity score model in which the exposure is regressed on the measured confounding variables. For a categorical exposure, such a model is commonly a logistic regression model in which the exposure is the outcome, and all confounding variables identified using the approach described in Section 6.2.2 are independent variables. The propensity score model must be correctly specified, implying that the functional form of the dependencies in the model is correct (i.e., the dependencies are in fact linear). In the second step, inverse-probability-of-exposure-weights are created for each individual. These are based on the probability of observed exposures values from the fitted propensity score model, which are inverted, and then multiplied across the time points per individual. The resulting weights are used to balance the original sample: Individuals with a low probability of scoring their observed exposure value have a higher weight, and are therefore over-represented in the pseudo-population, whereas individuals with a high probability of scoring their observed exposure have a lower weight, and are therefore underrepresented in the pseudo-population. The consequence of this weighting procedure is that in the pseudo-population the dependencies of the time-varying exposure on the time-varying varying-covariates—the paths "Time-varying covariates$_{-1}$" → "Smoking cessation$_0$"; "Time-varying covariates$_{-1}$" → "Smoking cessation$_1$"; 'Time-varying covariates$_0$" → "Smoking cessation$_0$"; 'Time-varying covariates$_0$" → "Smoking cessation$_1$"; and "Time-varying covariates$_1$" → "Smoking cessation$_1$"—are broken, such that these covariates are not confounders anymore for the effect of smoking cessation on end-of-study body weight. In the third step, estimates of the targeted effects are obtained by fitting a weighted regression model to the pseudo-population in which the outcome is regressed on both exposure-times. If the parametric assumptions (e.g., linearity, the absence of interaction effects) of this outcome model hold, then this procedure leads to unbiased estimates of the effect of exposure at time point 0 not going through later exposures, and exposure at time point 1, on the outcome. These effects of smoking cessation at each time point are also sometimes referred to as controlled direct effects, where the term "controlled"

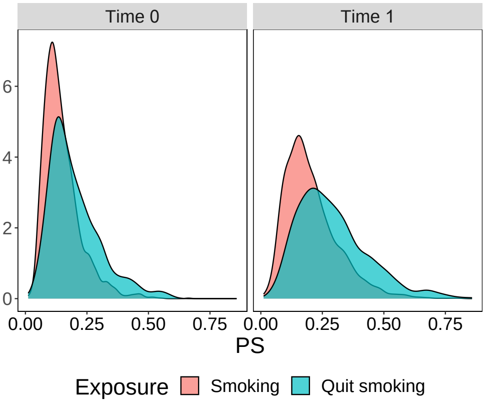


Figure 6.2: Density of propensity scores for individuals who quit smoking versus individuals who continued smoking at time points 0 and 1 (before weighting). Propensity scores were computed using all covariates.

refers to controlling for future exposures, and “direct” refers to the fact that the intermediate process by which smoking cessation leads to body weight is not modeled (Daniel et al., 2013). The sum of both controlled direct effects is our estimate of the joint effect.

The IPW regression method was applied to our empirical example. In contrast to path analysis, we include both categorical and continuous time-varying covariates here. A propensity score model was fitted by regressing the exposure variables on covariate history and previous exposure status. Positivity was evaluated by a visual inspection of overlap of the distributions of propensity scores of exposed and non-exposed at each time point, see Figure 6.2. No violation to positivity was detected. Stabilized IPWs were computed from the propensity score model using the R package WeightIt (version 0.14.0; Greifer, 2023b). Balance of the confounding variables in the propensity score model was assessed by comparing the standardized means of covariates for those who quite smoking, and those who continued smoking, using the R package cobalt (version 4.5.0; Greifer, 2023a). This comparison was done in both the unweighted sample and the weighted sample (i.e., the pseudo-population), and at both exposure-times, see Figure 6.3. Absolute standardized mean differences indicated well-balanced data based on a recommended threshold value of 0.2 (Stuart, 2010).

A regression model was fitted to the pseudo-population, regressing body weight at $t = 2$ on smoking cessation at $t = 0$ and $t = 1$. The regression coefficients of smoking cessation at $t = 0$ and $t = 1$ are the controlled direct effects, the combination of which is our joint effect of interest. 95% confidence intervals were obtained using the nonparametric bootstrap with 999 replications. Bootstrapping was performed using

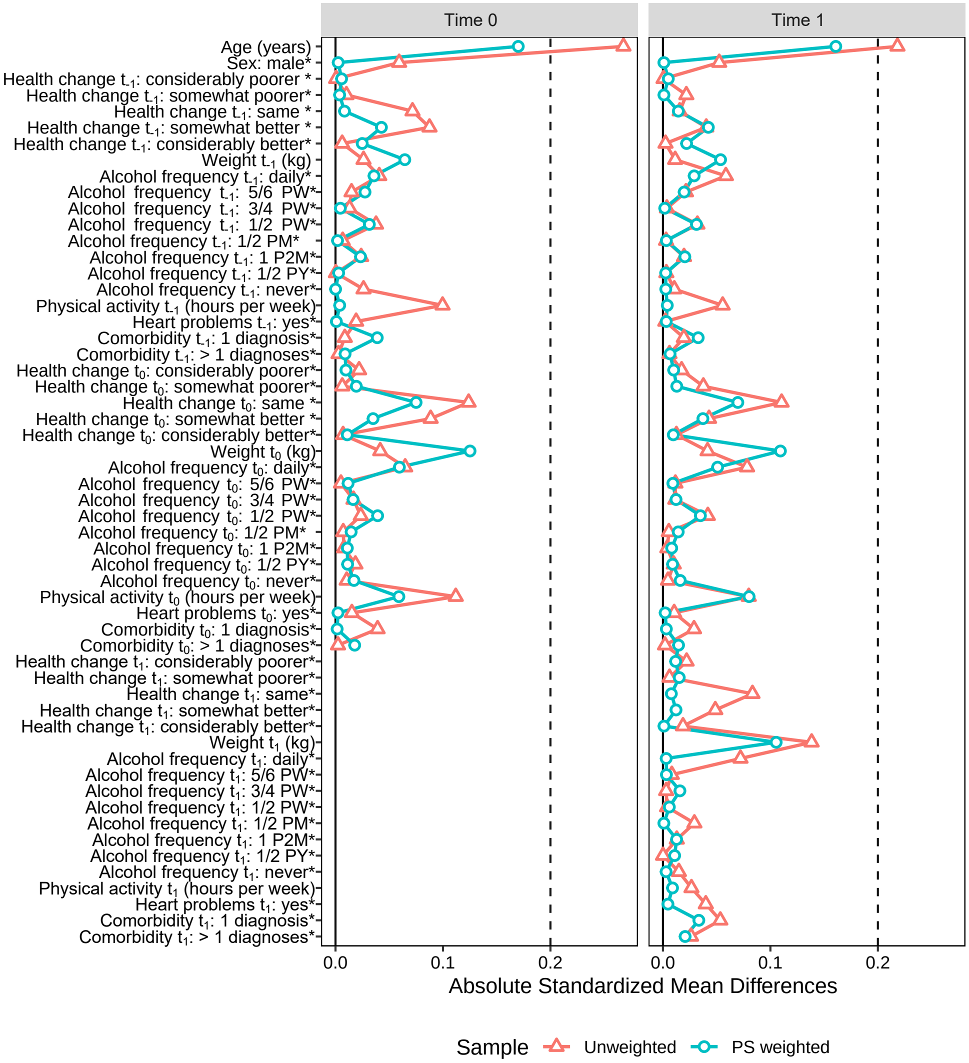


Figure 6.3: Visualization of covariate balance (standardized mean differences) before and after reweighing at time points 0 and 1. The asterisk * denotes binary covariates (or dummy variables) for which the displayed value is the raw (unstandardized) difference in means. PW = per week; PM = per month; P2M = per 2 months; PY = per year.

the R package boot (version 1.3-28; Canty & Ripley, 2022).

6.2.3.4 Results

Path analysis resulted in an estimated always-exposed versus never-exposed effect of 0.69, 95% CI [-0.01, 1.34], implying that there is no evidence of an effect of sustained smoking cessation on body weight a year later. Analysis with IPW regression resulted in a negative estimate of sustained smoking cessation, -1.87, [-4.29, 0.53], although it similarly was not significant at the $\alpha = .05$ level. The substantive conclusions drawn using the different analyses would thus be the same. Differences between the point estimates can be due to the different set of covariates that was adjusted for, and the different parametric assumptions that both methods rely on.

6.3 Simulation study

So far, we have given an elaborate illustration of the investigation of joint effects, specifically an always-exposed versus never-exposed effect, in the causal inference framework, and using path analysis and IPW linear regression as estimation methods. In the current section, we study the impact of violations of parametric assumptions in path analysis and IPW linear regression, particularly, violations of the linearity assumption. We performed a simulation study to compare the finite sample performance of both estimation methods in terms of bias and MSE under various scenarios of model misspecification. In line with the empirical example, we focused on investigating the always-exposed versus never-exposed effect. The scenarios considered here were further simplified compared to the empirical example (in terms of number of covariates), but are based on the same causal structure as the simplified DAG in Figure 6.1.

6.3.1 Data generation

Data were generated under five different data-generating mechanisms (DGMs). All DGMs contain a time-varying binary exposure A measured at time points $t = 0, 1$, a continuous end-of-study outcome Y_2, a continuous baseline confounder L_{-1}, and continuous time-dependent confounding variables L_t at time points $t = 0, 1$. The simulated data have a causal structure as visualized in Figure 6.4, with continuous variables following a normal distribution. Appendix 6.A contains a table with population values for all regression coefficients. In DGM 1, all dependencies are linear. In DGM 2, the dependencies of the time-dependent confounders L_0 and L_1 include a quadratic term (see Figure 6.5a). These terms were created by first grand mean centering the predictors before squaring them. The quadratic regression coefficients were

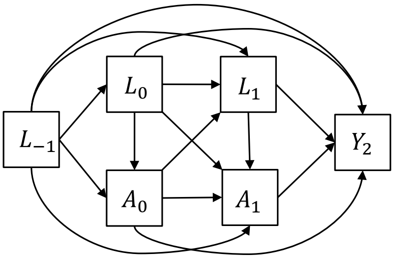


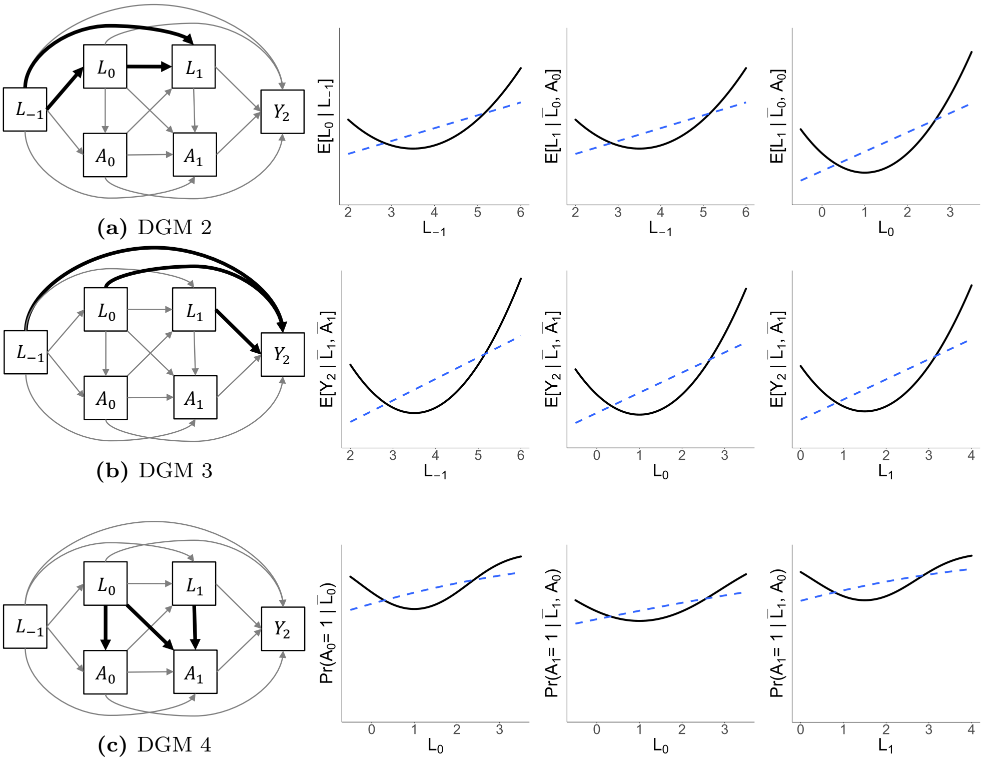
Figure 6.4: The causal structure of the data generating mechanisms used in the simulations.

equal to the linear regression coefficients. By grand mean centering the predictors, the population value of the always-exposed versus never-exposed effect does not change. In DGM 3, the dependencies of the outcome on the baseline and time-dependent covariates are quadratic (see Figure 6.5b). In DGM 4, the dependencies of the exposure on the time-varying covariates are quadratic (see Figure 6.5c). Finally, DGM 5 combines all quadratic dependencies of DGMs 2, 3 and 4. In all five DGMs, the population controlled direct effect of A_0 on Y_2 is 0.32, and the population controlled direct effect of A_1 on Y_2 is 0.40, such that, combined, the population always-exposed versus never-exposed effect is 0.72. Data generation was performed in base R (version 4.2.2; R Core Team, 2022).

6.3.2 Estimation

Five different estimation methods were used for investigating the always-exposed versus never-exposed effect: IPW linear regression, linear path analysis, IPW regression with both linear and quadratic terms in the propensity score model, path analysis with both linear and quadratic terms, and linear regression without confounding adjustment. These methods estimated the always-exposed versus sustained non-exposed effect as a combination of the controlled direct effects of A_0 and A_1. The propensity score model and outcome model of IPW linear regression were fitted using standard OLS regression in R version 4.2.2 (R Core Team, 2022). The path analysis models were fitted in Mplus version 8.9, with the probit link used for the regression models of the exposures, and robust maximum likelihood selected as the estimator (L. K. Muthén & Muthén, 2017).

The linear IPW estimation method was misspecified under DGMs 4 and 5: It wrongly assumed linear dependencies for the propensity score model. For path analysis, a linear path analysis model was specified, which was locally misspecified under DGMs 2, 3, 4, and 5. To get a sense for the impact of misspecification on performance of the method, we also estimated the joint effects without misspecificiation in the methods: For IPW, this was implemented using a propensity score model that



(a) DGM 2

(b) DGM 3

(c) DGM 4

Figure 6.5: Overview of data generating mechanisms (DGMs) 2, 3, and 4. Bold black arrows in the DAGs indicate nonlinear dependencies. These are visualized in the plots to the right of each respective DGM, with the solid black line representing the true (nonlinear) functional relationship between two variables, and the dashed blue line representing the linear projection. DGM 1 (not illustrated here) contains only linear dependencies. DGM 5 (not illustrated here) combines the nonlinear dependencies of DGMs 2, 3, and 4.

included quadratic terms for DGMs 4 and 5; for path analysis, a path analysis model was specified which included quadratic terms where relevant for DGMs 2, 3, 4, and 5. These latter two scenarios thus represent a "best-case scenario", in which no model misspecification occurs in the IPW regression, and path analysis methods. Finally, a linear regression model with Y_2 as the outcome and A_0 and A_1 as independent variables was specified, without any regression adjustment for confounding, or weighting. This method provides a benchmark for a "worst-case scenario" against which we can compare (misspecified) IPW regression and path analysis methods. Performance of these five methods under different simulation conditions was evaluated in terms of bias of the joint-effect point estimates, and mean square error (MSE).

In addition to varying the source of model misspecification, we varied sample size ($n = 300, 1000$) and proportion exposed at both time points ($p = 0.1, 0.5, 0.9$). Combined, this lead to thirty simulation conditions. For each condition, a thousand replications were simulated.

6.4 Results

Figure 6.6 visualizes the bias of point estimates for the always-exposed versus never-exposed effect across the five estimation methods. Here, we only present results for a sample size of $n = 1000$, and 10% and 50% exposed. Figure 6.7 contains the mean square error of these point estimates. The horizontal bars in both plots are 95% confidence intervals (CI), based on Monte Carlo standard errors, for the bias and MSE (Morris et al., 2019). For most estimates of bias and MSE, this CI is so narrow that it is not visible. Numerical results, as well as results for the other simulation conditions, are included in the online supplementary materials.

As expected under DGM 1, the linear IPW regression model and linear path analysis model performed well in terms of bias and MSE. Here, the IPW regression model and path analysis model with additional quadratic effects were equivalent, as all dependencies are in fact linear under DGM 1. For DGM 2, there was only slight upward bias for the linear path analysis model under the 10% exposed condition, which reduced to near 0 when exposure was balanced (it is barely visible in Figure 6.6, but shows in the numerical results in the online supplementary materials). This bias did not exist for the linear IPW regression model, although it had more variability of the estimates as reflected in the slightly increased MSE.

Results for DGM 3 and 10% exposed showed significant bias in the estimates of the linear path analysis model, and small bias for the linear IPW regression model. The higher bias for the linear path analysis model was also reflected in the MSE, which was now higher than that of the linear IPW regression model. When exposure was balanced, these biases disappeared and linear path analysis had a lower MSE again.

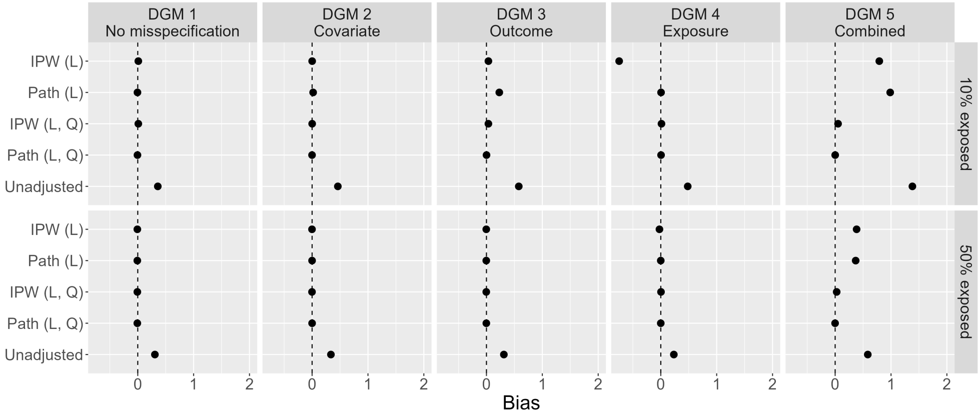


Figure 6.6: Bias in the point estimates of the always-exposed versus never-exposed effect across five methods: “IPW (L)” is linear IPW regression; “Path (L) is linear path analysis; “IPW (L, Q)” is IPW regression with linear and quadratic terms in DGMs 3, 4, and 5; “Path (L, Q)” is path analysis with linear and quadratic terms in DGMs 2, 3, 4, and 5; “Unadjusted” is a linear regression without confounding adjustment. Results are presented for the case of $n = 1000$, 10% and 50% exposed, and across five DGMs.

Results for DGM 4 with 10% exposed showed a large negative impact of an incorrectly specified propensity score model for IPW-based estimators (Hernán & Robins, 2020). There was considerable bias for the linear IPW regression model and increased MSE. When exposure was balanced in the sample, both bias and MSE were close to zero again, although some bias remained. Somewhat surprisingly, estimates of the effect of interest in the linear path analysis model appeared unaffected by the incorrectly modeled exposures, with no bias and low MSE for both the 10% exposed and 50% exposed conditions.

Finally, for DGM 5, both the linear IPW regression model and the linear path analysis model performed badly, with significant bias in the point estimates and high MSE. This was expected as there was considerable misspecification of functional forms in multiple locations of the models (i.e., numerous violations of parametric assumptions). Performance increased somewhat when the proportion exposed in the sample was balanced, but bias remained significant. In this situation, both methods performed almost as poorly as the naive, unadjusted method.

6.5 Discussion

While the use of SEM models for causal inference from longitudinal observational data is quite popular in psychology, this practice has been criticized in the causal inference literature for its high potential of model misspecification and, consequently,

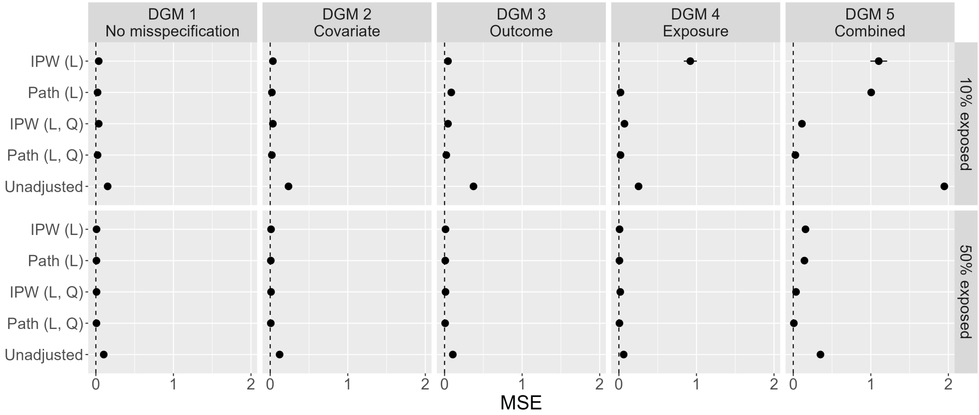


Figure 6.7: MSE of the point estimates of the always-exposed versus never-exposed effect across five methods, and five DGMS ($n = 1000$).

bias in the estimates of causal effects of interest (cf. Bollen & Pearl, 2013; Van der Laan & Rose, 2011; VanderWeele, 2012). To fully understand this critique, and to see why the alternative causal inference methods that have been proposed counter these problems, researchers need to be knowledgeable of the potential outcomes framework. Although SEM methods are compatible with the potential outcomes framework (e.g., Loeys et al., 2014; Moerkerke et al., 2015; B. O. Muthén et al., 2016), the literature on the potential outcomes framework comes predominantly from the disciplines of epidemiology and biostatistics; as such, the literature is targeted to research problems and common practices that psychological researchers are less familiar with, making it difficult to bridge the disciplinary gap. In this article, we first introduced SEM users from psychology (and related disciplines) to three core phases of the potential outcomes approach to causal inference (inspired by Goetghebeur et al., 2020; Petersen & Van der Laan, 2014). In particular, we compared path analysis from the SEM framework, to IPW estimation of MSMs when investigating an always-exposed versus never-exposed effect of a time-varying exposure on an end-of-study outcome, in the presence of baseline and time-varying confounding. Through the use of a simulation study, we assessed the finite-sample performance (in terms of bias and MSE) of both methods under varying violations of parametric assumptions.

Simulation results showed that, for the specific scenarios investigated in this study, path analysis generally had lower MSE than IPW estimation when estimating the time-varying exposure effect; the only exception here was for DGM 3, with misspecification in the relationships between the confounders and the outcome. The lower MSE obtained with path analysis was mainly due to higher efficiency, which compensated for the higher bias under particular forms of misspecification (see, for example, the lower MSE of path analysis in DGM 4, specifically for "IPW (L, Q)" and "Path (L, Q)"; and DGM 5, even while path analysis was as biased, or more biased than IPW regression; Vansteelandt & Sjolander, 2016). For misspecification of the covariate-outcome relations (i.e., in DGM 3, in which a linear relation was assumed in the fitted model whereas data were generated under a quadratic relation), results for an uneven distribution of exposed and non-exposed individuals (10% exposed) confirmed that path analysis was more prone to bias in the always-exposed versus never-exposed effect than IPW estimation. However, the bias appeared to be minor. For misspecification of the propensity score model (the covariate-exposure relationships in DGM 4), IPW estimation led to significant bias for the always-exposed versus never-exposed effects, whereas no bias was observed for path analysis in this scenario. When covariate, exposure, and outcome dependencies were all misspecified (DGM 5), then both path analysis and IPW regression performed almost as poorly (in terms of bias) as standard linear regression without any covariate adjustment. Interestingly, bias across all scenarios was significantly reduced when the proportion exposed was balanced.

Hence, our comparison of path analysis and IPW estimation across the three phases of causal inference has made insightful how SEM approaches fit within a principled approach to causal inference, the causal identification assumptions that both methods rely on, and the differences between them in terms of the parametric assumptions they make. Subsequently, our simulations have shown that violations of parametric assumptions unique to path analysis (i.e., concerning covariate-covariate relationships, investigated in DGM 2; and covariate-outcome relationships, investigated in DGM 3) did not always translate into substantial bias when estimating joint effects from finite samples. These results nuance the criticism of SEM for the purpose of causal inference, as expressed by VanderWeele (2012), for instance. Moreover, we find that in a setting without unmeasured confounding, path analysis actually performed better generally in terms of MSE, and showed no bias when the functional forms of the covariate-exposure relations are misspecified, in contrast to IPW estimation (see DGM 4).

However, this should not be interpreted to mean that SEM can be easily used for the purpose of causal inference. First, our illustrative example highlights that attempts to model the entire data generating mechanism (as with cross-lagged panel modeling approaches) complicates computations of joint effects when categorical time-varying covariates are included (e.g., combining linear regression coefficients with logit or probit coefficients). This is problematic as the inclusion of many time-varying covariates is required to make the causal identification assumption of conditional exchangeability plausible in the first place, and some covariates are likely to be categorical in practice (e.g., level of education, diagnoses of psychological disorders, relationship status, etc.). Second, our simulations focused only on a limited number of scenarios, and we may find different results when considering other scenarios, such as: The presence of unmeasured confounding variables; wrongfully omitting interactions and second-order lagged effects from the model; a different set of population parameter values; and more severe violations of parametric assumptions. In light of this uncertainty, it is still advisable to consider methods that relax the parametric assumptions as much as possible. Causal inference methods from the potential outcomes framework are advantageous in this respect.

Furthermore, we emphasize that these simulation results should certainly not be interpreted as an incentive to continue currently popular SEM modeling practices, when the actual goal is causal inference. While estimation of causal effects using SEM models *can* work well (as illustrated in the simulations), it requires very careful and elaborate consideration of the issues and topics in Phases 1 and 2 of causal research, as we have shown in this article. Fitting an off-the-shelf bivariate cross-lagged panel model (or a related SEM model) without inclusion of additional covariates (both baseline and time-invariant), and without consideration of lag-2 and further relationships, is inappropriate for investigating causal effects. While this paper focused on the investigation of joint effects, our conclusion equally applies when the interest is in cross-lagged effects. In fact, we estimated joint effects as combinations of CDEs, and under the causal DAGs in Figure 6.1 and 6.4, the CDE of exposure at time point 2 is the same as the cross-lagged effect of exposure at time point 2 to the end-of-study outcome. Psychological researchers are therefore well-advised to study the potential outcomes framework, and the proposed causal inference methods therein such that they can make better-informed decisions about which modeling approach is appropriate given their considerations in Phases 1 and 2.

In this article, we limited our simulations to misspecification of functional forms, and did not investigate the impact of unobserved confounding variables from the analysis, or the potential of latent variables to (partially) adjust for this (Usami et al., 2019). Unobserved confounding is, however, a fundamental issue in causal research. We also did not study the performance of path analysis and IPW estimation under violations of conditional independence assumptions—that is, when the causal DAG that acts as the basis for our analyses wrongfully omits one, or multiple, dependencies—which was an additional critique in VanderWeele (2012). Instead, in our simulations and in our illustrative example, both the path analysis model and IPW estimation included lag-0, lag-1, lag-2, and lag-3 relationships. Furthermore, missingness in the illustrative example was handled by single stochastic imputation for practical reasons. However, as the SEM framework and potential outcome framework have different techniques for missing data handling—IPW for censoring is more common in the potential outcomes framework, whereas the use of full information likelihood is widespread in SEM—it would be interesting to also investigate how differences in these techniques impact estimation performance.

In conclusion, psychological research has fully embraced the SEM framework for causal inference, whereas the uptake of the potential outcomes framework, and the causal inference methods developed herein, has been lagging behind. However, reduced reliance on parametric assumptions and the possibility to include a large set of (categorical) time-varying covariates, are good reasons to invest time in learning techniques such as IPW estimation of MSMs. We hope this comparison of IPW estimation and path analysis facilitates a better understanding of these methods for causal inference about time-varying exposure effects.

Author contributions: Conceptualization: JM, KL, BP, and EH - Data curation: JM - Formal analysis: JM – Investigation: JM and KL - Methodology: JM, KL, BP, and EH – Project administration: JM and KL - Software: JM - Visualization: JM and KL - Writing (original draft preparation): JM and KL – Writing (review and editing): JM, KL, BP, and EH

Funding: Jeroen D. Mulder and Ellen L. Hamaker are part of the Consortium on Individual Development (CID). CID is funded through the Gravitation Program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003).

Online supplementary materials: This study's online supplementary materials can be found at https://jeroendmulder.github.io/SEM-and-MSM.

Appendices

6.A Population values simulation study

Table 6A.1 contains the population values used for data generation. L_{-1} is normally distributed with mean 4 and variance 1. Residuals are standard-normally distributed. The intercepts of L_0 and L_1 are set to 1. The intercept of Y_2 is 0. These population values resulted in an always-exposed versus never-exposed effect of 0.72.

Table 6A.1: Population values used for data generation.

Causal effect	Population value
$L_{-1} \rightarrow ...$	0.1[a]
$L_0 \rightarrow A_0$	0.5
$L_0 \rightarrow L_1$	0.3
$L_0 \rightarrow A_1$	0.25
$L_0 \rightarrow Y_2$	0.15
$A_0 \rightarrow L_1$	0.4
$A_0 \rightarrow A_1$	0.8
$A_0 \rightarrow Y_2$	0.2
$L_1 \rightarrow A_1$	0.5
$L_1 \rightarrow Y_2$	0.3
$A_1 \rightarrow Y_2$	0.4

[a] This value applies to all effects of L_{-1}.

6.B Derivation statistical estimand

Here, we describe the derivation of the statistical estimand in Equation 6.4 from the causal estimand in Equation 6.1. In the derivation we make use of mathematical formalisms such as the law of iterated expectations, as well as the causal identification assumptions, conditional exchangeability, consistency, and positivity:

$$
\begin{aligned}
\text{causal estimand} :&= \mathbb{E}[Y_2^{\bar{1}_1}] - \mathbb{E}[Y_2^{\bar{0}_1}], \\
&\overset{(1)}{=} \mathbb{E}\big\{\mathbb{E}[Y_2^{\bar{1}_1} \mid L_0]\big\} - \mathbb{E}\big\{\mathbb{E}[Y_2^{\bar{0}_1} \mid L_0]\big\} \\
&\overset{(2)}{=} \mathbb{E}\big\{\mathbb{E}[Y_2^{\bar{1}_1} \mid A_0 = 1, L_0]\big\} - \mathbb{E}\big\{\mathbb{E}[Y_2^{\bar{0}_1} \mid A_0 = 0, L_0]\big\} \\
&\overset{(3)}{=} \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2^{\bar{1}_1} \mid A_0 = 1, \overline{L}_1] \bigm| A_0 = 1, L_0\big)\Big\} \\
&\qquad - \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2^{\bar{0}_1} \mid A_0 = 0, \overline{L}_1] \bigm| A_0 = 0, L_0\big)\Big\} \\
&\overset{(4)}{=} \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2^{\bar{1}_1} \mid \overline{A}_1 = \bar{1}_1, \overline{L}_1] \bigm| A_0 = 1, L_0\big)\Big\} \\
&\qquad - \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2^{\bar{0}_1} \mid \overline{A}_1 = \bar{0}_1, \overline{L}_1] \bigm| A_0 = 0, L_0\big)\Big\} \\
&\overset{(5)}{=} \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2 \mid \overline{A}_1 = \bar{1}_1, \overline{L}_1] \bigm| A_0 = 1, L_0\big)\Big\} \\
&\qquad - \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2 \mid \overline{A}_1 = \bar{0}_1, \overline{L}_1] \bigm| A_0 = 0, L_0\big)\Big\} \\
&=: \text{statistical estimand (g-formula representation).}
\end{aligned}
$$

Equality (1) follows from law of iterated expectations with regards to L_0. Equality (2) follows from conditional exchangeability of the form $Y_2^{\overline{a}_1} \perp\!\!\!\perp A_0 \mid L_0$ and positivity. Equality (3) follows from law of iterated expectations with regards to L_1, conditional on L_0 and A_0. As we now condition on both L_0 and L_1, we represent this as conditioning on covariate history $\bar{L}_1$. Equality (4) follows from conditional exchangeability of the form $Y_2^{\overline{a}_1} \perp\!\!\!\perp A_1 \mid \overline{L}_1, A_0 = a_0$; and positivity. Equality (5) relies on the consistency assumption.

This statistical estimand takes the form that is known in the causal inference literature as the standard g-formula for time-varying exposures (Naimi et al., 2016; Robins, 1986). It can be further rewritten to a form that is known in the causal inference literature as the IPW representation. Continuing from the statistical estimand in g-formula representation on the right-hand side of Equality (5), it yields:

$$
\begin{aligned}
&\mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2 \mid \overline{A}_1 = \overline{1}_1, \overline{L}_1] \mid A_0 = 1, L_0\big)\Big\} \\
&\quad - \mathbb{E}_{L_0}\Big\{\mathbb{E}_{L_1}\big(\mathbb{E}[Y_2 \mid \overline{A}_1 = \overline{0}_1, \overline{L}_1] \mid A_0 = 0, L_0\big)\Big\} \\
&\qquad \overset{(6)}{=} \mathbb{E}_{L_0}\left\{\mathbb{E}_{L_1}\left(\mathbb{E}\left[\frac{Y_2 \mathbb{1}(A_1 = 1)}{\Pr(A_1 = 1 \mid A_0 = 1, \overline{L}_1)} \,\middle|\, A_0 = 1, \overline{L}_1\right] \,\middle|\, A_0 = 1, L_0\right)\right\} \\
&\qquad\qquad - \mathbb{E}_{L_0}\left\{\mathbb{E}_{L_1}\left(\mathbb{E}\left[\frac{Y_2 \mathbb{1}(A_1 = 0)}{\Pr(A_1 = 0 \mid A_0 = 0, \overline{L}_1)} \,\middle|\, A_0 = 0, \overline{L}_1\right] \,\middle|\, A_0 = 0, L_0\right)\right\} \\
&\qquad \overset{(7)}{=} \mathbb{E}_{L_0}\left\{\mathbb{E}\left(\frac{Y_2 \mathbb{1}(A_1 = 1)}{\Pr(A_1 = 1 \mid A_0 = 1, \overline{L}_1)} \,\middle|\, A_0 = 1, L_0\right)\right\} \\
&\qquad\qquad - \mathbb{E}_{L_0}\left\{\mathbb{E}\left(\frac{Y_2 \mathbb{1}(A_1 = 0)}{\Pr(A_1 = 0 \mid A_0 = 0, \overline{L}_1)} \,\middle|\, A_0 = 0, L_0\right)\right\} \\
&\qquad \overset{(8)}{=} \mathbb{E}_{L_0}\left\{\mathbb{E}\left(\frac{Y_2 \mathbb{1}(\overline{A}_0 = \overline{1}_1)}{\Pr(A_0 = 1 \mid L_0)\Pr(A_1 = 1 \mid A_0 = 1, \overline{L}_1)} \,\middle|\, L_0\right)\right\} \\
&\qquad\qquad - \mathbb{E}_{L_0}\left\{\mathbb{E}\left(\frac{Y_2 \mathbb{1}(\overline{A}_1 = \overline{0}_1)}{\Pr(A_0 = 0 \mid L_0)\Pr(A_1 = 0 \mid A_0 = 0, \overline{L}_1)} \,\middle|\, L_0\right)\right\} \\
&\qquad \overset{(9)}{=} \mathbb{E}\left\{\frac{Y_2 \mathbb{1}(\overline{A}_1 = \overline{1}_1)}{\Pr(A_0 = 1 \mid L_0)\Pr(A_1 = 1 \mid A_0 = 1, \overline{L}_1)}\right\} \\
&\qquad\qquad - \mathbb{E}\left\{\frac{Y_2 \mathbb{1}(\overline{A}_1 = \overline{0}_1)}{\Pr(A_0 = 0 \mid L_0)\Pr(A_1 = 0 \mid A_0 = 0, \overline{L}_1)}\right\} \\
&\qquad \overset{(10)}{=} \mathbb{E}\big[WY_2 \mathbb{1}(\overline{A}_1 = \overline{1}_1)\big] - \mathbb{E}\big[WY_2 \mathbb{1}(\overline{A}_1 = \overline{0}_1)\big] \\
&\qquad \overset{(11)}{=} \mathbb{E}\big[WY_2 \Pr(\overline{A}_1 = \overline{1}_1) \mid \overline{A}_1 = \overline{1}_1\big] - \mathbb{E}\big[WY_2 \Pr(\overline{A}_1 = \overline{0}_1) \mid \overline{A}_1 = \overline{0}_1\big] \\
&\qquad \overset{(12)}{=} \mathbb{E}\left[\frac{W}{\mathbb{E}[W \mid \overline{A}_1]} Y_2 \,\middle|\, \overline{A}_1 = \overline{1}_1\right] - \mathbb{E}\left[\frac{W}{\mathbb{E}[W \mid \overline{A}_1]} Y_2 \,\middle|\, \overline{A}_1 = \overline{0}_1\right] \\
&\qquad \overset{(13)}{=} \mathbb{E}\big[W^* Y_2 \mid \overline{A}_1 = \overline{1}_1)\big] - \mathbb{E}\big[W^* Y_2 \mid \overline{A}_1 = \overline{0}_1\big] \\
&\qquad =: \text{statistical estimand (IPW representation).}
\end{aligned}
$$

Equality (6) follows from the law of iterated expectations with regard to A_1, conditional on A_0 and $\overline{L}_1$; and positivity. Here, $\mathbb{1}(\cdot)$ is an indicator function that equals one if $(\cdot)$ is true, and zero otherwise. Equality (7) follows from the law of iterated expectations with regard to L_1, conditional on A_0 and L_0. Equality (8) follows from the law of iterated expectations with regard to A_0, conditional on L_0; and positivity. Equality (9) follows from the law of iterated expectations with regard L_0. For Equality (10) we define inverse probability of exposure weights W as $W = W_0 W_1$, that is the product of the inverse probability weights at each exposure-time. With a dichotomous exposure taking on values $A = 0$ or $A = 1$, and for time points $t = 0, 1$, the time-specific weights are defined as

$$W_t = \frac{A_t}{\Pr(A_t = 1 \mid \overline{L}_t, \overline{A}_{t-1})} + \frac{1 - A_t}{1 - \Pr(A_t = 1 \mid \overline{L}_t, \overline{A}_{t-1})}.$$

Equality (10) then follows from this definition of the weights W. Equation (11) follows from the law of iterated expectations with regard to $\overline{A}_1$. To see that Equality (12) is true, suppose for simplicity that $\overline{L}_1$ is discrete, and observe that

$$\begin{aligned}
\mathbb{E}[W \mid \overline{A}_1 = \overline{a}_1] &= \mathbb{E}\left[\frac{1}{\Pr(A_0 = a_0 \mid L_0)\Pr(A_1 = a_1 \mid A_0, \overline{L}_1)} \,\middle|\, \overline{A}_1 = \overline{a}_1\right] \\
&= \sum_{l_0}\sum_{l_1} \frac{\Pr(\overline{L}_1 = \overline{l}_1 \mid \overline{A}_1 = \overline{a}_1)}{\Pr(A_0 = a_0 \mid L_0)\Pr(A_1 = a_1 \mid A_0 = a_0, \overline{L}_1 = \overline{l}_1)} \\
&= \frac{1}{\Pr(\overline{A}_1 = \overline{a}_1)} \sum_{l_0}\sum_{l_1} \frac{\Pr(\overline{A}_1 = \overline{a}_1, \overline{L}_1 = \overline{l}_1)}{\Pr(A_0 = a_0 \mid L_0)\Pr(A_1 = a_1 \mid A_0 = a_0, \overline{L}_1 = \overline{l}_1)} \\
&= \frac{1}{\Pr(\overline{A}_1 = \overline{a}_1)} \sum_{l_0}\sum_{l_1} \Pr(L_0 = l_0)\Pr(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0) \\
&= \frac{1}{\Pr(\overline{A}_1 = \overline{a}_1)} \sum_{l_0} \Pr(L_0 = l_0) \sum_{l_1} \Pr(L_1 = l_1 \mid A_0 = a_0, L_0 = l_0) \\
&= \frac{1}{\Pr(\overline{A}_1 = \overline{a}_1)}.
\end{aligned}$$

Equality (13) is a more succinct way of expressing the statistical estimand in IPW representation using stabilised weights $W^* = W/\mathbb{E}[W \mid \overline{A}_2]$. This derivation shows that a statistical estimand in g-formula representation can be rewritten into a statistical estimand in IPW representation. However, both representations suggest a different modeling approach: The g-formula representation suggest that a statistical model for L_1 and Y might need to be specified, whereas the IPW representation suggests that the time-varying exposures are modelled (via the inverse probability of exposure weights W).

CHAPTER 7

Joint effects in panel data using structural nested mean models: An introduction for psychologists familiar with cross-lagged panel modeling

Revised version submitted as:
Mulder, J. D., Usami, S., & Hamaker, E. L. (submitted). Joint effects in panel data using structural nested mean models: An introduction for psychologists familiar with cross-lagged panel research.

Abstract

A popular approach among psychological researchers for investigating causal relationships from panel data is cross-lagged panel modeling. However, structural equation models are critiqued in the causal inference literature for relying on an unnecessarily large number of parametric assumptions, thereby increasing the risk of model misspecification and bias. Instead, the use of structural nested mean models (SNMMs) with G-estimation are promoted as an approach that relies on fewer assumptions and therefore, in principle, leads to more valid causal conclusions. However, the uptake of SNMMs and G-estimation in the psychological literature is lacking, hampered by a disconnect between the causal inference literature, and the statistical concepts and modeling practices that psychological researchers are familiar with. In this paper, we aim to bridge this disciplinary divide by introducing joint effects, controlled direct effects, SNMMs, and G-estimation, and comparing these to cross-lagged panel modeling approaches. An empirical example from psychological practice is used throughout.

Across a wide range of disciplines, researchers analyze longitudinal, observational data to investigate prospective causal relationships between variables. In psychology, a signification portion of this kind of research is devoted to *lag-1 relationships*, which are investigated using cross-lagged panel modeling approaches within the framework of structural equation modeling (Gische et al., 2021; Usami et al., 2019; Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020). In contrast, in disciplines like epidemiology and biostatistics, research more typically focuses on *exposure regimes* and *joint effects*. These concern effects of a collection of repeatedly-measured exposures (i.e., the effect of an X-variable measured at time points 1, 2, 3, etc., combined) on an outcome. An interesting aspect of joint effects is that it comprises both short-term and long-term influences (i.e., effects at multiple time lags: lag-0, lag-1, lag-2, etc.), and that it goes beyond the individual, direct paths that are targeted by cross-lagged effects (Vansteelandt & Joffe, 2014). While a joint effect can be assessed within the structural equation modeling framework, they are traditionally investigated using a class of formal causal modeling approaches that were developed largely by James M. Robins (Daniel et al., 2013; Naimi et al., 2016). In this paper, we focus on one of these approaches, namely structural nested mean models (SNMMs) with G-estimation, as it is best equipped to analyze continuous exposures (common in psychological research) and has the most advantages from a causal inference point-of-view (Vansteelandt & Sjolander, 2016). The appeal of SNMMs with G-estimation is that it relies on fewer parametric assumptions than structural equation modeling approaches, thereby reducing the potential for model misspecification (e.g., wrongly assuming an effect is linear whereas, in fact, it is nonlinear) and leading, in principle, to more robust causal conclusions (Van der Laan & Rose, 2011; VanderWeele, 2012).

Despite this advantage, the interest in joint effects and the uptake of SNMMs with G-estimation is limited in the psychological literature. While there are many introductions to this approach for investigating joint effects (e.g., Goetghebeur et al., 2020; Hernán & Robins, 2020; Naimi et al., 2016; Petersen & Van der Laan, 2014), these are typically not targeted towards psychological researchers, and provide little to no connection to the modeling practices that they are familiar with. Such a disconnect between strands of literature hinders researchers from understanding how different kinds of causal hypotheses, and the modeling approaches for estimating causal effects, are related. Two important contributions in this regard are recently published papers by Loh and Ren (2023a, 2023b), who provide an introduction to SNMMs with G-estimation (based on Vansteelandt & Sjolander, 2016), and illustrate how SNMMs can be fitted to longitudinal data with G-estimation within the structural equation modeling framework. The current paper supplements both papers by (a) extending the use of SNMMs to continuous predictors, commonly used in psychological panel data; (b) providing a more conceptual explanation of the concepts that underlie this causal inference approach, such as joint effects, controlled direct effects, exposure regimes, and the essence of SNMMs and G-estimation; and most importantly (c) comparing the use of SNMMs explicitly to modeling practices SEM users are familiar with, in particular cross-lagged panel modeling. We introduce key concepts using a cross-lagged panel design, minimize technical details, and present an empirical psychological example regarding self-esteem and depression throughout.

This article is organized as follows. Section 7.1 provides the necessary background: We start with introducing a visual representation of a causal process (also referred to as a causal directed acyclical graph, or DAG); explain the difference between cross-lagged effects and joint effects; and end with discussing the causal identification assumptions needed for a causal interpretation of model estimates. This is followed, in Section 7.2, by the introduction and comparison of cross-lagged panel modeling approaches versus SNMMs using G-estimation. In Section 7.3, we illustrate both approaches for the investigation of joint effects with empirical data of self-esteem and depression. The discussion in Section 7.4 connects our treatment of joint effects and SNMMs to other modeling topics that are prominent in the psychological modeling literature, such as the decomposition of observed variance into within- and between-person variance, the inclusion of contemporaneous effects, and the inclusion of lag-2 effects to control for confounding. Annotated R code for the empirical analyses in this paper can be found in the online supplementary materials at https://jeroendmulder.github.io/joint-effects-using-SNMM.

7.1 Background

This section starts with an introduction of causal DAGs, using an empirical psychological example based on Kuster et al. (2012). Subsequently, joint effects are introduced and compared to the cross-lagged effects which tend to be the key focus in psychological research. We end with a discussion of causal identification assumptions; while these are not the focus of this article, they are needed for a causal interpretation of statistical results, regardless of the kind of causal effect that is targeted, or the modeling approach that was taken.

7.1.1 A causal DAG for self-esteem and depression

Causal DAGs are graphical tools that can be used to represent the causal structure of empirical phenomena that researchers want to study. It consists of a set of variables (nodes) and one-headed arrows representing the causal dependencies between them (edges; Pearl, 2009). All variables that are believed to play a role in the empirical phenomena should be included in the *causal* DAG. Thus, in addition to exposures and outcomes, causal DAGs usually also include a set of time-varying and time-invariant covariates (both observed and unobserved), and their causal connections (Hamaker et al., 2020; Pearl, 2009; Rohrer, 2018). These causal DAGs appear similar to path diagrams in the structural equation modeling framework, but there are three important differences: causal DAGs (a) do not necessarily imply linear relationships, that is, they represent dependencies between variables, without assuming a specific functional form of this dependency; (b) do not make any assumptions about the distribution underlying this system of variables; and (c) do not include two-headed arrows representing unexplained covariances between variables (Pearl, 2009).

Suppose we are interested in assessing causal relations between *self-esteem* and *depressive symptoms*. Let X_t be a measure of self-esteem and Y_t be a measure of depressive symptoms, both measured at time point t. Let L_t represent a time-varying covariate at time point t, for example *rumination*, and let C represent time-invariant baseline covariates such as gender, family social economic status, and maternal age (Boden et al., 2008). We can represent the causal structure underlying these variables over time in a causal DAG, as shown in Figure 7.1. It contains the four repeated measures of self-esteem, depressive symptoms and rumination, two (sets of) time-invariant covariates C and U, and the dependencies between these variables over time. The time-invariant covariates in C influence all variables at future time points; to avoid clutter, not every arrow is drawn in the DAG. The time-invariant variable U represents covariates that exist before $t = 1$, and that only has direct effects on X, Y, and L at the first time point. The existence of such a variable is often assumed in panel data, as measurements of X, Y, and L are obtained at random points in time in an ongoing process: U can then represent unobserved realisations of X, Y, and L before the start of measurement that results in covariances between X_1, Y_1, and L_1. This specific causal DAG represents a structure where time-varying variables influence all other time-varying variables at the next time point, but this influence does not extend beyond lag 1. This is also the predominant causal structure that is assumed in psychological cross-lagged panel research (Usami et al., 2019).

While working with the causal DAG in Figure 7.1, we make the implicit assumption that it correctly represents the underlying causal structure between depressive symptoms, self-esteem, rumination, and the time-invariant covariates (Imbens, 2019). Arguably, the lag-1 process as encoded in the DAG of Figure 7.1 is an oversimplification, as empirical processes might include effects that extend beyond a single time interval (i.e., lag-2, and further; Little, 2013, p. 203). Additionally, lag-0 effects can be added to the DAG to represent instantaneous effects. Such effects are commonly assumed in longitudinal biomedical research, and recently, B. O. Muthén and Asparouhov (2022a) argued that lag-0 effects may also be realistic in psychological research when data are collected with long time intervals, and measurements referring

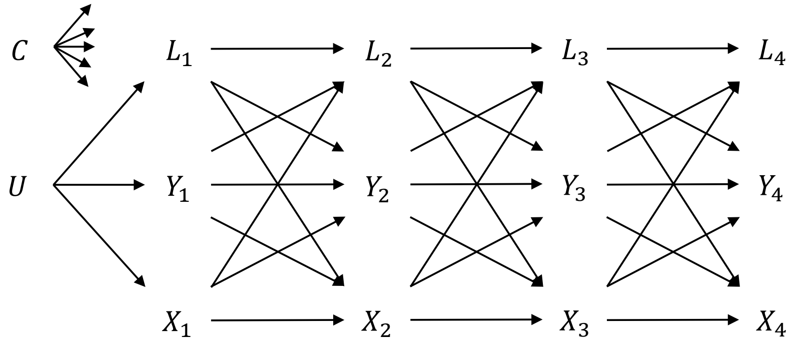


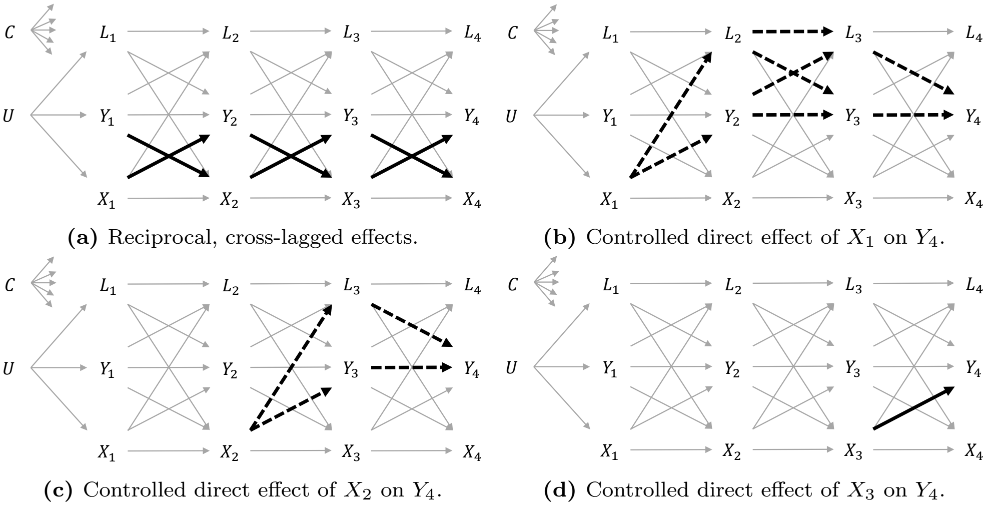
Figure 7.1: A causal DAG, representing how a time-invariant variable C, and time-varying variables X, Y, and L are causally related to each other across 4 repeated measurements. C is causally related to all other variables in the model, although not all arrows are included in the DAG to prevent clutter.

to past experiences. For didactical reasons, we start with the simplified DAG in Figure 7.1, but in Section 7.4, we discuss in more detail the addition of lag-0 and lag-2 causal dependencies to the DAG, and how this impacts the use of structural equation models (SEMs) and SNMMs.

7.1.2 Cross-lagged effects and joint effects

Figure 7.2a visualizes cross-lagged effects in the causal DAG of Figure 7.1. Characteristically, cross-lagged effects are bidirectional, implying that self-esteem and depression take on the role of both presumed cause and outcome: At the first wave self-esteem and depression are presumed causes, at the final wave self-esteem and depression are outcomes, and at the intermediate waves self-esteem and depression are both. In psychological research oftentimes each dependency (i.e., arrow) independently is a target of inference. That is, when interested is in cross-lagged effects and assuming the causal structure of Figure 7.1, we target six causal effects: Three cross-lagged effects from self-esteem to depression, and three cross-lagged effects from depression to self-esteem. Typically, these cross-lagged effects concern lag-1 relationships.

In the epidemiological and biostatistical literature, rather than focusing on path-specific effects, it is more common to investigate effects of exposure regimes (also sometimes referred to as exposure sequences or exposure history; Wallace et al., 2017). Regimes are predetermined rules that determine the value of a time-varying exposure *for all time points jointly.* One example is a regime in which individuals are made to have a self-esteem score of, say, five at each of the four measurement occasions, that is $\{X_1 = 5, X_2 = 5, X_3 = 5, X_4 = 5\}$. Another example would be a regime in which the self-esteem score of individuals is set at 2 at the first occasion, set to 1 at the second



(a) Reciprocal, cross-lagged effects.

(b) Controlled direct effect of X_1 on Y_4.

(c) Controlled direct effect of X_2 on Y_4.

(d) Controlled direct effect of X_3 on Y_4.

Figure 7.2: Representation of the causal dependencies that are targeted by research questions on reciprocal, cross-lagged effects, and joint effects.

occasion, and set to 0 thereafter, $\{X_1 = 2, X_2 = 1, X_3 = 0, X_4 = 0\}$. For simplicity of notation, we will write such regimes as $\{5, 5, 5, 5\}$ and $\{2, 1, 0, 0\}$, respectively. Contrasting end-of-study outcomes that follow from two different treatment regimes then allows researchers to assess the average causal effect (ACE) of being exposed to one specific regime over another specific regime. Such contrasts are also referred to as *joint effects*, where "joint" refers to the exposures at multiple time points combined. Moreover, in biomedical research, the exposures are oftentimes dichotomous (e.g., an individual either did attend a therapy or not; an individual either was diagnosed to be depressed or not) such that regimes only concern zeros and ones. One particularly popular joint effect, especially in the pharmacoepidemiologic research, is the always-treated versus never-treated effect. This is represented as a contrast of regimes $\{1, 1, 1, 1\}$ versus $\{0, 0, 0, 0\}$, in which the exposure is binary with $1 =$ treatment, and $0 =$ no treatment.

The joint effect of X can be decomposed into multiple *controlled direct effects* (CDEs) of X, specifically (1) the effect of X_1 on end-of-study Y_4, which does not through later versions of X; (2) the effect of X_2 on end-of-study Y_4 which does not go through later versions of X; and (3) the effect of X_3 on end-of-study Y_4 which does not go through later X (Daniel et al., 2013). These three CDEs are visualized in Figures 7.2b, 7.2c, and 7.2d, respectively. Any single CDE captures the total effect of increasing self-esteem *at a particular point in time* on end-of-study depression, while controlling for the future self-esteem scores. The term "controlled" in CDE thus refers to the fact that values of later self-esteem are held constant at a particular value (or set of values), whereas the term "direct" refers to the fact that the underlying intermediate process by which self-esteem at a particular time point affects end-of-study depression is not modeled, but that rather a single estimate summarizing this intermediate process is obtained (Tompsett et al., 2022; Wallace et al., 2017). For researchers familiar with structural equation modeling techniques, this might be confusing terminology as the intermediate process would be regarded as a set of indirect effects, rather than direct. To accentuate the fact that for CDEs the intermediate process is not our target of inference, the intermediate dependencies for the CDEs of X_1 and X_2 in Figures 7.2b and 7.2c appear as dotted arrows.

Let us zoom in on the CDE of X_1 in Figure 7.2b. This can alternatively be represented as a contrast of outcomes following the regimes $\{x_1 + 1, x_2, x_3\}$ versus $\{x_1, x_2, x_3\}$—i.e., the effect of a one-point increase in self-esteem at the first measurement occasion on end-of-study depression, while keeping future levels of self-esteem constant at values x_2, and x_3. These values can be anything, but for interpretational reasons, researchers might set x_2 and x_3 to the mean self-esteem score, or the lowest possible score on the self-esteem scale. Even more generally, the CDE can be represented as a contrast of outcomes following the regimes $\{x_1^*, x_2, x_3\}$ versus $\{x_1, x_2, x_3\}$ to represent the effect of an arbitrary increase of self-esteem at the first measurement occasion. We ignore X_4 here as, based on the causal DAG in Figure 7.1, it has no causal effect on the end-of-study outcome. Similarly, the CDE of X_2 can be regarded as a contrast of outcomes following the regimes $\{x_1, x_2^*, x_3\}$ versus $\{x_1, x_2, x_3\}$. In this contrast, we control for exposure before time point 2 ($X_1 = x_1$), and future exposure ($X_3 = x_3$). The CDE of X_3 can be represented as a contrast of $\{x_1, x_2, x_3^*\}$ and $\{x_1, x_2, x_3\}$. Representing the CDEs as contrasts of exposure regimes is useful later for understanding how SNMMs are build up. Finally, for this particular DAG, the CDE of X_3 and the cross-lagged effect of X_3 concern the same dependency in the causal DAG, $X_3 \rightarrow Y_4$. However, note that this (causal) equivalence does not hold generally (e.g., when lag-0 effects of the time-varying covariate to the outcome are added to the causal DAG; this is further discussed in Section 7.4.2.

7.1.3 Conceptual differences between cross-lagged effects and joint effects

There are multiple conceptual differences between cross-lagged effects and joint effects. These not only affect the interpretation of the effects, but also have some statistical ramifications. First, research questions about cross-lagged effects in a psychological context are typically bidirectional in nature: Researchers investigate if effects between variables go from X to Y, from Y to X, if both processes are at work, and if so, which process is causally dominant (Rogosa, 1980). Instead, investigations of joint effects in the literature are predominately unidirectional, with researchers deciding a priori which specific causal process (i.e., which "causal direction") is studied. However, in theory, joint effects could be studied in both directions as well (e.g., Li et al., 2016).

Second, the role variables take on in a causal process depends on the causal effect that is targeted. For cross-lagged effects, six variables are exposures, namely X_1, Y_1, X_2, Y_2, X_3, and Y_3, and six variables are outcomes, namely X_2, Y_2, X_3, Y_3, X_4, and Y_4. Instead, for joint effects, the exposure is a single variable measured at multiple time points, X_t. Moreover, the majority of the studies investigating joint effects concern a single outcome, usually measured at the end of a study (e.g., Y_4). However, when an outcome is measured repeatedly (as done in a cross-lagged panel design), SNMMs can be extended to include time-varying outcomes as well (Vansteelandt & Sjolander, 2016).

The role of time-varying covariates also changes depending on whether one targets cross-lagged, or joint effects. For example, the cross-lagged effect $X_3 \rightarrow Y_4$ is confounded by L_2 via the paths $X_3 \leftarrow L_2 \rightarrow L_3 \rightarrow Y_4$ and $X_3 \leftarrow L_2 \rightarrow Y_3 \rightarrow Y_4$. This implies that rumination at time point 2 should be controlled for in a statistical analysis. In contrast, for the joint effect of X, L_2 is both a confounder and a mediator: It is a confounder for the CDE of X_3 (i.e., it is a common cause on the paths $X_3 \leftarrow L_2 \rightarrow Y_3 \rightarrow Y_4$ and $X_3 \leftarrow L_2 \rightarrow L_3 \rightarrow Y_4$), and it is mediator for the CDE of X_1 (it lies on the paths $X_1 \rightarrow L_2 \rightarrow L_3 \rightarrow Y_4$ and $X_1 \rightarrow L_2 \rightarrow Y_3 \rightarrow Y_4$). Such a "double role" complicates statistical analyses, as attempts to estimate the joint effect with standard regression methods—for example, a linear regression of Y_4 on all exposures X_1, X_2, X_3, and all confounders *simultaneously*—is incorrect: Controlling for L_2 leads to overcontrol bias for the CDE of X_1, whereas not controlling for L_2 leads to confounder bias in the CDE of X_3. In the causal inference literature, this problem is referred to as *exposure-confounder feedback*, and the causal inference approaches by Robins have been developed specifically to tackle this problem (Robins & Greenland, 2000). In Section 7.2, we discuss how exposure-confounder feedback is dealt with in a SEM and in a SNMM with G-estimation.

Third, cross-lagged effects and joint effects relate to different time lags at which the causal process operates. In general, estimates of causal effects depend critically on the size of the time interval between subsequent measures (Gollob & Reichardt, 1987; Kuiper & Ryan, 2018; Voelkle et al., 2012). Therefore, estimates of cross-lagged effects are interpreted as causal effects that take one time-lag to materialize. For our empirical example, we make use of data from Kuster et al. (2012), with measures of self-esteem, depressive symptoms, and rumination collected on a bimonthly basis. Hence, an estimate of the cross-lagged effect of self-esteem on depression is the expected change in depressive symptoms *two months later* for a one-unit increase in self-esteem. In contrast, the joint effect is a combination of causal effects at varying time-lags: The CDE of X_1 relates to six months, the CDE of X_2 relates to four months, and the CDE of X_3 relates to two months. It can be regarded as a mix of short- and longer-term effects, describing the effect of repeated bimonthly interventions on self-esteem across a six-month period. This makes it possible, at least in principle, that the separate CDEs that make up the joint effect all have significant effects on the outcome independently, but when considered jointly, they cancel each other out (e.g., when the longer-term effects are in the opposite direction as short-term effects, or vice versa).

7.1.4 Causal identification assumptions

When estimating the effects discussed above from empirical data, a causal interpretation thereof relies critically on both *causal identification assumptions* and *parametric assumptions*. While the focus of this article is on a comparison of the parametric assumptions that a SEM and a SNMM with G-estimation make, causal identification assumptions are fundamental to a causal interpretation of estimates. Therefore, we briefly introduce two central causal identification assumptions here, namely conditional exchangeability and consistency. The plausibility of these assumptions for our empirical example is elaborated upon in the Discussion section; for the purpose of this article, we continue as if these assumptions hold. Introductions to causal identification assumptions are given by Hernán and Robins (2020) and Imbens and Rubin (2015).

Both exchangeability and consistency concern *potential outcomes* and observed variables. A potential outcome, denoted by Y^x, is an outcome for a particular individual that would be observed if the individual had that exposure $X = x$. For example, suppose that we are only looking at self-esteem at time point 3 X_3, then Y^5 would be the end-of-study depression if an individual had a self-esteem score of five at time point 3, and Y^1 would be the end-of-study depression if an individual had a self-esteem score of one. In reality, an individual has only a single self-esteem score at time point 3, and thus we can only observe one potential outcome (also referred to as the factual), the others remain unknown (referred to as the counterfactuals). Similarly, we can have potential outcomes for exposure regimes, $Y^{\{x_1,x_2,x_3\}}$, which is the outcome for a particular individual that would be observed if the individual had the exposure regime $\{x_1, x_2, x_3\}$. Potential outcomes are the fundamental building blocks of much of the causal inference literature as they are used to define causal effects. In fact, we have already implicitly used these above to explain joint effects as differences between end-of-study outcomes that follow from two different regimes (i.e., as a contrast of two potential outcomes). What causal identification assumptions do, is link the causal effect of interest (in terms of potential outcomes) to the data from which we attempt to estimate this effect.

The assumption of (conditional) exchangeability states that the potential outcomes are independent from their observed value on the exposure X (conditional on a set of covariates).[1] It is a condition that is reasonable in the context of a randomized controlled trial, but is likely to be violated to some degree in nonexperimental settings. To make the assumption plausible, researchers condition on covariates that confound the targeted effect. The set of covariates to be adjusted for can be determined using the *d-separation rules* by Pearl (1995).[2] In practice, the major challenge is making sure that all identified confounders have actually been measured. Unfortunately, this cannot be tested with data, but should be evaluated by the researcher based on theory, existing literature, and/or expert opinion (Goetghebeur et al., 2020; Petersen & Van der Laan, 2014). Note that the assumption of exchangeability merely concerns *which* confounders should be accounted for, not *how* they should be accounted for. The latter concerns estimation rather than identification, and which is where SEMs and SNMMs with G-estimation show some key differences.

The consistency assumption states that the potential outcomes can be tied to observed variables, meaning that, for example, the potential outcome $Y^{\{5,5,5\}}$ is the same as the observed Y for individuals with exposure regime $\{5,5,5\}$ (Hernán & Robins, 2020). In practice, this assumption implies that these constructs are well-defined, including being specific about the (hypothetical) intervention that could set an individual's exposure regime to $\{5,5,5\}$ (even if the intervention is impractical, unethical, or impossible to carry out; Robins & Greenland, 2000). For our example, changing an individual's self-esteem can be accomplished by having participants partake in some form of therapy, or by giving them a compliment. If multiple versions of an intervention on self-esteem have different effects, then observed outcomes might not necessarily equal the potential outcomes, and it remains unclear how numerical estimates of "the effect" relate to the "the effect" as formulated in the research question (Hernán, 2016; Pearl, 2018).

7.1.5 Conclusion

Different disciplines investigate different kind of prospective causal effects using longitudinal observational data, with psychological researcher focused largely on cross-lagged effects, and biomedical researcher more focused on joint effects. However, there is no inherent reason why joint effects would not be interesting for psychology, and we are of the opinion that an exclusive focus on cross-lagged effects is unnecessarily limiting. Once researchers have decided which causal effect is interesting for their particular research project, and have evaluated the plausibility of the causal

[1] Researchers from other scientific disciplines might be more familiar with closely-related assumptions such unconfounded assignment, unconfoundedness, no unmeasured confounding, ignorability, (conditional) independence of treatment and potential outcomes, and exogeneity (cf. Angrist & Pischke, 2009; Hernán & Robins, 2020; Imbens & Rubin, 2015).

[2] We do not provide an introduction to these graphical rules here, but the interested reader is referred to Hernán and Robins (2020) and Pearl (2009).

identification assumptions, they can estimate the effect one of several approaches.

7.2 Estimation approaches

We focus on two estimation approaches: The use of cross-lagged panel models (CLPMs) within the framework of structural equation modeling, and the use of a SNMM with G-estimation. We discuss the statistical specification of CLPMs and SNMMs: Which dependencies of a causal DAG need to be correctly specified, and how are differences herein across approaches (dis)advantageous when the goal is to estimate the targeted causal effect? To help clarify some key characteristics of SNMMs with G-estimation, we also briefly discuss a repeated multiple regression approach for estimating joint effects.

7.2.1 CLPMs in the structural equation modeling framework

One of the most popular classes of SEMs in psychology for assessing prospective causal relations between variables is CLPMs (Usami et al., 2019; Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020). In this section, we outline some of the defining characteristics of this specific structural equation modeling approach for estimating causal effects, and discuss its advantages and disadvantages.

7.2.1.1 The basic idea

Cross-lagged panel models typically attempt to model the entire causal structure of the process under study. In a longitudinal context, this includes specifying a model for (a) the outcome, modeling how the outcome depends on previous exposure and covariates; (b) the time-varying exposure, modeling how the exposure depends on previous exposures and covariates; and (c) the time-varying covariates, modeling how the covariate depends on previous exposures and covariates. For our example, this modeling approach implies that the causal DAG in Figure 7.1 would be interpreted as a path diagram, with all individual dependencies (arrows) specified. In practice, covariances between the residuals at the same wave are usually added to the model to capture the direct effects of unobserved time-varying confounders whose effects are limited to a single time point, and who themselves show no dependencies over time. Such confounding variables are not assumed in the causal DAG of Figure 7.1, which implies that estimation of residual covariances would be redundant.

Once all causal dependencies are estimated, estimates of cross-lagged effects can be read off directly as the regression coefficients of the boldfaced paths in Figure 7.2a. Instead, CDEs can be obtained as combinations of the paths that make up a particular CDE (as visualized in boldface in Figures 7.2b to 7.2d). For example, using the path tracing rules by Wright (1934), the CDE of X_1 is a combination of the regression coefficients on the paths $X_1 \rightarrow L_2 \rightarrow L_3 \rightarrow Y_4$; $X_1 \rightarrow L_2 \rightarrow Y_3 \rightarrow Y_4$; $X_1 \rightarrow Y_2 \rightarrow L_3 \rightarrow Y_4$; and $X_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_4$. This CDE is thus the effect of X_1 on Y_4 that is mediated by all covariates in L and previous depression symptoms Y, and does not go through future X's. The same principle applies for the specification of the CDEs of X_2 and X_3. With estimates of the CDEs, we can predict individuals' outcomes under various exposure regimes. Comparisons of two predicted outcomes that follow from different regimes are then estimates of particular joint effects.

7.2.1.2 Advantages

One of the advantages of CLPMs is that they allow for the estimation of multiple causal effects of interest simultaneously. For example, a single CLPM can estimate all cross-lagged effects, as well as additionally specified joint effect parameters in a single model, allowing researchers to investigate multiple hypotheses at the same time. Furthermore, as CLPMs are commonly based on the specification of all dependencies in a causal DAG, the problem of exposure-confounder feedback is not applicable: If the assumed causal structure in the causal DAG is correct, and all parametric assumptions underlying the CLPM are true (i.e., all effects are linear, and the residuals are normally distributed), then the CLPM results in unbiased estimates.

Other advantages are related to some of the powerful statistical techniques that have been incorporated in the structural equation modeling framework. One major advantage is the ability to include latent variables in models. This is not only useful for measuring unobserved constructs using multiple indicators (Loeys et al., 2014), but also has advantages from a causal perspective. For example, it can be used to control for (unobserved) time-invariant confounders that have a time-invariant effect across time (Usami, 2021) or that have a time varying effect (if you free the factor loadings; Kenny & Zautra, 2001), as well as measurement error (Kenny & Zautra, 1995). Second, SEMs are relatively easy to use as many software packages have implemented structural equation modeling techniques—for example, the R packages lavaan (Rosseel, 2012) and OpenMx (Neale et al., 2016), Mplus (L. K. Muthén & Muthén, 2017), or Stata (StataCorp, 2023)—and syntax for specifying various SEMs is widely available. Many of these software packages have implemented multiple estimators (e.g., maximum likelihood, weighted least squares, Bayesian), as well as model fit indices that can be used to evaluate the fit of the specified SEM to the data. Third, many structural equation modeling software packages can handle various types of incomplete data through the use of full information maximum likelihood (FIML; Arbuckle, 1996). This is convenient as missing data are the norm rather than the exception in non-experimental longitudinal settings (van Buuren, 2018, p. 7). With FIML, all available data of individuals in the analysis are used assuming the missing data are missing completely at random (MCAR), or at random (MAR). This is a big advantage compared to listwise deletion which, especially for longitudinal data, can result in (unnecessary) loss of large portions of a dataset.

7.2.1.3 Disadvantages

The cross-lagged panel modeling approach also has several disadvantages from a causal inference point-of-view. First, it implies that parts of the causal DAG are modeled that are not necessary for identification and estimation of targeted causal effects. This is a risk, as parametric misspecication of any dependency in the SEM, such as wrongly assuming a causal effect to be linear, whereas, in fact, it is nonlinear, can lead to bias that propagates to other effects in the model as well (VanderWeele, 2012). Take the effect $X_3 \rightarrow Y_4$ for example, which is of interest as both a cross-lagged effect, and as the CDE of X_3. Obtaining an unbiased estimate requires, amongst other things, correctly adjusting for covariates that could confound this relationship (i.e., the conditional exchangeability assumption). Based on the causal DAG in Figure 7.1 and using the d-separation rules, it can be shown that adjustment for covariates L_3, Y_3, and C is enough to block all noncausal pathways between X_3 and Y_4: It does not require modeling how these covariates themselves depend on previous covariates. However, since structural equation modeling is concerned with modeling a data generating mechanism in its entirety, the causal structure of these covariates is typically modeled as well. This is often required to achieve desirable levels of model fit for *the SEM as a whole*; yet, it is redundant if the researcher is exclusively interested in obtaining unbiased estimates of *specific causal dependencies.* Similar arguments apply when estimating other cross-lagged effects or CDEs. Van der Laan and Rose (2011) point out that such unnecessary modeling only increases the potential for model misspecification, and ultimately results in bias for the estimates of the targeted causal effects (see also Naimi et al., 2016). This point has been made before in the context of cross-lagged panel models (Allison et al., 2017; Bollen, 1989), but does not appear to have been picked up in current structural equation modeling practices.

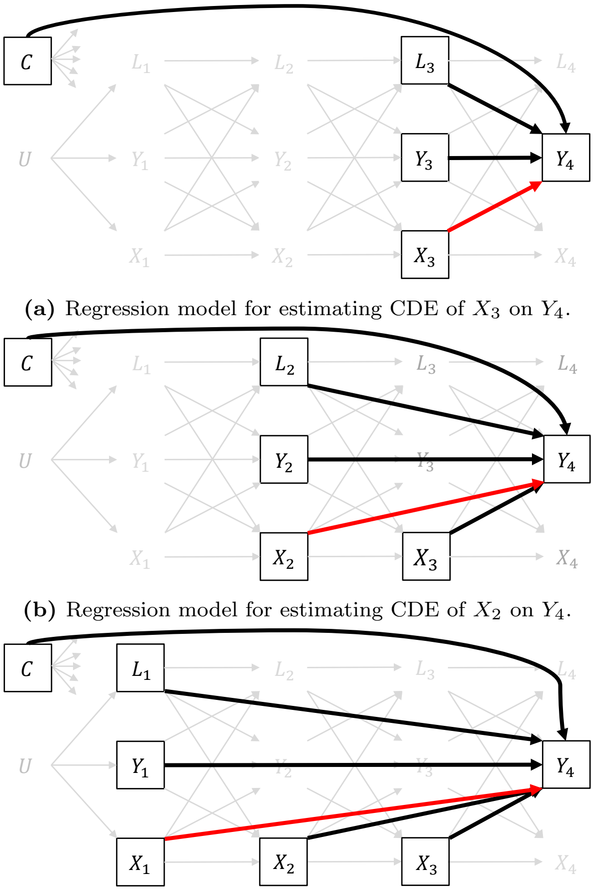
A second disadvantage of cross-lagged panel modeling approaches to causal inference is that the incorporation of multiple time-varying covariates in a SEM can quickly become unwieldy. This also applies if bidirectional lag-0, or lag-2 effects (or further) are to be included, or if quadratic terms are added to the model to specify nonlinear dependencies (B. O. Muthén & Asparouhov, 2022a). Such extensions (and many others) of basic linear SEMs can dramatically increase the number of parameters that need to be estimated, and can steeply increase the size of the covariance matrices that need to be modeled, thereby requiring increasingly large sample sizes to find a stable solution for the parameter estimates. For our example, if we were to interpret the causal DAG in Figure 7.1 as a path diagram, it would include (at least, excluding covariances, and residual covariances) 65 parameters, that is: 39 regression parameters, 1 variance, and 12 residual variances, 1 mean, and 12 intercepts. The inclusion of 1 additional time-varying covariate with a similar lag-1 causal structure adds 21 regression coefficients, 4 residual variances, and 4 intercepts to the model. As psychological mechanisms can involve a plethora of time-varying covariates that researchers (should) want to adjust for, attempts to model the entire causal system can quickly become practically prohibitive.

Third, including categorical variables as covariates in CLPMs is challenging, as the estimated regression coefficients are then on different scales, making it difficult to combine coefficients to compute CDEs. Suppose that the time-varying covariate L is categorical, for example use of antidepressants. This implies that regressions of L on other variables, for example, for the path $X_2 \rightarrow L_3$, concern logistic or probit regressions, resulting in logistic (e.g., odds ratios) and probit regression coefficients, respectively (B. O. Muthén et al., 2016). It becomes challenging to combine these coefficients with linear regression coefficients from other paths in the SEM, for instance $L_3 \rightarrow Y_4$, to compute the CDEs of interest. While these computations are possible for relatively simple situations with a single categorical time-varying covariate, this process becomes increasingly involved when the number of time-varying categorical covariates increases (B. O. Muthén et al., 2016; Nguyen et al., 2016).

7.2.2 Repeated multiple regression

In the causal inference literature, the causal inference approaches are usually presented in the context of exposure-confounder feedback (VanderWeele, 2021). In the presence of this problem, standard regression methods that attempt to simultaneously estimate all CDEs that make up a particular joint treatment effect—for example, by regressing the outcome on all exposures and covariates—are inadequate, leading to biased estimates of joint effects. However, it is possible to use standard regression methods in a "repeated" manner: Multiple standard regression models are then fitted, one for the estimation of each CDE separately. This makes it possible to work with distinct sets of covariates to adjust for confounding, thereby preventing the problem of exposure-confounder feedback. We explore this method as a first step towards the explanation of SNMMs with G-estimation.

Figure 7.3 illustrates the three regression models that must be specified to estimate the joint effect of X on end-of-study Y_4 (assuming the causal DAG in Figure 7.1). Again, the set of covariates to condition on in each model can be determined from the causal DAG in Figure 7.1 and using the d-separation rules by Pearl (1995). For example, to estimate the CDE of X_3, we need to adjust for L_3, Y_3, and C, as shown in Figure 7.3a. Under the causal identification assumptions (see Section 7.1.4) and the parametric assumptions of this regression model (i.e., the functional form of the



(a) Regression model for estimating CDE of X_3 on Y_4.

(b) Regression model for estimating CDE of X_2 on Y_4.

(c) Regression model for estimating CDE of X_1 on Y_4.

Figure 7.3: Overview of the regression models that need to be correctly specified for estimating the joint effect of X on Y using standard regression methods.

modeled dependencies is correct), the regression coefficient of X_3 obtained with this regression model is an unbiased estimate of the CDE of X_3 on Y_4.

To estimate the CDE of X_2, we fit a second regression model, this time adjusting for L_2, Y_2, C, and X_3, as shown in Figure 7.3b. Adjustment for X_3 is required to block the effect of X_2 on Y_4 that goes through the future exposure (by definition of a CDE, this is not allowed). In this regression model, we can block the effect of X_2 on Y_4 which goes through X_3 simply by including X_3 as an additional covariate in the model. Under the causal identification assumptions and the parametric assumptions of this regression model, the regression coefficient of X_2 is an unbiased estimate of the CDE of X_2 on Y_4.

Finally, the CDE of X_1 can be estimated by the regression model illustrated in Figure 7.3c. Here we include L_1, Y_1, and C as covariates to control for confounding.

Future exposures X_2 and X_3 are included in the regression model to block the effect of X_1 on Y_4 through X_2 and X_3. Similarly, under the causal identification assumptions and the parametric assumptions of this regression model, the regression coefficient of X_1 is an unbiased estimate of the CDE of X_1 on Y_4.

Compared to a CLPM, the regression models in this repeated procedure rely on the specification of fewer dependencies to get unbiased estimates of the targeted causal effects. Specifically, no model is specified for time-varying covariates L and Y (before the end-of-study), and the CDEs are specified directly rather than indirectly through the individual dependencies underlying them. For this reason, this approach has a lower risk of parametric model misspecification. This difference becomes even starker when lag-0 or lag-2 effects are added to the causal DAG, implying that CLPMs and repeated multiple regression models need to condition on a larger set of covariates to adjust for confounding. Furthermore, these regression models can also be fitted within the structural equation modeling framework. As such, researchers can combine advantages of structural equation modeling techniques (e.g., the use of FIML for missing data handling, the ability to impose constraints over time on parameters, control for measurement error), with the advantages of this sequential regression approach. A disadvantage is that this approach does not estimate the CDEs simultaneously, requiring researchers to fit multiple models themselves.

The use of SNMMs with G-estimation shows some resemblance with the repeated approach here, in that the CDEs are estimated separately as well (i.e., sequentially, each with a different set of covariates to adjust for), and that G-estimation of the CDEs of X_2 and X_1 requires adjustment for future exposures as well. However, with SNMMs, adjustment for future exposures is done differently; G-estimation methods are derived from a different principle than the repeated regression methods here; and G-estimation methods are doubly-robust, implying that estimates of causal effects converge to the true value (as sample size increases) even if part of the model is misspecified. This latter characteristic is hugely appealing from a causal inference point of view.

7.2.3 SNMMs using G-estimation

SNMMs with the associated method of G-estimation are described as a flexible and robust method for investigating joint effects in the presence of exposure-confounder feedback. What makes this approach challenging for psychological researchers to learn about is that (a) its use in the literature is described for diverse research problems, for instance for assessing both joint effects, for mediation analysis, or for survival analysis; (b) there exist multiple different G-estimation methods for fitting SNMMs to data; (c) these different methods each have different features that make them (dis)advantageous for specific research settings; and (d) there is little comprehensive software that has implemented all these methods. Therefore, our goal in this subsection is to provide the reader with a basic understanding of what a SNMM is, what the essence of G-estimation is, and what the (dis)advantages of this approach are compared to CLPMs. We focus specifically on the G-estimation method as described by Vansteelandt and Sjolander (2016). Like the repeated regression approach, this method is repeated in nature, and has the advantage that it can be implemented with standard regression methods, but also within the structural equation modeling framework (Loh & Ren, 2023b). The key additional advantage that it has over repeated multiple regression is that it is doubly-robust.

7.2.3.1 The basic idea

We have already seen that joint effects are a collection of CDEs (Daniel et al., 2013), and that CDEs can be represented as contrasts of end-of-study outcomes that follow from two different regimes. An SNMM is a model for these contrasts, where each CDE is equated to a causal parameter ψ_t. G-estimation is a sequential process that estimates the ψ_t's, starting with the last CDE, and then working backwards through time.

The joint effect can be represented as a comparison of the regimes $\{x_1, x_2, x_3\}$ with $\{x_1 + 1, x_2 + 1, x_3 + 1\}$. We start with the CDE of X_3 on Y_4, which can be parameterized as

$$\mathbb{E}(Y_4^{\{x_1,x_2,x_3+1\}} - Y_4^{\{x_1,x_2,x_3\}}|C = c, L_3 = l_3, Y_3 = y_3) = \psi_3. \tag{7.1}$$

The term on the left-hand side is the difference in the expected outcome of end-of-study Y_4 if all individuals followed the regime $\{x_1, x_2, x_3 + 1\}$ versus if all followed the regime $\{x_1, x_2, x_3\}$; hence, the only difference is in the exposure at the third time point. We condition on those covariates that are sufficient to block all noncausal paths between X_3 and Y_4, that is, $C = c$, $L_3 = l_2$, and $Y_3 = y_3$ according to the causal DAG in Figure 7.1. The causal effect is equated to the parameter ψ_3. For the purpose of this paper, we start with a basic SNMM here (e.g., no interaction term is included here implying an absence of moderation), although Equation 7.1 can be extended.

To estimate ψ_3, we make use of G-estimation, which is any estimation procedure that can be derived from the conditional exchangeability assumption (Vansteelandt & Joffe, 2014). As discussed in subsection 7.1.4, conditional exchangeability states that the potential outcomes are independent from observed exposure conditional on covariates. For didactical reasons, we assume linearity here such that we can write this independence assumption as

$$Cov(Y_4^{\{x_1,x_2,x_3^*\}},\ X_3 \mid C, L_3, Y_3) = 0, \tag{7.2}$$

where $Y^{\{x_1,x_2,x_3^*\}}$ represents the potential outcome for the treatment regimes with x_1 and x_2 set to their actual observed values, while x_3^* is set to a specific value, for instance zero, for all people.

This expression in Equation 7.2 may at first appear unrelated to our parameter of interest ψ_3, and also rather impractical, as the potential outcome term $Y^{\{x_1,x_2,x_3^*\}}$ is not actually observed (Naimi et al., 2016). However, through the SNMM, we can connect ψ_3 and Equation 7.2 (Vansteelandt & Joffe, 2014). To see this, suppose we want to compute the expected end-of-study depression score for each individual if their self-esteem score at the third wave had been set zero, that is, $Y^{\{x_1,x_2,0\}}$; however, we only have $Y^{\{x_1,x_2,x_3\}}$. But recall that ψ_3 is the difference in the (expected) potential outcomes, when there is a one unit difference in x_3 (when going from the observed x_3 to $x_3 + 1$). Hence, when going from the actual observed x_3 to $x_3^* = 0$, the (expected) change in the potential outcomes is $Y_4^{\{x_1,x_2,x_3\}} - Y_4^{\{x_1,x_2,0\}} = \psi_3 x_3$. Since, under consistency, $Y_4^{\{x_1,x_2,x_3\}} = Y_4$ (i.e., our observed end-of-study outcome), this implies we can write

$$Y_4^{\{x_1,x_2,0\}} = Y_4 - \psi_3 X_3. \tag{7.3}$$

Plugging Equation 7.3 into Equation 7.2 then leads to

$$Cov(Y_4 - \psi_3 X_3,\ X_3 \mid C, L_3, Y_3) = 0 \tag{7.4}$$

This shows the essence of G-estimation: Finding a value for ψ_3 such that Equation 7.4 holds.[3]

Multiple methods have been developed for finding ψ_3. For example, Hernán and Robins (2020) describe (for didactical reasons) a grid search, simply plugging in a range of values for ψ_3 until you find the value such that Equation 7.2 holds. However, we continue with the method by Vansteelandt and Sjolander (2016). It relies on fitting regression models for both the exposures and the outcome; a model for the covariates is not required. How this procedure can be derived from the conditional exchangeability assumption is shown in their appendix.

The method consists of three steps. First, a regression model for the exposure X_3 is specified, conditional on a set of covariates for blocking all noncausal paths, L_3, Y_3, and C. Figure 7.4a illustrates this exposure model, which Vansteelandt and

[3] Note that, for continuous measures, the potential outcome for an exposure score of zero, $Y^{\{x_1,x_2,0\}}$, might not be substantively meaningful on itself as zero may lay outside the measurement range. However, for dichotomous exposures (commonly used for applications of SNMMs) a zero-score can represent a "no treatment" condition.

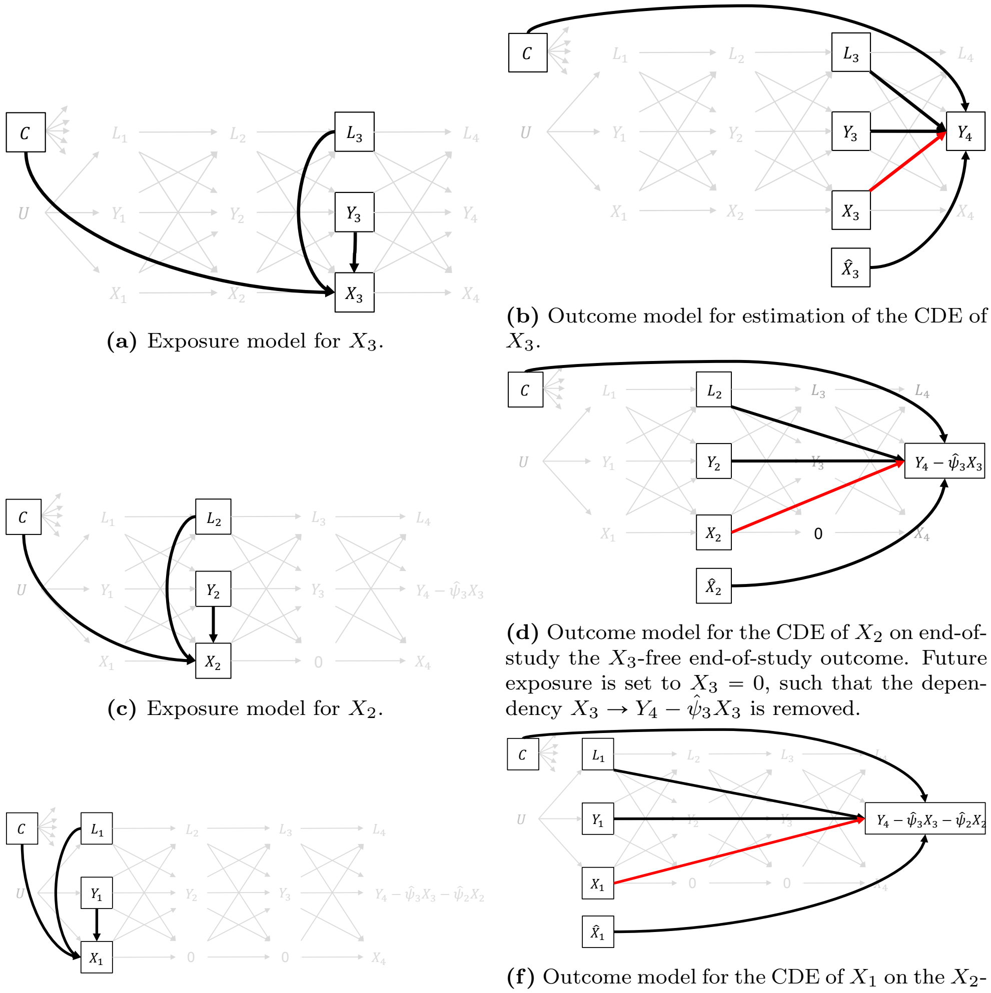
Sjolander (2016) also refer to generally as the propensity score (PS) model. Second, from the exposure model, predicted values for the exposure X_3 are calculated, which we denote by $\hat{X}_3$. These values would be referred to as "propensity scores" if the exposure was dichotomous, but work essentially the same for continuous exposures. The idea of this score is that it contains all information from variables that are needed to block noncausal paths (Imbens & Rubin, 2015). Third, a regression model for the outcome is specified conditional on the observed exposure X_3, the covariates L_3, Y_3, C, and $\hat{X}_3$. By conditioning on the PS $\hat{X}_3$ and the covariates L_3, Y_3, C, we attempt to block all noncausal pathways by conditioning on both the PS, and a set of covariates. If only the exposure model in step 1 is correctly specified, then this procedure is comparable to regression adjustment on the propensity score, and the additional covariates only increase precision (Vansteelandt & Daniel, 2014). If only the covariate-outcome relations in the outcome model are correctly specified, then we block all noncausal paths akin to the repeated multiple regression approach, and the additional PS covariate merely leads to an overfitted outcome model. The regression coefficient of X_3 is the G-estimate of ψ_3, that is $\hat{\psi}_3$: It is unbiased with both the exposure model and the outcome model correctly specified, and consistent when only one of both models is correct (i.e., this method is doubly-robust). Both the exposure model to obtain the PS, and outcome model to obtain an estimate of ψ_3 can be fitted using standard OLS regression, or using maximum likelihood within the structural equation modeling framework (Loh & Ren, 2023b).

This entire procedure works similarly for estimating the CDEs of X_2 and X_1. In the SNMM, the CDE of X_2 is parameterized as

$$\mathbb{E}(Y_4^{\{x_1,x_2+1,0\}} - Y_4^{\{x_1,x_2,0\}}|C = c, L_2 = l_2, Y_2 = y_2) = \psi_2, \tag{7.5}$$

The term on the left-hand side represents the difference in the expected outcome of end-of-study Y if all individuals were exposed to the regime $\{x_1, x_2 + 1, 0\}$ versus if all individuals were exposed to $\{x_1, x_2, 0\}$. Again, we condition on covariates to block all noncausal paths, $C = c$, $L_2 = l_2$, and $Y_2 = y_2$. To get the unique effect of X_2 (i.e., not going through future exposure), we additionally need to adjust for future exposure. Unlike the repeated multiple regression approach—in which we included future exposure as an additional covariate in the model—this method relies on computing a new outcome variable as if everyone had the same value on future exposures: When future exposures are a constant, they cannot have a causal effect on the outcome. In practice, future exposure value is commonly set to zero for all individuals such that the new outcome can be computed by

$$Y_{\text{blipped-down}} = Y_4 - \hat{\psi}_3 X_3. \tag{7.6}$$



(a) Exposure model for X_3.

(b) Outcome model for estimation of the CDE of X_3.

(c) Exposure model for X_2.

(d) Outcome model for the CDE of X_2 on end-of-study the X_3-free end-of-study outcome. Future exposure is set to $X_3 = 0$, such that the dependency $X_3 \rightarrow Y_4 - \hat{\psi}_3 X_3$ is removed.

(e) Exposure model for G-estimation of Equation 7.7.

(f) Outcome model for the CDE of X_1 on the X_2-and-X_3-free end-of-study outcome. Future exposures are set to $X_2 = 0$ and $X_3 = 0$.

Figure 7.4: Overview of the regression models that need to be correctly specified in the fitting procedure of an SNMM.

This equation is similar to Equation 7.3, except that we plug $\hat{\psi}_3$ into ψ_3. The new outcome is also referred to as the "blipped-down version of Y" or the "candidate counterfactual", and represents the outcome if unaffected by the exposure at occasion 3. To estimate ψ_2, we first estimate an exposure model again, regressing X_2 on the covariates L_2, Y_2, and C (see Figure 7.4c). Second, we compute the predicted values of exposure at time point 2, the PS score $\hat{X}_2$. Third, we fit a regression model for the blipped-down outcome conditional on the covariates, observed exposure X_2, and predicted exposure $\hat{X}_2$ (see Figure 7.4d, in which future exposure X_3 is set to 0). The regression coefficient of X_2 is then an estimate of ψ_2.

Finally, in the SNMM, the CDE of X_1 is parameterized as

$$\mathbb{E}(Y_4^{\{x_1+1,0,0\}} - Y_4^{\{x_1,0,0\}} | C = c, L_1 = l_1, Y_1 = y_1) = \psi_1. \tag{7.7}$$

The term on the left-hand side represents the difference in the expected outcome of end-of-study Y_4 if all individuals were exposed to the regime $\{x_1 + 1, 0, 0\}$ versus if all individuals were exposed to $\{x_1, 0, 0\}$. We control for a those covariates that block noncaual pathsways between X_1 and the outcome, C, L_1, Y_1, and additionally for future exposures by setting them to $X_2 = 0$ and $X_3 = 0$. As such, new blipped-down versions of Y are computed by

$$Y_{\text{blipped-down}} = Y_4 - \hat{\psi}_3 X_3 - \hat{\psi}_2 X_2. \tag{7.8}$$

To estimate ψ_1, first fit an PS model for X_1 given C, L_1, and Y_1 (see Figure 7.4e). Second, computed the predicted exposure $\hat{X}_1$. Third, fit a regression model for the outcome given the covariates C, L_1, Y_1, observed exposure X_1, and predicted exposure $\hat{X}_1$ (see Figure 7.4f, in which future exposures X_2 and X_3 are set to 0). The regression coefficient of X_1 is then an estimate of ψ_1. This completes the procedure for obtaining point estimates for the CDEs of self-esteem on end-of-study depression using a basic SNMM via G-estimation. To obtain confidence intervals for these estimates, Vansteelandt and Sjolander (2016) recommend the use of non-parametric bootstrapping.

7.2.3.2 Advantages

Like the repeated multiple regression approach, SNMMs with G-estimation have a lower risk of model misspecification compared to cross-lagged panel modeling approaches as no model needs to be specified for time-varying covariates L and Y (before the end-of-study), and the CDEs are obtained directly (Naimi et al., 2016). Furthermore, because this procedure is doubly-robust, the reliance on parametric assumptions being correctly specified is further reduced. Additionally, forgoing the need to model the covariates L means that researchers can more easily adjust for multiple time-varying covariates, and it provides them with increased flexibility for specifying functional forms of dependencies in the exposure model and regression model (compared to CLPMs).

The basic SNMM that we introduced here can be extended such that researchers can explore a wider range of research questions. For example, Tompsett et al. (2022) and Vansteelandt and Sjolander (2016) discuss extensions including interactions (to investigate effect modification) and time-varying outcomes. Finally, Loh and Ren (2023b) illustrate how the SNMM fitting procedure can be performed within the structural equation modeling framework. As such, researchers can combine advantages of structural equation modeling techniques (e.g., the use of FIML for missing data handling, the ability to impose constraints over time on parameters, control for measurement error etc.), with the advantages of SNMMs with G-estimation.

7.2.3.3 Disadvantages

Vansteelandt and Sjolander (2016) describe the G-estimation procedure using standard regression methods, and recommend the use of inverse probability weighting to account for missing data. A disadvantage of such a missing data handling approach is that it only supports right-censored missing data: Once an individual has a missing value at a particular time point, all future observed values of this individual are deleted as well. In practice, there are many possible missing data patterns, and this restriction is likely to result in the deletion of observed information. An alternative would be to use a multiple imputation procedure (van Buuren, 2018), or to fit the required models in the structural equation modeling framework and rely on FIML (Loh & Ren, 2023b). Furthermore, the implementation of G-estimation of SNMM in software, for example in R packages such as gestTools (Tompsett et al., 2022) and DTRreg (Wallace et al., 2017), is currently still inflexible. Specifically, these packages are often tailored to situations which include lag-0 effects, and it can be challenging to adjust the data file, and the input for required arguments in such a way that these R packages work for situations without contemporaneous effects (as for our empirical example). Alternatively, researchers can code the G-estimation procedure themselves, which requires a basic knowledge of coding.

7.2.4 Conclusion

Cross-lagged panel modeling and structural nested mean modeling can both be used to investigate research questions of joint effects. We highlight some key differences. First, cross-lagged panel modeling approaches attempt to model the entire data generating mechanism, which requires the correct specification of the functional form (i.e., potential nonlinearities and/or interactions between variables) of all dependencies of the exposure, outcome, and time-varying covariates. The specification of a large number of dependencies increases the risk of model misspecification and consequently invalid inferences. SNMMs with the G-estimation, however, do not require postulating a model for time-varying covariates. Furthermore, the structural nested mean modeling procedure by Vansteelandt and Sjolander (2016) is doubly-robust: It requires the specification of a model for the exposures and the outcome, and still results in consistent estimates of CDE when either model is misspecified, thereby further reducing reliance on parametric assumptions. Second, by forgoing the specification of a covariate model, researchers have increased flexibility for including a large set of covariates compared to cross-lagged panel approaches. This is advantageous as the addition of multiple covariates is usually warranted to make the causal assumption of conditional exchangeability plausible. Third, estimates of joint effects using a cross-lagged panel modeling approach are obtained as linear combinations of path-specific coefficients. However, in the presence of both binary and continuous covariates, some path-coefficients represent linear regression coefficients, whereas others are interpreted as logit or probit regression coefficients. It can be challenging to combine these paths to obtain the CDEs of interest. Instead, in SNMMS, the CDEs are obtained directly.

There are additional differences between cross-lagged panel modeling and structural nested mean modeling that are not inherent to the models themselves, but rather concern their application in practice. For example, missing data handling in the G-estimation approach is predominantly done via inverse probability weighting (Hernán & Robins, 2020; Vansteelandt & Sjolander, 2016). This can be disadvantageous as it requires missing data to have a right-censoring structure. In practice, many other missing data patterns may occur, resulting in unnecessary loss of data when a right-censoring structure is enforced. However, when SNMMs with G-estimation is used within the structural equation modeling framework (e.g., Loeys et al., 2014; Loh & Ren, 2023b), then researchers can take advantage of structural equation modeling techniques for missing data handling, as well as latent variables. Such applications of SNMMs within the structural equation modeling framework seem promising, but are rare in practice.

Finally, joint effects can be investigated alternatively by using repeated multiple regression methods. The major difference between the standard regression approach and the structural nested mean modeling approach by Vansteelandt and Sjolander (2016), is that repeated multiple regression is not doubly-robust. Causal inference advocates would argue that this is shortcoming of this method as it is therefore more reliant on correct model specification than structural nested mean modeling approaches.

7.3 Empirical example: Joint effect of self-esteem on depression

To illustrate the approaches for assessing joint effects in a psychological context, we reanalyze self-esteem and depression data from Kuster et al. (2012). Using an online survey, five repeated measures (at two-month intervals) of self-esteem, rumination, and depression symptoms were collected in a German-speaking convenience sample of $N = 663$ individuals largely residing in Switzerland (96%). The original study fitted a bivariate cross-lagged panel model (CLPM) to the data to study cross-lagged effects between self-esteem and depression, and a trivariate CLPM to assess whether the relationship between self-esteem and depression was mediated by rumination. Instead, we will attempt to estimate joint effects of self-esteem on depression, with depression at time point 5 as our end-of-study outcome of interest Y_5. For pedagogical purposes, we restrict ourselves to the inclusion of rumination as the sole time-varying covariate although, arguably, many more baseline and time-varying covariates should be included to make the causal assumption of exchangeability plausible.

Self-esteem was measured with the ten-item Rosenberg Self-Esteem Scale, with responses measured on a five-point scale ranging from one (strongly disagree) to five (strongly agree; Rosenberg, 1965; Von Collani & Herzberg, 2003). Depression symptoms were assessed using the twenty-item Center for Epidemiologic Studies Depression Scale in which participants were asked to assess how frequently they had experienced each symptom within the preceding thirty days. Participants responses were measured on a four-point scale from zero (rarely or none of the time) to three (most or all of the time; Hautzinger & Bailer, 1993; Radloff, 1977). Rumination was measured using the eight-item rumination subscale of the Rumination-Reflection Questionnaire, with responses measured on five-point scales ranging from one (strongly disagree) to five (strongly agree; Trapnell & Campbell, 1999). Self-esteem and depression measures were made publicly available by Orth et al. (2021). Rumination measures were made available upon request by Kuster et al. (2012).

7.3.1 Statistical analyses

For all analyses, it is assumed that the causal DAG in Figure 7.1 corresponds to the causal process by which the data in the sample were generated. To prevent the results from being influenced by differences in how missing data are handled, a complete data set was created first by single imputation using the R package mice (van Buuren, 2018).

For the cross-lagged panel modeling approach, we use the the causal DAG as the basis for a path diagram, and extended it with covariances amongst the variables at the first wave, and residuals at waves 2 and later. This is essentially tantamount to the trivariate cross-lagged panel model (CLPM) as in the original study. It was fitted to the complete data using the R package lavaan (Rosseel, 2012). The CDEs of self-esteem at time points 1, 2, 3, and 4 on Y_5 were specified as linear combinations of paths in the model, and computed as additional parameters (i.e., quantities) in the model.

For the SNMM with G-estimation, the procedure by Vansteelandt and Sjolander (2016) was followed. For completeness, we also fitted repeated multiple linear regression models to estimate joint effects. For all analysis approaches, 95% confidence intervals were created based on the nonparametric bootstrap with 999 bootstrap samples using the R package boot (version 1.3-28; Canty & Ripley, 2022). All analyses were performed in base R (version 4.2.2; R Core Team, 2022). Annotated code can be found in the online supplementary materials.

7.3.2 Results

The column "CLPM" of Table 7.1 contains the CLPM estimates and 95% confidence intervals of the joint effect of self-esteem on end-of-study depression. Overall, fit indices indicated bad model fit, $\chi^2(54) = 755.962$, $p < .001$, CFI = .926, TLI = .857, RMSEA = .140, SRMR = 0.078 (Browne & Cudeck, 1992; Hu & Bentler, 1999; Little, 2013).[4] All estimated CDEs are negative, with the CDEs at time points 1, 2, and 4 denoting significance at the $\alpha = 0.05$ level. For example, the CDE at time point 1 implies that an increase in self-esteem reduces depression 8 months later even if self-esteem at time points 2, 3, and 4 are held constant.

Results for the SNMM are presented in the column "SNMM" of Table 7.1. The estimates of the CDEs at time points 2, 3, and 4 are similar to those of the CLPM in terms of sign and significance. However, the CDE at time point 1 is not significant in the SNMM (whereas it is in the CLPM).

Results for the repeated multiple linear regressions are presented in the column "Rep. regr." of Table 7.1. In contrast to results from the CLPM and the SNMM, the estimate of the CDE at time point 1 is positive and significant, implying that an increase in self-esteem is expected to lead to an increase in depression symptoms eight months later. The CDE at time point 2, however, is nonsignificant. Akin to the CLPM and the SNMM, the CDE at time point 3 is nonsignificant as well. The CDE

[4] In the SEM literature, this is commonly interpreted as a sign of model misspecification, warranting changes to the model (e.g., the inclusion of lag-2 effects or a random intercept factor). However, in the causal inference literature, some researchers argued that the importance of model fit for causal inference is greatly reduced for multiple reasons: (1) Lüdtke and Robitzsch (2022) and Orth et al. (2021) argue that model fit is uninformative about the appropriateness of a SEM in relation to a research question; (2) Tomarken and Waller (2005) argues that model fit is uninformative about the plausibility of the conditional exchangeability assumptions as encoded in an assumed causal DAG.

CDE	CLPM	SNMM	Rep. regr.
$SE_1 \rightarrow DE_5$	-0.027* [-0.042, -0.014]	0.004 [-0.037, 0.050]	0.195* [0.119, 0.263]
$SE_2 \rightarrow DE_5$	-0.092* [-0.124, -0.063]	-0.072* [-0.128, -0.021]	0.015 [-0.063, 0.094]
$SE_3 \rightarrow DE_5$	-0.014 [-0.043, 0.015]	0.017 [-0.035, 0.066]	0.093 [-0.002, 0.177]
$SE_4 \rightarrow DE_5$	-0.129* [-0.185, -0.073]	-0.129* [-0.189, -0.077]	-0.129* [-0.183, -0.077]

Table 7.1: Point estimates and 95% bootstrap confidence intervals (in square brackets) of the controlled direct effects of self-esteem on end-of-study depression, estimated using cross lagged panel modeling ("CLPM"), structural nested mean modeling ("SNMM"), and repeated multiple linear regression ("Rep. regr."). Analyses are based on the causal DAG of Figure 7.1. Asterisks (*) denote significance at the $\alpha = .05$ level.

at time point 4 is equivalent to that of the CLPM, and similar to that of the SNMM in terms of sign and significance.

In general, the results from the CLPM approach and SNMM approach are similar for this particular example. Differences in the effect estimates across approaches (and their significance) can be due to numerous factors. First, these approaches rely on different (sets of) parametric assumptions. For example, violations of the linearity assumption of the dependencies of the variable rumination do not impact the validity of the SNMM (and the repeated multiple linear regression) estimates, whereas they are expected to bias estimates in the CLPM. Moreover, the SNMM here is doubly-robust, implying that potential misspecification in the outcome models still results in consistent estimates from the SNMM if the exposure model is correct (and vice versa). It is unknown which exact parametric assumptions are incorrect, and to what degree, but given the complex nature of the phenomenon under study, some degree of violation is expected. Second, it is likely that there are numerous confounding covariates, both time-varying and time-invariant, that have not been taken into account here, violating the causal conditional exchangeability assumption (this is further elaborated upon in the Discussion). Such violations might impact both modeling approaches differently; future studies are needed to gain more insight in this, under various settings of violation.

7.4 Discussion

Cross-lagged panel modeling is widely used by psychological researchers as a structural equation modeling approach for assessing lag-1 relationships between two variables over time. While some (e.g., Bollen & Pearl, 2013) argue that SEM is a good framework for causal inference, there is critique in the causal inference literature that this popular modeling practice is not a viable option if the goal is to investigate causal relationships. One of the main points of concern is that attempts to model a causal process in its entirety has a high potential of model misspecification, and is unnecessary if interest is limited to a set of well-defined causal effects. This problem is only exacerbated with the inclusion of multiple time-invariant and time-varying covariates, which researchers would want to do to make the causal identification assumption of conditional exchangeability plausible in nonexperimental data.

In this article, we explored this concern using an empirical psychological example. Taking inspiration from disciplines such as epidemiology and biostatistics, we introduced joint effects as an alternative causal effect that can be interesting for psychologists to target. While these effects can be specified akin to a cross-lagged panel modeling approach within a structural equation modeling framework, they are traditionally estimated with SNMMs using G-estimation. This is an appealing method as it does not require the specification of a model for covariates, and is flexible in accommodating a large set of (time-varying) covariates, and lag-0 and lag-2 (or further) effects. Furthermore, the implementation of G-estimation by Vansteelandt and Sjolander (2016) is robust to misspecification in either the exposure model or the outcome model, further reducing this method's reliance on parametric assumptions. These properties provide a motivation for psychological researchers to seriously consider the use of SNMM with G-estimation to investigate causal relationships between variables in panel data.

To further support integration of formal causal inference methods with literature on psychological research methods, we discuss some overlap between these strands of literature next. We also consider some limitations of empirical example, and extend our analyses of the empirical data.

7.4.1 Controlling for stable, between-person differences

A much-discussed idea in psychology, and the social sciences more generally, is the separation of longitudinal data into stable, between-person differences, and temporal, within-person fluctuations (Asparouhov & Muthén, 2019; Hamaker et al., 2015; Kreft et al., 1995). The idea has been discussed extensively in the context of cross-lagged effects, but equally applies to the investigation of joint effects. The appeal is that a decomposition of observed variance allows researchers to better align effect estimates from statistical analyses with their research questions about (causal) effects at the within-person level (Raudenbush & Bryk, 2022). This line of thinking has inspired many researchers in the social sciences, and led to the development of many (cross-lagged) panel models in the structural equation modeling framework (Usami et al., 2019). One particularly popular model is the random intercept cross-lagged panel model (Hamaker et al., 2015): By including a random intercept factor to separate the two sources of variance, the lagged effects can be interpreted as pertaining to effects at within-person level. Usami (2021) describes how the inclusion of the random intercept factor has the additional advantage of controlling for unobserved heterogeneity.

While this idea has sparked much excitement (and debate) in the psychological literature, it has passed the epidemiological and biostatistics literature relatively unnoticed. Only recently, Usami (2022) introduced a method for combining the random intercept cross-lagged panel model with structural nested mean modeling approaches for estimating CDEs. This development combines strengths of analysis approaches from different strands of literature. Work on making these developments broadly applicable for applied researchers is ongoing (Usami, 2023).

7.4.2 Lag-0 effects

The causal DAG of Figure 7.1 does not include direct effects of variables on other variables *at the same time point* (contemporaneous effects). While the vast majority of the SEM literature on cross-lagged panel modeling makes this (implicit) assumption (merely controlling for relationships between contemporaneous variables through the inclusion of a residual covariance), causal DAGs in epidemiological and biomedical literature do commonly include contemporaneous effects. In (bio)medical settings, the decision to give an individual a treatment X at a particular time point often depends on a range of previous covariates, as well as current values of covariates (e.g., blood results). The addition of contemporaneous effects in a causal DAG reflects this, and it has consequences for the interpretation of causal effects in the DAG. Subsequent statistical analyses, whether it is a cross-lagged panel modeling approach in the structural equation modeling framework, or an SNMM approach, then also need to take this contemporaneous effects into account. Ignoring these effects (i.e., wrongly assuming the DAG in Figure 7.1 is correct) can create a mismatch between the targeted causal effect, and the estimated effect.

Recently, this issue has been brought up in the SEM literature by B. O. Muthén and Asparouhov (2022a). They state that the addition of contemporaneous effects to cross-lagged panel models (replacing residual covariances at the same wave) may be warranted based on the timing of measurements in datasets, especially when there are long time intervals between subsequent measurements. Based on a reanalysis of five empirical datasets using cross-lagged panels both with and without lag-0 effects, they also argue that there might not be enough information in the data to make an informed decision about whether or not the contemporaneous effect can be safely ignored. As the omission of such effects from the causal DAG is a stronger assumption than their inclusion (i.e., it amounts to constraining these paths to zero; Bollen & Pearl, 2013), it is therefore advisable to always include these effects whenever there is doubt about whether or not they exist, and to clarify if these causal paths are of substantive interest. B. O. Muthén and Asparouhov (2022a) recommend reporting results from models both with and without lag-0 effects.

7.4.3 Lag-2 effects

Different rationales for inclusion of lag-2 effects in statistical models have been provided in the SEM literature and the causal inference literature. In cross-lagged panel modeling, the addition of lag-2 autoregressive effects is sometimes discussed in the context of achieving adequate model fit (Hamaker et al., 2015; B. O. Muthén & Asparouhov, 2022b). This is related to the discussion on controlling for stable, between-person differences, with lag-2 autoregressive effects interpreted as the stabilizing influences underlying trait-like differences between individuals (Asendorpf, 2021). In the causal inference literature, however, lag-2 (and further) autoregressive and cross-lagged effects are usually considered for confounding control. Whenever exposures or covariates have effects that span multiple lags, it is possible that confounding cannot be adjusted for by merely controlling for immediately prior variables in statistical analyses. This is the case when, for example, in the causal DAG of Figure 7.1 L_1 directly effects X_3 (a lag-2 cross-lagged effect) and Y_4 (a lag-3 cross-lagged effect). Then, to unbiasedly estimate the CDE of X_3 on end-of-study Y_4, additional lagged covariates need to be included as controls in the analyses. So while the control of immediately prior (i.e., lag-1) exposures and covariates is usually important for control of confounding of CDEs, it might be advisable to also consider lag-2 (and longer) effects in causal DAGs, and adjust the statistical analyses based on this (VanderWeele, 2021). Others, such as Daniel et al. (2013) and Vansteelandt and Sjolander (2016), advise to condition on the entire exposure and covariate history in analyses.

7.4.4 Limitations of the empirical example

For this article, we have used an empirical example that is close to the cross-lagged panel modeling practices that many psychological researchers are familiar with. However, from a causal inference point-of-view, there are some serious concerns. First, the causal assumption of (conditional) exchangeability is compromised, as we have not included any time-invariant covariates that have been found to confound the relationship between self-esteem and depression, such as gender, social economic status, or personality traits like neuroticism (Mu et al., 2019). There are also likely to be a numerous time-varying covariates, such as substance use, relationship status, relationship satisfaction, job success, and academic performance, that have not been included in the analyses (Boden et al., 2008). Second, we argue that the causal assumption of consistency is compromised as well. There are numerous options for a(n) (hypothetical) intervention on self-esteem, each of which might have a different effect on the outcome. Information on how self-esteem was increased was also not present in the empirical data. As such, our research question is ill-defined making it difficult to link our theoretical interest to the observed data (Hernán, 2016).

CDE	CLPM (lag-1,2)	SNMM (lag-1,2,3)
$SE_1 \rightarrow DE_5$	-0.070* [-0.113, -0.030]	0.037 [-0.044, 0.117]
$SE_2 \rightarrow DE_5$	-0.084* [-0.137, -0.034]	-0.154* [-0.243, -0.061]
$SE_3 \rightarrow DE_5$	0.020 [-0.066, 0.107]	0.016 [-0.086, 0.118]
$SE_4 \rightarrow DE_5$	-0.110* [-0.193, -0.032]	-0.096* [-0.182, -0.017]

Table 7.2: Point estimates and 95% bootstrap confidence intervals (in square brackets) of the controlled direct effects of self-esteem on end-of-study depression, estimated using a CLPM and structural nested mean models. Compared to the analyses in Table 7.1, the models are extended with lag-2 (and lag-3) effects. Asterisks (*) denote significance at the $\alpha = .05$ level.

One aspect that can be improved using the available data is the conditional independence assumptions that are represented in the causal DAG in Figure 7.1, and that serve as the basis for our statistical analyses. The omission of lag-2 (and longer) effects are conditional independence assumptions that are regularly made in cross-lagged panel modeling, but that can negatively affect the validity of estimates when violated (VanderWeele, 2012). To prevent making these assumptions at all, we extend the CLPM with lag-2 effects, and SNMM with lag-2 and lag-3 effects. The results are presented in Table 7.2.

The inclusion of lag-2 effects significantly improved model fit compared to the CLPM with lag-1 effects, $\Delta\chi^2(27) = 471.598$ with $p < .001$, although overall model fit remains subpar, $\chi^2(27) = 284.364$, $p < .001$, CFI = .973, TLI = .895, RMSEA = .120, SRMR = 0.035. Numerical results from the CLPM changes somewhat with the inclusion of lag-2 effects, most significantly the CDE of self-esteem point 1 (the effect of which is now stronger): The conclusions drawn would be the same. Results from the SNMM also changed somewhat numerically, but not substantively. The choice of which model's results to report might not be obvious in practice. The consistency of these results across methods, gives some degree of confidence that the results are not unduly reliant on parametric assumptions specific to any one analysis approach. If the results were to differ significantly across methods, then one could argue that the SNMM with lag-1, lag-2, and lag-3 effects makes fewest causal and parametric assumptions, and hence produced most reliable results. At the same time, violations of the causal identification assumptions imply that the results of this empirical example from any method should not be interpreted causally.

7.5 Conclusion

We discussed joint effects as an alternative causal effect to cross-lagged effects, and discussed the use of SNMM with G-estimation as an alternative modeling approach. We hope that this introduction and the empirical example allows psychological researchers to make better informed decisions about which kind of causal effect is interesting to target, while also managing the number of parametric assumptions that one needs to make during the statistical analyses. While explicit causal reasoning is not unique to causal inference methods in the epidemiological and biostatistical literature, the statistical (parametric) advantages of an SNMM approach should be a motivation for psychological researchers to gain experience with this modeling approach. This article aids in developing an intuition for some of the concepts that this modeling approach builds on. We recommend the recent work of Loh and Ren (2023b) and the work of Loeys et al. (2014) as introductions to the G-estimation procedure itself. The works of Daniel et al. (2013), Hernán and Robins (2020), and Naimi et al. (2016) are useful as more detailed introductions to other causal inference methods from an biomedical perspective.

Acknowledgments: We would like to thank prof. dr. Ulrich Orth for making the “rumination” measurements of Kuster et al. (2012) available to us.

Author contributions: Conceptualization: JM - Data curation: JM - Formal analysis: JM – Investigation: JM, SU, and EH - Methodology: JM, SU, and EH – Project administration: JM - Software: JM - Supervision: SU and EH - Visualization: JM - Writing (original draft preparation): JM – Writing (review and editing): JM, SU, and EH

Funding: Jeroen D. Mulder and Ellen L. Hamaker are part of the Consortium Individual Development, which was supported by a Gravitation grant awarded by the Dutch Ministry of Education, Culture, & Science, and the Netherlands Organization for Scientific Research [NWO grant number 024.001.003; 2013].

Online supplementary materials: This study’s online supplementary materials can be found at https://jeroendmulder.github.io/joint-effects-using-SNMM.

CHAPTER 8

Discussion

For some time now, the use of structural equation modeling (SEM) has been fully established as a statistical modeling framework in psychological research. A common area of application is for the analysis of longitudinal observational data (i.e., panel data), in which relations between variables over time are investigated. In this dissertation I studied and applied popular classes of longitudinal SEM-models for descriptive, predictive, and causal research questions. In Chapter 2, a relatively straight-forward multivariate regression model was specified in the SEM framework to investigate the associations between individual development of social emotion regulation, and later social well-being. In Chapter 3 I used k-fold cross-validation to compare the out-of-sample prediction performance of three latent growth curve models (LGCMs) for predicting patients' reduction of posttraumatic stress disorder symptoms four weeks after completion of a clinical treatment program. Chapters 4 and 5 studied and extended the random intercept cross-lagged panel model, which is a popular method amongst psychologists for investigating relations between variables over time. In Chapters 6 and 7 cross-lagged panel modeling approaches were compared to analysis methods from the potential outcomes framework, in particular the use of inverse probability weighting (IPW) estimation of marginal structural models (MSMs), and structural nested mean models (SNMMs) with the associated method of G-estimation.

The popularity and accessibility of SEM in the social sciences is also associated with an increasing concern about misconceptions of SEM, and ignorance about its constraints and limitations (Tomarken & Waller, 2005). Chapters 6 and 7 discussed some concerns in relation to cross-lagged panel modeling approaches and causal inference, specifically the critiques of Van der Laan and Rose (2011) and VanderWeele (2012) who point out that SEM models depend heavily on conditional independence and parametric assumptions; since these are likely to be violated—at least to some degree—in practice, they state that SEM models are prone to bias when one attempts to infer causal effects from observational data. As argued in Chapter 6, these claims should be of concern whenever SEM models are used for applied causal research. In practice, however, empirical researchers appear unbothered by these concerns, and many estimated relationships between variables in SEM models are frequently interpreted causally, albeit implicitly (Grosz et al., 2020; Hamaker et al., 2020). Furthermore, adaptation of alternative causal inference methods that have been developed in disciplines like epidemiology and biostatistics (and within the potential outcomes framework) is still lacking in the psychological literature, despite the fact that these methods rely on fewer parametric assumptions, offer a principled approach for reasoning explicitly about causality, and therefore, in principle, should lead to more robust causal conclusions.

In this chapter, I continue the discussion in Chapters 6 and 7 on the use of SEM for causal inference in psychology. Specifically, I consider three lessons that I have drawn about how the SEM approach and potential outcomes approach to causal inference can complement each other. The focus in this discussion on Chapters 6 and 7 is because these topics have been of primary interest to me in the latter half of my PhD project, and will play a central role in my research as a postdoc. This does not imply that I value the work done in other chapters of this dissertation less. Instead, I consider collaborations with applied researchers as vital means for explaining novel and complex statistical methods to the end-users of the methods that we, as statisticians, develop. Similarly, didactical treatments of existing methods, such as the random intercept cross-lagged panel model discussed in Chapters 4 and 5, are important as well to inform applied researchers about how these methods are best applied, as well as their limitations.

8.1 Lesson 1: Increase focus on Phases 1 and 2 of causal research in psychology

Chapter 6 describes three phases of causal research, namely (1) the *formulation* of a causal research question using potential outcomes, that is, formulate a causal estimand; (2) the *identification* of the causal estimand, translating it into a statistical estimand; and (3) *estimation* of the statistical estimand from a finite random sample using a statistical model (inspired by Goetghebeur et al., 2020; Petersen & Van der Laan, 2014). While SEM models are compatible with the potential outcomes-framework (De Stavola et al., 2015; Moerkerke et al., 2015; B. O. Muthén et al., 2016), typical applications of SEM for causal inference ignore Phases 1 and 2, and are mainly concerned with estimation of statistical models (Kunicki et al., 2023). This is problematic, as a research question needs to be well-defined in Phase 1 to be able to derive something from it that is statistically testable. Furthermore, identification of a causal estimand in Phase 2 is needed to link the theoretical constructs in a research question to observable data. The steps taken in both phases are fundamental for estimation in Phase 3, and for interpretation of the estimates. If the research question is ill-defined (i.e., ambiguous), and not carefully identified, then it remains unclear how the numerical estimates that are obtained in Phase 3 relate to the research question, and how valid these estimates actually are.

However, implementing a change to increased focus on Phases 1 and 2 in the causal research practices of psychological researchers is no mean feat. It requires effort on behalf of empirical researchers themselves—who are well-advised to get acquainted with the potential outcomes approach for causal inference—and of methodologists who would need to write accessible papers that outline how potential outcomes approaches work in a psychological context. I hope that Chapters 6 and 7 are useful first steps in this regard. Moreover, I believe that it is important that methodologists and statisticians also become actively involved in empirical research to aid in implementing the potential outcomes approach in psychological research, and performing the, sometimes rather complex, potential outcomes methods. Doing so also helps to set good examples for other empirical researchers, and can inform methodologists about practical (or conceptual) issues of potential outcomes methods that empirical researchers run into. Finally, increased awareness of the importance of Phases 1 and 2 for causal research also requires adjustments to our statistical education. Currently many SEM-related courses, and statistics courses in general, appear to focus predominantly on estimation methods, with relatively little attention devoted to how one carefully formulates a well-defined causal estimand, and how to identify it. Inclusion of these topics in methodological and statistical education, along with practical, and didactical examples about how Phases 1 and 2 work in psychological settings, might greatly contribute to increase focus on Phases 1 and 2 in causal psychological research.

8.2 Lesson 2: Avoid unnecessary modeling

Once a research question has been carefully formulated in Phase 1 of causal research (resulting in a causal estimand), researchers should evaluate if the causal estimand can be estimated from observable data (i.e., Phase 2). A commonly used instrument for this is the use of directed acyclical graphs (DAGs, a visualization of the data generating mechanism) in combination with the d-separation rules, to determine which set of covariates needs to be adjusted for to obtain unbiased estimates of specific effects (i.e., which covariates need to be adjusted for to close backdoor-paths; Pearl, 2009; Petersen & Van der Laan, 2014). Typically, from these DAGs it can be derived that a causal estimand can be identified without specifying a statistical model for time-varying covariates in the adjustment set. This point is described in more detail in Chapter 6. However, SEM approaches, and in particular cross-lagged panel modeling approaches, typically model the entire data generating mechanism, including specifying a statistical model for the time-varying covariates. Such SEM models thus model more than strictly necessary for identification of the causal estimand. This increases the risk of violating parametric assumptions that the model makes, and while such violations do not always lead to significant bias (as demonstrated for specific scenario's in Chapter 6), it might be generally advisable to use methods that relax the parametric assumptions as much as possible (VanderWeele & Vansteelandt, 2010). Especially when including many covariates, the number of parametric assumptions, and the risk of bias associated with it, can increase dramatically, and unnecessarily so from a purely causal perspective (VanderWeele, 2012).

Here I also want to critically reflect on my own work in Chapter 4. This chapter addresses some of the statistical modeling questions that applied researchers have voiced about the random intercept cross-lagged panel model (RI-CLPM), and presents three extensions of this model. The RI-CLPM has rapidly increased in popularity among psychological researchers in recent years for studying causal processes between variables over time, and in the online supplementary materials of this chapter I describe how the RI-CLPM can be extended with a third time-varying variable. Researchers might be interested in this extension as a way to include an additional covariate for confounding-control. Specifically, I describe how a third time-varying variable L can be modeled similarly to the two time-varying variables X and Y: That is, decomposing L into a between-person component and within-person components, and including lagged relationships between the within-person components. Looking back, I find this advice incomplete from a causal inference point-of-view, as it forgoes any discussion about the necessity of specifying specific lagged effects for the within-components of L. In fact, using the d-separation rules it can be shown that modeling the dependencies of L are unnecessary if the interest is solely in the cross-lagged effects between X and Y.

8.3 Lesson 3: The SEM framework can play an important role in causal research in psychology

The advantages of the potential outcomes framework for causal inference in relation to SEM approaches are clearly outlined in Chapters 6 and 7, and related papers (De Stavola et al., 2015; Loeys et al., 2014; Loh & Ren, 2023b; Moerkerke et al., 2015; B. O. Muthén et al., 2016). The potential outcomes framework provides researchers with a principled approach for formulating an explicit causal question, and identifying the causal estimand. Furthermore, the causal inference methods that have been derived from the potential outcomes framework rely on fewer parametric assumptions compared to popular SEM models. Therefore, I have occasionally asked myself the question if there is still a role for SEM in causal inference?

My answer is “yes”, but I think it is important to clearly distinguish between the general SEM framework, and specific (classes of) SEM models. The work done in Chapters 6 and 7 has made me more critical of the use of cross-lagged panel models as a *class of SEM models* in causal research: Methods from the potential outcomes framework such as IPW esimation of MSMs and SNMMs with G-estimation simply rely less on conditional independence and parametric assumptions than cross-lagged panel models. However, there are some clear reasons why such methods from the potential outcomes framework cannot be readily adapted in psychological research. One reason is that many of the psychological constructs of interest are latent in nature, for example attitudes, emotions, and psychopathology. To study these constructs in quantitative research, measurements models are required to capture the latent constructs statistically, and the general SEM framework has been incredibly successful in providing researchers with the tools to do so. In epidemiology—the discipline that much of the methodological research into potential outcomes methods takes place in—latent variables are encountered only rarely. In such biomedical disciplines, the exposures and outcomes are more commonly directly observable and well-defined, pertaining to, for example, some treatment that individuals did (or did not) receive, cell counts, and/or the occurrence of some event (e.g., death, diagnosis), etc. I believe that this mismatch between the (latent) practice of psychological researchers and the potential outcomes framework is one of the reasons why psychological researchers have not widely taken up potential outcomes approaches yet.

Luckily, the potential outcomes approach and SEM are not polar opposites that detract. In contrast many of the causal inference methods derived from the potential outcomes framework, such as IPW estimation and G-estimation, can actually be performed within the SEM framework (e.g., see Loeys et al., 2014; Loh et al., 2020; Loh & Ren, 2023b). This allows researchers to combine the desirable properties of the SEM framework, such as the incorporation of latent variables, the possibility to easily impose parameter constraints, (relatively) easy control for measurement error, and flexible techniques for dealing with missing data handling (e.g, the use of full information maximum likelihood), with the advantages of potential outcomes methods for causal inference. I am convinced that the combination of methods from the potential outcomes framework and the latent variable capabilities of the SEM framework is a promising avenue for future methodological research, and can much improve causal inference in psychology.

References

Achterberg, M., Mulder, J. D., Dobbelaar, S., Heunis, S., & Crone, E. (2022). *Individual differences in developmental trajectories of social emotion regulation from childhood to emerging adolescence* (Preregistration) [Publisher: Open Science Framework]. Retrieved August 10, 2023, from https://doi.org/10.17605/OSF.IO/HDRZC

Achterberg, M., & Van der Meulen, M. (2019). Genetic and environmental influences on MRI scan quantity and quality. *Developmental Cognitive Neuroscience*, *38*, 100667. https://doi.org/10.1016/j.dcn.2019.100667

Achterberg, M., Van Duijvenvoorde, A. C. K., Bakermans-Kranenburg, M. J., & Crone, E. A. (2016). Control your anger! The neural basis of aggression regulation in response to negative social feedback. *Social Cognitive and Affective Neuroscience*, *11*(5), 712–720. https://doi.org/10.1093/scan/nsv154

Achterberg, M., Van Duijvenvoorde, A. C. K., Van Der Meulen, M., Bakermans-Kranenburg, M. J., & Crone, E. A. (2018). Heritability of aggression following social evaluation in middle childhood: An fMRI study. *Human Brain Mapping*, *39*(7), 2828–2841. https://doi.org/10.1002/hbm.24043

Achterberg, M., Van Duijvenvoorde, A. C. K., Van IJzendoorn, M. H., Bakermans-Kranenburg, M. J., & Crone, E. A. (2020). Longitudinal changes in DLPFC activation during childhood are related to decreased aggression following social rejection. *Proceedings of the National Academy of Sciences*, *117*(15), 8602–8610. https://doi.org/10.1073/pnas.1915124117

Achterberg, M., Van Duijvenvoorde, A. C., Van der Meulen, M., Euser, S., Bakermans-Kranenburg, M. J., & Crone, E. A. (2017). The neural and behavioral correlates of social evaluation in childhood. *Developmental Cognitive Neuroscience*, *24*, 107–117. https://doi.org/10.1016/j.dcn.2017.02.007

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, *60*(1), 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Akwa GGZ. (2020). *Psychotrauma- en stressorgerelateerde stoornissen* (tech. rep.). Akwa GGZ. https://www.ggzstandaarden.nl/zorgstandaarden/psychotrauma-en-stressorgerelateerde-stoornissen/introductie/introductie

Allison, P. D., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius: Sociological Research for a Dynamic World*, *3*, 237802311771057. https://doi.org/10.1177/2378023117710578

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (Fifth Edition). https://doi.org/10.1176/appi.books.9780890425596

American Psychological Association. (2017). *Clinical practice guideline for the treatment of posttraumatic stress disorder (PTSD) in adults* (tech. rep.). American Psychological Association. https://www.apa.org/ptsd-guideline/ptsd.pdf

Andersen, H. K. (2022). Equivalent approaches to dealing with unobserved heterogeneity in cross-lagged panel models? Investigating the benefits and drawbacks of the latent curve model with structured residuals and the random intercept cross-lagged panel model. *Psychological Methods*, *27*(5), 730–751. https://doi.org/10.1037/met0000285

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. https://doi.org/10.1515/9781400829828

Apps, M. A., Rushworth, M. F., & Chang, S. W. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron*, *90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete cata. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced Structural Equation Modeling: Issues and Techniques*. Lawrence Erlbaum Associates.

Asendorpf, J. B. (2021). Modeling developmental processes. In Rauthmann (Ed.), *The Handbook of Personality Dynamics and Processes* (pp. 815–835). Elsevier. https://doi.org/10.1016/B978-0-12-813995-0.00031-5

Asparouhov, T., & Muthén, B. O. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142. https://doi.org/10.1080/10705511.2018.1511375

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling research. In *Structural equation modeling: A second course* (2nd, pp. 625–666). Information Age Publishing.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Bianconcini, S., & Bollen, K. A. (2018). The latent variable-autoregressive latent trajectory model: A general framework for longitudinal data analysis. *Struc-*

tural Equation Modeling: A Multidisciplinary Journal, *25*(5), 791–808. https://doi.org/10.1080/10705511.2018.1426467

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., & Keane, T. M. (1995). The development of a clinician-administered PTSD scale. *Journal of Traumatic Stress*, *8*(1), 75–90. https://doi.org/10.1002/jts.2490080106

Blakemore, S.-J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience*, *9*(4), 267–277. https://doi.org/10.1038/nrn2353

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The posttraumatic stress disorder checklist for *DSM-5* (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, *28*(6), 489–498. https://doi.org/10.1002/jts.22059

Boden, J. M., Fergusson, D. M., & Horwood, L. J. (2008). Does adolescent self-esteem predict later life outcomes? A test of the causal role of self-esteem. *Development and Psychopathology*, *20*(1), 319–339. https://doi.org/10.1017/S0954579408000151

Boeschoten, M. A., Bakker, A., Jongedijk, R. A., Van Minnen, A., Elzinga, B. M., Rademaker, A. R., & Olff, M. (2014). Clinician administered PTSD scale for DSM-5—Nederlandstalig versie.

Boeschoten, M. A., Bakker, A., & Jongedijk, R. A. (2014). *PTSS checklist for DSM-5: Nederlandstalige versie*. Stichting Centrum '45, Arq Psychotrauma Expert Groep.

Boeschoten, M. A., Van der Aa, N., Bakker, A., Ter Heide, F. J. J., Hoofwijk, M. C., Jongedijk, R. A., Van Minnen, A., Elzinga, B. M., & Olff, M. (2018). Development and Evaluation of the Dutch clinician-administered PTSD scale for DSM-5 (CAPS-5). *European Journal of Psychotraumatology*, *9*(1), 1546085. https://doi.org/10.1080/20008198.2018.1546085

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, *89*(1), 1–34. https://doi.org/10.1353/sof.2010.0072

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons Literaturverzeichnis: Seite 471-487 Hier auch später erschienene, unveränderte Nachdrucke.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models [Series Title: Handbooks of Sociology and Social Research]. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301–328). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_15

Borghuis, J., Bleidorn, W., Sijtsma, K., Branje, S., Meeus, W. H. J., & Denissen, J. J. A. (2020). Longitudinal associations between trait neuroticism and negative daily experiences in adolescence. *Journal of Personality and Social Psychology*, *118*(2), 348–363. https://doi.org/10.1037/pspp0000233

Bottenhorn, K. L., Cardenas-Iniguez, C., Mills, K. L., Laird, A. R., & Herting, M. M. (2023). Profiling intra- and inter-individual differences in brain development across early adolescence. *NeuroImage*, *279*, 120287. https://doi.org/10.1016/j.neuroimage.2023.120287

Bou, J. C., & Satorra, A. (2018). Univariate versus multivariate modeling of panel data: Model specification and goodness-of-fit testing. *Organizational Research Methods*, *21*(1), 150–196. https://doi.org/10.1177/1094428117715509

Bradley, R., Greene, J., Russ, E., Dutra, L., & Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for PTSD. *American Journal of Psychiatry*, *162*(2), 214–227. https://doi.org/10.1176/appi.ajp.162.2.214

Brechwald, W. A., & Prinstein, M. J. (2011). Beyond homophily: A decade of advances in understanding peer influence processes. *Journal of Research on Adolescence*, *21*(1), 166–179. https://doi.org/10.1111/j.1532-7795.2010.00721.x

Brett, M., Anton, J. L., Valabregue, R., & Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox. *NeuroImage*, *16*, 372–373. https://doi.org/10.1016/S1053-8119(02)90013-3

Brookhart, M. A. (2015). Counterpoint: The treatment decision design. *American Journal of Epidemiology*, *182*(10), 840–845. https://doi.org/10.1093/aje/kwv214

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Burns, R. A., Crisp, D. A., & Burns, R. B. (2020). Re-examining the reciprocal effects model of self-concept, self-efficacy, and academic achievement in a comparison of the cross-lagged panel and random-intercept cross-lagged panel frameworks. *British Journal of Educational Psychology*, *90*(1), 77–91. https://doi.org/10.1111/bjep.12265

Byllesby, B. M., Dickstein, B. D., & Chard, K. M. (2019). The probability of change versus dropout in veterans receiving cognitive processing therapy for posttraumatic stress disorder. *Behaviour Research and Therapy*, *123*, 103483. https://doi.org/10.1016/j.brat.2019.103483

Canty, A., & Ripley, B. D. (2022). Boot: Bootstrap R (S-plus) functions.

Casey, B., Jones, R. M., Levita, L., Libby, V., Pattwell, S. S., Ruberry, E. J., Soliman, F., & Somerville, L. H. (2010). The storm and stress of adolescence: Insights from human imaging and mouse genetics. *Developmental Psychobiology*, n/a–n/a. https://doi.org/10.1002/dev.20447

Cheng, T. W., Vijayakumar, N., Flournoy, J. C., Op De Macks, Z., Peake, S. J., Flannery, J. E., Mobasser, A., Alberti, S. L., Fisher, P. A., & Pfeifer, J. H. (2019). *Feeling left out or just surprised? Neural correlates of social exclusion and over-inclusion in adolescence* (preprint). Neuroscience. https://doi.org/10.1101/524934

Chester, D. S. (2019). Beyond the aggregate score: Using multilevel modeling to examine trajectories of laboratory-measured aggression. *Aggressive Behavior*, *45*(5), 498–506. https://doi.org/10.1002/ab.21837

Chester, D. S., Eisenberger, N. I., Pond, R. S., Richman, S. B., Bushman, B. J., & DeWall, C. N. (2014). The interactive effect of social pain and executive functioning on aggression: An fMRI experiment. *Social Cognitive and Affective Neuroscience*, *9*(5), 699–704. https://doi.org/10.1093/scan/nst038

Cocosco, C. A., Kollokian, V., Kwam, R. K.-S., & Evans, A. C. (1997). BrainWeb: Online Interface to a 3D MRI Simulated Brain Database. *Proceedings of 3rd International Conference on Functional Mapping of the Human Brain*, *5*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (0th ed.). Routledge. https://doi.org/10.4324/9780203771587

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological Methods*, *10*(1), 3–20. https://doi.org/10.1037/1082-989X.10.1.3

Consortium on Individual Development. (2023). Consortium on Individual Development. https://individualdevelopment.nl/

Constantin, M. A., Schuurman, N. K., & Vermunt, J. K. (2023). A general Monte Carlo method for sample size analysis in the context of network models. *Psychological Methods*. https://doi.org/10.1037/met0000555

Cooper, A. A., Clifton, E. G., & Feeny, N. C. (2017). An empirical review of potential mediators and mechanisms of prolonged exposure therapy. *Clinical Psychology Review*, *56*, 106–121. https://doi.org/10.1016/j.cpr.2017.07.003

Copeland, W. E., Wolke, D., Angold, A., & Costello, E. J. (2013). Adult psychiatric outcomes of bullying and being bullied by peers in childhood and adolescence.

JAMA Psychiatry, *70*(4), 419. https://doi.org/10.1001/jamapsychiatry.2013.504

Crone, E. A., Achterberg, M., Dobbelaar, S., Euser, S., Van den Bulk, B., Van der Meulen, M., Van Drunen, L., Wierenga, L. M., Bakermans-Kranenburg, M. J., & Van IJzendoorn, M. H. (2020). Neural and behavioral signatures of social evaluation and adaptation in childhood and adolescence: The Leiden consortium on individual development (L-CID). *Developmental Cognitive Neuroscience*, *45*, 100805. https://doi.org/10.1016/j.dcn.2020.100805

Crone, E. A., & Steinbeis, N. (2017). Neural perspectives on cognitive control development during childhood and adolescence. *Trends in Cognitive Sciences*, *21*(3), 205–215. https://doi.org/10.1016/j.tics.2017.01.003

Cui, L., Colasante, T., Malti, T., Ribeaud, D., & Eisner, M. P. (2016). Dual trajectories of reactive and proactive aggression from mid-childhood to early adolescence: Relations to sensation seeking, risk taking, and moral reasoning. *Journal of Abnormal Child Psychology*, *44*(4), 663–675. https://doi.org/10.1007/s10802-015-0079-7

Dale, A. M. (1999). Optimal experimental design for event-related fMRI. *Human Brain Mapping*, *8*(2-3), 109–114. https://doi.org/10.1002/(SICI)1097-0193(1999)8:2/3⟨109::AID-HBM7⟩3.0.CO;2-W

Dalgleish, T., Walsh, N. D., Mobbs, D., Schweizer, S., Van Harmelen, A.-L., Dunn, B., Dunn, V., Goodyer, I., & Stretton, J. (2017). Social pain and social gain in the adolescent brain: A common neural circuitry underlying both positive and negative social evaluation. *Scientific Reports*, *7*(1), 42010. https://doi.org/10.1038/srep42010

Daniel, R., Cousens, S., De Stavola, B., Kenward, M. G., & Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, *32*(9), 1584–1618. https://doi.org/10.1002/sim.5686

Davidson, J., Smith, R., & Kudler, H. (1989). Validity and reliability of the DSM-III criteria for posttraumatic stress disorder: Experience with a structured interview. *The Journal of Nervous and Mental Disease*, *177*(6), 336–341. https://doi.org/10.1097/00005053-198906000-00003

De Jonckere, J., & Rosseel, Y. (2022). Using bounded estimation to avoid nonconvergence in small sample structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 412–427. https://doi.org/10.1080/10705511.2021.1982716

De Stavola, B. L., Daniel, R. M., Ploubidis, G. B., & Micali, N. (2015). Mediation analysis with intermediate confounding: Structural equation modeling viewed through the causal inference lens. *American Journal of Epidemiology*, *181*(1), 64–80. https://doi.org/10.1093/aje/kwu239

De Vries, G.-J., & Olff, M. (2009). The lifetime prevalence of traumatic events and posttraumatic stress disorder in the Netherlands. *Journal of Traumatic Stress*, *22*(4), 259–267. https://doi.org/10.1002/jts.20429

Dietvorst, E., Hiemstra, M., Hillegers, M. H., & Keijsers, L. (2018). Adolescent perceptions of parental privacy invasion and adolescent secrecy: An illustration of Simpson's paradox. *Child Development*, *89*(6), 2081–2090. https://doi.org/10.1111/cdev.13002

Dobbelaar, S., Achterberg, M., Van Drunen, L., Van Duijvenvoorde, A. C., Van IJzendoorn, M. H., & Crone, E. A. (2022). Development of social feedback processing and responses in childhood: An fMRI test-replication design in two age cohorts. *Social Cognitive and Affective Neuroscience*, nsac039. https://doi.org/10.1093/scan/nsac039

Dobbelaar, S., Achterberg, M., Van Duijvenvoorde, A. C., Van IJzendoorn, M. H., & Crone, E. A. (2023). Developmental patterns and individual differences in responding to social feedback: A longitudinal fMRI study from childhood to adolescence. *Developmental Cognitive Neuroscience*, *62*, 101264. https://doi.org/10.1016/j.dcn.2023.101264

Dodge, K. A., Lansford, J. E., Burks, V. S., Bates, J. E., Pettit, G. S., Fontaine, R., & Price, J. M. (2003). Peer rejection and social information-processing factors in the development of aggressive behavior problems in children. *Child Development*, *74*(2), 374–393. https://doi.org/10.1111/1467-8624.7402004

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods*, *20*(4), 489–505. https://doi.org/10.1037/met0000041

Edwards, J. K., Hester, L. L., Gokhale, M., & Lesko, C. R. (2016). Methodologic issues when estimating risks in pharmacoepidemiology. *Current Epidemiology Reports*, *3*(4), 285–296. https://doi.org/10.1007/s40471-016-0089-1

Ehlers, A., Clark, D. M., Dunmore, E., Jaycox, L., Meadows, E., & Foa, E. B. (1998). Predicting response to exposure treatment in PTSD: The role of mental defeat and alienation. *Journal of Traumatic Stress*, *11*(3), 457–471. https://doi.org/10.1023/A:1024448511504

Ehlers, A., Grey, N., Wild, J., Stott, R., Liness, S., Deale, A., Handley, R., Albert, I., Cullen, D., Hackmann, A., Manley, J., McManus, F., Brady, F., Salkovskis, P., & Clark, D. M. (2013). Implementation of cognitive therapy for PTSD in routine clinical care: Effectiveness and moderators of outcome in a consecutive sample. *Behaviour Research and Therapy*, *51*(11), 742–752. https://doi.org/10.1016/j.brat.2013.08.006

Ehlers, A., Hackmann, A., Grey, N., Wild, J., Liness, S., Albert, I., Deale, A., Stott, R., & Clark, D. M. (2014). A randomized controlled trial of 7-day intensive and standard weekly cognitive therapy for PTSD and emotion-focused supportive therapy. *American Journal of Psychiatry*, *171*(3), 294–304. https://doi.org/10.1176/appi.ajp.2013.13040552

Emerson, D., Sharma, R., Chaudhry, S., & Turner, J. (2009). Trauma-sensitive yoga: Principles, practice, and research. *International Journal of Yoga Therapy*, *19*(1), 123–128. https://doi.org/10.17761/ijyt.19.1.h6476p8084l22160

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

European Medicines Agency. (2020). ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Retrieved October 12, 2023, from https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf

Fair, D. A., Dosenbach, N. U., Moore, A. H., Satterthwaite, T. D., & Milham, M. P. (2021). Developmental cognitive neuroscience in the era of networks and big data: Strengths, weaknesses, opportunities, and threats. *Annual Review of Developmental Psychology*, *3*(1), 249–275. https://doi.org/10.1146/annurev-devpsych-121318-085124

Finkel, S. (1995). *Causal analysis with panel data*. SAGE Publications, Inc. https://doi.org/10.4135/9781412983594

Fisher, R. A. (1935). *The design of experiments*. Oliver; Boyd.

Fite, P. J., Colder, C. R., Lochman, J. E., & Wells, K. C. (2008). Developmental trajectories of proactive and reactive aggression from fifth to ninth grade. *Journal of Clinical Child & Adolescent Psychology*, *37*(2), 412–421. https://doi.org/10.1080/15374410801955920

Foa, E. B., Riggs, D. S., Massie, E. D., & Yarczower, M. (1995). The impact of fear activation and anger on the efficacy of exposure treatment for posttraumatic stress disorder. *Behavior Therapy*, *26*(3), 487–499. https://doi.org/10.1016/S0005-7894(05)80096-6

Forbes, D., Creamer, M., Hawthorne, G., Allen, N., & Mchugh, T. (2003). Comorbidity as a predictor of symptom change after treatment in combat-related posttraumatic stress disorder. *The Journal of Nervous and Mental Disease*, *191*(2), 93–99. https://doi.org/10.1097/01.NMD.0000051903.60517.98

Foulkes, L., & Blakemore, S.-J. (2018). Studying individual differences in human adolescent brain development. *Nature Neuroscience*, *21*(3), 315–323. https://doi.org/10.1038/s41593-018-0078-4

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of “theory of mind”. *Trends in Cognitive Sciences*, *7*(2), 77–83. https://doi.org/10.1016/S1364-6613(02)00025-6

Gische, C., & Voelkle, M. C. (2022). Beyond the mean: A flexible framework for studying causal effects using linear models. *Psychometrika*, *87*(3), 868–901. https://doi.org/10.1007/s11336-021-09811-z

Gische, C., West, S. G., & Voelkle, M. C. (2021). Forecasting causal effects of interventions versus predicting future outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 475–492. https://doi.org/10.1080/10705511.2020.1780598

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E. E., & Waernbaum, I. (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, *39*(30), 4922–4948. https://doi.org/10.1002/sim.8741

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*(1), 80. https://doi.org/10.2307/1130293

Gracia-Tabuenca, Z., Moreno, M. B., Barrios, F. A., & Alcauter, S. (2021). Development of the brain functional connectome follows puberty-dependent nonlinear trajectories. *NeuroImage*, *229*, 117769. https://doi.org/10.1016/j.neuroimage.2021.117769

Green, K. H., Van De Groep, S., Van der Cruijsen, R., Polak, M. G., & Crone, E. A. (2023). The Multidimensional Wellbeing in Youth Scale (MWYS): Development and psychometric properties. *Personality and Individual Differences*, *204*, 112038. https://doi.org/10.1016/j.paid.2022.112038

Greifer, N. (2023a). Cobalt: Covariate balance tables and plots. https://CRAN.R-project.org/package=cobalt

Greifer, N. (2023b). WeightIt: Weighting for covariate balance in observational studies. https://CRAN.R-project.org/package=WeightIt

Griliches, Z., & Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics*, *31*(1), 93–118. https://doi.org/10.1016/0304-4076(86)90058-8

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Guina, J., Rossetter, S. R., DeRHODES, B. J., Nahhas, R. W., & Welton, R. S. (2015). Benzodiazepines for PTSD: A systematic review and meta-analysis. *Journal of Psychiatric Practice*, *21*(4), 281–303. https://doi.org/10.1097/PRA.0000000000000091

Gunther Moor, B., Van Leijenhorst, L., Rombouts, S. A., Crone, E. A., & Van der Molen, M. W. (2010). Do you like me? Neural correlates of social evaluation and developmental trajectories. *Social Neuroscience*, *5*(5-6), 461–482. https://doi.org/10.1080/17470910903526155

Guyer, A. E., Choate, V. R., Pine, D. S., & Nelson, E. E. (2012). Neural circuitry underlying affective response to peer feedback in adolescence. *Social Cognitive* *and Affective Neuroscience*, *7*(1), 81–92. https://doi.org/10.1093/scan/nsr043

Haber, N. A., Wieten, S. E., Rohrer, J. M., Arah, O. A., Tennant, P. W. G., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., Howcutt, S. J., Simmons, A. E., Leyrat, C., Schoenegger, P., Booman, A., Dufour, M.-S. K., O'Donoghue, A. L., Baglini, R., Do, S., . . . Fox, M. P. (2022). Causal and associational language in observational health research: A systematic evaluation. *American Journal of Epidemiology*, *191*(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

Hamaker, E. L., Schuurman, N. K., & Zijlmans, E. A. O. (2017). Using a few snapshots to distinguish mountains from waves: Weak factorial invariance in the context of trait-state research. *Multivariate Behavioral Research*, *52*(1), 47–60. https://doi.org/10.1080/00273171.2016.1251299

Hamaker, E. L. (2005). Conditions for the equivalence of the autoregressive latent trajectory model and a latent growth curve model with autoregressive disturbances. *Sociological Methods & Research*, *33*(3), 404–416. https://doi.org/10.1177/0049124104270220

Hamaker, E. L., & Dolan, C. V. (2009). Idiographic data analysis: Quantitative methods—From simple to advanced. In J. Valsiner, P. C. M. Molenaar, M. C. Lyra, & N. Chaudhary (Eds.), *Dynamic Process Methodology in the Social and Developmental Sciences* (pp. 191–216). Springer US. https://doi.org/10.1007/978-0-387-95922-1_9

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Hamaker, E. L., Mulder, J. D., & van IJzendoorn, M. H. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental Cognitive Neuroscience*, *46*, 100867. https://doi.org/10.1016/j.dcn.2020.100867

Hamaker, E. L., & Muthén, B. O. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hancock, G. R., & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 117–159). Information Age Publishing.

Harter, S. (1988). Developmental processes in the construction of the self. In *Integrative processes and socialization: Early to middle childhood* (pp. 45–78). Lawrence Erlbaum Associates, Inc.

Hautzinger, M., & Bailer, M. (1993). *Allgemeine depressions Skala: ADS: Manual.* Beltz.

Heise, D. R. (1970). Causal inference from panel data. *Sociological Methodology*, *2*, 3. https://doi.org/10.2307/270780

Hendriks, L., Kleine, R. A. D., Broekman, T. G., Hendriks, G.-J., & Minnen, A. V. (2018). Intensive prolonged exposure therapy for chronic PTSD patients following multiple trauma and multiple treatment attempts. *European Journal of Psychotraumatology*, *9*(1), 1425574. https://doi.org/10.1080/20008198.2018.1425574

Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, *26*(10), 674–680. https://doi.org/10.1016/j.annepidem.2016.08.016

Hernán, M. A. (2018). The C-Word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, *108*(5), 616–619. https://doi.org/10.2105/AJPH.2018.304337

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764. https://doi.org/10.1093/aje/kwv254

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if.* Chapman & Hall/CRC.

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., Platt, R., & Shrier, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, *79*, 70–75. https://doi.org/10.1016/j.jclinepi.2016.04.014

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development*, *58*(1), 93. https://doi.org/10.2307/1130294

Hesser, H., Hedman-Lagerlöf, E., Andersson, E., Lindfors, P., & Ljótsson, B. (2018). How does exposure therapy work? A comparison between generic and gastrointestinal anxiety–specific mediators in a dismantling study of exposure therapy for irritable bowel syndrome. *Journal of Consulting and Clinical Psychology*, *86*(3), 254–267. https://doi.org/10.1037/ccp0000273

Horvath, A. O., & Symonds, B. D. (1991). Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, *38*(2), 139–149. https://doi.org/10.1037/0022-0167.38.2.139

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Imbens, G. W. (2019). *Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics* (tech. rep. w26104). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w26104

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Imel, Z. E., Laska, K., Jakupcak, M., & Simpson, T. L. (2013). Meta-analysis of dropout in treatments for posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *81*(3), 394–404. https://doi.org/10.1037/a0031474

International Society of Traumatic Stress Studies. (2018). New ISTSS prevention and treatment guidelines. http://www.istss.org/treating-trauma/new-istssguidelines.aspx

Ioannidis, K., Askelund, A. D., Kievit, R. A., & Van Harmelen, A.-L. (2020). The complex neurobiology of resilient functioning after childhood maltreatment. *BMC Medicine*, *18*(1), 32. https://doi.org/10.1186/s12916-020-1490-7

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Jak, S., Jorgensen, T. D., Verdam, M. G. E., Oort, F. J., & Elffers, L. (2021). Analytical power calculations for structural equation modeling: A tutorial and Shiny app. *Behavior Research Methods*, *53*(4), 1385–1406. https://doi.org/10.3758/s13428-020-01479-0

Jaycox, L. H., Foa, E. B., & Morral, A. R. (1998). Influence of emotional engagement and habituation on exposure therapy for PTSD. *Journal of Consulting and Clinical Psychology*, *66*(1), 185–192. https://doi.org/10.1037/0022-006X.66.1.185

Jongh, A. d., & Broeke, E. t. (2020). *Handboek EMDR: Een geprotocolleerde behandelmethode voor de gevolgen van psychotrauma* (Zevende editie, 2e oplage) [OCLC: 1286026287]. Pearson.

Jorgensen, T. D., Pornprasertmanit, S., & Schoemann, A. M. (2022). semTools: Useful tools for structural equation modeling. https://CRAN.R-project.org/package=semTools

Karatzias, A., Power, K., McGoldrick, T., Brown, K., Buchanan, R., Sharp, D., & Swanson, V. (2007). Predicting treatment outcome on three measures for post-traumatic stress disorder. *European Archives of Psychiatry and Clinical Neuroscience*, *257*(1), 40–46. https://doi.org/10.1007/s00406-006-0682-2

Keijsers, L. (2016). Parental monitoring and adolescent problem behaviors: How much do we really know? *International Journal of Behavioral Development*, *40*(3), 271–281. https://doi.org/10.1177/0165025415592515

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*(1), 52–59. https://doi.org/10.1037/0022-006X.63.1.52

Kenny, D. A., & Zautra, A. (2001). Trait–state models for longitudinal data. In *New methods for the analysis of change* (pp. 243–263). American Psychological Association.

Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H.-U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States: Anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research*, *21*(3), 169–184. https://doi.org/10.1002/mpr.1359

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00513

Kim, C. J., & Nelson, C. R. (1999). *State-space models with regime switching.* The MIT Press.

Kim, D., Bae, H., & Chon Park, Y. (2008). Validity of the subjective units of disturbance scale in EMDR. *Journal of EMDR Practice and Research*, *2*(1), 57–62. https://doi.org/10.1891/1933-3196.2.1.57

Kim, J., Christy, A. G., Schlegel, R. J., Donnellan, M. B., & Hicks, J. A. (2018). Existential ennui: Examining the reciprocal relationship between self-alienation and academic amotivation. *Social Psychological and Personality Science*, *9*(7), 853–862. https://doi.org/10.1177/1948550617727587

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1

Kuiper, R. M., & Ryan, O. (2018). Drawing conclusions from cross-lagged relationships: Re-considering the role of the time-interval. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 809–823. https://doi.org/10.1080/10705511.2018.1431046

Kunicki, Z. J., Smith, M. L., & Murray, E. J. (2023). A primer on structural equation model diagrams and directed acyclic graphs: When and how to use each in psychological and epidemiological research. *Advances in Methods and Practices in Psychological Science*, *6*(2), 251524592311560. https://doi.org/10.1177/25152459231156085

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuster, F., Orth, U., & Meier, L. L. (2012). Rumination mediates the prospective effect of low self-esteem on depression: A five-wave longitudinal study. *Personality and Social Psychology Bulletin*, *38*(6), 747–759. https://doi.org/10.1177/0146167212437250

Ladd, G. W. (2006). Peer rejection, aggressive or withdrawn behavior, and psychological maladjustment from ages 5 to 12: An examination of four predictive models. *Child Development*, *77*(4), 822–846.

Lamblin, M., Murawski, C., Whittle, S., & Fornito, A. (2017). Social connectedness, mental health and the adolescent brain. *Neuroscience & Biobehavioral Reviews*, *80*, 57–68. https://doi.org/10.1016/j.neubiorev.2017.05.010

Lash, T. L., Fox, M. P., & Fink, A. K. (2009). *Applying quantitative bias analysis to epidemiologic data* [OCLC: ocn310400811]. Springer.

Leary, M. R., Twenge, J. M., & Quinlivan, E. (2006). Interpersonal rejection as a determinant of anger and aggression. *Personality and Social Psychology Review*, *10*(2), 111–132. https://doi.org/10.1207/s15327957pspr1002_2

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences*, *39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

Lee, S. (2015). Implementing a simulation study using multiple software packages for structural equation modeling. *SAGE Open*, *5*(3), 215824401559182. https://doi.org/10.1177/2158244015591823

Lek, K., Oberski, D., Davidov, E., Cieciuch, J., Seddig, D., & Schmidt, P. (2018). Approximate measurement invariance. In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods* (pp. 911–929). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118884997.ch41

Lewis, C., Roberts, N. P., Andrew, M., Starling, E., & Bisson, J. I. (2020). Psychological therapies for post-traumatic stress disorder in adults: Systematic review and meta-analysis. *European Journal of Psychotraumatology*, *11*(1), 1729633. https://doi.org/10.1080/20008198.2020.1729633

Li, S., Okereke, O. I., Chang, S.-C., Kawachi, I., & VanderWeele, T. J. (2016). Religious service attendance and lower depression among women: A prospective cohort study. *Annals of Behavioral Medicine*, *50*(6), 876–884. https://doi.org/10.1007/s12160-016-9813-9

Liker, J. K., Augustyniak, S., & Duncan, G. J. (1985). Panel data and models of change: A comparison of first difference and conventional two-wave models. *Social Science Research*, *14*(1), 80–101. https://doi.org/10.1016/0049-089X(85)90013-4

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Loeys, T., Moerkerke, B., Raes, A., Rosseel, Y., & Vansteelandt, S. (2014). Estimation of controlled direct effects in the presence of exposure-induced confounding and latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 396–407. https://doi.org/10.1080/10705511.2014.915372

Loh, W. W., Moerkerke, B., Loeys, T., Poppe, L., Crombez, G., & Vansteelandt, S. (2020). Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomized studies. *Multivariate Behavioral Research*, *55*(5), 763–785. https://doi.org/10.1080/00273171.2019.1681251

Loh, W. W., & Ren, D. (2023a). Estimating time-varying treatment effects in longitudinal studies. *Psychological Methods*. https://doi.org/10.1037/met0000574

Loh, W. W., & Ren, D. (2023b). A tutorial on causal inference in longitudinal data with time-varying confounding using G-estimation. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231174029. https://doi.org/10.1177/25152459231174029

Lüdtke, O., & Robitzsch, A. (2022). A comparison of different approaches for estimating cross-lagged effects from a causal inference perspective. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(6), 888–907. https://doi.org/10.1080/10705511.2022.2065278

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of cognitive processes from late childhood to adulthood. *Child Development*, *75*(5), 1357–1372. https://doi.org/10.1111/j.1467-8624.2004.00745.x

Luna, B., Padmanabhan, A., & O'Hearn, K. (2010). What has fMRI told us about the development of cognitive control through adolescence? *Brain and Cognition*, *72*(1), 101–113. https://doi.org/10.1016/j.bandc.2009.08.005

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., . . . Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *68*(3), 438–450. https://doi.org/10.1037/0022-006X.68.3.438

Matsueda, R. L. (2023). A brief history of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd).

Mavranezouli, I., Megnin-Viggars, O., Grey, N., Bhutani, G., Leach, J., Daly, C., Dias, S., Welton, N. J., Katona, C., El-Leithy, S., Greenberg, N., Stockton, S., & Pilling, S. (2020). Cost-effectiveness of psychological treatments for post-traumatic stress disorder in adults (S. McDonald, Ed.). *PLOS ONE*, *15*(4), e0232245. https://doi.org/10.1371/journal.pone.0232245

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122. https://doi.org/10.1007/BF02294746

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/-Taylor & Francis Group.

Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *The Academy of Management Review*, *26*(4), 530. https://doi.org/10.2307/3560240

Moerkerke, B., Loeys, T., & Vansteelandt, S. (2015). Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. *Psychological Methods*, *20*(2), 204–220. https://doi.org/10.1037/a0036368

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mu, W., Luo, J., Rieger, S., Trautwein, U., & Roberts, B. W. (2019). The relationship between self-esteem and depression when controlling for neuroticism (S. Vazire & S. Vazire, Eds.). *Collabra: Psychology*, *5*(1), 11. https://doi.org/10.1525/collabra.204

Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 638–648. https://doi.org/10.1080/10705511.2020.1784738

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69. https://doi.org/10.2307/1913646

Muthén, B. O., & Asparouhov, T. (2022a). Can cross-lagged panel modeling be relied on to establish cross-lagged effects? https://www.statmodel.com/download/WT5.pdf

Muthén, B. O., & Asparouhov, T. (2022b). Recent advances in modeling short and long longitudinal data. https://www.statmodel.com/download/ASA2022.pdf

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (2009). Categorical latent variable modeling using Mplus: Cross-sectional data. https://statmodel.com/download/Topic%205.pdf

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide*. Muthén & Muthén.

Naimi, A. I., Cole, S. R., & Kennedy, E. H. (2016). An introduction to G methods. *International Journal of Epidemiology*, *46*(2), 756–762. https://doi.org/10.1093/ije/dyw323

Narmandakh, A., Roest, A. M., Jonge, P. D., & Oldehinkel, A. J. (2020). The bidirectional association between sleep problems and anxiety symptoms in adolescents: A TRAILS report. *Sleep Medicine*, *67*, 39–46. https://doi.org/10.1016/j.sleep.2019.10.018

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., Estabrook, R., Bates, T. C., Maes, H. H., & Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535–549. https://doi.org/10.1007/s11336-014-9435-8

Nesdale, D., & Lambert, A. (2007). Effects of experimentally manipulated peer rejection on children's negative affect, self-esteem, and maladaptive social behavior. *International Journal of Behavioral Development*, *31*(2), 115–122. https://doi.org/10.1177/0165025407073579

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, *54*(2), 638. https://doi.org/10.2307/3109770

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*(7), 771–785. https://doi.org/10.1177/0146167201277001

Nguyen, T. Q., Webb-Vargas, Y., Koning, I. M., & Stuart, E. A. (2016). Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: Simulations and application to an alcohol intervention. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 368–383. https://doi.org/10.1080/10705511.2015.1062730

Nijdam, M. J., Gersons, B. P. R., Reitsma, J. B., De Jongh, A., & Olff, M. (2012). Brief eclectic psychotherapy *v.* eye movement desensitisation and reprocessing therapy for post-traumatic stress disorder: Randomised controlled trial. *British Journal of Psychiatry*, *200*(3), 224–231. https://doi.org/10.1192/bjp.bp.111.099234

Nolan, C. R. (2016). Bending without breaking: A narrative review of trauma-sensitive yoga for women with PTSD. *Complementary Therapies in Clinical Practice*, *24*, 32–40. https://doi.org/10.1016/j.ctcp.2016.05.006

Oertzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 627–646. https://doi.org/10.1348/000711010X486633

Ormel, J., Rijsdijk, F. V., Sullivan, M., van Sonderen, E., & Kempen, G. I. J. M. (2002). Temporal and reciprocal relationship between IADL/ADL disability and depressive symptoms in late life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(4), P338–P347. https://doi.org/10.1093/geronb/57.4.P338

Ormel, J., & Schaufeli, W. B. (1991). Stability and change in psychological distress and their relationship with self-esteem and locus of control: A dynamic equilibrium model. *Journal of Personality and Social Psychology*, *60*(2), 288–299. https://doi.org/10.1037/0022-3514.60.2.288

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034. https://doi.org/10.1037/pspp0000358

O'Shaughnessy, E. S., Berl, M. M., Moore, E. N., & Gaillard, W. D. (2008). Pediatric functional magnetic resonance imaging (fMRI): Issues and applications. *Journal of Child Neurology*, *23*(7), 791–801. https://doi.org/10.1177/0883073807313047

Ousey, G. C., Wilcox, P., & Fisher, B. S. (2011). Something old, something new: Revisiting competing hypotheses of the victimization-offending relationship among adolescents. *Journal of Quantitative Criminology*, *27*(1), 53–84. https://doi.org/10.1007/s10940-010-9099-1

Ozkok, O., Vaulont, M. J., Zyphur, M. J., Zhang, Z., Preacher, K. J., Koval, P., & Zheng, Y. (2022). Interaction rffects in cross-lagged panel models: SEM with latent interactions applied to work-family conflict, job satisfaction, and gender. *Organizational Research Methods*, *25*(4), 673–715. https://doi.org/10.1177/10944281211043733

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology*, *21*(6), 872–875. https://doi.org/10.1097/EDE.0b013e3181f5d3fd

Pearl, J. (2018). Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference*, *6*(2), 20182001. https://doi.org/10.1515/jci-2018-2001

Petersen, M. L., & Van der Laan, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, *25*(3), 418–426. https://doi.org/10.1097/EDE.0000000000000078

Price, M., Szafranski, D. D., Van Stolk-Cooke, K., & Gros, D. F. (2016). Investigation of abbreviated 4 and 8 item versions of the PTSD Checklist 5. *Psychiatry Research*, *239*, 124–130. https://doi.org/10.1016/j.psychres.2016.03.014

Prinstein, M. J., & Aikins, J. W. (2004). Cognitive moderators of the longitudinal association between peer rejection and adolescent depressive symptoms. *Journal of Abnormal Child Psychology*, *32*(2), 147–158. https://doi.org/10.1023/B:JACP.0000019767.55592.63

Prinstein, M. J., & La Greca, A. M. (2004). Childhood peer rejection and aggression as predictors of adolescent girls' externalizing and health risk behaviors: A 6-year longitudinal study. *Journal of Consulting and Clinical Psychology*, *72*(1), 103–112. https://doi.org/10.1037/0022-006X.72.1.103

R Core Team. (2022). R: A language and environment for statistical computing.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Ragsdale, K. A., Watkins, L. E., Sherrill, A. M., Zwiebach, L., & Rothbaum, B. O. (2020). Advances in PTSD treatment delivery: Evidence base and future directions for intensive outpatient programs. *Current Treatment Options in Psychiatry*, *7*(3), 291–300. https://doi.org/10.1007/s40501-020-00219-7

Raudenbush, S. W., & Bryk, S. W. R. A. S. (2022). *Hierarchical linear models: Applications and data analysis methods*. SAGE Publications.

Riva, P., Romero Lauro, L. J., DeWall, C. N., Chester, D. S., & Bushman, B. J. (2015). Reducing aggressive responses to social exclusion using transcranial direct current stimulation. *Social Cognitive and Affective Neuroscience*, *10*(3), 352–356. https://doi.org/10.1093/scan/nsu053

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: Application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9-12), 1393–1512. https://doi.org/10.1016/0270-0255(86)90088-6

Robins, J. M., & Greenland, S. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, *95*(450), 431. https://doi.org/10.2307/2669381

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. http://www.jstor.org/stable/3703997

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*(2), 245–258. https://doi.org/10.1037/0033-2909.88.2.245

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, S., Vancampfort, D., Steel, Z., Newby, J., Ward, P. B., & Stubbs, B. (2015). Physical activity in the treatment of Post-traumatic stress disorder: A systematic review and meta-analysis. *Psychiatry Research*, *230*(2), 130–136. https://doi.org/10.1016/j.psychres.2015.10.017

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. Van de Schoot & M. Miocevic (Eds.), *Small sample size solutions* (pp. 226–238). Routledge.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (Ed.). (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rudolph, K. D., Davis, M. M., Skymba, H. V., Modi, H. H., & Telzer, E. H. (2021). Social experience calibrates neural sensitivity to social feedback during adolescence: A functional connectivity approach. *Developmental Cognitive Neuroscience*, *47*, 100903. https://doi.org/10.1016/j.dcn.2020.100903

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*(1), 83–90. https://doi.org/10.1007/BF02294150

Scherpenzeel, A. C. (2018). “True” longitudinal and probability-based internet panels: Evidence from the netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet* (1st ed., pp. 77–104). Routledge. https://doi.org/10.4324/9780203844922-4

Schottenbauer, M. A., Glass, C. R., Arnkoff, D. B., Tendick, V., & Gray, S. H. (2008). Nonresponse and dropout rates in outcome studies on PTSD: Review and methodological considerations. *Psychiatry: Interpersonal and Biological Processes*, *71*(2), 134–168. https://doi.org/10.1521/psyc.2008.71.2.134

Schriber, R. A., & Guyer, A. E. (2016). Adolescent neurobiological susceptibility to social context. *Developmental Cognitive Neuroscience*, *19*, 1–18. https://doi.org/10.1016/j.dcn.2015.12.009

Seddig, D. (2020). Individual attitudes toward deviant behavior and perceived attitudes of friends: Self-stereotyping and social projection in adolescence and emerging adulthood. *Journal of Youth and Adolescence*, *49*(3), 664–677. https://doi.org/10.1007/s10964-019-01123-x

Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: Concept and application with panel data [Artwork Size: 29-41 Pages Publisher: European Survey Research Association]. *Survey Research Methods*, *Vol 12*, 29–41 Pages. https://doi.org/10.18148/SRM/2018.V12I1.7210
SeriesInformation Survey Research Methods, Vol 12, No 1 (2018)

Shapiro, F. (2018). *Eye movement desensitization and reprocessing (EMDR) therapy: Basic principles, protocols, and procedures* (Third edition) [OCLC: 1007506753]. The Guilford Press.

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

Somerville, L. H., & Casey, B. (2010). Developmental neurobiology of cognitive control and motivational systems. *Current Opinion in Neurobiology*, *20*(2), 236–241. https://doi.org/10.1016/j.conb.2010.01.006

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006). Anterior cingulate cortex responds differentially to expectancy violation and social rejection. *Nature Neuroscience*, *9*(8), 1007–1008. https://doi.org/10.1038/nn1728

Speyer, L. G., Ushakova, A., Blakemore, S.-J., Murray, A. L., & Kievit, R. (2023). Testing for within × within and between × within moderation using random intercept cross-lagged panel models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(2), 315–327. https://doi.org/10.1080/10705511.2022.2096613

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–480. https://doi.org/10.1214/ss/1177012031

Sripada, R. K., Ready, D. J., Ganoczy, D., Astin, M. C., & Rauch, S. A. (2020). When to change the treatment plan: An analysis of diminishing returns in VA patients undergoing prolonged exposure and cognitive processing therapy. *Behavior Therapy*, *51*(1), 85–98. https://doi.org/10.1016/j.beth.2019.05.003

StataCorp. (2023). Stata Statistical Software: Release 18.

Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Developmental Review*, *28*(1), 78–106. https://doi.org/10.1016/j.dr.2007.08.002

Steinberg, L., & Morris, A. S. (2001). Adolescent development. *Annual Review of Psychology*, *52*(1), 83–110. https://doi.org/10.1146/annurev.psych.52.1.83

Steyer, R., Ferring, D., & Schmnitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*, 79–98.

Stoel, R. D., Garre, F. G., Dolan, C., & Van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*(4), 439–455. https://doi.org/10.1037/1082-989X.11.4.439

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Suissa, S. (2008). Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, *167*(4), 492–499. https://doi.org/10.1093/aje/kwm324

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136. https://doi.org/10.1177/0146167298242002

Tarrier, N., Sommerfield, C., Pilgrim, H., & Faragher, B. (2000). Factors associated with outcome of cognitive-behavioural treatment of chronic post-traumatic stress disorder. *Behaviour Research and Therapy*, *38*(2), 191–202. https://doi.org/10.1016/S0005-7967(99)00030-3

Telzer, E. H., McCormick, E. M., Peters, S., Cosme, D., Pfeifer, J. H., & Van Duijvenvoorde, A. C. (2018). Methodological considerations for developmental longitudinal fMRI research. *Developmental Cognitive Neuroscience*, *33*, 149–160. https://doi.org/10.1016/j.dcn.2018.02.004

Tian, J., Venn, A., Otahal, P., & Gall, S. (2015). The association between quitting smoking and weight gain: A systemic review and meta-analysis of prospective cohort studies. *Obesity Reviews*, *16*(10), 883–901. https://doi.org/10.1111/obr.12304

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*(1), 31–65. https://doi.org/10.1146/annurev.clinpsy.1.102803.144239

Tompsett, D., Vansteelandt, S., Dukes, O., & De Stavola, B. (2022). Gesttools: General purpose g-estimation in R. *Observational Studies*, *8*(1), 1–28. https://doi.org/10.1353/obs.2022.0003

Trapnell, P. D., & Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology*, *76*(2), 284–304. https://doi.org/10.1037/0022-3514.76.2.284

Tuerk, P. W., Yoder, M., Grubaugh, A., Myrick, H., Hamner, M., & Acierno, R. (2011). Prolonged exposure therapy for combat-related posttraumatic stress disorder: An examination of treatment effectiveness for veterans of the wars in Afghanistan and Iraq. *Journal of Anxiety Disorders*, *25*(3), 397–403. https://doi.org/10.1016/j.janxdis.2010.11.002

Usami, S. (2021). On the differences between general cross-lagged panel model and random-intercept cross-lagged panel model: Interpretation of cross-lagged parameters and model choice. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 331–344. https://doi.org/10.1080/10705511.2020.1821690

Usami, S. (2022). Within-person variability score-based causal inference: A two-step estimation for joint effects of time-varying treatments. *Psychometrika*. https://doi.org/10.1007/s11336-022-09879-1

Usami, S. (2023). *A two-step robust estimation approach for inferring within-person relations in longitudinal design: Tutorial and simulations*. Retrieved March 27, 2023, from 10.31234/osf.io/vkq7s

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, *24*(5), 637–657. https://doi.org/10.1037/met0000210

Vahedi, S. (2010). World Health Organization Quality-of-Life Scale (WHOQOL-BREF): Analyses of their item response theory properties based on the graded responses model. *Iranian journal of psychiatry*, *5*(4), 140–153.

Van de Groep, I. H., Bos, M. G., Jansen, L. M., Achterberg, M., Popma, A., & Crone, E. A. (2021). Overlapping and distinct neural correlates of self-evaluations and self-regulation from the perspective of self and others. *Neuropsychologia*, *161*, 108000. https://doi.org/10.1016/j.neuropsychologia.2021.108000

Van de Groep, I. H., Bos, M. G., Jansen, L. M., Kocevska, D., Bexkens, A., Cohn, M., Van Domburgh, L., Popma, A., & Crone, E. A. (2022). Resisting aggression in social contexts: The influence of life-course persistent antisocial behavior on behavioral and neural responses to social feedback. *NeuroImage: Clinical*, *34*, 102973. https://doi.org/10.1016/j.nicl.2022.102973

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. O. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00770

Van der Laan, M. J., & Rose, S. (2011). *Targeted Learning*. Springer New York. https://doi.org/10.1007/978-1-4419-9782-1

Van Harmelen, A.-L., Blakemore, S. J., Goodyer, I. M., & Kievit, R. A. (2021). The interplay between adolescent friendship quality and resilient functioning following childhood and adolescent adversity. *Adversity and Resilience Science*, *2*(1), 37–50. https://doi.org/10.1007/s42844-020-00027-1

Van Harmelen, A.-L., Kievit, R. A., Ioannidis, K., Neufeld, S., Jones, P. B., Bullmore, E., Dolan, R., The NSPN Consortium, Fonagy, P., & Goodyer, I. (2017). Adolescent friendships predict later resilient functioning across psychosocial domains in a healthy community cohort. *Psychological Medicine*, *47*(13), 2312–2322. https://doi.org/10.1017/S0033291717000836

Van Lissa, C. J., Keizer, R., Van Lier, P. A. C., Meeus, W. H. J., & Branje, S. (2019). The role of fathers' versus mothers' parenting in emotion-regulation development from mid–late adolescence: Disentangling between-family differences from within-family effects. *Developmental Psychology*, *55*(2), 377–389. https://doi.org/10.1037/dev0000612

Van Minnen, A., Arntz, A., & Keijsers, G. (2002). Prolonged exposure in patients with chronic PTSD: Predictors of treatment outcome and dropout. *Behaviour Research and Therapy*, *40*(4), 439–457. https://doi.org/10.1016/S0005-7967(01)00024-9

Van Minnen, A., & Hagenaars, M. (2002). Fear activation and habituation patterns as early process predictors of response to prolonged exposure treatment in PTSD. *Journal of Traumatic Stress*, *15*(5), 359–367. https://doi.org/10.1023/A:1020177023209

Van Minnen, A., Hendriks, L., Kleine, R. D., Hendriks, G.-J., Verhagen, M., & De Jongh, A. (2018). Therapist rotation: A novel approach for implementation of trauma-focused treatment in post-traumatic stress disorder. *European Journal of Psychotraumatology*, *9*(1), 1492836. https://doi.org/10.1080/20008198.2018.1492836

Van Minnen, A., Voorendonk, E. M., Rozendaal, L., & De Jongh, A. (2020). Sequence matters: Combining Prolongued Exposure and EMDR therapy for PTSD. *Psychiatry Research*, *290*, 113032. https://doi.org/10.1016/j.psychres.2020.113032

Van Woudenberg, C., Voorendonk, E. M., Bongaerts, H., Zoet, H. A., Verhagen, M., Lee, C. W., Van Minnen, A., & De Jongh, A. (2018). Effectiveness of an intensive treatment programme combining prolonged exposure and eye movement desensitization and reprocessing for severe post-traumatic stress disorder. *European Journal of Psychotraumatology*, *9*(1), 1487225. https://doi.org/10.1080/20008198.2018.1487225

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/9780429492259

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

VanderWeele, T. J. (2012). Invited commentary: Structural equation models and epidemiologic analysis. *American Journal of Epidemiology*, *176*(7), 608–612. https://doi.org/10.1093/aje/kws213

VanderWeele, T. J., & Vansteelandt, S. (2010). VanderWeele and Vansteelandt respond to "Decomposing with a lot of supposing" and "Mediation". *American Journal of Epidemiology*, *172*(12), 1355–1356. https://doi.org/10.1093/aje/kwq331

VanderWeele, T. J. (2018). On well-defined hypothetical interventions in the potential outcomes framework. *Epidemiology*, *29*(4), e24–e25. https://doi.org/10.1097/EDE.0000000000000823

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, *34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

VanderWeele, T. J. (2021). Causal inference with time-varying exposures. In T. L. Lash, T. J. VanderWeele, S. Haneuse, & K. J. Rothman (Eds.), *Modern epidemiology* (pp. 605–618). Wolters Kluwer.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, *167*(4), 268. https://doi.org/10.7326/M16-2607

Vangeel, L., Vandenbosch, L., & Eggermont, S. (2018). The multidimensional self-objectification process from adolescence to emerging adulthood. *Body Image*, *26*, 60–69. https://doi.org/10.1016/j.bodyim.2018.05.005

Vansteelandt, S., & Daniel, R. (2014). On regression adjustment for the propensity score. *Statistics in Medicine*, *33*(23), 4053–4072. https://doi.org/10.1002/sim.6207

Vansteelandt, S., & Joffe, M. (2014). Structural nested models and G-estimation: The partially realized promise. *Statistical Science*, *29*(4). https://doi.org/10.1214/14-STS493

Vansteelandt, S., & Sjolander, A. (2016). Revisiting g-estimation of the effect of a time-varying exposure subject to time-varying confounding. *Epidemiologic Methods*, *5*(1), 37–56. https://doi.org/10.1515/em-2015-0005

Vijayakumar, N., Pfeifer, J. H., Flournoy, J. C., Hernandez, L. M., & Dapretto, M. (2019). Affective reactivity during adolescence: Associations with age, puberty and testosterone. *Cortex*, *117*, 336–350. https://doi.org/10.1016/j.cortex.2019.04.024

Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17*(2), 176–192. https://doi.org/10.1037/a0027543

Von Collani, G., & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *24*(1), 3–7. https://doi.org/10.1024//0170-1789.24.1.3

Wallace, M. P., Moodie, E. E. M., & Stephens, D. A. (2017). Dynamic treatment regimen estimation via regression-based techniques: Introducing R package DTRreg. *Journal of Statistical Software*, *80*(2). https://doi.org/10.18637/jss.v080.i02

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592091825. https://doi.org/10.1177/2515245920918253

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The PTSD Checklist for DSM-5 (PCL-5). https://www.ptsd.va.gov

Welsh, M., & Peterson, E. (2014). Issues in the conceptualization and assessment of hot executive functions in childhood. *Journal of the International Neuropsychological Society*, *20*(2), 152–156. https://doi.org/10.1017/S1355617713001379

Wichstraum, L. (1995). Harter's Self-Perception Profile for adolescents: Reliability, validity, and evaluation of the question format. *Journal of Personality Assessment*, *65*(1), 100–116. https://doi.org/10.1207/s15327752jpa6501_8

Winkens, B., Schouten, H. J., Van Breukelen, G. J., & Berger, M. P. (2006). Optimal number of repeated measures and group sizes in clinical trials with linearly divergent treatment effects. *Contemporary Clinical Trials*, *27*(1), 57–69. https://doi.org/10.1016/j.cct.2005.09.005

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data* (6th). MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th). South-Western Cengage Learning.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215. https://doi.org/10.1214/aoms/1177732676

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*(5), 737–757. https://doi.org/10.1177/0013164404264853

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*(3), 405–426. https://doi.org/10.1037/met0000080

Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, n/a–n/a. https://doi.org/10.1111/j.1750-8606.2012.00246.x

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, *46*(4), 1184–1198. https://doi.org/10.3758/s13428-013-0424-0

Zhang, Z., & Liu, H. (2018). Sample size and measurement occasion planning for latent change score models through Monte Carlo simulation. In E. Ferrer, S. M. Boker, & K. J. Grimm (Eds.), *Longitudinal Multivariate Psychology* (1st ed., pp. 189–211). Routledge. https://doi.org/10.4324/9781315160542-10

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From data to causes I: Building a general cross-lagged panel model (GCLM). *Organizational Research Methods*, *23*(4), 651–687. https://doi.org/10.1177/1094428119847278

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E.

(2020). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods*, *23*(4), 688–716. https://doi.org/10.1177/1094428119847280

Nederlandse samenvatting

Structureel vergelijkingsmodeleren (in het Engels “structural equation modeling”, kortweg “SEM”) is een breed toepasbare techniek voor statistische analyses. Het is voornamelijk populair als analysemethode onder psychologen en sociale- en gedragswetenschappers, en vele verschillende SEM-modellen zijn reeds ontwikkeld voor een breed scala aan onderzoeksvragen en toepassingen. Eén van de unieke kenmerken van SEM is de mogelijkheid om “latente variabelen” te definiëren. Dit zijn variabelen die niet direct geobserveerd kunnen worden, maar worden afgeleid op basis van één, of meerdere gemeten “indicatoren”. Deze latente variabelen kunnen worden gebruikt om inhoudelijke concepten te meten —bijvoorbeeld de mate van depressie bij een persoon aan de hand van diens antwoorden op een depressie-vragenlijst—maar ze worden ook gebruikt voor puur statistische concepten, zoals meetfout, clusters, of groei-componenten.

Dit proefschrift gaat over het toepassen, het verder ontwikkelen, en het kritisch evalueren van SEM modellen, specifiek voor de analyse van longitudinale observationele data. De term “longitudinaal” refereert hier naar een studie-opzet waarbij participanten herhaaldelijk zijn gemeten (bijvoorbeeld, elke maand, voor een jaar lang). De term “observationeel” refereert naar de niet-experimentele aard van de metingen: Er is geen interventie toegepast door onderzoekers om de score van participanten op bepaalde variabelen te manipuleren, er is slechts passief geobserveerd/-gemeten. Om een goede afweging te kunnen maken over welke specifiek longitudinaal SEM model gebruikt kan worden, moet eerst duidelijk zijn of een onderzoeksvraag *beschrijvend*, *voorspellend*, of *oorzakelijk* van aart is. Dit onderscheid is belangrijk omdat de onderliggende problemen/zorgen bij statistische analyses verschillen per type onderzoeksvraag.

Bij beschrijvend onderzoek is het voornaamste doel om eigenschappen van de data, van groepen, of van personen, samen te vatten. Vaak wordt hierbij gekeken naar (cor)relaties tussen variabelen, maar worden deze niet geïnterpreteerd als oorzakelijk. Bij voorspellend onderzoek is het primaire doel om nieuwe voorspellingen te doen op basis van reeds gemeten data. Deze voorspellingen kunnen worden gebruikt om bepaalde individuen te selecteren, te monitoren, of te screenen. Een groot punt van aandacht is hier het minimaliseren van de voorspellingsfout *bij nieuwe data* ("out-of-sample"). Bij oorzakelijk onderzoek is het doel om inzicht te krijgen in het onderliggende causale mechanisme (of een specifiek deel daarvan) van een bepaald proces. Om dit te doen op basis van observationele data is er gedegen theorie nodig over het proces wat onderzocht wordt. Een van de grootste zorgen bij dit type onderzoek is het voorkomen van schijnverbanden, dat zijn misleidende relaties tussen variabelen die niet daadwerkelijk oorzakelijk zijn.

In dit proefschrift worden SEM modellen toegepast, verder ontwikkeld, en geëvalueerd voor zowel beschrijvend, voorspellend, als oorzakelijk onderzoek. Verder zijn de hoofdstukken een mix van toegepast en methodologisch onderzoek, en betreft het onderzoek in verschillende disciplines: van neurowetenschappen en klinische psychologie, tot epidemiologie en biostatistiek.

Hoofdstuk 2 is een samenwerking met dr. Michelle Achterberg en dr. Simone Dobbelaar, en betreft de emotieregulatie van kinderen. Dit toegepaste longitudinale onderzoek is tweedelig. Allereerst beschreven wij de algemene ontwikkeling van neurale- en gedragsresponse op sociale afwijzing bij kinderen in de leeftijd van 7 tot 13 jaar oud, evenals individuele verschillen in de ontwikkeling. Hiervoor is gebruik gemaakt van een Bayesiaans multilevel model. Ten tweede onderzochten wij de relaties tussen deze individuele verschillen in ontwikkeling van kinderen en hun sociaal-welbevinden in hun vroege adolescentie. Hiervoor hebben wij gebruik gemaakt van een multivariaat regressiemodel binnen SEM.

Hoofdstuk 3 is een toegepast, longitudinaal project samen met Valentijn Alting van Geuseau en dr. Suzy Matthijsen. Het betreft een recent ontwikkeld, tweeweken durend klinisch behandelprogramma voor patiënten met een posttraumatische stressstoornis (PTSS) bij Altrecht, een instelling voor geestelijke gezondheidszorg. Wegens kostenoverwegingen en de hoge mate van patiënten-uitval was er interesse in het vroegtijdig voorspellen of een patiënt in het behandelprogramma baat zou hebben bij het afmaken van de behandeling. Het doel van deze studie was dus om PTSS-afname vier weken *na* het behandelprogramma te voorspellen op basis van de PTSS-symptoomontwikkeling *gedurende* het programma. Hiervoor is gebruik gemaakt van k-voudige kruisvalidatie om de "out-of-sample"-voorspellingsfout van vijf verschillende latente groeicurve-modellen te vergelijken.

Hoofdstuk 4 is een samenwerking met prof. dr. Ellen Hamaker waarin we drie uitbreidingen van het *random intercept cross-lagged panel model* (RI-CLPM) hebben besproken, namelijk (a) het includeren van een stabiele tussenpersoon-variabele als voorspeller of uitkomst in het model; (b) het specificeren van een meervoudige-groep-extensie; en (c) het includeren van meerdere indicatoren voor latente constructen binnen het model. Een belangrijk onderdeel van dit project is een website die als online bijlage dient. Hierop kunnen lezers R code en Mplus syntax vinden, evenals voorbeelddata en antwoorden op veelgestelde vragen.

Hoofdstuk 5 introduceert een strategie voor poweranalyse specifiek voor het RI-CLPM. Deze strategie is ontworpen om zo gebruiksvriendelijk mogelijk te zijn, en is tevens geïmplementeerd in de R package powRICLPM.

Hoofdstuk 6 is een samenwerking met dr. Kim Luijken, dr. Bas Penning de Vries, en dr. Ellen Hamaker. Het doel van deze studie was tweeledig. Ten eerste diende de studie om inzicht te geven in kritiek vanuit de causale inferentie literatuur op het gebruik van SEM-modellen voor causaal onderzoek. Hiertoe hebben we eerst beschreven hoe het gebruik van SEM past binnen het veelgebruikte *potential outcomes* kader, en hoe het in verhouding staat tot een andere analysemethode, namelijk het gebruik van *marginal structural models* met *inverse probability weighting* (IPW-MSM). Ten tweede, hebben we in deze studie onderzocht wat de *finite-sample performance* is van padanlayse (een SEM-methode) en IPW-MSM bij schendingen van parametrische aannames waar de methoden op berusten. Hierbij is gebruik gemaakt van simulaties.

Hoofdstuk 7 is een samenwerking met dr. Satoshi Usami en dr. Ellen Hamaker. We hebben cross-lagged effecten met zogenaamde *joint effecten* vergeleken, evenals twee methodes om deze effecten te schatten: cross-lagged panel modellen (SEM-modellen), en *structural nested mean modellen* in combinatie met *G-estimation* vanuit de *potential outcomes* literatuur. Hierbij hebben we gebruik gemaakt van een empirisch psychologisch voorbeeld over zelfvertrouwen en depressie. Daarnaast leveren we een bijdrage aan de integratie tussen de SEM literatuur en de *potential outcomes* literatuur door onderwerpen te bespreken die voornamelijk óf in de SEM literatuur, of in de *potential outcomes* literatuur worden besproken, maar niet in beide: Deze onderwerpen betreffen (a) het scheiden van tussen-persoon- en binnen-persoonvariantie, (b) het gebruik van lag-0 relaties, en (c) het gebruik van lag-2 relaties.

Curriculum vitae

2023 – present **Universiteit Utrecht, Utrecht**
Humboldt-Universität, Berlin

Postdoctoral researcher at the department of Methodology and Statistics as part of the Dynamic Modeling Lab of Dr. Ellen Hamaker. Furthermore, from March 2024 through August 2024, Jeroen is a visiting researcher at the Methodegruppe at the Humboldt-Universität lead by dr. Manuel Voelkle. His postdoc project involves (a) the use machine learning techniques to estimate propensity scores in longitudinal models, which can then be used in causal inference in psychological research; and (b) the comparison of causal inference methods from epidemiology and biostatistics (e.g., g-methods such as inverse probability weighting estimation of marginal structural models, and structural nested mean models), to structural equation modeling methods popular in psychological research.

2019 – 2023 **Universiteit Utrecht, Utrecht**

PhD candidate at the department of Methodology and Statistics under supervision of Dr. Ellen Hamaker and Dr. Satoshi Usami. Jeroen was funded by the Consortium on Individual Development, which was supported by a Gravitation grant awarded by the Dutch Ministry of Education, Culture, & Science, and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003; 2013). His project concerned the development and evaluation of longitudinal structural equation models for investigating prospective causal relationships between variables. During this time, he was also involved in education for bachelors, masters, and professionals. For his research, he collaborated with researchers from clinical psychology, neuroscience, as well as epidemiology and biostatistics.

2020 – 2021 NPO Radio 2 (BNNVARA), Hilversum

Jeroen was involved as an editor, producer, and presenter of the radio program WILDGROEI on NPO Radio 2, and as a substitute for the shows De Staat van Stasse (KRO-NCRV) and Giel (BNNVARA).

2018 – 2019 Universiteit Utrecht, Utrecht

After completion of his master's program, Jeroen worked as a junior teacher at the Methods and Statistics department of the Faculty of Social and Behavioural Sciences at Utrecht University. He was involved in statistical education for both bachelor and master students.

2016 – 2018 Universiteit Utrecht, Utrecht

Jeroen obtained his master's degree in methodology and statistics for the behavioural, biomedical, and social sciences at Utrecht University. For his elective, he completed courses on econometric methods, data analysis and visualization, and mark-up languages. He wrote a dissertation on the impact of restriction of range in measurements (e.g., floor effects) in intensive longitudinal data for the estimation of autoregressive effects.

2013 – 2016 Universiteit Twente, Enschede

Jeroen obtained his bachelor's degree in communication science (cum laude) at the University of Twente, and with a minor in technical computer science. For his bachelor's thesis he investigated the usability of a newly developed website, https://www.ikstopnu.nl, which aims to support individuals with smoking cessation.

Publications

Published manuscripts

Mulder, J. D., Luijken, K., Penning-de Vries, B. B. L., & Hamaker, E. L. (2024). Causal effects of time-varying exposures: A comparison of structural equation modeling and marginal structural models in cross-lagged panel research. *Structural Equation Modeling: A Multidisciplinary Journal.*https://doi.org/10.1080/10705511.2024.2316586

Mulder, J. D., Dobbelaar, S., & Achterberg, M. (2024). Behavioral and neural responses to social rejection: Individual differences in developmental trajectories across childhood and adolescence. *Developmental Cognitive Neuroscience*, *66*, 101365. https://doi.org/10.1016/j.dcn.2024.101365

Mulder, J. D. (2023). Power analysis for the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(4), 645-658. https://doi.org/10.1080/10705511.2022.2122467

Alting van Geusau, V. V. P., **Mulder, J. D.**, & Matthijssen, S. J. M. A. (2021). Predicting outcome in an intensive outpatient PTSD treatment program using daily measures. *Journal of Clinical Medicine*, *10*, 4152. https://doi.org/10.3390/jcm10184152

Mulder, J. D, & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 638-648. https://doi.org/10.1080/10705511.2020.1784738

Hamaker, E. L., **Mulder, J. D.**, & Van IJzendoorn, M. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental Cognitive Neuroscience*, *46*, 100867. https://doi.org/10.1016/j.dcn.2020.100867

Submitted manuscripts

Mulder, J. D., Usami, S., & Hamaker, E. L. (submitted). Joint effects in panel data using structural nested mean models: An introduction for psychologists familiar with cross-lagged panel research.

Acknowledgements (dankwoord)

It was Saturday December 8, 2018, in Berlin. Still slightly hungover, Kim, Anna Sophia, Dario, and I were having breakfast at Dario's home after a night clubbing. The four of us had graduated from the master Methodology and Statistics for the Behavioural, Biomedical and Social Sciences that summer, and we were discussing our future plans. I mentioned that Ellen Hamaker, our structural equation modeling teacher, and my master thesis supervisor, had offered me a PhD position: There was no funding yet for a PhD project, but she expected that in the near future some money from a large "Gravitation grant" would become available to her. I was uncertain about whether or not to accept the offer. On the one hand it was a great, and unique possibility to continue learning about statistics, and to work with someone who was renowned for her contributions to the statistical literature. On the other hand, the prospect of committing myself to a four-year long research project daunted me. What if I didn't like it? On top of that, some friends and former colleagues did not think I was the kind of person who would "sit behind a desk the entire day". Ultimately, it was something that Kim said that morning in Berlin, that made me make a decision: "In the end, a PhD position is a job, right? If it doesn't make you happy, you can just quit."

Now, after four and a halve years of doing a PhD, I can confidently say that there wasn't a single moment in which I regretted my decision. A major reason for this is that I have always felt right at home at the Methodology and Statistics department at Utrecht University. There is such a diverse group of professionals working here, each with their own area of expertise, whether it is research-, education-, or support-related. I have found every colleague to be willing to help whenever I had a question, to listen whenever I had to get something off my chest, and to engage in academic discussion whenever we were discussing complex statistical and methodological topics. I am happy that I can continue working here in the future. Therefore, to all my colleagues, and my colleagues from the Dynamic Modeling Lab in particular: Thank you.

I would also like to extend a word of thanks to my friends from the research master. At the start of the study, in September 2016, I never expected that together we would have gone on so many adventures. We have been to Geneva (Switzerland), Berlin (Germany), Mérida (Mexico), Bogotá (Colombia), Amboise (France), Santander (Spain), and many other places; Not to mention our activities in the Netherlands such as joining in the Batavierenrace, or going to Oerol. What makes this so special is that at the same time, we share a passion for research and methodology and statistics. This regularly led to discussions over dinner about causal inference and statistical methods, and we could share our experiences of starting a career in academia. A discussion I remember vividly is one we were having in Colombia, as we were hurrying to get back to Bogotá in time before all regional borders were closed: How to best get an accurate estimate of the number of people infected with COVID-19? I look forward to our future (academic) adventures.

Uiteraard wil ik ook mijn familie bedanken. Alhoewel jullie niet *direct* betrokken zijn bij deze dissertatie, helpt een goede familieband absoluut om het beste uit jezelf te halen op het werk. Ik heb me altijd gesteund gevoeld door iedereen, en vind het bijzonder dat we samen, nog altijd, zoveel plezier hebben tijdens verjaardagen, vakanties, verbouwingen, feesten, enz. Ook wil ik graag kort aandacht geven aan familie die helaas niet meer onder ons is. Van beide kanten mijn opa en oma die, met hun persoonlijke interesse in de ander en de geweldige vakanties samen, aan de wieg hebben gestaan van de familie zoals ik die nu heb. En uiteraard ook Olav, mijn oom. Ik had hem graag dit dankwoord willen laten lezen. Hij had geheid een aantal goede schrijftips weten aan te dragen.

A special word of thanks goes out to my co-promotor Satoshi Usami. His work combining the use of longitudinal SEM models with techniques from the causal inference literature has been a huge inspiration for me. I think that his research at the USAMI Lab at the University of Tokyo is absolutely essential for the development of SEM as an approach for causal inference. I look forward to potential collaborations in the future.

Maar boven alles wil ik Ellen bedanken. Vierenhalf jaar aan samenwerking is moeilijk samen te vatten in één alinea. Als ik onze samenwerking omschrijf aan anderen, dan benoem ik altijd dat jij mij het gevoel geeft dat je me *serieus* neemt. Zo neem je me mee in academische discussies waarin je verwikkeld bent geraakt, mag ik met jou een workshop geven in Finland, ben je open over hoe het is om als vrouw hoogleraar te zijn in een werkveld waarin de meerderheid van de professoren man is, en vraag je naar mijn mening over recent gepubliceerde artikelen en statistische analyses. Dat alles was niet alleen zeer leerzaam voor mij als promovendus, maar zorgde er ook voor dat ik me altijd gezien heb gevoeld. En dan heb ik het nog geeneens gehad over je enorme hoeveelheid vakinhoudelijke kennis, je schrijfvaardigheden, en je academische netwerk waarvan ik heb mogen profiteren. Daar ben ik je zeer dankbaar voor.