FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

Statistical models for causal mediation in within-subject designs:

Dealing with unmeasured confounders and interactions

Haeike Josephy

Supervisor: Prof. Dr. Tom Loeys
Co-supervisor: Prof. Dr. Stijn Vansteelandt

A dissertation submitted to Ghent University in partial fulfillment of the requirements for the degree of Doctor in Psychology

Academic Year 2017-2018

GHENT
UNIVERSITY

The most beautiful experience we
can have is the mysterious.
It is the fundamental emotion that
stands at the cradle of true art and
true science.

Albert Einstein

It's the questions we can't answer
that teach us the most. They teach
us how to think. If you give a man
an answer, all he gains is a little fact.
But give him a question and he'll
look for his own answers.

Patrick Rothfuss
The Wise Man's Fear

Science, my lad, is made up of
mistakes, but they are mistakes
which it is useful to make, because
they lead little by little to the truth.

Jules Verne
A Journey to the Center of the Earth

Table of Contents

Acknowledgments

Although I do enjoy a bit of writing now and then, I am slightly less keen on writing thanks, because acknowledgements will never quite live up to the people who helped me along the way or what they meant to me at the time. But I suppose no book is complete without its proper thanks-giving, so I'l have a go at it anyway (don't laugh, but yes, I do feel like I have produced some kind of extremely obscure book in writing this PhD).

First of all, I would like to thank all the authors, artists, and other creative minds that helped me lose myself in fantasy and fiction. The stories in books, movies, comics, and games allowed me to ride out every bump along the way, because when I read, the world around me fades into nonexistence. And when only the world in the story matters, if only for a few hours, I manage to put everything back into perspective and continue with my own adventure.

Next in line are my colleagues, collaborators, and proof-readers. I want to thank Tom for pitching me to the department, hauling me in, and seeing me all the way through. The road towards finishing this dissertation didn't always go smoothly, as the last six years were definitely not the easiest, nor the most straightforward ones I experienced up until now. But through mutual understanding and communication we found our way, and I think we both learned a great deal in doing so. Also a big thank you to Bieke, for entrusting me with your courses (even when failing to open the exercise subscriptions on time on several occasions, flooding you with the indignant mails of so many students), and to Rudi and Stijn for the much-appreciated advice and feedback during our meetings. I also want to express my enormous appreciation for Isabelle, I think no-one can quite figure out how you manage to organise an entire department within the blink of an eye. Thanks to all my fellow PhD students throughout the years, for making this arduous task that much brighter, with movies-guessing, games, plants, good stories, and the occasional drinks. A special thanks to Jacob, you truly surprised me in finding a friend at the office; you always knew the right time for talks, beer, or games when I needed them. And also to Justine, for making work and conferences that much more interesting, through leg-kicking in bars and way-finding in London when subways had

inconveniently closed. Finally, thank you Isabelle, Kris, and Danielle for reading the introduction of this PhD and helping me keep it 'readable' for non-nerds.

Of course, I also want to express my gratitude towards my friends, but since it seems like an impossible endeavour to mention everyone who deserves to be in here, I'll resort to some epic memories. A warm thanks to all the nights where we appreciated (both good and extremely bad) movies together, for the times when we slayed dragons, warred against undead fiends, or simply annoyed each other until exasperation during roleplaying sessions or (board)gaming afternoons. Thank you for all those evenings spent in bars with beer or whisky tastings, some of them ending in a notorious place we all know but dare not mention. An even warmer appreciation to all the exhausting yet satisfying training sessions in fencing, archery, and so many other sports. Thank you for the travels, the laughs, the talks, the silly dances, and your invaluable support.

I also wish to express the sincerest of thanks to my family. Not everyone who encouraged me in starting this endeavour is here today to see me finish it. But I am sure they would be extremely proud, and if they could, would throw easter eggs in the sky (with a white flurry of paws catching them), make me lobster bellevue (with someone watching it with undying interest), and drink a good glass of wine (or Amaretto) to my health (with someone knocking it over). Thank you dad for the French retreats when I was in dire need of them, 'de Van Neckskes' for always being up for a much-welcome gathering of minds, food-happenings, and other (often food-related) activities. Especially to Jonathan and Marieke, I'm so very grateful you guys moved here and I got to see more of the two of you. I hope you stick in Ghent, so we can continue the drinks, the food, the plant-, cat-, and house-sitting for a lot of years to come. Maxime, although you make up but a recent addition to this paragraph, you cannot begin to imagine what you meant in supporting me through the tough home stretch (even setting aside all of our bad jokes). I have to admit that I found a strangely gratifying comfort in you going through a similar experience while completing you Master's thesis. I am fondly looking forward towards a future with a profusion of eighties music, a perhaps slightly curtailed amount of explanations concerning synchronous machines, lots and lots of sandwiches with Nutella, and an ever higher abundance of bad jokes (although we should really try and limit those to Sundays).

Finally, the most noteworthy mention goes to my mom, who I want to thank from the bottom of my heart. With you in my life, backing me up every single step of the way (the occasional mishap included), I could move mountains if I needed to. You supported me through wondrous new experiences as well as heartaches, through achievements and setbacks, with all the joys, pain, and uncertainties those entail. I know you will be there for me, always, and for that I am truly, deeply grateful.

Haeike,
June 2018

1 Introduction

Although the pursuit of knowledge has been around for a long time, science as a popular vocation rather than an eccentric one, is a relatively recent phenomenon. Of course we know of notable individuals in our past who helped science get where we are today, but rather than surveying their job titles for the term 'scientist', history acknowledges such individuals in their quest for truth, improvement, and innovation. These unique qualities are still very astute today, as there is one major concept that will assumably drive you if you choose a career in science: your need to understand. You will want to grasp how everything works, to try and unravel how things are the way they are and why they do the things they do. In order to evaluate this, you will need to understand the hidden processes underlying the object that sparked your interest. Consider for example communication processes, where researchers not only seek to establish whether or not messages have an effect in a specific context, but would much rather try to understand *how* these messages influence their recipient. Alternatively, in medicinal research we may discover that a newly developed vaccine has a detrimental effect on the maturation of a specific disease. Although this, in and by itself, is groundbreaking and inspiring news, the experimenters won't feel satisfied until they truly fathom *how* the vaccine sways our immune system to fight off the infection. Likewise, many epidemiologists promote the opening of their 'black boxes' as to try and elucidate explanatory theories for *how* diseases arise, rather than solely taking comfort in identifying risk factors and leaving it at that. The same reasoning holds for research in psychology: we want to *understand* the psychological processes by which interventions affect our behaviour. When we observe a decline in ruminative thinking after depressed patients receive a subtle electric stimulation of a specific brain area, psychologists will want to figure out which underlying mechanisms are responsible. Or in other words, they wish to identify

the processes that *mediate* the relationship between an intervention and its effect.

1.1 Mediation

Mediation can be described as the collective processes that disclose an observed relation between an intervention and its response, be it in either communication, medicine, epidemiology, social sciences, or any other discipline. As such, the fundamental enquiry into how and why things work lies at the essence of any mediation analysis. At its core, it attempts to jump into the world of causality by trying to discern any intermediary steps between cause and effect.

In order to determine the effect that an independent variable may have on a dependent variable, experimenters irrevocably require the assistance of statistical or mathematical models. This dependent variable represents the object or measure that is being studied, while the independent variable represents a specific input or cause. As such, the message, vaccine, cause of infection, or electric stimulation of which we spoke earlier, are defined as independent variables or exposures X. This annotation describes our interest in the effect that these variables have, when someone is being *exposed* to them. Equivalently, the message impact, disease resistance, disease manifestation, or amount of rumination are defined as dependent variables or outcomes Y. These are appropriately labelled 'dependent', since we are interested in their *dependency* on the exposures. To evaluate and assess the relationship between exposure and outcome, researchers will gather information from a sample of individuals. This sample will be summarised in a data set, where each individual i ($i = 1...n$ with n representing the sample size) contributes one value for the exposure, X_i, and one for the outcome, Y_i. When the outcome is a continuous measure (such as weight, height or age), a linear regression model represents the most convenient way of capturing the relationship between exposure and outcome:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad \text{with } \epsilon_i \sim N(0, \sigma^2) \tag{1.1}$$

In such an equation, we attempt to model and predict the outcome Y_i in terms of the exposure X_i. This is achieved through the estimation of an intercept, β_0, and a slope or regression parameter, β_1. By means of these two parameters, we aim to transform any value for the exposure

into a corresponding measure for the outcome. Of course, this conversion will never be perfect, which is why the error terms ϵ_i are included; they represent the difference between the true values of the outcome (i.e., Y_i) and the predicted/transformed outcome measures (i.e., $\beta_0 + \beta_1 X_i$). These errors terms are assumed to be independently distributed (of each other and the exposure X_i), following a normal distribution with constant variance (σ^2) and *mean zero*. This implies that, *on average*, the outcome will equal $\beta_0 + \beta_1 X_i$ for a given value of the exposure. Consequently, the intercept can be interpreted as the mean value of the outcome, when the exposure equals zero (i.e., when X_i is zero then, on average, Y_i equals $\beta_0 + \beta_1 \cdot 0 = \beta_0$). Equivalently, the slope parameter β_1 can be interpreted as the association between exposure and outcome: if X_i changes from zero to one, the mean outcome increases by β_1 (i.e., when X_i is set to one then, on average, Y_i will equal $\beta_0 + \beta_1 \cdot 1 = \beta_0 + \beta_1$). If, however, the true value of β_1 equals zero, the exposure and outcome will not show any association (i.e., Y_i will, on average, equal $\beta_0 + 0 \cdot X_i = \beta_0$). In any case, the slope parameter β_1 describes the relationship between the exposure and the outcome: when the regression parameter equals zero, no relationship between both variables exists, but when β_1 differs from zero, it expresses the size and direction of the association between exposure on outcome.

1.1.1 Causal effects

When an exposure and an outcome are related (i.e., when $\beta_1 \neq 0$), we can conclude the presence of an association between both variables. In practice, however, researchers are most often interested in testing causal relations, rather than mere associations. In order to label the association between exposure and outcome in equation (1.1) a causal effect, two conditions need to be fulfilled: (1) the exposure X needs to temporally precede the outcome Y, and (2) any variable that confounds the relationship between exposure and outcome needs to be taken into account. The first rule is rather straightforward to understand, since the exposure cannot possibly cause the outcome if the former occurs after the latter. The second condition is less intuitive, but can be easily understood through an example. Consider the following well-documented association: people with yellow stains on their fingers tend to have a worse lung function and a greater risk of lung cancer, compared to people with unstained fingers. This negative association does *not* represent a causal effect, however, as painting your fingers will not inherently worsen your lung capacity, nor

will you suddenly develop lung cancer. Instead, the observed relationship between yellow fingers and lung cancer constitutes the byproduct of a variable that influences both X and Y simultaneously: smoking behaviour. Smoking will affect the colouring of your hands, as the nicotine in cigarette smoke can leave nasty-looking stains on your fingers. Additionally, people who smoke will diminish their lung function over time, as the chemicals found in cigarettes damage key genes that protect us against cancer. As such, smoking *confounds* the relationship between yellow fingers and lung function: the observed association between X and Y is due to a common cause (i.e., smoking), rather than a causal effect between both variables. Consequently, in order for association to become causation, the relation between X and Y must persist when all such confounding variables are taken into account.

As any mediation analysis attempts to discern intermediary steps between a cause and its effect, mediation is, in essence, a causal process. Consequently, we need to establish and investigate causal relations, rather than associations. Graphically, causal effects of an exposure on an outcome can be represented by an arrow originating from X and arriving at Y (see the left part of figure 1.1); this causal pathway is often labelled the *total effect* of an intervention on a specific outcome.

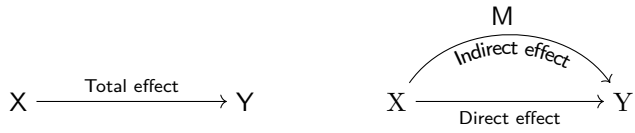

Figure 1.1 **Left.** The total effect of the exposure X on the outcome Y. **Right.** The total effect of X on Y is split up into two arrows: an indirect effect (upper arrow) and a direct effect (lower arrow).

Mediation amounts to deducing the mechanism through which the exposure influences the outcome of interest; any variables fitting this description are labelled a possible mediator M. The question of whether or not the causal effect of X on Y (partly) runs through a specific mediator, can be answered by decomposing the total effect into two parts (see right side of figure 1.1). One arrow is defined as the *indirect* or intervening effect (as it indirectly runs from the exposure to the outcome through the mediator), while the other is branded the *direct effect* (i.e. the remaining effect that does not pass through M). Consequently, we can assess mediation by evaluating the existence of the arrow that represents the intervening effect. An example of mediation in experimental psychology consists of figuring

out why ruminative thinking (the outcome Y) declines after depressed patients receive a subtle electric stimulation to the brain (the exposure X). A possible mediator might be found in cognitive control, as this mental process allows a person to override his or her impulses by making decisions based on goals, rather than habits or reactions (Vanderhasselt et al., 2013). Keeping this definition in mind, it seems very plausible that people with a high cognitive control may have an easier time in redirecting their attention from ruminative thoughts to more constructive ones. As such, a slight electric stimulation might decrease rumination *indirectly* by increasing the amount of cognitive control a person possesses.

Although the idea behind mediation has undoubtedly been around as long as curiosity itself, the origins of its conceptualisation are far more recent. As far as we know, Wright (1934) was the investigatory pioneer to put the first form of mediation analysis to paper. However, it wasn't until after the introduction of Baron and Kenny (1986)'s causal steps that mediation analysis began to flourish. Since their research has had such a great impact on social sciences in particular, let us take a closer look at their groundbreaking work.

1.1.2 The causal steps approach

So once you have a valid research hypothesis concerning mediation, how do you answer it? In response to this question, Baron and Kenny (1986) proposed a framework in which mediation is assessed in four consecutive steps. It relies on the following three regression equations:

$$\begin{aligned} Y &= i_{Y1} + cX + e_{Y1} \\ M &= i_M + aX + e_M \\ Y &= i_{Y2} + c'X + bM + e_{Y2} \end{aligned} \tag{1.2}$$

As we saw in the previous section, each equation can be represented by a number of arrows equal to the amount of independent variables it contains. As such, these three equations are captured by four distinct projectiles: the first equation translates into an arrow flowing from X to Y (see left part of figure 1.2), the second into an arrow going from X to M (right side of figure 1.2), and the third into two arrows converging at the outcome: one emanating from X and the other from M (also right side of figure 1.2). These arrows/effects can be summarised by the slope parameters belonging to the independent variables from which they originate (c, a,

c', and b, respectively). As before, these regression parameters describe the relationship between the dependent and independent variables: when a slope parameter equals zero, the corresponding variables will lack a connecting arrow, while a slope coefficient different from zero will encode the strength of their connection.

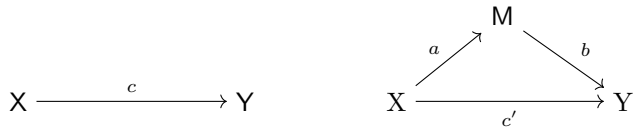

Figure 1.2 **Left.** The regression of the exposure X on the outcome Y is translated into an arrow flowing from X to Y. This arrow is marked by the regression coefficient c. **Right.** The regression of X on M can be depicted by an arrow going from X to M; this arrow is labelled by the slope parameter a. The simultaneous regression of X and M on Y can be represented by two arrows: one flowing from X to Y (symbolised by c') and one going from M to Y (symbolised by b).

Keeping this in mind, Baron and Kenny (1986) assess mediation in four consecutive steps:

- *The first step* tests the null hypothesis that c equals zero. If we accept this hypothesis, we lack evidence that supports the existence of a total effect (see left part of figure 1.2). If this is the case, our analysis ends here. If, however, we can reject this hypothesis, we hold statistical proof for an arrow going from X to Y; we can proceed to the next step.

- *The second step* tests the null hypothesis that a equals zero. Again, accepting this hypothesis halts our assessment procedure: we cannot find evidence for any kind of mediation. Only when we reject this hypothesis, that is, when there is proof of an arrow going from X to M, can we proceed to the next step.

- *The third step* tests the null hypothesis that b equals zero, or equivalently, whether there is an arrow going from M to Y (controlling for the exposure X). Once again, accepting this hypothesis disrupts our analysis: there is still not enough evidence to conclude mediation. However, if we reject this null hypothesis, we can surmise the presence of an indirect effect, or equivalently, that mediation has occurred.
  This means that we can *only* conclude the presence of mediation, when we reject the null hypotheses from *both* the second and the

third step. This can be understood by comparing figure 1.2 to figure 1.1: they are identical, except for the number of arrows that make up the intervening effect (in figure 1.2 it consists of two arrows, while there is only one in figure 1.1). Consequently, testing for an indirect effect will amount to checking the existence of these two arrows/regression parameters.

- When mediation is concluded, *the fourth step* categorises the type of mediation by testing the null hypothesis that c' equals zero. It tests whether there is evidence for an arrow going from X to Y , while at the same time controlling the mediator M. If we reject this null hypothesis, we find evidence for a direct effect (see right part of figure 1.2). In this case, we observe partial mediation: part of the total effect is mediated, and part of it is not. If we accept the null hypothesis, on the other hand, we lack evidence for a direct effect and researchers can claim complete mediation: the entire effect of exposure on outcome flows through the mediator M.

When all steps have been iterated through and mediation is concluded, this approach also lets you evaluate the intervening effect itself: it can either be estimated as the product of the a- and b-paths (i.e. $a \cdot b$), or as the difference between the c- and c'-paths (i.e. $c - c'$). Both the product-of-coefficients and the difference-of-coefficients approaches will always provide identical results in linear settings (as was considered here). As you can imagine, the straightforward way in which Baron and Kenny (1986) assess and estimate mediation ensured that, to this very day, their work remains one of the most cited papers in social science literature.

1.1.3 Criticism and improvements

Of course, as do most first attempts, the causal steps approach has received its share of criticism, and naturally, earned a vast array of suggestions, improvements and extensions.

For one, the first step mentioned above has proven superfluous, as it is entirely possible for a mediator to causally appear between the exposure and the outcome in the absence of a (detectable) association between both. Since the total effect aggregates all possible paths of influence (both direct and indirect), a lack of observable association might result from two or more (in)direct effects cancelling each other out (Collins et al., 1998; MacKinnon et al., 2000; Preacher et al., 2007). Alternatively, we

may lack adequate power to detect a significant total effect, which may in turn result in the incorrect acceptance of this null hypothesis (Preacher et al., 2007). These considerations suggest that a significant total effect does not necessarily imply mediation, while a nonsignificant *c*-path does not necessarily indicate a lack of it (Zhao et al., 2010).

Two, the approach is not based on a quantification of the very thing it is attempting to test: the intervening effect (Hayes, 2009). Given that the indirect effect is quantified as the product of its constituent paths, it seems only natural to base inference on tests of the product term, rather than on its composing parts. Generally, Sobel tests are counted on to complement Baron and Kenny (1986)'s causal steps, even though these wrongfully assume a normally, rather an an asymmetrically distributed indirect effect in finite samples. As the skewness and kurtosis of this sampling distribution often bring about a low power in detecting mediation, Shrout and Bolger (2002) suggest bootstrapping as a laudable alternative. As such, the main role for Baron and Kenny (1986)'s equations therefore dwindles down to deciding on the type of mediation (Zhao et al., 2010):

- *Complementary mediation*: both the direct and indirect effect are statistically significant and point in the same direction.
- *Competitive mediation*: the direct and indirect effect are both statistically significant, but their effects are facing in opposite directions.
- *Indirect-only mediation*: only the indirect effect is found to be statistically significant.
- *Direct-only nonmediation*: there is no mediation, only a significant direct effect is found.
- *No-effect nonmediation*: no significant effects are detected.

Three, the last null hypothesis test in Baron and Kenny (1986)'s causal steps approach also received several objections. Essentially, it claims that mediation is strongest when an indirect effect exists in the absence of a direct effect, even though the strength of mediation ought to be measured by the *size* of the intervening effect rather than by the lack of a direct one. Any unexplained part of the exposure-outcome relationship (rounded up into the direct effect), simply hints to the existence of other intermediary variables not (yet) included into the regression models (Zhao et al., 2010).

Four, among the methods for testing intervening variable effects, the causal steps approach ranks very low in terms of power (MacKinnon et al., 2002; Hayes, 2009). Due to possible type-II decision errors during each hypothesis test, it is entirely plausible for an indirect effect to be

detectably different from zero, even though one of its constituent path coefficients is not (Hayes, 2009). Moreover, the power associated with the b- and c'-paths is known to decrease as the exposure-mediator relationship becomes stronger, while the power to test the indirect effect maximises when b is equal to or slightly larger than the a-path (Kenny et al., 1998). Considering these fluctuations in power, minimising the total number of tests might prove profitable in increasing the overall detection rate for mediation.

Five, the causal steps approach requires continuous, as opposed to discrete, values for both mediator and outcome. Continuous data can take any value within its defined range; examples include a person's height, their intelligence, their age, weight, or the time they need to finish a test. In contrast, discrete or categorical measures can only take a few specific values. Examples include the number of cats in a litter (as you cannot have two and a half kittens), the result of a dice roll (the only possible values are one to six), a person's eye colour (brown, blue, green, ...), or a test result (pass or fail). Extending mediation from continuous to categorical measures redirects us from linear regression models to generalised linear models (GLM). Although this progression complicates the mediation analysis itself, it also infinitely increases the number of possible applications and research areas to which mediation analysis can be applied.

Six, the authors ignore the possible existence of (unmeasured) confounders of the exposure-mediator, exposure-outcome, and mediator-outcome relation. As already mentioned in section 1.1.1, disregarding the existence of such confounding variables removes any causal interpretation from regression equations (2.1).

Seven, Baron and Kenny (1986)'s work is also limited to additive effects on both the mediator and the outcome. As such, the approach does not support the inclusion of interactions (i.e. product terms), either between included covariates and the exposure or mediator, or between the exposure and the mediator themselves. In response, recent literature suggests the counterfactual framework, which correctly identifies the direct and indirect effects under a broad variety of conditions (Imai et al., 2010; Pearl, 2001, 2012; VanderWeele and Vansteelandt, 2009; VanderWeele, 2013). This framework provides a very broad applicability, as it also conveniently encompasses the product-of-coefficients approach when mediation is investigated in additive, linear settings.

Another stingy, yet important, subject entails the extension of mediation analysis to multilevel designs. Before we can expand mediation

to multilevel settings, however, we will first provide some background knowledge concerning this type of data structure.

1.2 Multilevel Data

In a lot of scientific areas, researchers are often confronted with hierarchically structured data. The key feature of such multilevel, nested, or clustered data is that the objects under study are assembled together in groups or clusters. These groups define the different levels present within the data structure; most often, multilevel data are gathered at two different levels, but theoretically, there is no limit to the number of cluster-levels you can define. In two-level designs, a lower-level (level-1) is always nested within an upper-level (or level-2): at the upper-level, information is gathered about the various groups present in the study, while at the lower-level, measurements are taken from within the groups themselves.

A typical example of clustered data structures is found in educational studies, where students (level-1) are observed within classrooms (level-2): at the upper-level, we gather information about the teachers or classrooms, while at the lower-level, we measure students nested within these classes (see upper part of figure 1.3). The most important feature of such nested data is that students who were taught by the same teacher will perform more alike, compared to students taught by different teachers. This might result from the specific way in which a teacher transfers knowledge, or because some subjects might have been focussed on or glossed over during a lecture, or maybe just because pupils from the same class tend to help each other out. When going up one level in the hierarchy, classrooms (level-2) within a school (level-3) will often be more similar, compared to classes from different schools (see lower part of figure 1.3).

The same reasoning can be applied to within-subject or longitudinal studies, where individuals (level-2) are measured over several occasions (level-1): at the upper-level, we gather information about our subjects, while at the lower-level, we measure multiple time points nested within these individuals (see figure 1.4). Naturally, measurements taken from within the same individual will prove more alike, compared to measurements taken from different subjects. Imagine for example that we are interested in modelling the number of hits fencers manage to set during a monthly competition. In this setting, a fencer can be seen as the upper-level (or level-2), while monthly measurement occasions constitute the lower-level (or level-1). As such, each monthly score is *nested* within a specific fencer.

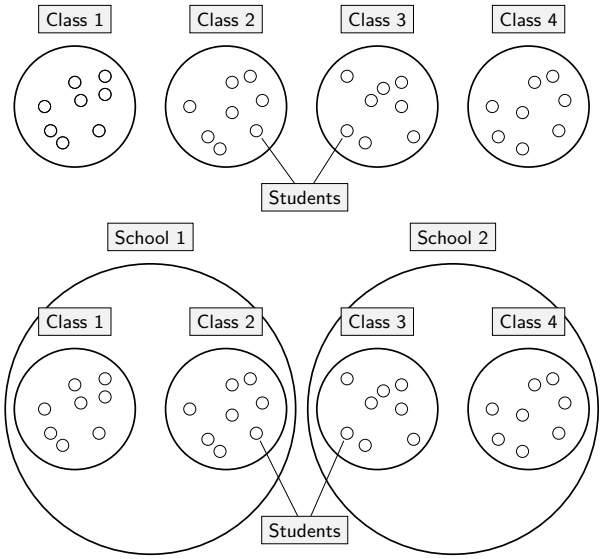

Figure 1.3 **Upper part.** In a two-level hierarchical data structure in educational research, students (level-1 units) are nested within classrooms (level-2 units). **Lower part.** In a similar three-level structure, students (level-1 units) are nested within classrooms (level-2 units), which are in turn nested within schools (level 3 units).

When you think about it, it seems very reasonable that the monthly hit rate from the same athlete will be more alike, compared to scores from different fencers. This might be due to a fencer's (lack of) experience or tactical insights, result from a chronic injury with which the fencer struggles, or it might simply be the effect of rigorous training. Put in short, the monthly scores of a specific fencer will be *correlated* with each other: these repeated measures will often show undeniable similarities across time.

Equivalently, twin studies collect measurements from both siblings (level-1) within a twin pair (level-2), resulting in very similar data for the twins, compared to random strangers. Other examples can be found in ophthalmology, where two eyes (level-1) form natural clusters within an individual (level-2), or in teratology, where data are gathered from all members (level-1) within a litter (level-2).

Unfortunately, such correlated data structures present a number of challenges in the construction, the estimation, as well as the interpretation of suitable statistical models. To confront these challenges, multilevel models were developed.

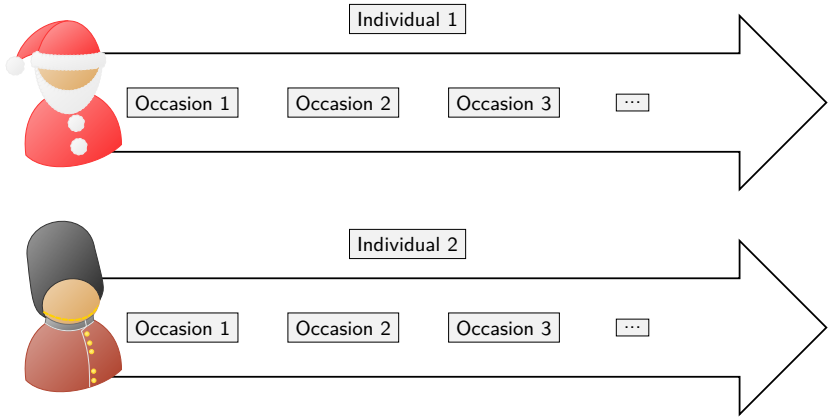

Figure 1.4 In a two-level clustered data structure in longitudinal research, measurement occasions (level-1 units) are nested within individuals (level-2 units).

1.2.1 Multilevel Models

In past literature, there have been four main strategies through which researchers attempted to deal with clustered data structures. A first strategy removes the dependencies from the observations by aggregating all lower-level measures within a cluster, which forces the data into a single-level structure (e.g., by summarising all time points from one individual into a single average). As you can imagine, condensing the data in this way may lead to a substantial loss of both information and power (Snijders and Bosker, 1999; Raudenbush and Bryk, 2002). A second possibility ignores any correlations between the lower-level measures and simply analyses the data through ordinary regression. In our fencing example, this line of thinking would ignore the dependencies between the scores from the same fencers and analyse the data as if all measures arose from different athletes. A third approach starts out like the second (i.e., it analyses the data through ordinary regression), but subsequently adjusts the estimated standard errors to correct for any correlation within the data (so-called Generalised Estimating Equations or GEE, Liang and Zeger (1986)). A fourth strategy was found in multilevel models, which attempt to model variation at the different levels of the data structure. Comparing and reviewing these four options resulted in a general consensus that some of these approaches hold a number of advantages over others.

First, GEE and multilevel models offer the opportunity to analyse all levels of the data simultaneously, in contrast to the first two strategies. As

a convenient result, they allow us to examine the influence of two types of covariates at the same time (Raudenbush and Bryk, 2002; Snijders and Bosker, 1999). The first type, a cluster-level or upper-level variable, displays the same value for all measurements within a cluster. When observing monthly tournament scores nested within fencers, an example of an upper-level variable can be found in a chronic injury, the fencer's handedness, or gender: these characteristics will remain fixed within a specific fencer (i.e., they are fixed over time). The second type of covariate fluctuates across all measurements within a cluster; it is conveniently referred to as a within-cluster or lower-level variable. When assessing a fencer's monthly competition scores, the number of hours of sleep just before the tournament, muscle spasms, or malfunctioning equipment represent examples of within-cluster covariates: these features will be present at some occasions, but not at others (i.e., they fluctuate over time).

Second, as opposed to ordinary regression, GEE and multilevel models provide correct standard errors and, consequently, appropriate confidence intervals and significance tests. Since observations from within the same cluster tend to be more alike than observations from different clusters, we observe a lower variation within a cluster, compared to data from a random sample. As traditional regression models have no way of realising that this decreased variation is due to similarities between lower-level measurements, they will irrevocably lead to an underestimation of the standard errors.

Third, in contrast to the other three approaches, multilevel models allow us to decompose the total variance of the outcome into portions associated with each level present within the data. For example, when observing tournament scores with fencers, we can quantify the proportion of variation in tournament scores caused by differences between individual fencers, and compare it to the percentage of variation in fencers' scores due to differences over time.

The collective weight of these advantages ensured that inappropriate aggregation of clusters and single level regression of multilevel data, have become less and less common. GEE, on the other hand, constitutes a popular alternative to multilevel models, even though both approaches tend to focus on different research questions: the former extracts population-averaged effects, while the latter estimates effects within clusters. To allow the estimation of such cluster-specific effects, several multilevel modelling frameworks have been developed. One such framework entails mixed-effects models, which model both the regression parameters common to all clusters

(i.e., the fixed effects), as well as any parameters specific to a certain group (i.e., the random effects). Because they integrate both random and fixed effects, they are conveniently termed *mixed*-effects models. A second framework that allows the analysis of multilevel outcomes are Structural Equation Models (SEM). Since mixed models are used more frequently and have been around the longest, and realising that, in most cases, SEM is entirely equivalent to its mixed-effects counterpart in the absence of latent variables, we will focus on the former. Multilevel models can be categorised into two different types, according to the distributional assumptions of the dependent variable. When the outcome variable is normally distributed, we end up with Linear Mixed Models (LMMs), while Generalised Linear Mixed Models (GLMMs) are called upon when it is not. Let us take a closer look at each of these models in turn.

1.2.1.1 Linear Mixed Models

Intuitively, Linear Mixed Models or LMMs can be seen as an extension of the linear regression model from section 1.1 to correlated outcomes. Let us demonstrate the LMM by formulating a statistical model for a variable measured within-subjects: a fencer's monthly tournament score, $score_{ij}$. In this notation, j ($j = 1...J$) indexes a specific fencer, while i ($i = 1...I$) represents the particular month at the end of which the tournament takes place. Suppose we aim to model and predict these scores in terms of a newly developed training program. At the beginning of every month, the trainer who implements this new program picks out a number of fencers that need to attend the training sessions this month. Since program participation varies from month to month (e.g., a specific fencer gets to participate one month but not the next), this predictor represents a time-varying or within-cluster covariate (see section 1.2.1). Consequently, it ought to be able to change over fencers and measurement moments: training program, $program_{ij}$, is indexed by both i and j. To specify whether an athlete follows the training program during a specific month, the variable $program_{ij}$ will equal one when fencer j participates during month i, and zero otherwise.

At the *lower-level* (i.e., the measurement occasion), the regression equation for tournament scores can be written as follows:

$$score_{ij} = \beta_{0j} + \beta_{1j} program_{ij} + \epsilon_{ij} \tag{1.3}$$

In this multilevel equation, β_{0j} encodes the intercept, while β_{1j} represents

the regression slope for $program_{ij}$. The lower-level residuals ϵ_{ij} are assumed to be normally distributed with mean zero and variance σ_e^2. As you can see, the major difference from a single level regression model (see section 1.1) is that we assume a different intercept and slope for each fencer. This is indicated by the subscript j (indexing fencers) present in both the intercept and slope parameters. Keeping this in mind, the intercept β_{0j} can be interpreted as the average score for fencer j when he or she *did not* partake in training (i.e., when $program_{ij}$ equals zero, the average hit score becomes $\beta_{0j} + \beta_{1j} \cdot 0 = \beta_{0j}$). Similarly, the average number of hits for fencer j when he or she *did* partake in training will equal $\beta_{0j} + \beta_{1j} \cdot 1 = \beta_{0j} + \beta_{1j}$). As such, the β_{1j}-slope coefficient represents the effect of the training program in fencer j. Since β_{0j} ($\beta_{0j} + \beta_{1j}$, respectively) represents the *average* score for fencer j for a month where the athlete did (respectively, did not) partake in training, the lower-level residual ϵ_{ij} indicates the deviation of that month's score from this average. For this reason, the variance of the lower-level residuals, σ_e^2, will express the amount of variation (i.e., deviations from the mean) that exists *within* fencers.

Because both the intercept and slope are allowed to change across fencers, the next equations explain their variation at *the upper-level* (i.e., the subject-level):

$$\beta_{0j} = \gamma_0 + u_{0j} \tag{1.4}$$

$$\beta_{1j} = \gamma_1 + u_{1j} \tag{1.5}$$

The upper-level error terms u_{0j} and u_{1j} are assumed to have a zero mean, with respective variances σ_{u0}^2 and σ_{u1}^2, and a covariance of σ_{u01}. Additionally, these are assumed to be independent of the lower-level residuals, as well as the lower-level predictor, $program_{ij}$. Since these upper-level residuals, u_{0j} and u_{1j}, are assumed to be *randomly* drawn from a multivariate normal distribution, equation (1.4) is often referred to as the *random intercept*, while equation (1.5) is designated a *random slope*.

Upper-level equation (1.4) predicts the random intercept β_{0j} (i.e., the average hit score for fencer j in months without training) by means of an intercept, γ_0, and an upper-level residual, u_{0j}. The γ_0-parameter represents the mean hit score during months without training, averaged *across* fencers. Consequently, for months without training, the upper-level residual u_{0j} denotes the deviation of fencer j's mean score from the global average γ_0. The variance of this residual, σ_{u0}^2, will hence express the amount of

variation in hit scores that exists *between* fencers.

Equivalently, upper-level equation (1.5) predicts the effect of training on fencer j's hit scores, by means of an intercept, γ_1, and an upper-level residual, u_{1j}. The γ_1-parameter represents the mean effect of training on tournament scores, averaged *across* fencers. Consequently, the upper-level residual u_{1j} denotes the deviation of the effect of training in fencer j from the global average γ_1. The variance of this upper-level residual, σ^2_{u1}, will hence express the amount of variation in the effect of training that occurs *between* fencers.

Combining both upper-level equations (1.4)-(1.5) with the lower-level equation for tournament scores (equation (1.3)), provides us with a more compact, composite model:

$$score_{ij} = \gamma_0 + u_{0j} + (\gamma_1 + u_{1j})program_{ij} + \epsilon_{ij} \tag{1.6}$$

We clearly see that the fencers' monthly tournament scores are predicted by a *fixed* component that is common to all athletes (the γ_0-parameter), alongside a fencer-specific *random* component that varies across individuals (u_{0j}). Additionally, the hit scores are predicted in terms of an independent variable $program_{ij}$, where the effect of this variable is partitioned into two pieces: an average program effect common to all fencers (γ_1) and a *random* fencer-specific slope that varies across individuals (u_{1j}). This flexibility in modelling the monthly scores in terms of the different levels of the data constitutes an undeniable perk of mixed-effect models (i.e., combining both fixed- and random effects).

1.2.1.2 Generalised Linear Mixed Models

While the above-introduced LMM represents an extension of the linear regression model, the Generalised Linear Mixed Model or GLMM similarly broadens the Generalised Linear Model (GLM) to multilevel data. For single level data, the GLM's major contributions to statistics include logistic- and probit-regression of binary outcomes, as well as Poisson regression for count data. Of course, similar to linear regression models, GLMs are only suited for the analysis of independent data structures. Hence, we introduce the GLMM, which is capable of modelling both random- and fixed effects for correlated categorical outcomes. In this thesis, we primarily focus on binary dependent variables, as this type of outcome is very popular amongst applied researchers, and as a result, has been studied most extensively.

Let us define such a binary outcome variable based on the illustrating example that we introduced before. Suppose that, rather than focussing on their monthly scores, we instead look at whether or not the fencers are happy about their performance that month. This new dependent variable, $happy_{ij}$, provides a prototypical example of a binary response, as it relays but two possible outcomes: either fencers are happy with their monthly performance (i.e., $happy_{ij} = 1$), or they are not (i.e. $happy_{ij} = 0$). Because this dependent variable can only ever obtain a value of zero *or* one, we cannot hope to appropriately fit a model where the outcome values are allowed to span the entire continuous scale (as in equation (1.6)). For this reason, we will not attempt to predict the binary outcome itself, but rather, we will model the monthly probability of a fencer being happy about his/her performance. At *the lower-level*, the GLMM equation for the happiness-outcome can be written as follows:

$$g[P(happy_{ij} = 1)] = \beta_{0j} + \beta_{1j} program_{ij} \tag{1.7}$$

In this generalised multilevel regression equation, the parameter coefficients are defined as they were before: β_{0j} encodes the intercept, while β_{1j} represents the regression slope for $program_{ij}$. The upper-level equations for β_{0j} and β_{1j} are again defined as in equations (1.4)-(1.5). Additionally, g represents the link function between the monthly probability of a fencer being happy, $P(happy_{ij} = 1)$, and the linear predictor on the right. Most often, for a binary outcome, g is defined by either the *logit*- or the *probit*-link. For the *logit*-link, equation (1.7) becomes:

$$\begin{aligned}
& logit[P(happy_{ij} = 1)] && = \beta_{0j} + \beta_{1j} program_{ij} \\
\iff & log\left[\frac{P(happy_{ij} = 1)}{P(happy_{ij} = 0)}\right] && = \beta_{0j} + \beta_{1j} program_{ij} \\
\iff & \frac{P(happy_{ij} = 1)}{P(happy_{ij} = 0)} && = e^{\beta_{0j} + \beta_{1j} program_{ij}} \\
\iff & Odds(happy_{ij} = 1) && = e^{\beta_{0j} + \beta_{1j} program_{ij}}
\end{aligned} \tag{1.8}$$

As such, the odds of fencers being happy with their monthly scores (with the odds defined as the probability of being happy, divided by the probability of disappointment), are modelled in terms of an exponential function. This rather complicated expression looks meaner than it is, since the odds of being happy when the fencer did not partake in training the previous month, will simply equal $e^{\beta_{0j} + \beta_{1j} \cdot 0} = e^{\beta_{0j}}$. Equivalently, the monthly odds of being happy when the fencer did partake in training, will

equal $e^{\beta_{0j}+\beta_{1j}\cdot 1} = e^{\beta_{0j}+\beta_{1j}}$.

Equivalently, for the *probit*-link, equation (1.7) becomes:

$$\begin{aligned} & probit[P(happy_{ij}=1)] &&= \beta_{0j} + \beta_{1j} program_{ij} \\ \iff \quad & \Phi^{-1}[P(happy_{ij}=1)] &&= \beta_{0j} + \beta_{1j} program_{ij} \\ \iff \quad & P(happy_{ij}=1) &&= \Phi[\beta_{0j} + \beta_{1j} program_{ij}] \end{aligned} \tag{1.9}$$

Where Φ represents the cumulative standard normal distribution. As such, a fencer's probability of being happy with his or her monthly performance is modelled in terms of a cumulative standard normal function. Again, this expression can be interpreted by filling in the two possible values for training program; when fencer j did not partake in training during month i (i.e., when $program_{ij} = 0$), the monthly probability of being happy with his/her performance will equal $\Phi(\beta_{0j} + \beta_{1j} \cdot 0) = \Phi(\beta_{0j})$. Equivalently, when fencer j did follow the training program in month i (i.e., when $program_{ij} = 1$), the monthly probability of being happy with his or her performance, will equal $\Phi(\beta_{0j} + \beta_{1j} \cdot 1) = \Phi(\beta_{0j} + \beta_{1j})$.

A more intuitive explanation of the GLMM considers the observed binary variable $happy_{ij}$ as a coarse representation of an underlying continuous variable $performance_{ij}$ (e.g., a score ranging from $-\infty$ to $+\infty$). In doing so, we can specify the relationship between the latent (i.e. unobserved) variable $performance_{ij}$ and the observed variable $happy_{ij}$, by defining a threshold model:

$$happy_{ij} = 1 \text{ if } performance_{ij} > 0 \tag{1.10}$$

In other words, $happy_{ij}$ will equal one when the unobserved monthly performance is positive, while $happy_{ij}$ will equal zero when $performance_{ij} \leqslant 0$. Because $performance_{ij}$, unlike $happy_{ij}$, is defined on a continuous scale, it *can* be expressed as a linear combination of the predictors (as in section 1.2.1.1):

$$performance_{ij} = \beta_{0j} + \beta_{1j} program_{ij} + \epsilon_{ij} \tag{1.11}$$

In this multilevel regression equation for the latent outcome variable, the interpretation of all parameters is similar to the ones explained in the previous section: β_{0j} encodes the random intercept, while β_{1j} represents the random slope for $program_{ij}$. Contrary to section 1.2.1.1, however, $performance_{ij}$ is defined as a hypothetical construct and is never actually observed. Because of this, the scale of the corresponding lower-level residu-

als, ϵ_{ij}, cannot be estimated and may therefore be chosen arbitrarily, where different choices will consequently lead to different modelling strategies. As it turns out, defining the distribution of ϵ_{ij} as standard logistical with variance $\sigma_\epsilon^2 = \frac{\pi}{3}$, proves equivalent to selecting the *logit*-link function in equation (1.7). Alternatively, defining ϵ_{ij} as standard normally distributed with variance $\sigma_\epsilon^2 = 1$, will effectively implement the *probit*-link.

1.2.2 Challenges to multilevel modelling

Now that we have introduced the basics to multilevel modelling, it seems only reasonable to mention several of the challenges these models experience. A first difficulty relates to the inclusion of more than two levels into the data hierarchy. Since the complexity of the modelling process increases with each level that is additionally included, researchers often limit the number of hierarchical structures to two. This strategy may force scientists to overlook some dependencies present within the data, leading to suboptimal estimation processes.

Another challenge involves the number of random effects included into a multilevel model: as this number increases (e.g. a random slope for each included predictor), the complexity of the model and its corresponding interpretation will inflate correspondingly. Fortunately, for a lot of studies, modelling a random intercept without any random slopes suits the data just fine. Also, when there are but a scant number of units within each cluster, the amount of random effects that can be identified is limited. In such cases researchers may often stick to a random-intercept model with fixed, rather than random slopes. For the example in section 1.2.1.1, this would imply that the effect of the training program will be the same for every fencer (i.e., a fixed slope parameter).

Another challenge that multilevel models face, involves the possible existence of unmeasured confounding at the upper-level. As such confounding presents difficulties not so easily addressed, we will discuss this issue next.

1.2.2.1 Unmeasured upper-level confounding

Let us get back to the example introduced in section 1.2.1.1, where we looked at the effect of a new training program on the monthly tournament scores of fencers. This one-month program assignment can be appointed in a systematic or a non-systematic way. The latter option would imply that training is randomly appointed every month, e.g. by flipping a coin

to decide whether or not a fencer is assigned to the one-month program. As training assignment is completely arbitrary in this case, fencers will not have any impact on program participation. In contrast, during a systematic designation of training, the fencers' characteristics will, in some way, influence the assignment process. Imagine for example that the athletes have to submit a letter of motivation to express their participation wishes. If this is the case, a fencer's overall motivation and general ambition towards reaching the Olympics may very well affect his or her efforts during this selection process. This would imply that being selected for training will implicitly depend on the fencer's general ambition (see the arrow pointing downwards to the left in figure 1.5).

Besides the effect that the fencers' ambition has on program participation, it seems equally likely that this internal motivation will permeate through to their efforts in general. As such, the fencers' ambition is likely to also influence their scores in the monthly competition (see the arrow pointing downwards to the right in figure 1.5). In summary, when program assignment occurs through motivational letters, the fencers' overall motivation and ambition ($motivation_j$, a level-2 variable) may influence both their participation in the program *as well as* their monthly tournament scores (see figure 1.5)[1].

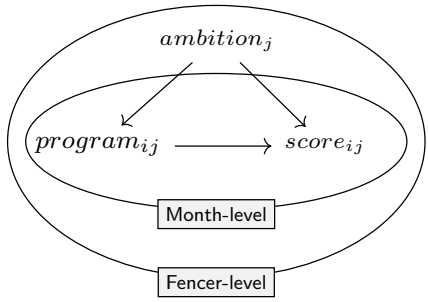

Figure 1.5 In a two-level data structure where monthly scores (level-1 units) are nested within fencers (level-2 units), both program participation $program_{ij}$ and monthly tournament scores $score_{ij}$ (measured at the lower-level), may be influenced by the fencers' ambition $ambition_j$ (a confounder at the upper-level). When this subject-level variable is not accounted for in a multilevel analysis, results may suffer from an omitted-variable bias.

This provides a straightforward example of an upper-level confounder: a level-2 variable that simultaneously influences both the exposure, $program_{ij}$,

[1]Note that this graph does *not* represent a mediation process, as this would imply an arrow pointing from $program_{ij}$ to $ambition_j$, instead of the other way round.

and the outcome, $score_{ij}$. When such confounding is present within the data, it can distort any causal pathway that exists between two variables. If this confounder is *measured* and included into the multilevel model, its distorting influence can be accounted for and poses no substantial problems. However, when the upper-level confounder remains *unmeasured*, it cannot be included into the model and will be assimilated into the upper-level error term of the random intercept (i.e., u_{0j} in the previous section). Here lies the true issue, since we stated in section 1.2.1 that the upper-level residuals ought to be independent of the predictors. Absorbing $ambition_j$ into the random intercept will induce a dependency between the upper-level residual, u_{0j}, and program participation, $program_{ij}$, thereby violating the independence assumption. This correlation may result in an omitted variable bias for the coefficient of training program, which may in turn lead to an over- or underestimation of the random intercept variance σ^2_{u0}.

1.2.2.2 Dealing with unmeasured upper-level confounding

Covariates correlated with error terms are often labelled endogeneous in econometrics (Wooldridge, 2010), suggesting that predictors correlated with the upper-level residuals can be appointed as 'upper-level endogeneous'. The most popular way of dealing with such upper-level endogeneity is to separate lower-level covariates into a within- and between-cluster part (Neuhaus and Kalbfleisch, 1998). The between-cluster component of a lower-level variable is defined by its cluster means, while its within-cluster part is outlined by the within-cluster deviations from these means. Scientists also refer to this approach as centring the data within clusters: each lower-level measure is centred according to the mean value in its cluster. In terms of our example, the between-cluster component of program assignment is defined as the mean value of $program_{ij}$, averaged within a fencer: $\overline{program}_j$. This variable is only indexed by fencer j, since it is averaged over all monthly measurements within the athlete. Since training represents a dichotomous variable (with value one or zero), this cluster mean can also be interpreted as the percentage of months where fencer j was engaged in training (i.e., the proportion of months where $program_{ij} = 1$). If this percentage is high (i.e., it is close to one), the athlete in question will have been assigned to the program during most months; if this percentage is close to zero, on the other hand, fencer j will have attended a limited number of one-month training sessions.

The cluster-mean deviations can in turn be interpreted as the difference

between fencer j's program assignment during month i, and the average training percentage within that fencer: $program_{ij} - \overline{program}_j$. Consequently, months where a fencer did not partake in training will exhibit a negative within-cluster deviation (i.e., $0 - \overline{program}_j$), while months that a fencer did follow training will produce positive values (i.e., $1 - \overline{program}_j$). For now, we focus on linear settings; when applying within-cluster centring to equations (1.3)-(1.4), we end up with:

$$score_{ij} = \beta_{0j} + \beta_{1j}(program_{ij} - \overline{program}_j) + e_{ij} \tag{1.12}$$

with:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}\overline{program}_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{1.13}$$

Since the within-subject deviations arise through subtraction of the between-cluster component, $\overline{program}_j$, from the original variable, $program_{ij}$, this computation will effectively remove any upper-level variation. As such, the within- and between parts of training are completely independent of each other. More importantly, the within-cluster part of training, $program_{ij} - \overline{program}_j$, will be uncorrelated with *all* upper-level variables, whether they are measured or remain unmeasured. Because of this, the fencer-centred scores of training will not be correlated with the random intercept, which implies that the regression slope β_{1j} will not show any bias in these linear settings, even in the presence of upper-level endogeneity (depending on the true underlying model). The same does *not* hold for the slope parameter γ_{01}, as the cluster means exist at the upper-level and can therefore still correlate with possible upper-level confounders. As such, in the presence of possible upper-level endogeneity, researchers advise to separate any lower-level variables into their respective within- and between parts, and to subsequently focus on the regression coefficient belonging to the within-cluster deviations. This way, the challenges introduced by unmeasured upper-level confounding can be adequately dealt with.

So far, we introduced two major statistical concepts, mediation and multilevel data. Combining both catapults mediation into a multilevel setting, where multilevel mediation faces its own challenges and issues.

1.3 Multilevel Mediation

When generalising mediation from independent to multilevel data structures, we can discern three different types based on the measurement level of all variables involved. The first type is termed *upper-level mediation*, because both the exposure and mediator are measured at the upper-level; the outcome is the sole variable to show variation at the lower-level (see upper panel of figure 1.6). Alternatively, it is also referred to as 2-2-1 mediation, corresponding to the measurement level of the exposure, the mediator, and the outcome, respectively. Since all the arrows in this mediation model originate from upper-level variables (and keeping in mind that random slopes are only possible for lower-level predictors), all path coefficients, a, b, and c', are fixed by design. An example of such a mediation analysis considers a program assignment that is fixed over time (i.e., a fencer is either assigned to the new program or not, irrespective of the measurement occasion); this way, the exposure is measured at the upper- instead of the lower-level (i.e., exposure $program_j$, instead of $program_{ij}$). This training program may in turn permantly increase the fencers' general tactical insights (the mediator $tactics_j$), thereby indirectly increasing their scores on the monthly tournament (the outcome $scores_{ij}$; see upper panel of figure 1.7).

A second type of mediation is referred to as *lower-level mediation of an upper-level effect* or 2-1-1 mediation. Here, the exposure is measured at the upper-level, while both the mediator and the outcome exist at the lower-level (see middle panel of figure 1.6). Now, the arrows in this mediation model no longer exclusively originate from upper-level variables; since the arrow pointing from the mediator to the outcome represents a lower-level effect, path coefficient b_j is allowed to vary across individuals. Referring back to our example, a training program that is fixed over time (the exposure $program_j$) may affect the fencers' monthly tournament scores (the outcome $score_{ij}$) indirectly by increasing the fencers' monthly precision (the lower-level mediator $precision_{ij}$). Such a mediation process can be summarised by the middle panel in figure 1.7.

Finally, the third type of mediation is termed *lower-level mediation of a lower-level effect* or, alternatively, 1-1-1 mediation. Here, the exposure, the mediator, as well as the outcome are all measured at the lower-level (see lower panel of figure 1.6). In this type of mediation model all arrows originate from lower-level variables, implying that all path coefficients, a_j, b_j, and c'_j, may vary across individuals. To illustrate this type of multilevel

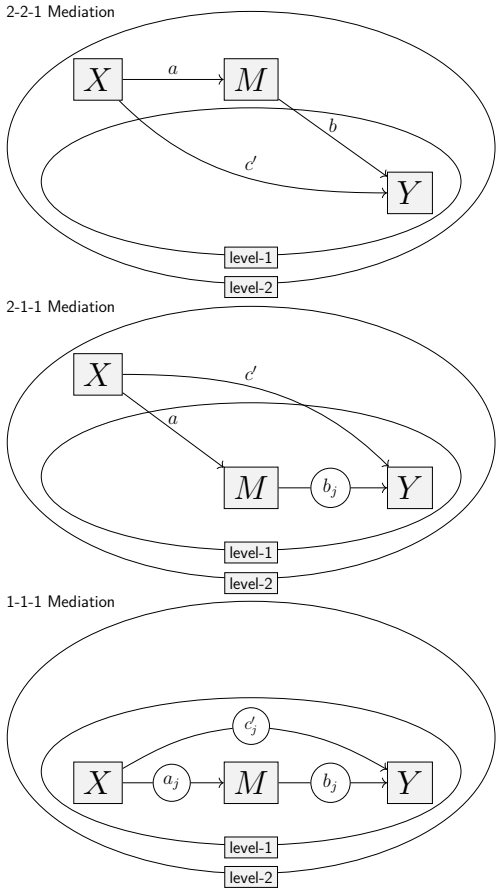

Figure 1.6 **Upper pannel:** Upper-level mediation considers a multilevel mediation setting where both the exposure and mediator are measured at the upper-level. The outcome is the only variable measured at the lower-level. **Middle panel:** Lower-level mediation of an upper-level effect exists when the exposure is measured at the upper-level, while the mediator and outcome are measured at the lower-level. **Lower panel:** Lower-level mediation of a lower-level effect occurs when all variables are measured at the lower-level. The circles around the path coefficients announce the possible existence of random slopes.

mediation, we need the exposure, mediator, as well as the outcome to exist at the lower-level and vary across time. As such, the training program again needs to be attributed monthly rather than staying fixed within fencers (the exposure $program_{ij}$). Consequently, this program may affect the fencers' monthly tournament scores (the outcome $score_{ij}$), by indirectly targeting their hand precision (the mediator $precision_{ij}$). This example of

lower-level mediation of a lower-level effect is rehashed in the lower panel of figure 1.7.

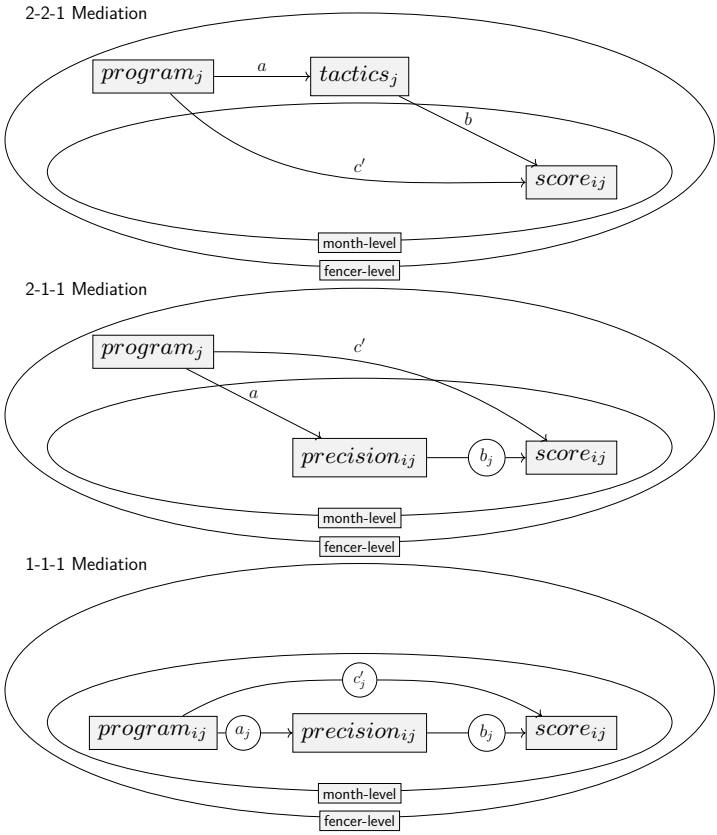

Figure 1.7 **Upper panel:** 2-2-1 mediation occurs when we aim to asses whether or not the effect of a time-invariant training program (the exposure $program_j$) on the monthly tournament scores (the outcome $score_{ij}$) is mediated by the fencers' technique (the mediator $technique_j$). **Middle panel:** Assessing if the effect of a time-invariant program (exposure $program_j$) on the fencers' monthly scores (the outcome $math_{ij}$) is mediated by their monthly hand precision (the mediator $presicion_{ij}$), exemplifies a 2-1-1-type mediation. **Lower panel:** 1-1-1 mediation occurs when the effect of a time-varying program (exposure $program_{ij}$) on the fencers' tournament scores (the outcome $score_{ij}$) is mediated by their monthly hand precision (the mediator $precision_{ij}$).

In order to try and estimate the intervening effect in 1-1-1 mediation settings (see lower panel of figure 1.6), Kenny et al. (2003) suggested to extend the product-of-coefficients approach to correlated data structures.

This entails a slight modification of the original formula:

$$\text{indirect effect} = E(a_j)E(b_j) + \sigma_{a_j b_j} \tag{1.14}$$

$$= ab + \sigma_{a_j b_j} \tag{1.15}$$

Here, a and b represent the mean values of the path coefficients from section 1.1.2, averaged across clusters. Additionally, a term is included that represents the covariance between both random effects, σ_{ab}. Conveniently, this formula can also be applied to other multilevel mediation settings, by realising that when *either* the a- or the b-arrows are fixed (i.e., they do not vary between clusters as random slopes), there will be no covariance between both coefficients and σ_{ab} will equal zero. However, when there is a random slope for *both* pathways, this term may differ from zero and needs to be included in order to correctly assess the intervening effect. Since 2-2-1 mediation precludes random slopes by design (see upper panel figure 1.5), the covariance between a and b will be zero and the intervening effect will revert back to the original formula ab. Similarly, since there may only ever exist a random slope for b_j in 2-1-1 mediation (see middle panel in figure 1.5), the covariance term also proves irrelevant; the mediated effect will again be defined by ab. In contrast, this covariance term may not necessarily equal zero for 1-1-1 mediation, since this setting may include a random slope for both the a- and b-paths (see lower panel of figure 1.5). In this setting, the intervening effect needs to account for the possible covariance of random slopes: the indirect effect will equal $ab + \sigma_{ab}$.

1.3.1 Challenges to lower-level mediation of a lower-level effect

Because of the additional complexity provided by 1-1-1 mediation models (from now on also referred to as *lower-level* or *within-subject mediation*), this type of multilevel mediation will be the major focus of this thesis. Although the above-mentioned extension of the product-of-coefficients approach to multilevel models provides an elegant solution to estimating the indirect effect, it seems unable to address several issues.

For one, causal mediation analysis has traditionally been formulated, understood, and implemented within a fixed set of linear (mixed) models. Consequently, the formula provided by Kenny et al. (2003) cannot offer a general definition of the direct and indirect effects beyond their specific set of linear mixed models. For this reason, we intend to rely upon a single

framework that enables the definition, identification, and estimation of the causal mediation effects, without reference to a specific statistical model: the counterfactual framework. Within this framework, it is possible to define and identify different types of effects, e.g. a controlled versus a natural direct effect. In this thesis, we focus on the latter; we will introduce and explain the concept of counterfactuals and natural effects within the next chapter.

Two, when random effects tend to covary, this suggests the presence of unmeasured upper-level confounders. In section 1.2.1 we saw that random effects represent unexplained variance at the upper-level, summarised through the upper-level residuals. These residuals can be interpreted as upper-level variables that influence the dependent variable, but are not (yet) included into the multilevel model. As such, a random intercept can be interpreted as an unmeasured upper-level variable that influences the dependent variable, while a random slope represents an interaction between an unmeasured upper-level variable and the lower-level covariate it belongs to. So when we are, for example, confronted with a random intercept for M, this suggests the existence of at least one unmeasured level-2 variable that influences the mediator. Equivalently, a random intercept for Y strongly implies the presence of one or several unmeasured upper-level variables for the outcome. As such, a covariance between both random intercepts can only be non-zero in the presence of an unmeasured upper-level variable that affects the mediator and outcome simultaneously (i.e., a confounder). As we saw in section 1.2.2.1, such upper-level endogeneity may result in an omitted variable bias for the coefficients in the model for the outcome. To account for such *upper-level endogeneity of the mediator-outcome relation*, we need appropriate estimation techniques that allow unbiased estimation of the mediation effects in the presence of such confounding.

Three, equivalent to the product-of-coefficients approach in single level settings, its extension to clustered data does not provide clear guidelines on how to tackle *interaction terms*. This imposes a major limitation to assessing mediation in multilevel settings, as interactions are as common here as they are in their single level counterparts. Hence, most researchers would more than welcome practical instructions on how to assess and estimate the intervening effect for multilevel mediation in the presence of both upper-and lower-level interactions.

Four, the formula depicted in equation (1.14) assumes continuous values for both the mediator and the outcome. Extending multilevel mediation from continuous to categorical measures redirects us from linear mixed

models (LMMs) to generalised linear mixed models (GLMMs), as discussed in sections 1.2.1.1 and 1.2.1.2. To this day, methodological research concerning multilevel mediation analysis for categorical, and more specifically for binary mediators and outcomes, is relatively sparse. However, since a lot of research and applications rely on categorical measures, reliable assessment of *mediation in binary settings* would make a refreshing addition to current literature.

Five, although often ignored, the *validity of estimation models* and numerical optimisation techniques often rely on a specific set of assumptions (e.g., the absence of unmeasured upper-level confounders) or approximations (e.g., the Laplace approximation). If those assumptions are not met or if the approximations prove suboptimal, the corresponding modelling procedures may provide biased and/or inefficient parameter estimates. Consequently, figuring out which estimation models (and their respective implementations in various software packages) provide valid estimates for the indirect effect under which circumstances, makes up an important part of evaluating existing approaches that assess multilevel mediation.

Six, when reviewing multilevel data structures, we have not yet considered sample sizes. In clustered data, we have to define sample size at all levels of the hierarchy: we have to look at the number of upper-level units, as well as the number of lower-level units nested within each cluster. Whereas a small upper-level sample size exhibits the same limitations found in single level analysis, lower-level sample sizes will introduce problems more specific to multilevel data. The smallest possible cluster size entails but two observations within each group, and although this seems extreme, such data structures are very frequently encountered in practice. The most common applications are seen in dyadic family studies, ophthalmology, twin studies, or when analysing measurements from a 2-period - 2-treatment crossover design. These multilevel designs deserve special attention as they often introduce difficulties during the estimation of multilevel models. More specifically, this translates in issues with model convergence, in identification difficulties concerning the random effects, or even in bias for the parameter coefficients. As such, studying *small cluster-sizes* deserves a special mention in multilevel data analysis.

1.4 Goal of this thesis

In this thesis, we want to provide applied researchers with a concrete set of guidelines on how to assess within-subject mediation from a counter-

factual point of view, when confronted with one of several issues. These issues include (1) dealing with unmeasured upper-level confounding of the mediator-outcome relation, (2) appropriate inclusion and assessment of multilevel mediation in the presence of interaction terms, (3) assessing mediation in multilevel settings with binary measures, and (4) exploring which estimation technique provides the best overall performance (i.e., in terms of bias and efficiency) under a broad variety of settings (with a special focus on small sample sizes).

Chapter 2 addresses multilevel mediation in linear settings from a counterfactual point of view, where a mere two observations for each cluster are observed within AB/BA crossover designs. We compare different estimations models, as well as the assumptions they rely on, and demonstrate that the intervening effect can be identified in some models, but *not* in others, in the presence of unmeasured upper-level confounding of the mediator-outcome relationship.

In chapter 3, we continue exploring mediation analysis within linear multilevel settings, but now additionally focus on the inclusion of lower-level interaction terms. To this end, we compare several estimation techniques that differ in their centring of lower-level interactions, in terms of bias and efficiency. We observe that unmeasured upper-level endogeneity of the mediator-outcome relation can lead to biased parameter coefficients for the interaction coefficient, when the lower-level variables are centred inappropriately. In addition, some centring approaches provide all-round more precise estimation, compared to others.

In chapter 4 we temporarily digress from multilevel mediation in preparation for chapter five. In contrast to the precious two chapters, we venture into the world of binary multilevel outcomes. We attempt to assess which estimation models (and implementation procedures) perform best when modelling binary outcomes in small clusters, under a vast array of settings. The estimation procedure that best survives our performance assessment is consequently chosen to lay the basis of the last chapter.

In chapter 5 we redirect our attention to multilevel mediation, but now we focus on binary instead of continuous measures for the outcome. We compare the performance of several estimation models for binary data that allow assessment of multilevel mediation from a counterfactual point of view , while considering small cluster sizes and varying settings. Again, we evaluate the impact of unmeasured upper-level confounding of the mediator-outcome relation on the estimation of the intervening effect.

In a final chapter, we recap and summarise the previous chapters

in a brief discussion. Here, we aim to postulate specific instructions and warnings to practical researchers, as to which multilevel mediation methods to rely on under which circumstances.

Bibliography

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Collins, L. M., Graham, J. J., and Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, 33(2):295–312.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4):408–420.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Kenny, D. A., Kashy, D. A., and Bolger, N. (1998). Data analysis in social psychology. In *Handbook of Social Psychology*, chapter 6, pages 233–265. McGraw-Hill, New York, 4 edition.

Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8(2):115–128.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika Trust*, 73(1):13–22.

MacKinnon, D. P., Krull, J. L., and Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4):173–181.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Pearl, J. (2012). The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–36.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1):185–227.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and data analysis methods.* Sage, Thousand Oaks, CA, second edition.

Shrout, P. E. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7(4):422–445.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Sage, Thousand Oaks, CA.

Vanderhasselt, M.-A., Brunoni, A. R., Loeys, T., Boggio, P. S., and De Raedt, R. (2013). Nosce te ipsum–Socrates revisited? Controlling momentary ruminative self-referent thoughts by neuromodulation of emotional working memory. *Neuropsychologia*, 51(13):2581–2589.

VanderWeele, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* The MIT Press, Cambridge, MA.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

Zhao, X., Lynch Jr., J. G., and Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2):197–206.

2 Within-subject mediation analysis in AB/BA crossover designs

Abstract. Crossover trials are widely used to assess the effect of a reversible exposure on an outcome of interest. To gain further insight into the underlying mechanisms of this effect, researchers may be interested in exploring whether or not it runs through a specific intermediate variable: the mediator. Mediation analysis in crossover designs has received scant attention so far and is mostly confined to the traditional Baron and Kenny approach. We aim to tackle mediation analysis within the counterfactual framework and elucidate the assumptions under which the direct and indirect effects can be identified in AB/BA crossover studies. Notably, we show that both effects are identifiable in certain statistical models, even in the presence of unmeasured time-independent (or upper-level) confounding of the mediator-outcome relation. Employing the mediation formula, we derive expressions for the direct and indirect effects in within-subject designs for continuous outcomes that lend themselves to linear modelling, under a large variety of settings. We discuss an estimation approach based on regressing differences in outcomes on differences in mediators and show how to allow for period effects as well as different types of moderation. The performance of this approach is compared to other existing methods through simulations and is illustrated with data from a neurobehavioral study. Lastly, we demonstrate how a sensitivity analysis can be performed that is able to assess the robustness of both the direct and indirect effect against violation of the "no unmeasured lower-level mediator-outcome confounding" assumption.

This chapter is based on Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., & Loeys, T. (2015). Within-subject mediation analysis in AB/BA crossover designs. *International Journal of Biostatistics*, 11(1): 1-22.

2.1 Introduction

The concept of mediation has received a great deal of attention during the last couple of decades, with Baron and Kenny (1986) among the first to scratch the surface of this vast realm. These authors presented a causal-steps approach to establish whether or not a variable serves as the generative mechanism, through which an independent variable (subsequently referred to as the 'exposure' X) influences a dependent variable of interest (the 'outcome' Y). Any such 'mediator' variable may help to clarify the nature of the relationship between exposure and outcome. The question of whether or not the causal effect of exposure X on outcome Y (partly) runs through a mediator M, can be verified by decomposing the total effect of X on Y into a direct and an indirect effect (see figure 2.1). The traditional Baron and Kenny framework, which assumes independent observations of X, M and Y (with the latter two measured at the interval level), relies on three regression equations:

$$\begin{aligned} Y &= i_{Y1} + cX + e_{Y1} \\ M &= i_M + aX + e_M \\ Y &= i_{Y2} + c'X + bM + e_{Y2} \end{aligned} \tag{2.1}$$

Figure 2.1 Decomposition of a total effect (of exposure X on outcome Y) into a direct (not through the mediator M) and an indirect effect (through the mediator M), by means of mediation analysis.

The above mentioned direct effect is conventionally captured by c', which is justified provided all relations in the path diagram in figure 2.1 are linear and satisfy a specific set of 'no unmeasured confounders'-assumptions (Loeys et al., 2013). The indirect effect, on the other hand, can be obtained either by means of the product-of-coefficients approach (as a product of its constituent path coefficients, ab) or by means of the difference-of-coefficients approach ($c - c'$). Both estimators are equivalent in linear models (Mackinnon and Dwyer, 1993).

In AB/BA crossover studies, where each participant is observed exactly twice (once under exposure A and once under B, see figure 2.2), observations are no longer independent and exhibit a multilevel structure

(where the subject is considered the upper-level and the measurement moment the lower-level). Such AB/BA crossover trials are ubiquitous, as they can effectively eliminate between-patient variation from the data (Senn, 2002). Unfortunately, when it comes to decomposing the total effect, the mediation analysis literature has almost exclusively relied on extending the product-of-coefficients approach to multilevel settings (Judd et al., 2001; Kenny et al., 2003; Bauer et al., 2006; Raykov and Mels, 2007; Preacher et al., 2010; Pituch and Stapleton, 2012), without due attention to the interpretation of the effects as direct and indirect effects, and to the underlying assumptions needed to identify these.

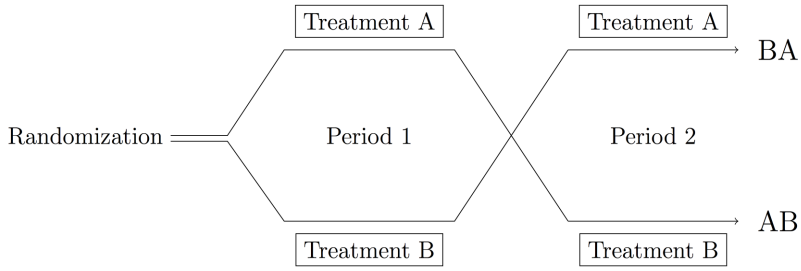

Figure 2.2 The mechanism of a simple AB/BA crossover design. Each treatment (treatment A and B) is administered during one of two periods, the sequence of which (sequence AB or BA) is determined during randomisation.

To surmount these limitations, this paper will tackle mediation analysis in crossover designs from a counterfactual perspective. This framework has proven useful in explicating the assumptions underlying mediation analysis, and in identifying the direct and indirect effects of interest (Pearl, 2001; VanderWeele and Vansteelandt, 2009; Imai et al., 2010; VanderWeele, 2010; Pearl, 2012). In section 2.2, we start by defining counterfactual outcomes in the AB/BA design, introduce non-parametric expressions for the direct and indirect effect, and discuss the assumptions needed to identify both effects. Subsequently, we derive expressions for these effects under a simple data-generating mechanism that satisfies these assumptions. In section 2.3, we discuss Judd et al. (2001)'s difference approach alongside three multilevel approaches for the assessment of within-subject mediation in this simple setting. Next, we consider more complex data-generating mechanisms involving a variety of interactions (section 2.4), introduce an extension of the difference approach to accommodate such moderation and compare the relative performance of the different estimation approaches in a simulation study (section 8). Furthermore, employing data from a crossover experiment

that evaluates the effect of neurostimulation on ruminative thinking, we illustrate how the different estimation techniques may lead to contrasting conclusions about the indirect effect running through the working memory (section 2.6). Additionally, since we will show in the subsequent section that mediation analysis in crossover settings relies on the assumption of 'no unmeasured lower-level M-Y confounding', we develop a sensitivity analysis method. This analysis appraises the robustness of the estimated direct and indirect effect against violations of this assumption, and can easily be embedded within our proposed estimation framework. We end with a discussion.

2.2 Specification of the natural direct and indirect effect in within-subject mediation models

2.2.1 The counterfactual framework

In order to formalise the notion of direct and indirect effects, we introduce counterfactual outcomes in AB/BA crossover settings. A 'counterfactual' or 'potential outcome' $Y_{it}(x)$ denotes the outcome that we would (possibly contrary to fact) have observed for individual i at the end of period t, had the exposure X_{it} been set to a value x through some manipulation (Rubin, 1978). Since the AB/BA design dictates a dichotomous exposure (with a value 0 for baseline exposure or no exposure, and 1 otherwise), each subject is tied to exactly two potential outcomes during a specific period: $Y_{it}(0)$ and $Y_{it}(1)$. With these definitions, the individual period-specific total effect of X on Y is defined as the difference between both counterfactuals: $Y_{it}(1) - Y_{it}(0)$. Since only one of both potential outcomes is actually observed for each individual during period t, the period-specific individual total effect is unobserved. In contrast, the population average of the total causal effect $E[Y_{it}(1) - Y_{it}(0)]$ can be identified under specific assumptions (cf. infra).

Similarly, counterfactuals for the mediator, $M_{it}(0)$ and $M_{it}(1)$ can be defined. These represent the mediator values for an individual during period t, under exposure 0 and 1 respectively. Relying on these definitions, a nested counterfactual $Y_{it}(x, M_{it}(x^*))$ can be devised (Robins and Greenland, 1992; Pearl, 2001). It represents the value for the outcome Y_{it}, when X_{it} is set to x and M_{it} is fixed at the value it would obtain when $X_{it} = x^*$. Nested

counterfactuals allow us to rephrase the average period-specific total effect of X on Y, to include the mediator: $E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0))] = E[Y_{it}(1) - Y_{it}(0)]$. This moreover allows the partitioning of a total causal effect into a direct and indirect effect. Such effect decomposition can occur in two ways: one possibility is to decompose the total causal effect (TCE) into a total natural indirect effect ($TNIE$) and a pure natural direct effect ($PNDE$); the other decomposition yields a pure natural indirect effect ($PNIE$) and a total natural direct effect ($TNDE$) (Hafeman and Schwartz, 2009; VanderWeele, 2013):

$$\begin{aligned} TCE =& E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0))] \\ =& E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0)) + Y_{it}(1, M_{it}(0)) - Y_{it}(0, M_{it}(0))] \\ =& TNIE + PNDE \\ =& E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(1)) + Y_{it}(0, M_{it}(1)) - Y_{it}(0, M_{it}(0))] \\ =& TNDE + PNIE \end{aligned} \tag{2.2}$$

We will focus on the first decomposition ($TCE = TNIE + PNDE$) from now on.

2.2.2 Causal and modelling assumptions

To identify the (pure) natural direct and (total) natural indirect effect in multilevel settings, the standard set of 'no unmeasured confounding'-assumptions for simple settings with independent observations, has been generalised as follows (VanderWeele, 2010):

(i) There are no unmeasured upper- or lower-level confounders of the association between exposure and mediator.
(ii) There are no unmeasured upper- or lower-level confounders of the association between mediator and outcome.
(iii) There are no unmeasured upper- or lower-level confounders of the association between exposure and outcome.
(iv) There are no confounders of the association between mediator and outcome, caused by exposure (i.e. no intermediate confounding).

In crossover settings, the upper-level refers to the individual i, while the lower-level refers to the period t at which measurements were taken. Crossover designs render several of these assumptions obsolete. Since the sequence of exposure is by definition randomised, the first and third assumption are redundant. Also, as crossover studies are able to eliminate

between-subject variation, we will show that the second assumption can sometimes be weakened to: (iib) There are no unmeasured lower-level confounders of mediator and outcome. In addition to these four confounding assumptions, we add the following assumption:

(v) There is no causal transience (no carry-over effect): exposure, mediator and outcome measures from the first period cannot affect mediator and outcome measures from the second period.

Assumption (v) is plausible in crossover designs if the wash-out period is sufficiently long.

These assumptions related to the AB/BA design can be summarised by the (lack of) arrows in the directed acyclic graph of figure 2.3, which we will interpret as a nonparametric structural equation model with independent errors (Robins and Richardson, 2010). In this causal diagram, M_{i0} and Y_{i0} represent the mediator and outcome values for subject i, measured during the first period ($t = 0$), while M_{i1} and Y_{i1} denote these values during the second period ($t = 1$).

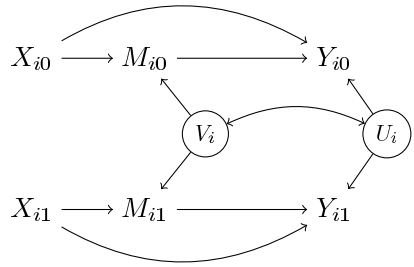

Figure 2.3 Causal diagram, graphically representing the assumptions regarding mediation in AB/BA crossover designs. The variables X_{i0}, M_{i0} and Y_{i0} represent the respective values of the exposure, mediator and outcome for subject i during the first measurement period. X_{i1}, M_{i1} and Y_{i1} on the other hand, reflect these variables assessed during the second measurement period. The unmeasured upper-level confounders V_i of the mediator and U_i of the outcome allow for upper-level M-Y confounding. Absence of a unidirectional arrow between two variables indicates the absence of a direct causal effect between them, while a bidirectional arrow captures an unmeasured common cause.

Note that in view of assumptions (iv) and (v), we do not allow the exposure, mediator and outcome measurements from the first period to causally affect the outcome and mediator values of the second period, respectively. Also, assumption (iv) dictates that measured within-subject confounders in the second period should not be affected by the exposure

(or the mediator) of the first period. We do, on the other hand, allow for unmeasured subject-specific, period-independent common causes U of the outcome to correlate with unmeasured subject-specific and period-independent common causes V of the mediator (relaxation of assumption (ii) into (iib)), provided that assumptions (vi) and (vii) hold. Note that V can be expressed as a function of U (i.e. $g(U)$) without loss of generality, rendering the unmeasured upper-level confounding of the M-Y relationship more explicit.

In addition to the above-mentioned causal assumptions, we will make the following modelling assumptions throughout this paper:

(vi) Unmeasured upper-level confounders of the association between mediator and outcome exert an additive effect on both the mediator and the outcome.

(vii) There is no unmeasured heterogeneity among subjects in the effect of exposure on mediator and in the effect of exposure and mediator on outcome.

Unlike path diagrams in the structural equation modelling framework, the lack of interactions implied by assumptions (vi) and (vii) (i.e. no interactions with unmeasured confounders U_i and V_i) cannot be represented on a causal diagram. These assumptions are therefore not depicted in figure 2.3.

Throughout the paper, we will make no assumptions regarding temporal stability and accordingly allow for period effects, implying that the outcome and mediator values can depend on the measurement moment. This is important because period effects are quite common in crossover studies (Tucker-Drob, 2011) (e.g. seasonal effects, changes in conditions of measurements, disease progression, habituation) and ignoring them would be disadvantageous for two reasons. First, if the exposure sequence were allocated in an unbalanced way, ignoring a period effect would bias the estimate of the exposure effect (Palta et al., 1994; Senn, 2002). Second, if such a trend exists but was not taken into account, its influence would be attributed to random variation instead of systematic changes, resulting in an inflated variance of the effect of X (Senn, 2002).

2.3 Estimating direct and indirect effects in simple settings with no interactions

Based on the above stipulated set of assumptions, we start with a simple data-generating mechanism for the mediator and outcome (for subject i during period t):

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + g(U_i) + \epsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \kappa_Y t + U_i + \epsilon_{Yit} \end{aligned} \tag{2.3}$$

Under this mechanism, in correspondence to figure 2.3, the mediator value of subject i during period t may be affected by exposure X_{it}, as well as by unmeasured individual level confounders $V_i = g(U_i)$. Similarly, the outcome of individual i during period t may be affected by the exposure X_{it}, the mediator M_{it} and any unmeasured subject level confounders U_i. In the equations above, the parameters δ_M and δ_Y represent the respective intercepts for M and Y, while α, β and ζ' represent the effects of exposure on mediator, mediator on outcome and exposure on outcome, respectively. Note that we assume that those effects are homogeneous across subjects (in accordance with assumption (vii)). The parameters κ_M and κ_Y define the respective period effects of M and Y. The presence of U_i and $g(U_i)$ in the models for Y and M, allows for unmeasured time-independent, subject-specific confounding of the M-Y relationship, without making strong parametric assumptions about their effects. Furthermore, U_i is independent of both exposure and period, and exhibits additive effects on mediator and outcome (in accordance with assumption (vi)). Finally, ϵ_{Mit} and ϵ_{Yit} represent the lower-level error terms, which are assumed to have mean zero and to be independent from the model predictors, as well as from one another.

Starting from this simple setting, the next subsection will first describe the identification of the natural direct and indirect effect. Next, we will summarise four existing approaches that can assess within-subject mediation in AB/BA crossover designs (sections 2.3.2 - 5.2).

2.3.1 Identification of the direct and indirect effect

Under the above described data-generating mechanism, the assumptions introduced in section 3.2 are met. This enables us to operate Pearl's mediation formula (Pearl, 2001, 2010) in deriving the total, pure natural direct and the total natural indirect effect for each subject i (conditional on

i) during period t. Based on equation (2.3), the subject- and period-specific total causal effect equals $\alpha\beta + \zeta'$, the individual- and period-specific total natural indirect effect equals $\alpha\beta$ and the pure natural direct effect in turn equals ζ' (detailed calculations can be found in appendix A.1).

2.3.2 The difference approach for the AB/BA design

Judd et al. (2001) proposed a straightforward method to evaluate mediation specifically in AB/BA crossover designs. They suggested an approach in which they perform regression on the differences of mediator and outcome values under both exposures, hereby eliminating between-subject variability. Following their approach, mediation can be assessed in three consecutive steps:

- *The first step* determines whether or not there is evidence for an overall effect of exposure on outcome, by performing a paired t-test on the outcomes under both exposures. Let $Y_i^{x=1}$ and $Y_i^{x=0}$ represent the outcome variables for subject i under exposure ($X = 1$) or no exposure ($X = 0$), respectively. When modelling the outcome differences through linear regression, the average effect of the exposure X on the outcome Y is estimated by the intercept c.

$$Y_i^{Dif} = Y_i^{x=1} - Y_i^{x=0} = c + e_{Yi1}^*$$

  with error terms e_{Yi1}^*. If there is evidence of such a total effect c different from zero, one can proceed to the next step.

- *The second step* tests whether or not there is evidence for an effect of X on M, by performing a paired t-test on the mediator values under both exposures. Let $M_i^{x=1}$ and $M_i^{x=0}$ represent the mediator variables for subject i under exposure ($X = 1$) or not ($X = 0$), respectively. The average effect of X on M can be estimated by the intercept a, from a linear regression model for these differences.

$$M_i^{Dif} = M_i^{x=1} - M_i^{x=0} = a + e_{Mi}^*$$

  with error terms e_{Mi}^*. If there is evidence of an effect of exposure on mediator, one can proceed to the next step.

- *The final step* assesses mediation itself. In the absence of moderation, mediation is evaluated by regressing the outcome differences (Y_i^{Dif}) on the mediator differences (M_i^{Dif}):

$$Y_i^{Dif} = c' + b_{Dif} M_i^{Dif} + e_{Yi2}^*$$

with error terms e^*_{Yi2}. Now, the intercept c' captures the direct effect, while the coefficient of M_i^{Dif} describes the effect of the mediator on the outcome. When it is found that the effect of M_i^{Dif} on the outcome differences is significantly different from zero, one can conclude that there is indeed mediation. Judd et al. (2001) argue that the type of mediation can subsequently be determined by the significance of the intercept in this equation: if it differs significantly from zero, partial mediation has occurred, if not, researchers can claim complete mediation.

Although this method elegantly bypasses the need for multilevel modelling approaches (which we will discuss from section 2.3.3 onwards) and eliminates between-subject variation (and hence any unmeasured confounders of the M-Y relationship that have additive effects at the subject-level), it has several drawbacks. Besides the frequently raised criticism concerning the necessity of each of the different steps (Collins et al., 1998; MacKinnon et al., 2000; Preacher et al., 2007; Hayes, 2009; Zhao et al., 2010), a first shortcoming is that the approach is not based on a quantification of the very thing it is attempting to test - the indirect effect (Hayes, 2009). A second drawback is that it does not account for period effects (Senn, 2002; Tucker-Drob, 2011).

2.3.3 Standard multilevel mediation analysis

Another approach for mediation analysis in the AB/BA design relies on multilevel modelling of the mediator and outcome (MacKinnon, 2008). Allowing for a period effect, the following lower-level equations would typically be considered:

$$\begin{aligned} M_{it} &= d_{Mi} + aX_{it} + k_M t + e_{Mit} \\ Y_{it} &= d_{Yi} + c'X_{it} + bM_{it} + k_Y t + e_{Yit} \end{aligned} \tag{2.4}$$

alongside the following upper-level (e.g. individual-level) equations:

$$\begin{aligned} d_{Mi} &= d_M + u_{Mi} && \text{with } u_{Mi} \amalg X_{it} \\ d_{Yi} &= d_Y + u_{Yi} && \text{with } u_{Yi} \amalg X_{it}, M_{it} \end{aligned}$$

Here, d_{Mi} and d_{Yi} represent the random intercepts, while e_{Mit} and e_{Yit} encode the lower-level error terms. The terms u_{Mi} and u_{Yi}, on the other hand, represent the upper-level error terms (subject level) for the random intercepts, assumed to be independent (as depicted by the symbol $\amalg$) of

the predictors in their respective equations. Both upper- and lower-level error terms are assumed to be independent of one another and normally distributed with mean zero.

Throughout this paper, maximum likelihood estimators for the parameters from the working models are denoted with ' $\hat{}$ '. The total natural indirect is estimated from (2.4) as $\hat{a}\hat{b}$, while the pure natural direct effect is estimated from (2.4) as $\hat{c}'$. Unfortunately, since u_{Yi} reflects unmeasured subject-specific variability in Y_{it} and is assumed to be independent of M_{it}, it is unable to capture the unmeasured subject-specific confounding of the M-Y relationship, under data-generating mechanism (2.3). This may result in an 'omitted variable bias' (Tofighi et al., 2013) for β and ζ' when such confounding is indeed present (as is the case in data-generating mechanism (2.3)), resulting in biased estimators for direct and indirect effect, ζ' and $\alpha\beta$, respectively. To this end, we will henceforth refer to this approach as the *Naive modelling approach*.

2.3.4 Approaches separating within-subject and between-subject effects

Many scholars have recently commented on the importance of separating within-subject from between-subject effects in multilevel settings (Louis, 1988; Neuhaus and Kalbfleisch, 1998; Begg and Parides, 2003; Zhang et al., 2009; Kenward and Roger, 2010; Preacher et al., 2010; Pituch and Stapleton, 2012). Within-effects (effects of the deviations from the subject means) and between-effects (effects of the subject means) can be different and even opposite in sign (Davis et al., 1961; Zhang et al., 2009). This can result from unmeasured upper-level confounding, which is absorbed in the between-subject effect (Goetgeluk and Vansteelandt, 2008). In view of this, allowing both effects in the outcome equation, will not dictate a 'forced average' of within- and between-effects, as demanded by the single parameter coefficient when no centring of the mediator is used. In this paper, we will focus on the within-subject effects, as these are of primary interest in crossover studies.

Following MacKinnon (2008), the within-subject effect can be estimated by regressing Y_{it} on the subject-mean centred mediator $(M_{it} - \overline{M}_{it})$. Here $\overline{M}_{it}$ denotes the subject-specific average of the M_{it} scores for subject i across periods. This modelling approach can be described by the following

set of linear mixed models:

$$\begin{aligned} M_{it} &= d_{Mi} + aX_{it} + k_M t + e_{Mit} && \text{with } d_{Mi} = d_M + u_{Mi} \\ Y_{it} &= d_{Yi} + c'X_{it} + b(M_{it} - \overline{M}_{it}) + k_Y t + e_{Yit} && \text{with } d_{Yi} = d_Y + u_{Yi} \end{aligned} \quad (2.5)$$

The lower-level residuals ϵ_{Mit} and ϵ_{Yit}, as well as the upper-level error terms u_{Mi} and u_{Yi}, are again assumed to be independently distributed with mean zero. Under the data-generating mechanism (equation (2.3)), the unmeasured confounder U_i is uncorrelated with $M_{it} - \overline{M}_{it}$. That is to say, while (un)measured individual level confounders of the outcome might correlate with the time-dependent M_{it} scores, these subject-mean centred mediators will no longer correlate with U_i. Hence, subtraction of the individual mean from period specific M_{it} scores will effectively eliminate any additive upper-level confounding of the M-Y relation, in contrast to the Naive modelling approach. We will refer to this procedure as the *Separate W(ithin)-only modelling approach.*

Assuming data-generating mechanism (2.3) holds, the total natural indirect can be estimated unbiasedly from equation (2.5) as $\hat{a}\hat{b}$ and the pure natural direct effect as $\hat{c'}$.

A second centring approach not only models the effect of the subject-mean centred mediator on the outcome, but also the effect of the subject mean of the mediator itself (MacKinnon, 2008). By doing so, two separate estimates for the effect of M on Y are obtained: a within-subject effect and a between-subject effect. This approach is equivalent to the Separate W-only approach for the estimation of within-subject effects in linear models, because of $\overline{M}_i$ being uncorrelated with $(M_{it} - \overline{M}_i)$ and is, as such, not considered any further.

2.3.5 A joint modelling approach

Another multilevel approach, described by Bauer et al. (2006), models the mediator and the outcome jointly, in a way that allows for unmeasured subject-specific common causes of M and Y, by incorporating a covariance term for the two random intercepts. Technically, this can be achieved by creating a new outcome variable Z which stacks M and Y for each period t within individual i. Next, two dummy variables are defined as follows: $S_M = 1$ when $Z = M$ and $S_M = 0$ otherwise, and similarly $S_Y = 1$ when $Z = Y$ and $S_Y = 0$ otherwise.

$$Z_{it} = S_M(d_{Mi} + aX_{it} + k_M t) + S_Y(d_{Yi} + c'X_{it} + bM_{it} + k_Y t) + e_{Zit} \quad (2.6)$$

This enables fitting a multivariate model, using univariate multilevel software (e.g. PROC MIXED in SAS). In contrast to the Naive modelling approach where u_{Mi} and u_{Yi} are assumed to be independent, this approach assumes the random intercepts to be bivariate normally distributed. Unmeasured upper-level M-Y confounding may therefore be captured by the correlation between both random effects. As such, it allows assessment of the viability of the assumption required in the Naive modelling approach, namely that no upper-level M-Y confounding is present. This method will be referred to as the *Joint modelling approach*.

The total effect under data-generating mechanism (2.3) is estimated from equation (2.6) as $\hat{a}\hat{b}$ and the pure natural direct effect as $\hat{c}'$. Since, in contrast to the separate modelling approaches, the estimation of fixed effects in the joint modelling approach relies on a bivariate normal distribution of the random intercepts, violation of this assumption may lead to biased fixed effects even if the mean is correctly specified. As shown in appendix A.2, one may expect bias when (a) M_{it} is non-normally distributed (because of non-normal random effects or residual errors), when (b) the distribution of u_{Yi} is non-normal or when (c) u_{Yi} moderates (i.e. modifies) the effect of X_{it} in the mediator model.

2.4 Estimating direct and indirect effects in more complex settings involving interactions

The data-generating mechanism that we considered so far (equation 2.3)), assumed no moderating effects of exposure. This section, allows for an interaction between exposure and mediator, moderation of the exposure effect by measured upper-level confounders D_i for both the mediator and outcome, as well as moderation of the mediator-outcome relationship by D_i. These effects can be jointly represented by the following data-generating mechanism:

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + \omega_M D_i + \nu_M X_{it} D_i + g(U_i) + \epsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \phi X_{it} M_{it} + \kappa_Y t + \omega_Y D_i + \nu_Y X_{it} D_i \\ &\quad + \eta M_{it} D_i + U_i + \epsilon_{Yit} \end{aligned} \tag{2.7}$$

Here, D_i is a vector of measured confounders at the subject level. Without loss of generality, we can assume that $E[D_i] = 0$. The parameter ϕ encodes the moderating effect of the mediator on the effect of exposure on outcome, while ν_M and ν_Y represent moderation of confounders D_i on the effect of

X on M and of X on Y, respectively. The parameter η in turn encodes the moderating effect of confounders D_i on the effect of the mediator on the outcome. As before, we assume additive effects for the unmeasured confounders U_i and $g(U_i)$.

2.4.1 Identification of the direct and indirect effect in complex settings

Under data-generating mechanism (2.7) the assumptions introduced in section 3.2 continue to apply, which enables use of the mediation formula to derive the total causal, pure natural direct and total natural indirect effect for subject i during period t. Here, the subject- and period-specific *total natural indirect effect* equals (detailed calculations can be found in appendix A.3):

$$\begin{aligned}
&E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0))|D_i = d, U_i] \\
&\quad = \sum_m \{E[Y_{it}|X_{it} = 1, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = 1, D_i = d, U_i) \\
&\qquad - E[Y_{it}|X_{it} = 1, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = 0, D_i = d, U_i)\} \\
&\quad = (\alpha + \nu_M d)(\beta + \phi + \eta d), \qquad (2.8)
\end{aligned}$$

Since this expression does not depend on U_i (see appendix A.3), the subject and period-specific total natural indirect effect can be marginalised over U_i ($E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0))|D_i = d, U_i] = E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0))|D_i = d]$).

This does not always hold for the pure natural direct and total causal effect; appendix A.3 demonstrates their dependence on unmeasured upper-level confounders U_i, when $\phi \neq 0$. This dependency on unmeasured upper-level confounders can be dealt with in one of two ways. A first possibility is to consider the *pure natural direct effect* at $g(U_i) = 0$:

$$\begin{aligned}
&E[Y_{it}(1, M_{it}(0)) - Y_{it}(0, M_{it}(0))|D_i = d, U_i] \\
&\quad = \sum_m \{E[Y_{it}|X_{it} = 1, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = 0, D_i = d, U_i) \\
&\qquad - E[Y_{it}|X_{it} = 0, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = 0, D_i = d, U_i)\} \\
&\quad = \zeta' + \nu_Y d + \phi(\delta_M + \kappa_M t + \omega_M d) \qquad (2.9)
\end{aligned}$$

and *total causal effect* at $g(U_i) = 0$:

$$\begin{aligned}
&E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0))|D_i = d, U_i] \\
&\quad = \sum_m \{E[Y_{it}|X_{it} = 1, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = 1, D_i = d, U_i) \\
&\qquad - E[Y_{it}|X_{it} = 0, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = 0, D_i = d, U_i)\} \\
&\quad = (\alpha + \nu_M d)(\beta + \phi + \eta d) + \zeta' + \nu_Y d + \phi(\delta_M + \kappa_M t + \omega_M d) \qquad (2.10)
\end{aligned}$$

respectively. However, these lack a clear interpretation due to the fact that U_i is unmeasured and the subgroup $g(U_i) = 0$ therefore unknown.

Alternatively, one may estimate the total causal effect marginally over U_i, for example by regressing the outcome Y_{it} on X_{it} and D_i. The pure natural direct effect marginalised over U_i can subsequently be estimated by subtracting the total natural indirect effect from this estimated total effect.

Finally note that for these settings the period effect comes into play, as the direct as well as the total effect now show time-dependency. The indirect effect however remains constant over time.

2.4.2 A more flexible Difference approach

The previously discussed difference approach by Judd et al. (2001) explicitly allows testing for one specific type of moderated mediation: moderation of the relation between exposure and outcome by the mediator itself. Technically, this is done by using the sum of the two mediator values M_i^{Sum}, as a predictor in addition to the difference M_i^{Dif}, in the model for the outcome: $Y_i^{Dif} = c' + b_{Dif}M_i^{Dif} + b_{Sum}M_i^{Sum} + e^*_{Yi2}$. Moderation is then assessed by testing whether or not b_{Sum} equals zero, but again no indirect effect estimators are derived. Interactions including external moderators are not allowed for by this approach either, but may often occur in practice (Edwards and Lambert, 2007; Fairchild and MacKinnon, 2009; Preacher et al., 2007). To allow for such moderation, as well as the above mentioned period effects, we will extend the approach proposed by Judd et al. (2001) as follows:

$$\begin{aligned}
M_i^{Dif} &= a + k_M t_i^{Dif} + v_M D_i + e_{Mi} \\
Y_i^{Dif} &= c' + bM_i^{Dif} + fXM_i^{Dif} + k_Y t_i^{Dif} + v_Y D_i + nD_i M_i^{Dif} + e_{Yi} \qquad (2.11)
\end{aligned}$$

In equation (2.11), XM_i^{Dif} equals $M_i^{x=1}$ and $t_i^{Dif} = t_i^{x=1} - t_i^{x=0}$, where $t_i^{x=0}$ and $t_i^{x=1}$ represent the measurement moments ($t = 0$ or 1) when no treatment and treatment were administered to individual i, respectively.

The error terms e_{Mi} and e_{Yi} are once again assumed to be normally and independently distributed with mean zero. We will refer to this approach as the *Difference approach* from now on. Since under data-generating mechanism (2.7):

$$\begin{aligned} M_i^{Dif} &= \alpha + \kappa_M t_i^{Dif} + \nu_M D_i + \epsilon_{Mi} \\ Y_i^{Dif} &= \zeta' + \beta M_i^{Dif} + \phi X M_i^{Dif} + \kappa_Y t_i^{Dif} + \nu_Y D_i + \eta D_i M_i^{Dif} + \epsilon_{Yi}, \end{aligned} \quad (2.12)$$

the difference approach will allow unbiased estimation of the indirect effect in this setting. The indirect effect can be estimated from equation (2.11) as $\hat{a}(\hat{b} + \hat{f})$ when $D_i = 0$, or by $(\hat{a} + \hat{v}_M d)(\hat{b} + \hat{f} + \hat{n}d)$ when $D_i = d$. Estimation of the direct effect, on the other hand, is more complicated. When there is no X-M interaction ($\phi = 0$), all parameters that constitute the direct effect can be unbiasedly estimated with the Difference approach under the assumed data-generating mechanism. However, when $\phi \neq 0$, we suggest the above-mentioned approach based on subtracting the estimated indirect effect from the total effect (both marginalised over U_i).

A final remark can be made regarding equation (2.11), which assumes the subject-level confounders D_i are measured. If these confounders remain unmeasured, however (and are therefore not included in the estimating equations), assumption (vi) will be violated and the Difference approach may yield biased estimates for the parameters of interest. There may be bias in the estimates for the parameters in the outcome equation (2.12) when $\nu_Y \neq 0$ or $\eta \neq 0$, as the interaction term between exposure and D_i is correlated with M_i^{Dif}, and the interaction term between the mediator and D_i is correlated with $M_i^{x=1}$. However, when $\nu_Y = 0$ and $\eta = 0$, but $\nu_M \neq 0$, the estimator $\hat{a}_d(\hat{b}_d + \hat{f}_d)$ (where the subscript d refers to the estimates in the model ignoring D_i) will still provide an unbiased estimate of the indirect effect at average levels of D_i. Indeed, when $\nu_Y = \eta = 0$, omitting D_i introduces no bias for $\hat{b}$ and $\hat{f}$, and as residuals are assumed to have a zero mean, the intercept $\hat{a}_d$ reflects the effect of exposure on mediator at average values of D_i.

2.4.3 The Naive, Separate W-only and Joint modelling approach in complex settings

The Naive and Joint modelling approach can incorporate the moderation effects present in data-generating mechanism (2.7), by adding the respective

interaction terms to the models. These become:

$$\begin{aligned} M_{it} &= d_{Mi} + aX_{it} + k_M t + w_M D_i + v_M X_{it} D_i + e_{Mit} \\ Y_{it} &= d_{Yi} + c' X_{it} + bM_{it} + fX_{it} M_{it} + k_Y t + w_Y D_i + v_Y X_{it} D_i \\ &\quad + nM_{it} D_i + e_{Yit} \end{aligned} \tag{2.13}$$

where the Joint modelling approach additionally models a covariance term for both random effects.

The method that separates between- from within-subject effects can also incorporate such moderation, by adding interaction terms with the subject-mean deviation scores of the mediator. The model becomes:

$$\begin{aligned} M_{it} &= d_{Mi} + aX_{it} + k_M t + w_M D_i + v_M X_{it} D_i + e_{Mit} \\ Y_{it} &= d_{Yi} + c' X_{it} + b(M_{it} - \overline{M}_{it}) + f(X_{it} M_{it} - \overline{X_{it} M}_{it}) + k_Y t \\ &\quad + w_Y D_i + v_Y X_{it} D_i + n(M_{it} - \overline{M}_{it}) D_i + e_{Yit} \end{aligned} \tag{2.14}$$

Note that the X-M interaction is modelled as the difference between the product of the individual, time-specific exposure and mediator values, and the average of this product over periods, within individuals; modelling it any other way (e.g. as $X_{it}(M_{it} - \overline{M}_{it})$), might lead to bias in the presence of unmeasured upper-level M-Y confounding.

All three approaches will produce unbiased estimates of the direct and indirect effect under data generating mechanism (2.7), but in contrast to the Difference approach, the Separate W-only approach and Joint modelling approaches also require modelling (and hence potentially correct specification) the main effect of D_i.

2.5 Simulation study

To gain insight into the finite sample performance of all four modelling approaches represented by equations (2.11), (2.13) and (2.14), we compare them through simulations. For simplicity, we assume no measured subject-specific confounders D that moderate the treatment or mediator effect.

We consider three different simulation settings, which are defined as special cases of a general data-generating mechanism, specified by the following models for M and Y:

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + \nu_M V_i X_{it} + V_i + \epsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \phi X_{it} M_{it} + \kappa_Y t + U_i + \epsilon_{Yit} \end{aligned} \tag{2.15}$$

Here, V_i and U_i represent zero-mean bivariate normally distributed un-

measured individual-level confounders, with a variance of 4 and covariance σ_{V_i,U_i}. We generated independently and normally distributed error terms ϵ_{Mit} and ϵ_{Yit}, with mean zero and variance 9 and 16 respectively. As deviations from normality for either the individual-level confounders or the lower-level error terms have little or no effect (Verbeke and Lesaffre, 1997; McCulloch and Neuhaus, 2011), we kept these distributions fixed. Results based on misspecified random effects in the data-generating mechanism confirmed that such linear mixed models are very robust against any such incorrect specifications (results not shown). We also generate period effects for both the mediator and outcome (κ_M and κ_Y respectively), and an exposure-mediator interaction for the outcome (ϕ) as well as an exposure-'unmeasured confounder' interaction for the mediator (ν_M). Note that when $\nu_M \neq 0$, assumption (vi) is violated.

For the first simulation setting, the M and Y values are generated according to equation (5.7), but with ν_M and σ_{V_i,U_i} both set to zero (thus satisfying the assumptions in section 3.2). The other parameters are fixed, with $\delta_M = 1$, $\delta_Y = 1.5$, $\alpha = 3$, $\zeta' = 2$, $\beta = -1$, $\phi = 2$, $\kappa_M = 0.1$ and $\kappa_Y = 0.2$. For the second simulation setting, we allowed for a non-zero covariance term between both upper-level confounders U_i and V_i, with $\sigma_{V_i,U_i} = 0.50$. In the third simulation setting, we also considered $\sigma_{V_i,U_i} = 0.50$ but in addition set the parameter value of ν_M to 1 (thus violating assumption (vi) in section 3.2).

To get an indication of how the four different modelling approaches are affected by sample size, we considered samples of size $N = 50$ and $N = 200$. Together, these varying factors yield 6 conditions (3 simulation settings and 2 sample sizes) for which 500 data sets were generated.

For each method, the average value, empirical standard error and the coverage of the 95% confidence intervals of the β, ζ' and ϕ estimators are provided. Note that the respective estimators for β, ζ' and ϕ are given by $\hat{b}$, $\hat{c'}$ and $\hat{f}$, for all approaches (equations (2.11), (2.13) and (2.14)). Additionally, the square root of the Mean Squared Error is provided for these estimators. The results for $N = 50$ and $N = 200$ are shown in table 2.1. The parameters in this table that show significant bias (as indicated by a significant deviation of the empirical mean from the true mean) are marked in boldface.

N	Method	$\beta = -1.00$			$\zeta' = 2.00$			$\phi = 2.00$		
		mean (se)	cov	$\sqrt{MSE}$	mean (se)	cov	$\sqrt{MSE}$	mean (se)	cov	$\sqrt{MSE}$
50	Naive	-1.00 (0.18)	93.40	0.18	1.93 (1.02)	96.40	1.02	2.00 (0.23)	94.60	0.23
	Sep-W	-1.01 (0.24)	96.40	0.24	1.94 (1.18)	95.20	1.18	2.00 (0.27)	95.40	0.27
	Joint	-1.01 (0.24)	84.60	0.24	1.94 (1.11)	94.20	1.11	2.00 (0.23)	94.80	0.23
(1)	Diff	-1.01 (0.24)	96.40	0.24	1.94 (1.18)	95.20	1.18	2.00 (0.27)	95.40	0.27
50	Naive	**-0.86** (0.18)	86.00	0.23	**1.51** (1.03)	94.20	1.14	2.00 (0.23)	94.60	0.23
	Sep-W	-1.01 (0.24)	96.40	0.24	1.94 (1.18)	95.20	1.18	2.00 (0.27)	95.40	0.27
	Joint	-1.01 (0.23)	84.80	0.23	1.95 (1.12)	94.20	1.12	2.00 (0.23)	94.80	0.23
(2)	Diff	-1.01 (0.24)	96.40	0.24	1.94 (1.18)	95.20	1.18	2.00 (0.27)	95.40	0.27
50	Naive	**-0.85** (0.18)	85.20	0.23	**1.50** (0.92)	94.00	1.05	2.00 (0.20)	94.20	0.20
	Sep-W	-1.01 (0.26)	96.40	0.26	1.94 (1.02)	95.80	1.02	2.00 (0.23)	95.40	0.23
	Joint	-1.03 (0.25)	81.00	0.25	1.89 (0.99)	94.60	1.00	2.03 (0.20)	94.00	0.20
(3)	Diff	-1.01 (0.26)	96.40	0.26	1.94 (1.02)	95.80	1.02	2.00 (0.23)	95.40	0.23
200	Naive	-1.00 (0.08)	96.40	0.08	1.98 (0.54)	93.60	0.54	2.00 (0.11)	96.60	0.1
	Sep-W	-1.00 (0.11)	95.80	0.11	1.98 (0.62)	93.60	0.62	2.00 (0.13)	96.40	0.13
	Joint	-1.00 (0.11)	88.20	0.11	1.98 (0.59)	91.60	0.59	2.00 (0.11)	96.40	0.11
(1)	Diff	-1.00 (0.11)	95.80	0.11	1.98 (0.62)	93.60	0.62	2.00 (0.13)	96.40	0.13
200	Naive	**-0.86** (0.08)	64.60	0.16	**1.57** (0.54)	88.80	0.69	2.00 (0.11)	97.40	0.11
	Sep-W	-1.00 (0.11)	95.80	0.11	1.98 (0.62)	93.60	0.62	2.00 (0.13)	96.40	0.13
	Joint	-1.00 (0.11)	88.20	0.11	1.98 (0.59)	91.00	0.59	2.00 (0.11)	97.40	0.11
(2)	Diff	-1.00 (0.11)	95.80	0.11	1.98 (0.62)	93.60	0.62	2.00 (0.13)	96.40	0.13
200	Naive	**-0.86** (0.08)	61.80	0.16	**1.56** (0.49)	85.60	0.66	2.00 (0.10)	96.20	0.10
	Sep-W	-1.00 (0.12)	95.80	0.12	1.98 (0.53)	93.80	0.53	2.00 (0.11)	95.60	0.11
	Joint	**-1.02** (0.11)	83.60	0.11	**1.93** (0.53)	92.60	0.53	**2.03** (0.10)	94.60	0.10
(3)	Diff	-1.00 (0.12)	95.80	0.12	1.98 (0.53)	93.80	0.53	2.00 (0.11)	95.60	0.11

Table 2.1 Results of fitting each of the four within-subject modelling approaches for the simulated data for $N = 50$ (upper table) and for $N = 200$ (lower table). Simulations (1) to (3) represent the three possible data-generating mechanisms, while 'Naive', 'Joint', 'Sep-W' and 'Diff' represent the four different modelling approaches: the Separate modelling, Joint modelling, Separate W-only modelling and Difference approach, respectively. For each method applied to each setting, the average value (*mean*), empirical standard error (*se* = the standard deviation of the estimates over the 500 replications) and the coverage of the 95% confidence intervals (*cov* = percentage of the 500 simulations in which the 95% Wald confidence intervals of the estimators contain the true parameter value) for β, ζ' and ϕ over the 500 simulations are provided. On top of this, the square root of the Mean Squared Error is also provided for these parameters ($\sqrt{MSE}$). The means of the estimates over the 500 simulations that show bias (when the true parameter value is not included in the empirical 95% confidence interval) are written in bold text.

2.5.1 Parameter estimates

As expected, the Naive modelling approach provides unbiased estimates of the β, ζ' and ϕ parameters for both sample sizes, as long as no unmeasured confounding of the M-Y relation is present (first data-generating mechanism). As soon as a non-zero covariance between V_i and U_i is present, we observe bias in the parameter estimators $\hat{b}$ and $\hat{c}'$, but not in $\hat{f}$. The Difference and Separate W-only modelling approaches yield unbiased estimators for all three effects of interest, irrespective of the data-generating mechanism or sample size (so even when assumption (vi) is violated). Moreover, both methods provide identical estimators for β, ζ' and ϕ. This equivalence in estimates form the Difference and Separate W-only modelling approach is expected when there are no random slopes (Goetgeluk and Vansteelandt, 2008). Lastly, the Joint modelling approach performs rather well in terms of bias, except for the estimators obtained under the third simulation setting (where assumption (vi) is violated) for the larger sample size, where we find significant bias for all three parameters. This bias follows from the arguments provided in section 5.2, as the assumption of no moderation of the effect of X_{it} on M_{it} by u_{Yi} (condition (c)) is violated under the third setting ($\nu_M \neq 0$).

2.5.2 Coverage and mean squared error

First of all, as long as the parameter estimates themselves are unbiased, we observe good coverages for all modelling approaches and all parameters. One exception is the low coverage of the estimator for β, obtained by the Joint modelling approach. However, this undercoverage improves as the sample size increases. The Separate W-only and Difference approach differ slightly in their estimated standard errors (even though the empirical standard error is the same), with the Separate W-only approach yielding the largest, and the Difference approach providing the smallest.

Secondly, as long as there is no unmeasured confounding of the M-Y relation, we observe the lowest Mean Squared Error (MSE) for the Naive modelling approach. As soon as such confounding is introduced however, the MSE of the Naive approach increases to a level at least as high as the MSE's of the other three approaches. Overall, of these methods the Joint modelling approach seems to provide the lowest MSE's, while the other two collectively yield the highest. This is not surprising, considering that the Joint modelling approach is based on maximum likelihood under a more restrictive model.

2.6 Analysis of a neurostimulation experiment

We applied the different estimation approaches discussed in this paper to data from a recent crossover study in behavioural neuroscience (Vanderhasselt et al., 2013). This crossover study evaluates the effect of anodal transcranial direct current stimulation (tDCS) over the dorsolateral prefrontal cortex on the occurrence of self-referent thoughts, in 32 healthy participants. This neuromodulatory technique applies a weak electric current during 20 minutes (through the use of electrodes), which induces polarisation-shifts in the resting membrane potential (Brunoni et al., 2012). It was postulated that tDCS-exposure ($X = 1$ for tDCS stimulation, $X = 0$ for placebo stimulation) affected the outcome (self-referent thoughts) by inducing changes in the ability to shift from negative representations in the working memory (the mediator). The wash-out period lasted for a minimum of 48 hours, and since current research suggests an intersession interval of 48 hours after a long stimulation is more than sufficient (Nitsche et al., 2008), the absence of carry-over effects is very plausible here. With respect to the assumptions, we mentioned earlier that some of the estimation approaches can deal with unmeasured M-Y confounding at the upper-level, this in contrast to such confounding at the lower-level. To assess robustness against violations of this 'no unmeasured confounding'-assumption at the lower-level, a sensitivity analysis will be presented at the end of this section.

For all approaches, we start from simple estimation without interactions but accounting for period effects. The direct and indirect effect estimates tied to these models, alongside their 95% confidence intervals (these estimates and 95% percentile-based confidence intervals were obtained through bias-corrected bootstrapping, based on 1000 bootstrap samples), can be found at the left panel of figure 2.4 (**A.**). The estimates of the Naive multilevel modelling method stand out clearly, which may imply that there is indeed unmeasured individual-level confounding present. Such confounding can be captured by the correlation between both random intercepts from the Joint modelling approach, and is estimated at -0.43 ($p = 0.31$). If such unmeasured confounding is indeed present, the Naive method produces biased estimates, in contrast to the other three approaches (provided the corresponding models hold). As we have not yet included any nonlinearities in our models at this point, the three other methods yield almost identical results.

In a second step, we check for moderation effects of a centred subject-

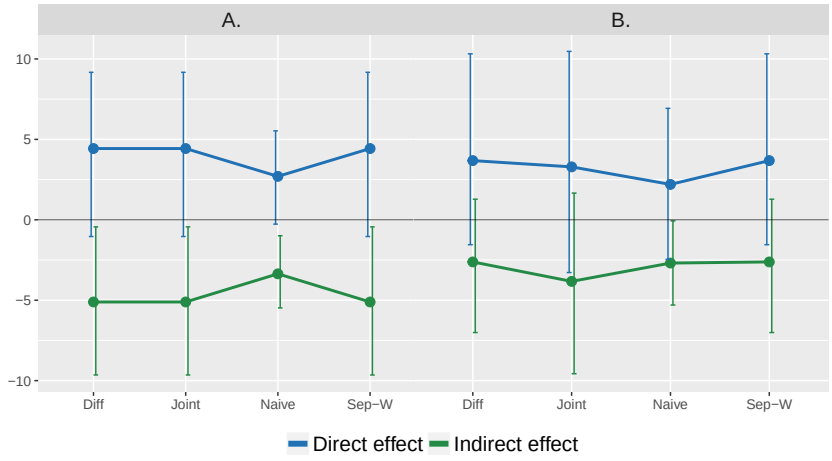

Figure 2.4
A. Results of fitting each of the four within-subject modelling approaches for the neurostimulation data, according to the settings summarised by equation (2.3). 'Naive', 'Joint', 'Sep-W' and 'Diff' represent the four different approaches: the Separate, Joint, Separate W-only and Difference approach, respectively. For each method applied to each setting, the estimated direct and indirect effect are given, alongside their 95% confidence intervals (these estimates and confidence intervals are obtained by percentile-based bias-corrected bootstrapping, based on 1000 samples).
B. Results of fitting each within-subject modelling approach for the neurostimulation data, according to the settings summarised by equation (2.7) (with $\phi = 0$).

specific baseline confounder D (representing trait rumination, a stable subject-specific measure), in the models for the mediator and outcome, as well as an X-M interaction in the model for the outcome. From the models based on the difference approach, we find evidence for an interaction between X and D in both the mediator ($p = 0.018$) and outcome equations ($p = 0.054$), as well as an interaction between M and D in the outcome equation ($p = 0.0057$). In contrast, the exposure-mediator interaction in the outcome model is not significant ($p = 0.92$), and for this reason excluded from the model. As mentioned before, the direct and indirect effect under this assumed data-generating mechanism, do not depend on the main effects of D, nor on the period effect and the unobserved U_i. An estimate of the direct and indirect effect at average values of trait rumination ($d = 0$), accompanied by their 95% confidence intervals is provided at the right side of figure 2.4 (**B.**). We observe no significant indirect and direct effect at $d = 0$, except for the Naive modelling approach;

once again, the estimates from this approach stand out. Note that the Separate W-only and Difference approach again yield identical results for both causal effects, while the Joint modelling approach provides somewhat different estimates in the presence of the above mentioned interactions. This most likely results from the presence of X-D interactions, as the linearity assumptions for u_{Yi} might be violated in the Joint modelling approach (see supporting appendix A.2). It is also worth mentioning that with the additional inclusion of D and its interactions, the Joint modelling approach now yields a correlation between both random intercepts of -0.21 (p=0.58), which is already closer to zero.

Additionally, a plot is provided for the direct and indirect effects obtained from the difference approach, over the total range of values for trait rumination (ranging from -11.87 to 37.13, see upper panel in figure 2.5 (**A.**)). As hypothesised, significant indirect effects are only observed for high levels of trait rumination (Vanderhasselt et al., 2013). An average direct effect on the other hand, remains absent over the entire range of D-values.

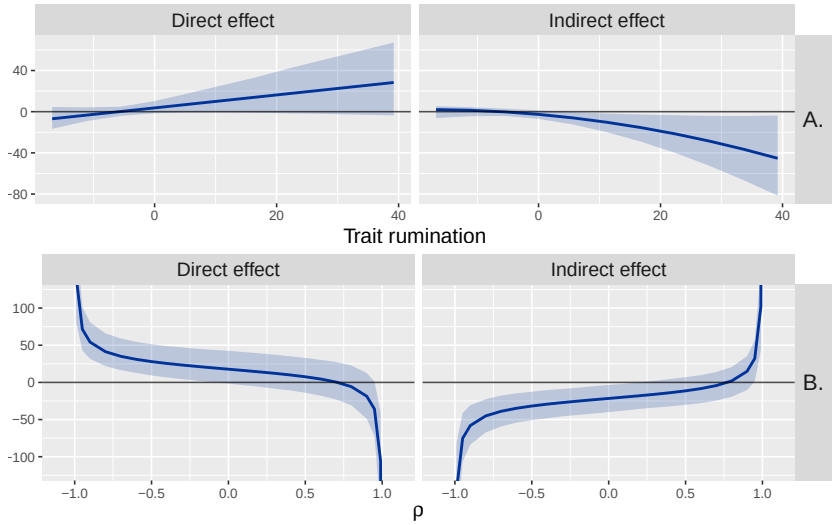

Figure 2.5
A. The average direct (on the left) and indirect effects (on the right) and their 95% confidence intervals (95% CI)) obtained from the difference approach (by percentile-based bias-corrected bootstrapping, based on 1000 samples), over the total range of values for trait rumination $D \in [-11.87, 7.13]$.
B. The average direct (on the left) and indirect effects (on the right) and their 95% confidence intervals (95% CI)) at $D = 22.38$ (= 2 standard deviations above the mean), over a range of values for the sensitivity parameter $\rho \in [-1, 1]$ (the estimates, alongside their 95% confidence intervals were obtained trough percentile-based bias-corrected bootstrapping, based on 1000 samples).

2.6.1 A sensitivity analysis for omitted lower-level M-Y confounding

While the assumption of 'no unmeasured upper-level M-Y confounding' is not necessary, the absence of unmeasured lower-level confounding of the M-Y relation remains essential for unbiased estimation of the direct and indirect effect. In this section, we present a sensitivity analysis that is able to assess the impact of such lower-level M-Y confounding. More precisely, we assume the following extension of data-generating mechanism (2.7):

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + \omega_M D_i + \nu_M D_i X_{it} + g(U_i) + \epsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \phi X_{it} M_{it} + \kappa_Y t + \omega_Y D_i + \nu_Y D_i X_{it} + \eta D_i M_{it} \\ &\quad + U_i + \theta \epsilon_{Mit} + \epsilon_{Yit} \end{aligned} \tag{2.16}$$

where θ represents the influence of the lower-level residuals from the mediator equation on the outcome. Values of θ different from zero imply violation of the 'no unmeasured lower-level M-Y confounding' assumption. Note that in our neurobehavioral example, ϕ is assumed to be zero, but is included here to allow for generalisation. In terms of differences it follows from (2.16) that:

$$\begin{aligned} M_i^{Dif} &= \alpha + \kappa_M t_i^{Dif} + \nu_M D_i + \epsilon_{Mi} \\ Y_i^{Dif} &= \zeta' + \beta M_i^{Dif} + \phi X M_i^{Dif} + \kappa_Y t_i^{Dif} + \nu_Y D_i + \eta D_i M_i^{Dif} \\ &\quad + \theta \epsilon_{Mi} + \epsilon_{Yi} \end{aligned} \tag{2.17}$$

with ϵ_{Mi} and ϵ_{Yi} encoding the difference in M- and Y-residuals from (2.16). When ϵ_{Mi} is substituted by $M_i^{Dif} - \alpha - \kappa_M t_i^{Dif} - \nu_M D_i$ in the outcome equation, we obtain:

$$\begin{aligned} Y_i^{Dif} &= (\zeta' - \theta\alpha) + (\beta + \theta) M_i^{Dif} + \phi X M_i^{Dif} + (\kappa_Y - \theta\kappa_M) t_i^{Dif} \\ &\quad + (\nu_Y + \theta\nu_M) D_i + \eta D_i M_i^{Dif} + \epsilon_{Yi} \end{aligned} \tag{2.18}$$

which can be rewritten as:

$$\begin{aligned} Y_i^{Dif} &= \zeta'^* + \beta^* M_i^{Dif} + \phi (XM)_i^{Dif} + \kappa_Y^* t_i^{Dif} + \nu_Y^* D_i + \eta D_i M_i^{Dif} + \epsilon_{Yi}\ , \\ &\text{where:} \begin{cases} \zeta' &= \zeta'^* + \theta\alpha \\ \beta &= \beta^* - \theta \\ \kappa_Y &= \kappa_Y^* + \theta\kappa_M \\ \nu_Y &= \nu_Y^* + \theta\nu_M \end{cases} \end{aligned} \tag{2.19}$$

As such, under data-generating mechanism (2.16), the difference approach based on model (2.11) would result in biased parameter estimators for the effects on the outcome, with bias depending on the value of θ. To simplify interpretation, we will use a sensitivity parameter ρ, representing the correlation between the residual error terms in equation (2.17) (ϵ_{Mi} and $\theta\epsilon_{Mi} + \epsilon_{Yi}$), rather than θ (Imai et al., 2010). It can be shown that:

$$\theta = \frac{\rho}{\sqrt{1-\rho^2}} \frac{\sigma_{\epsilon Y}}{\sigma_{\epsilon M}} \tag{2.20}$$

Under the above setting, the sensitivity analysis then proceeds as follows. First, all parameters in the difference equations (2.11) are estimated and the estimates for the residual error variances, $\sigma_{\epsilon M}$ and $\sigma_{\epsilon Y}$, are determined, assuming that $\theta = 0$. Next, a plausible range of values of ρ (varying between -1 and 1) is considered while keeping $\sigma_{\epsilon M}$ and $\sigma_{\epsilon Y}$ fixed, so that θ can be calculated by applying expression (2.20). Then, relying on the estimate for

θ, the estimated parameters from (2.11) and the equalities on the right side of (2.19), estimates for the true parameters can be obtained. Additionally, precision of the resulting direct and indirect effects at different values of ρ can be assessed by bootstrapping procedures.

We will now illustrate the above proposed sensitivity analysis on our neurobehavioral data. More specifically, we look at the estimated indirect effect for large values of trait rumination (at $D = 2$ standard deviations above the mean = 22.38), since figure 2.5 (**A.**) revealed that this effect exists but for high values of D. At this value for D, we investigate how extensive the amount of unmeasured M-Y confounding at the lower-level must become in order for the indirect effect to vanish. For values of ρ ranging from -1.00 to 1.00, we estimate the direct and indirect effect (results are shown in lower panel, figure 2.5 (**B.**). We observe that the indirect effect disappears when unmeasured lower-level covariates induce a residual correlation between M and Y larger than 0.20.

2.7 Discussion

In this paper we presented and compared different modelling strategies for the estimation of the direct and indirect effect in crossover studies. First and foremost, since the absence of unmeasured upper-level M-Y confounding can never be guaranteed, we do not recommend the Naive modelling approach in any setting. Furthermore, we showed that the Joint modelling method relies on stronger modelling assumptions than the Difference or Separate W-only modelling approaches. The latter two approaches yield identical estimators in the absence of upper-level heterogeneity. In the presence of both exposure-mediator interactions and interactions with measured subject level confounders D, we have shown how to obtain unbiased direct and indirect effects at specific levels of D, even when D is correlated with unmeasured confounders. In general, the difference approach is simpler to apply and might for this reason prove more accessible to researchers unfamiliar with mixed effects models.

From a practical perspective, it is important to have clarified the underlying assumptions of each of the different approaches here. Note that easily accessible software for mediation analysis in the multilevel setting, such as the R *mediation* package (Tingley et al., 2014)), relies on separate linear mixed models for the mediator and outcome (if both are measured at the interval level). Considering our findings, these will only yield valid inference under unmeasured upper-level M-Y confounding,

when the subject-specific deviation scores for the mediator are used in the outcome equation model. While we focused on linear settings in this paper, the aforementioned *mediation* package additionally tackles non-linear multilevel settings. The question of whether or not the approach of Imai et al. (2010) yields unbiased estimators for the direct and indirect effect in the presence of unmeasured subject-level confounders in non-linear settings, remains to be explored. However, separating within- and between-effects in mixed models with log- or logit-links may yield inconsistent within-subject effects in the presence of unmeasured subject-specific confounders (Goetgeluk and Vansteelandt, 2008). We conjecture that the *mediation* package approach in the multilevel level setting may require assumptions that are too stringent, even if centred predictors were used. Other estimation approaches may thus be indicated, e.g. conditional generalised estimating equations (CGEE) provide a more general framework for sheltering the estimation of within-subject effects from unmeasured between-subject confounding factors (Goetgeluk and Vansteelandt, 2008).

Throughout this paper, we remained silent about the incorporation of measured lower-level confounders. Although at first sight it may seem very straightforward to incorporate such confounders in the four approaches we discussed, their inclusion requires additional thought. Assumption (iv), for example, dictates that measured within-subject confounders from the second period ought to be unaffected by the exposure (or the mediator) of the first period. If violated, we end up with time-dependent or intermediate confounding. It remains to be investigated how techniques such as inverse probability weighting (Robins, 1999) or G-estimation (Goetgeluk et al., 2009), that can deal with intermediate confounding concerning the estimation of the controlled direct effect in single level settings, could be applied to multilevel settings.

Bibliography

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Bauer, D. J., Preacher, K. J., and Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel

models: New procedures and recommendations. *Psychological Methods*, 11(2):142–163.

Begg, M. D. and Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, 22(16):2591–2602.

Brunoni, A. R., Nitsche, M. a., Bolognini, N., Bikson, M., Wagner, T., Merabet, L., Edwards, D. J., Valero-Cabre, A., Rotenberg, A., Pascual-Leone, A., Ferrucci, R., Priori, A., Boggio, P. S., and Fregni, F. (2012). Clinical research with transcranial direct current stimulation (tDCS): Challenges and future directions. *Brain Stimulation*, 5(3):175–195.

Collins, L. M., Graham, J. J., and Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, 33(2):295–312.

Davis, J. A., Spaeth, J. L., and Huson, C. (1961). A technique for analyzing the effects of group composition. *American Sociological Review*, 26(2):215–225.

Edwards, J. R. and Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, 12(1):1–22.

Fairchild, A. J. and MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, 10(2):87–99.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780.

Goetgeluk, S., Vansteelandt, S., and Goetghebeur, E. (2009). Estimation of controlled direct effects. *Journal of the Royal Statistical Society: Series B*, 70(5):1049–1066.

Hafeman, D. M. and Schwartz, S. (2009). Opening the Black Box: A motivation for the assessment of mediation. *International Journal of Epidemiology*, 38(3):838–845.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4):408–420.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115–134.

Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8(2):115–128.

Kenward, M. G. and Roger, J. H. (2010). The use of baseline covariates in crossover studies. *Biostatistics*, 11(1):1–17.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., and Vansteelandt, S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivariate Behavioral Research*, 48(6):871–894.

Louis, T. A. (1988). General methods for analyzing repeated measures. *Statistics in Medicine*, 7(1-2):29–45.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Taylor & Francis Group, LLC, New York.

Mackinnon, D. P. and Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, 17(2):144–158.

MacKinnon, D. P., Krull, J. L., and Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4):173–181.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. *Statistical Science*, 26(3):388–402.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

Nitsche, M. a., Cohen, L. G., Wassermann, E. M., Priori, A., Lang, N., Antal, A., Paulus, W., Hummel, F., Boggio, P. S., Fregni, F., and Pascual-Leone, A. (2008). Transcranial direct current stimulation: State of the art 2008. *Brain Stimulation*, 1(3):206–223.

Palta, M., Yao, T. J., and Velu, R. (1994). Testing for omitted variables and non-linearity in regression models for longitudinal data. *Statistics in Medicine*, 13(21):2219–2231.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):Article 7.

Pearl, J. (2012). The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–36.

Pituch, K. a. and Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research*, 41(4):630–670.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1):185–227.

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3):209–233.

Raykov, T. and Mels, G. (2007). Lower level mediation effect analysis in two-level studies: A note on a multilevel structural equation modeling approach. *Structural Equation Modeling*, 14(4):636–648.

Robins, J. M. (1999). Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In Glymour, C. and Cooper, G., editors, *Computation, Causation, and Discovery*, pages 349–405. AAAI Press/The MIT Press, Menlo Park, CA, Cambridge, MA.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):153–155.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Senn, S. (2002). *Cross-over trials in clinical research*. John Wiley & Sons, Chichester.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Tofighi, D., West, S. G., and Mackinnon, D. P. (2013). Multilevel mediation analysis: The effects of omitted variables in the 1-1-1 model. *The British Journal of Mathematical and Statistical Psychology*, 66(2):290–307.

Tucker-Drob, E. M. (2011). Individual differences methods for randomized experiments. *Psychological Methods*, 16(3):298–318.

Vanderhasselt, M.-A., Brunoni, A. R., Loeys, T., Boggio, P. S., and De Raedt, R. (2013). Nosce te ipsum–Socrates revisited? Controlling momentary ruminative self-referent thoughts by neuromodulation of emotional working memory. *Neuropsychologia*, 51(13):2581–2589.

VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, 38(4):515–544.

VanderWeele, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23(4):541–556.

Zhang, Z., Zyphur, M. J., and Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models. *Organizational Research Methods*, 12(4):695–719.

Zhao, X., Lynch Jr., J. G., and Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2):197–206.

A Appendix

A.1 Identification of the causal effects in simple settings

We start from a more general data generating mechanism (compared to model (2.3)), which relaxes assumption (vii):

$$\begin{aligned} M_{it} &= \delta_M + \alpha_i X_{it} + \kappa_{Mi} t + g(U_i) + \epsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta_i' X_{it} + \beta_i M_{it} + \kappa_{Yi} t + U_i + \epsilon_{Yit} \end{aligned} \tag{2.21}$$

Note that the AB/BA design with a single measurement in each of the two conditions, does not allow identification of such heterogeneous effects. This set of equations (2.21), however, encompasses the restrictions dictated by this design (and therefore also equation (2.3)). Based on this generalised data generating mechanism summarised in expression (2.21), the "*it*-th"-specific *total natural indirect effect* can be identified, when the assumptions (i)-(vi) from section 3.2 are satisfied:

$$\begin{aligned} &E[Y_{it}(x, M_{it}(x)) - Y_{it}(x, M_{it}(x^*))|\alpha_i, \kappa_{Mi}, \beta_i, \zeta_i', \kappa_{Yi}, U_i] \\ &= \sum_m \Big(E[Y_{it}|X_{it} = x, M_{it} = m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it} = m|X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\ &\quad - E[Y_{it}|X_{it} = x, M_{it} = m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it} = m|X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \Big) \\ &= \sum_m (d_Y + \zeta_i' x + \beta_i m + \kappa_{Yi} t + U_i) \cdot \Big(P(M_{it} = m|X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\ &\quad - P(M_{it} = m|X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \Big) \\ &= \beta_i \Big(\sum_m m P(M_{it} = m|X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\ &\quad - \sum_m m P(M_{it} = m|X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \Big) \\ &= \beta_i \Big(E[M_{it}|X_{it} = x, \alpha_i, \kappa_{Mi}, U_i] - E[M_{it}^*|X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \Big) \\ &= \beta_i \Big(d_M + \alpha_i x + \kappa_{Mi} t + g(U_i) - d_M - \alpha_i x^* - \kappa_{Mi} t - g(U_i) \Big) \\ &= \alpha_i \beta_i (x - x^*) \end{aligned} \tag{2.22}$$

Similarly, the "*it*-th"-specific *pure natural direct effect* can be identified (based on equation (2.21)):

$$E[Y_{it}(x, M_{it}(x^*)) - Y_{it}(x^*, M_{it}(x^*))|\alpha_i, \kappa_{Mi}, \beta_i, \zeta_i', \kappa_{Yi}, U_i]$$

$$
\begin{aligned}
&= \sum_m \Big(E[Y_{it}|X_{it}=x, M_{it}=m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it}=m|X_{it}=x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it}|X_{it}=x^*, M_{it}=m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it}=m|X_{it}=x^*, \alpha_i, \kappa_{Mi}, U_i) \Big) \\
&= \sum_m P(M_{it}=m|X_{it}=x^*, \alpha_i, \kappa_{Mi}, U_i) \cdot \\
&\quad \Big(\delta_Y + \zeta_i' x + \beta_i m + \kappa_{Yi} t + U_i - \delta_Y - \zeta_i' x^* - \beta_i m - \kappa_{Yi} t - U_i \Big) \\
&= \sum_m P(M_{it}=m|X_{it}=x^*, i)(x-x^*)(\zeta_i') \\
&= \zeta_i'(x-x^*)
\end{aligned} \tag{2.23}
$$

Finally, the "*it*-th"-specific *total causal effect* can be identified as (based on equation (2.21)):

$$
\begin{aligned}
&E[Y_{it}(x, M_{it}(x)) - Y_{it}(x^*, M_{it}(x^*))|\alpha_i, \kappa_{Mi}, \beta_i, \zeta_i', \kappa_{Yi}, U_i] \\
&= \sum_m \Big(E[Y_{it}|X_{it}=x, M_{it}=m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it}=m|X_{it}=x, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it}|X_{it}=x^*, M_{it}=m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it}=m|X_{it}=x^*, \alpha_i, \kappa_{Mi}, U_i) \Big) \\
&= \sum_m (\delta_Y + \beta_i m + \zeta_i' x + \kappa_{Yi} t + U_i) P(M_{it}=m|X_{it}=x, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - \sum_m (\delta_Y + \beta_i m + \zeta_i' x^* + \kappa_{Yi} t + U_i) P(M_{it}=m|X_{it}=x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&= \zeta_i'(x-x^*) + \beta_i \sum_m m P(M_{it}=m|X_{it}=x, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - \beta_i \sum_m m P(M_{it}=m|X_{it}=x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&= \zeta_i'(x-x^*) + \beta_i(d_M + \alpha_i x + \kappa_{Mi} t + g(U_i)) - \beta_i(d_M + \alpha_i x^* + \kappa_{Mi} t + g(U_i)) \\
&= (\alpha_i \beta_i + \zeta_i')(x-x^*)
\end{aligned} \tag{2.24}
$$

Consequently, the subject- and period-specific *total causal effect* equals:

$$
\begin{aligned}
&E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0))|\alpha_i, \kappa_{Mi}, \beta_i, \zeta_i', \kappa_{Yi}, U_i] \\
&= \sum_m \{ E[Y_{it}|X_{it}=1, M_{it}=m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it}=m|X_{it}=1, \alpha_i, \kappa_{Mi}, U_i) \\
&- E[Y_{it}|X_{it}=0, M_{it}=m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it}=m|X_{it}=0, \alpha_i, \kappa_{Mi}, U_i) \} \\
&= \alpha_i \beta_i + \zeta_i'
\end{aligned} \tag{2.25}
$$

The individual- and period-specific *total natural indirect effect* and the *pure natural direct effect* in turn equals:

$$
\begin{aligned}
&E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0))|\alpha_i, \kappa_{Mi}, \beta_i, \zeta_i', \kappa_{Yi}, U_i] \\
&= \sum_m \Big(E[Y_{it}|X_{it} = 1, M_{it} = m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it} = m|X_{it} = 1, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it}|X_{it} = 1, M_{it} = m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it} = m|X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \Big) \\
&= \alpha_i \beta_i
\end{aligned} \tag{2.26}
$$

and

$$
\begin{aligned}
&E[Y_{it}(1, M_{it}(0)) - Y_{it}(0, M_{it}(0))|\alpha_i, \kappa_{Mi}, \beta_i, \zeta_i', \kappa_{Yi}, U_i] \\
&= \sum_m \Big(E[Y_{it}|X_{it} = 1, M_{it} = m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it} = m|X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it}|X_{it} = 0, M_{it} = m, \beta_i, \zeta_i', \kappa_{Yi}, U_i] P(M_{it} = m|X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \Big) \\
&= \zeta_i'
\end{aligned} \tag{2.27}
$$

These effects are in line with results from traditional lower-level mediation analysis in linear settings (Kenny et al., 2003). When we marginalise these effects over individuals, we obtain a total natural indirect effect of $E[\alpha_i \beta_i] = E[\alpha_i]E[\beta_i] + Cov(\alpha_i, \beta_i) = E[\alpha_i]E[\beta_i] + \sigma_{\alpha_i, \beta_i}$, a pure natural direct effect of $E[\zeta_i']$ and a total causal effect of $E[\alpha_i]E[\beta_i] + \sigma_{\alpha_i, \beta_i} + E[\zeta_i']$. However, in an AB/BA design with only two repeated measurements, not all subject-specific effects in model (2.3) can be identified. This is why we will assume homogeneous effects across subjects, i.e. $\alpha_i = \alpha$, $\beta_i = \beta$, $\zeta_i' = \zeta'$, $\kappa_{Mi} = \kappa_M$ and $\kappa_{Yi} = \kappa_Y$, resulting in data-generating mechanism (2.3). When subject-specific slopes are absent, the indirect and direct effect then simplify to $\alpha\beta$ and ζ'. If there were more repeated measurements per individual (e.g. four, with two observations within each measurement period), these heterogeneous effects across individuals could become identifiable.

A.2 Limitations of the Joint modelling approach

The Joint modelling approach may provide biased estimates under some circumstances, even if the fixed effects part of the model is correctly specified. To understand this, note that the joint modelling approach

implies that:

$$E(Y_{it}|M_i^{x=0}, M_i^{x=1}, X_{it}) = d_{Yi} + c'X_{it} + bM_{it} + K_Y t + E(u_{Yi}|M_i^{x=0}, M_i^{x=1}, X_{it})$$

where:

$$\begin{aligned} E(u_{Yi}|M_i^{x=0}, M_i^{x=1}, X_{it}) &= E(u_{Yi}|M_i^{Dif}, M_i^{Sum}, X_{it}) \\ &= E(u_{Yi}|M_i^{Sum}, X_{it}) \end{aligned}$$

because M_i^{Dif} is independent of u_{Mi} and therefore also of u_{Yi}. When u_{Yi} and M_i^{Sum} have bivariate normal distributions (given X_{it}), this implies that $E(u_{Yi}|M_i^{Sum}, X_{it})$ is linear in M_i^{Sum}. It thus follows that the joint modelling approach is equivalent with fitting GEE to a marginal model that, besides linear terms in X_{it}, M_{it} and t, also involves a linear term in M_i^{Sum}. The assumption that u_{Yi} and M_i^{Sum} have bivariate normal distributions (given X_{it}) implies in particular that (a) u_{Yi} is normal, given X_{it}; (b) M_{it} is normal given X_{it} and u_{Yi}; and (c) u_{Yi} has a linear, additive effect on M_{it} (no interactions). When these conditions are not satisfied, then $E(u_{Yi}|M_i^{Sum}, X_{it})$ may be nonlinear in M_i^{Sum}, in which case the joint modelling approach amounts to fitting a misspecified marginal model.

Violations of condition (c) arise in the third data generating mechanism in our simulations: here, V_i in equation (2.7) depends linearly on U_i, but non-linearly on M_{it} (through the interaction between V_i and X_{it}), thereby inducing a nonlinear dependence between u_{Yi} and M_i^{Sum}.

A.3 Identification of the causal effects in complex settings

For the more complex data generating mechanism, summarised by equation (2.7), the "*it*-th"-specific *total natural indirect effect* can also be identified, when the assumptions (i)-(vii) from section 3.2 are satisfied:

$$\begin{aligned} &E[Y_{it}(x, M_{it}(x)) - Y_{it}(x, M_{it}(x^*))|D_i = d, U_i] \\ &= \sum_m \Big(E[Y_{it}|X_{it} = x, M_{it} = m, D_i = d, U_i] P(M_{it} = m|X_{it} = x, D_i = d, U_i) \\ &\quad - E[Y_{it}|X_{it} = x, M_{it} = m, D_i = d, U_i] P(M_{it} = m|X_{it} = x^*, D_i = d, U_i) \Big) \\ &= \sum_m (\delta_Y + \zeta' x + \beta m + \phi x m + \kappa_Y t + \omega_Y d + \nu_Y dx + \eta dm + U_i) \cdot \end{aligned}$$

$$
\begin{aligned}
&\Big(P(M_{it}=m|X_{it}=x,D_i=d,U_i)-P(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\Big)\\
=&(\beta+\phi x+\eta d)\Big(\sum_m mP(M_{it}=m|X_{it}=x,D_i=d,U_i)\\
&-\sum_m mP(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\Big)\\
=&(\beta+\phi x+\eta d)\Big(E(M_{it}|X_{it}=x,D_i=d,U_i)-E(M^*_{it}|X_{it}=x^*,D_i=d,U_i)\Big)\\
=&(\beta+\phi x+\eta d)(\delta_M+\alpha x+\kappa_M t+\omega_M d+\nu_M dx+g(U_i)\\
&-\delta_M-\alpha x^*-\kappa_M t-\omega_M d-\nu_M dx^*-g(U_i))\\
=&(\alpha+\nu_M d)(\beta+\phi x+\eta d)(x-x^*) \qquad (2.28)
\end{aligned}
$$

Similarly, the "*it*-th"-specific *pure natural direct effect* can be identified (based on equation (2.7)):

$$
\begin{aligned}
&E[Y_{it}(x,M_{it}(x^*))-Y_{it}(x^*,M_{it}(x^*))|D_i=d,U_i]\\
&=\sum_m\Big(E[Y_{it}|X_{it}=x,M_{it}=m,D_i=d,U_i]P(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\\
&\quad-E[Y_{it}|X_{it}=x^*,M_{it}=m,D_i=d,U_i]P(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\Big)\\
&=\sum_m P(M_{it}=m|X_{it}=x^*,D_i=d,U_i)(\delta_Y+\zeta' x+\beta m+\phi xm+\kappa_Y t+\omega_Y d\\
&\quad+\nu_Y dx+\eta dm+U_i-\delta_Y-\zeta' x^*-\beta m-\phi x^* m-\kappa_Y t-\omega_Y d-\nu_Y dx^*\\
&\quad-\eta dm-U_i)\\
&=\Big((\zeta'+\nu_Y d)\sum_m P(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\\
&\quad+\phi\sum_m mP(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\Big)(x-x^*)\\
&=(\zeta'+\nu_Y d+\phi E[M_{it}=m|X_{it}=x^*,D_i=d,U_i])(x-x^*) \qquad (2.29)\\
&=(\zeta'+\nu_Y d+\phi(\delta_M+\alpha x^*+\kappa_M t+\lambda_M c+\omega_M d+\nu_M dx^*+g(U_i))(x-x^*)
\end{aligned}
$$

Finally, the "*it*-th"-specific *total causal effect* can be identified as (based on equation (2.7)):

$$
\begin{aligned}
&E[Y_{it}(x,M_{it}(x))-Y_{it}(x^*,M_{it}(x^*))|D_i=d,U_i]\\
&=\sum_m\Big(E[Y_{it}|X_{it}=x,M_{it}=m,D_i=d,U_i]P(M_{it}=m|X_{it}=x,D_i=d,U_i)\\
&\quad-E[Y_{it}|X_{it}=x^*,M_{it}=m,D_i=d,U_i]P(M_{it}=m|X_{it}=x^*,D_i=d,U_i)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_m (\delta_Y + \zeta' x + \beta m + \phi x m + \kappa_Y + \omega_Y d + \nu_Y d x + \eta d m + U_i) \cdot \\
&\quad P(M_{it} = m | X_{it} = x, D_i = d, U_i) \\
&\quad - \sum_m (\delta_Y + \zeta' x^* + \beta m + \phi x^* m + \kappa_Y t + \omega_Y d + \nu_Y d x^* + \eta d m + U_i) \cdot \\
&\quad P(M_{it} = m | X_{it} = x^*, D_i = d, U_i) \\
&= (\zeta' + \nu_Y d)(x - x^*) + (\beta + \phi x + \eta d) \sum_m m P(M_{it} = m | X_{it} = x, D_i = d, U_i) \\
&\quad - (\beta + \phi x^* + \eta d) \sum_m m P(M_{it} = m | X_{it} = x^*, D_i = d, U_i) \\
&= (\zeta' + \nu_Y d)(x - x^*) + (\beta + \phi x + \eta d)(\delta_M + \alpha x + \kappa_M t + \omega_M d + \nu_M d x + g(U_i)) \\
&\quad - (\beta + \phi x^* + \eta d)(\delta_M + \alpha x^* + \kappa_M t + \omega_M d + \nu_M d x^* + g(U_i)) \\
&= \Big((\alpha + \nu_M d)(\beta + \eta d) + \zeta' + \nu_Y d + \phi (d_M + \kappa_M t + \omega_M d + g(U_i)) \Big) (x - x^*) \\
&\quad + \phi (\alpha + \nu_M d)(x^2 - x^{*2}) \qquad (2.30)
\end{aligned}
$$

3 More precise estimation of lower-level interaction effects in multilevel models

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

This chapter is based on Loeys, T., Josephy, H. & Marieke Dewitte (2018). More precise estimation of lower-level interaction effects in multilevel models. *Multivariate Behavioral Research*, 53(3): 335-347.

1 Introduction

When measures are collected repeatedly over time in individuals (e.g., in daily diary studies), such data can yield much more information compared to a cross-sectional sample. For example, when studying the relationship between intimacy and positive relationship feelings in a daily diary study, a between-person effect can be disentangled from a within-person effect (Curran and Bauer, 2011; Wang and Maxwell, 2015). In this example, the between-person effect reflects the extent to which individuals with higher intimacy differ in their positive relational feelings from individuals with a lower intimacy. The within-person effect on the other hand, reflects the extent to which an individual exhibits higher (or lower) positive relational feelings when (s)he had more (or less) intimacy on a particular day, as compared to other days.

During the last two decades, the behavioural science literature has increasingly focused on separating within- from between-effects in multilevel models (Curran and Bauer, 2011; Enders and Tofighi, 2007; Hofmann and Gavin, 1998; Kreft et al., 1995; Raudenbush and Bryk, 2002). Two important issues can be highlighted when disaggregating those effects within longitudinal data: centering and detrending (Curran and Bauer, 2011). The former refers to subtracting a constant from every value of a variable, while the latter refers to removing the time trend from time series. The centering issue is relevant for disaggregation, even when neither the predictor nor the outcome exhibits any trend over time, whereas the detrending issue is only relevant when at least one of those variables exhibits some trend over time (Wang and Maxwell, 2015). In this paper, we assume no time effects on either the predictor or the outcome and consequently limit our focus to the centering issue.

The multilevel literature typically considers two levels: the lower-level or level 1 (e.g. the daily measurements within the individual), and the upper-level or level 2 (e.g. the individuals in a diary study). Within such two-level data structures, three types of centering can be distinguished: no centering (i.e., the raw scores are used), grand-mean centering (i.e., subtraction of the overall average across individuals and time points) and cluster-mean centering (i.e., subtraction of a person-specific mean, averaged across time points within the individual). There is a general consensus that cluster-mean centering (also referred to as 'CWC', centering within clusters) is deemed most appropriate when lower-level predictors are of primary substantive interest (Enders and Tofighi, 2007). More specifically,

CWC may solve potential confounding issues in estimating the effect of a predictor on an outcome. A detailed explanation on why is discussed in the next section. When unmeasured upper-level common causes of the predictor-outcome relationship are present, we refer to such causes as unmeasured upper-level confounders. In the econometrics literature, this type of unmeasured confounding at the subject- or cluster-level is referred to as upper-level endogeneity (Wooldridge, 2010). Here we argue that confounding at the upper-level is very common in many contexts. In our illustration, for example, it is not unlikely that the daily measurements of intimacy and positive relational feelings are both affected by unmeasured stable (personality) traits of the individual. We will show that under a specific set of assumptions for the unmeasured upper-level confounder, CWC allows unbiased estimation of the within-subject effect of a lower-level predictor on an outcome.

Unfortunately, discussions on the role of centering are mostly limited to the assessment of main effects in multilevel models (MLM) and ignore the centering of interactions. An issue of particular importance entails the centering of interactions in the $1 \times (1 \rightarrow 1)$ design, where the first '1' corresponds to the level at which the moderator is measured, the second '1' represents the level of the predictor, and the last '1' defines the level of the outcome (Preacher et al., 2016; Ryu, 2015). We will refer to such interactions as 'lower-level interactions'. When cluster-mean centering such interactions, the question arises whether the predictor and moderator should be centred first and multiplied next (hereafter labeled as 'C1P2', centre-first and product-second), or whether it should be the other way around (labeled hereafter as 'P1C2'). In contrast to cluster-mean centering an interaction between an upper- and a lower-level variable, or between two upper-level variables, C1P2 and P1C2 produce different predictors when cluster-mean centering a lower-level interaction. Some scholars favoured P1C2 (Josephy et al., 2015), while others advised against it and promoted C1P2 instead (Preacher et al., 2016). In this paper we investigate how these two approaches deal with unmeasured upper-level confounding and whether they can unbiasedly estimate the moderated within-subject effect.

While Josephy et al. (2015) considered the traditional multilevel modelling (MLM) framework (also referred to as mixed modelling), Preacher et al. (2016) relied on Structural Equation Modelling (SEM). In contrast to the traditional MLM-framework, in which the within- and between-cluster decomposition of a predictor relies on the observed cluster means, latent cluster means are generally used in SEM. The latent means in a multilevel

SEM-framework avoid bias due to sampling error, which is typically associated with the observed cluster means in the MLM-framework (Lüdtke et al., 2008). And although the impossibility of the MLM-framework to deal with measurement error is a serious limitation, this does not pose an issue when the interest lies with the within-cluster effects. When the lower-level variables are assumed to be measured without error, (Lüdtke et al., 2008) have shown that the estimator of the within-effect is unbiased (we will not repeat their proof here). Additionally, Lüdtke et al. (2008) reported a similar performance in terms of standard errors for the estimated within-effects, when using the observed mean versus the latent mean. Unfortunately, the MLM-approach can result in substantially biased estimates of between-effects, as well as severely underestimate the associated standard errors in the presence of upper-level measurement error. However, since Nesselroade and Molenaar (2016) have recently re-emphasised the importance of studying within-subject processes in lower-level designs (Molenaar, 2004, 2009), we will primarily focus on the estimation of these effects. Given that MLM and SEM perform similarly for within-cluster effects, we limit our exposition to the MLM-framework.

In the following sections, we first introduce our illustrating example and describe cluster-mean centering within the MLM-framework for main effect models. Next, we consider MLMs with lower-level interaction effects and enumerate the various existing modelling strategies proposed for estimating such effects. We demonstrate how different estimates (and standard errors) are found for the moderating effect, when applying these strategies to the diary data on intimacy and relationship feelings. In a next step, we explore why and when those centering approaches perform differently by means of a simulation study. Finally, we discuss the interpretation of the parameters for the different modelling strategies and end with a short discussion.

2 Illustrating example

We consider longitudinal diary data on sexual behaviour from a Flemish study in 66 heterosexual couples (Dewitte et al., 2015). Every morning during three weeks, participants were asked about their sexual and intimate behaviour since the last time they had filled out their morning diary (i.e., sexual behaviour over the past 24 hours). Every evening, the participants were asked to report on their individual, relational, and partner-related feelings and behaviour, experienced during that day. In this manuscript, we limit our focus to the reports of the 66 male partners. Because the

diary reports were not always completed meticulously over the course of the 21 days, the number of observations per participant ranges from 5 to 21, with a median cluster size of 18. In total, we have 1127 observations clustered within 66 men, implying a missing rate of about 19%. The variables of interest are the extent (on a 7-point scale from 'not at all' to 'very much') to which they report that intimate acts had occurred with their partner (described as the amount of kissing, cuddling and caressing), the men's daily reports of masturbation (defined as any sexual act that involved self-stimulation in the absence of their partner), as well as their daily evening reports on positive relationship feelings. The latter were obtained by averaging the scores (on a seven-point scale) on nine items (the extent to which they felt happy, satisfied, understood, supported, accepted, loved, in love, connected, and close). The research question we will focus on, considers the contribution of intimacy to next-day positive relationship feelings within a man, and to what extent that the occurrence of masturbation during the previous day (yes or no) changes this effect.

3 Centering of main effects in multilevel models

In this section we first explain why a difference in within- and between-subject effects may result from omitted variable bias at the subject-level. Let X_{ij} denote the predictor and Y_{ij} the outcome of individual j ($j = 1, \ldots, N$) at time i ($i = 1, \ldots, n_j$). In our example, X_{ij} and Y_{ij} represent the daily measurements of intimacy and next day's positive relational feelings, respectively. As mentioned in the introduction, it is not unlikely that the daily measurements of intimacy and positive relational feelings are both affected by unmeasured stable (personality) traits of the individual. We referred to such unmeasured common causes of X_{ij} and Y_{ij} as an unmeasured upper-level confounder, which we will from now denote by b_j.

Consider a simple causal model for the effect of X_{ij} on Y_{ij} that takes an unmeasured subject-level confounder b_j into account:

$$E(Y_{ij} \mid X_{ij}, b_j) = \beta_0 + \beta X_{ij} + b_j, \tag{3.1}$$

where we assume that the unmeasured confounder has an additive effect on the outcome. The left panel of Figure 3.1 represents the corresponding data-generating process. For a given subject, the β-parameter reflects the average increase in the outcome for a one-unit increase in the predictor. As such, this parameter can be interpreted as the within-person effect

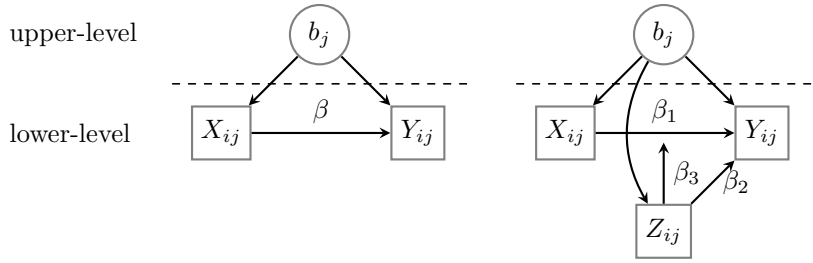

Figure 3.1 Left panel: Unmeasured subject-level confounding of the $X_{ij} - Y_{ij}$ relationship. Right panel: lower-level interaction model with unmeasured subject-level confounding.

of X_{ij} on Y_{ij}. Several remarks deserve some additional attention. First, in order for β to have a causal interpretation, the predictor X_{ij} should temporally precede Y_{ij}. For example, in our illustration we aim to estimate the causal effect of intimacy on next day's positive relationship feelings, implying a clear temporal ordering. Second, we assume a time-constant effect of X_{ij} on Y_{ij}; there is no reason to assume that the effect on day one is any different from the effect on day two. Third, we assume the absence of any unmeasured lower-level confounders of the $X_{ij} - Y_{ij}$ relationship. That is, given the unmeasured personality traits for example, we do not allow for further occasion-specific unmeasured common causes of X and Y. The question that we want to address now is: how can β be unbiasedly estimated, despite the presence of the unmeasured upper-level confounder b_j?

Naively, we could consider the following multilevel model:

$$E(Y_{ij} \mid X_{ij}, b_j) = \gamma_0 + \gamma X_{ij} + u_j, \tag{3.2}$$

Note that to clearly contrast estimation model (3.2) to data-generating process (3.1), we rely on different notations here, as well as in the remainder of the paper. Fixed effect parameters will be denoted by γ's and random effects by u_j in estimation models, while β's and b_j will represent these effects in the true causal models. To fit model (3.2), we could simply rely on standard multilevel modelling approaches. Unfortunately, an important but often ignored assumption in hierarchical linear modelling requires the random effect u_j in (3.2) to be uncorrelated with the predictor X_{ij} (McNeish et al., 2016). This assumption is violated in case of upper-level endogeneity. Consequently, the naive MLM-estimator (based on maximum

likelihood, restricted maximum likelihood, or feasible generalised least squares, abbreviated FGLS) that aims to estimate β in model (3.1), and which we will refer to as $\hat{\gamma}^{RE}$, will suffer from omitted variable bias (Raudenbush and Bryk, 2002; Castellano et al., 2014).

In a similar vein, it is important to stress that lagged variables should not be added to model (3.2), i.e.:

$$E(Y_{ij} \mid X_{ij}, Y_{i-1,j}, u_j) = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 Y_{i-1,j} + u_j, \tag{3.3}$$

Since model (3.3) applies to all time points, u_j has a direct effect on $Y_{i-1,j}$. However, if u_j affects $y_{i-1,j}$, it can't be statistically independent of $Y_{i-1,j}$ at the same time. The violation of this independence assumption in traditional hierarchical linear modelling can bias both the coefficient for the lagged dependent variable, as well as the coefficients for the other variables (Allison, 2015).

One possible way to deal with omitted variable bias in model (3.2) is to rely on the fixed effects approach (Mundlak, 1978), where u_j is treated as fixed rather than random. This approach is very popular within the econometrics literature (Wooldridge, 2010) and has recently resurfaced in behavioural science literature (Castellano et al., 2014). In practice, N dummy variables dn_j (i.e. one for each subject) are created in a way that $dn_j = 1$ if $n = j$, and $dn_j = 0$ when $n \neq j$ $(n = 1, \ldots, N)$. Consequently, Y_{ij} is regressed on $d1_j, \ldots, dN_j$ and x_{ij}:

$$E(Y_{ij} \mid X_{ij}, d1_j, \ldots, dN_j) = \gamma_1^* d1_j + \gamma_2^* d2_j + \ldots + \gamma_N^* dN_j + \gamma_{FE} X_{ij} \tag{3.4}$$

Under causal data-generating model (3.1), the OLS-estimator for γ_{FE}, denoted by $\hat{\gamma}_{FE}$ (obtained from estimation model (3.4)), represents an unbiased estimator for β (Wooldridge, 2010). Intuitively, this can be understood by the fact that the predictors in (3.4) are allowed to be correlated (in contrast to the predictor and random intercept in model (3.2)). One side effect of the fixed effects approach is that it cannot be used to investigate between-subject effects, as between-subject characteristics are perfectly collinear with the dummies.

A possible alternative that can deal with omitted variable bias and additionally allows the estimation of both within- and between-effects, is to rely on group-mean centering (i.e. the CWC-approach). That is, the predictor X_{ij} is separated into a between- (i.e. $\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$) and a within-subject (*i.e.*, $X_{ij} - \overline{X}_j = X_{ij}^c$) component within the MLM-framework. As such, we consider the following model, originally proposed

by Neuhaus and Kalbfleisch (1998):

$$E(Y_{ij} \mid X_{ij}, u_j) = \gamma_0 + \gamma_W X_{ij}^c + \gamma_B \overline{X}_j + u_j \tag{3.5}$$

with u_j assumed i.i.d. $\sim N(0, \tau^2)$ and independent of the predictors. Goetgeluk and Vansteelandt (2008) prove that the estimator $\hat{\gamma}_W$ from model (3.5) is consistent (i.e. asymptotically unbiased) for β in model (3.1), even in the presence of unmeasured upper-level confounding of X_{ij} and Y_{ij}. The rationale behind this is that by subject-mean centering the predictor, any subject-specific effects are effectively eliminated. Relying on simple OLS-estimators for γ_W and γ_B, we see that $\hat{\gamma}_W = \frac{\text{cov}(Y_{ij}, X_{ij} - \overline{X}_j)}{\text{var}(X_{ij} - \overline{X}_j)}$ and $\hat{\gamma}_B = \frac{\text{cov}(\overline{Y}_j, \overline{X}_j)}{\text{var}(\overline{X}_j)}$, which will converge to β and $\beta + \frac{\text{cov}(b_j, \overline{X}_j)}{\text{var}(\overline{X}_j)}$, respectively, under model (3.1). These two expressions clearly illustrate two important points. First, cluster-mean centering the predictor permits unbiased estimation of the within-person effect under upper-level endogeneity in causal model (3.1). Second, when b_j is a confounder of the $X_{ij} - Y_{ij}$ relationship, this implies that $\text{cov}(b_j, \overline{X}_j) \neq 0$, and that $\hat{\gamma}_B$ will no longer converge to β. In other words, upper-level endogeneity elicits differences in the between- and within-subject effects. Only in the absence of upper-level endogeneity in model (3.1) (i.e. $\text{cov}(b_j, \overline{X}_j) = 0$), will $\hat{\gamma}_B$ be equal to $\hat{\gamma}_W$.

Note that the naive MLM-estimator $\hat{\gamma}^{RE}$ actually represents a weighted combination of $\hat{\gamma}_W$ and $\hat{\gamma}_B$ (Raudenbush & Bryk, 2002, p.137); in balanced designs (i.e., with $n_j = n$ for all j), we see that:

$$\hat{\gamma}^{RE} = \frac{W_1 \hat{\gamma}_B + W_2 \hat{\gamma}_W}{W_1 + W_2} \quad \text{, with } W_1 = \widehat{\text{var}}(\hat{\gamma}_B)^{-1} \text{ and } W_2 = \widehat{\text{var}}(\hat{\gamma}_W)^{-1},$$

making $\hat{\gamma}^{RE}$ an uninterpretable blend of both effects. Also note that, since $\text{cov}(X_{ij} - \overline{X}_j, \overline{X}_j) = 0$, the within- and between-subject predictors are independent. As such, the cluster means can be dropped from estimation model (3.5) when the within-effect is the only quantity of interest:

$$E(Y_{ij} \mid X_{ij}, u_j) = \gamma_0 + \gamma_W X_{ij}^c + u_j \tag{3.6}$$

Furthermore, it is interesting to note that the fixed effect estimator $\hat{\gamma}^{FE}$ and the within-subject estimator $\hat{\gamma}_W$ are identical in balanced designs (Wooldridge, 2010).

Similar to Greenland (2002), Goetgeluk and Vansteelandt (2008), and Brumback et al. (2010), we argue that models (3.5) and (3.6) cannot be considered valid causal models. For example, in the longitudinal setting

considered here, model (3.5) would imply that the future causes the past (i.e. future X_{ij} would cause past Y_{i_0j} for $i > i_0$, since X_{ij} is contained within $\overline{X}_j$). Also, when model (3.5) is interpreted as a causal model for the manipulated effect of X_{ij} for a single i, it would conflict with causal model (3.1) unless $\gamma_W = \gamma_B = \beta$. As mentioned before, the individual causal effect of a one-unit increase in X_{ij} is represented by β in model (3.1), while this is $\gamma_W(1 - 1/n_j) + 1/n_j\gamma_B$ in model (3.5); the latter expression only equals β when $\gamma_W = \gamma_B = \beta$. This remark does not degrade the usefulness of model (3.5), but it emphasises that model (3.5) should be viewed as an estimation model rather than a causal one.

What are the principal implications for substantive researchers? Most importantly, that model (3.5) can be used as the vehicle to estimate the parameter of interest. In our example, we want to determine the effect of a one-unit increase in intimacy on next day's positive relationship feelings within a person. Unlike γ in model (3.2), the parameter γ_W in model (3.5) will target that quantity of interest, even in the presence of unmeasured time-constant subject-specific confounders. As such, we look at settings in which model (3.1) (graphically represented in Figure 3.1) rather than model (3.5) represents the true causal model. However, in these settings, model (3.5) still correctly describes the conditional association of Y_{ij} given X_{ij} and the independent subject effect u_i. In other words, while both models might be valid at the same time, model (3.1) constitutes the causal model, whereas model (3.5) represents an estimation model invoked to circumvent the issue that b_i is associated with X_{ij} (so that we can unbiasedly estimate β).

4 Centering of lower-level interactions in multilevel models

Researchers' interest is often not limited to assessing main effects only. In our illustrating example, researchers may want to know if the effect of intimacy on the following day's positive relational feelings differs according to whether or not the participant has masturbated during the previous day. Instead of model (3.1), we now assume a causal model in which an interaction effect is included:

$$E(Y_{ij} \mid X_{ij}, Z_{ij}, b_j) = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_3 X_{ij} Z_{ij} + b_j, \qquad (3.7)$$

with Z_{ij} the moderator at time i in individual j. Since both X_{ij} and Z_{ij} are measured at the lower-level, we have a setting with a lower-level interaction. The right panel of Figure 3.1 graphically represents the assumed data-generating process. Note that an arrow-on-arrow notation was used to indicate the moderating effect of Z.

The parameter β_3 in model (3.7) reflects the moderating effect for a given subject, i.e. the extent to which the effect of X_{ij} on Y_{ij} varies for different values of Z_{ij}. In our example, such an effect might translate into: how does the effect of intimacy on next day's positive relationship feelings change within a participant when the man has masturbated versus when he has not? The interpretation of the main effects β_1 and β_2 in (3.7) on the other hand, depends on whether X_{ij} and/or Z_{ij} are grand mean centred. When X_{ij} is grand mean centred, β_2 reflects the effect of masturbation on next day's positive relationship feelings within a subject at the sample average level of intimacy. If X_{ij} were not grand mean centred, β_2 would capture the effect of masturbation at the zero-level of intimacy. This, however, would not provide a very useful interpretation, since intimacy is measured on a 1-7 scale. Similarly, when Z_{ij} is grand mean centred (i.e., in our example, the sample proportion of days with masturbation is subtracted from the raw scores), β_1 reflects the effect of a one-unit increase in intimacy on the next day's positive relationship feelings within a participant, averaged over days with and without masturbation. When both X_{ij} and Z_{ij} are grand mean centred, the intercept β_0 can be interpreted as the average positive relationship feelings over all participants and days. As such, grand-mean centering of both continuous and binary predictors in interaction models provides useful interpretations of the main effects; we will therefore assume that X and Z are grand mean centred during the remainder of this manuscript. However, in order to avoid notational burden, we will not introduce any new notation to indicate this.

The researcher's primary interest now lies in estimating β_3. But how should β_3 be estimated? Naively, we may again consider a traditional MLM-approach:

$$E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 Z_{ij} + \gamma_3 X_{ij} Z_{ij} + u_j, \qquad (3.8)$$

Note that here too, we define the parameters γ and u within the estimation model, whereas β and b are used in the causal model. Given the standard assumption of independence of the random effect and predictors in model (3.8), the naive MLM-estimator of the interaction effect (which we

denote $\hat{\gamma}_3^{RE}$) will once again suffer from omitted variable bias. To address such unmeasured upper-level confounding, we may - similar to the main effects model - rely on separating within- from between-effects.

As was already mentioned in the introduction, two different strategies for centering lower-level interactions have been suggested. The first approach, advocated by Josephy et al. (2015), first multiplies X_{ij} with Z_{ij}, after which the cluster mean average of this product term is subtracted. As such, the "P1C2" estimation model amounts to:

$$E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) = \gamma_0 + \gamma_1 X_{ij}^c + \gamma_2 Z_{ij}^c + \gamma_3 (XZ)_{ij}^c + u_j \qquad (3.9)$$

with $X_{ij}^c = X_{ij} - \overline{X}_j$, $Z_{ij}^c = Z_{ij} - \overline{Z}_j$ and $(XZ)_{ij}^c = X_{ij}Z_{ij} - \overline{XZ}_j$ (where $\overline{XZ}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}Z_{ij}$). Under data-generating model (3.7), $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$ of the P1C2-approach consistently (i.e., asymptotically unbiased) estimate β_0, β_1, β_2 and β_3 (Goetgeluk and Vansteelandt, 2008; Josephy et al., 2015). As such, the estimators $\hat{\gamma}_0$, $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$ share the same interpretation as β_0, β_1, β_2 and β_3 (see below model (3.7)).

It is also possible to add all corresponding between-effects to estimation model (3.9), i.e.:

$$\begin{aligned} E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) = \quad & \gamma_0 + \gamma_1 X_{ij}^c + \gamma_2 Z_{ij}^c + \gamma_3 (XZ)_{ij}^c \\ & + \gamma_4 \overline{X}_j + \gamma_5 \overline{Z}_j + \gamma_6 \overline{XZ}_j + u_j \end{aligned} \qquad (3.10)$$

We will refer to estimation model (3.10) as the 'P1C2+' approach. Interestingly, since the within-predictors are independent of the between-predictors in this model, the estimated within-effects $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$ from models (3.9) and (3.10) are identical in balanced designs.

The second approach is suggested by Preacher et al. (2016), who are very explicit on their centering convictions in multilevel SEM (MSEM) models. If we ignore the distinction between centering at the observed versus the latent cluster means (Lüdtke et al., 2008), Preacher et al. (2016) distinctly argue that $X_{ij}Z_{ij}$ should not be separated into a within-part $X_{ij}Z_{ij} - \overline{XZ}_j$ and a between-part $\overline{XZ}_j$. These authors reason that "using these as predictors does not lead to interpretable effects, because researchers are not interested in the effects of product terms" (p.191). When solely focusing on within-effects, the multilevel model proposed by Preacher et al. (2016) (with observed rather than latent cluster means), can be written as:

$$E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) = \gamma_0 + \gamma_1 X_{ij}^c + \gamma_2 Z_{ij}^c + \gamma_3 X_{ij}^c Z_{ij}^c + u_j \qquad (3.11)$$

We refer to estimation model (3.11) as the 'C1P2'-approach. In their paper, Preacher et al. (2016) also describe a more complete model that additionally includes cross- and between-level effects:

$$\begin{aligned} E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) \quad &= \gamma_0 + \gamma_1 X^c_{ij} + \gamma_2 Z^c_{ij} + \gamma_3 X^c_{ij} Z^c_{ij} + \gamma_4 \overline{X}_j + \gamma_5 \overline{Z}_j \\ &\quad + \gamma_6 \overline{X}_j \overline{Z}_j + \gamma_7 \overline{X}_j Z^c_{ij} + \gamma_8 \overline{Z}_j X^c_{ij} + u_j \end{aligned} \tag{3.12}$$

which we will refer to as the 'C1P2++' approach. Model (3.12) contains four different interaction effects: a within- subject interaction (captured by the parameter γ_3), a between-subject interaction (captured by γ_6) and two cross-level interactions (captured by γ_7 and γ_8). Note that since $\text{Cov}\left(X^c_{ij} Z^c_{ij}, \overline{X}_j \overline{Z}_j\right)$ is not necessarily zero, the estimated parameters of the within-effects in the C1P2 and C1P2++ approaches are no longer identical in balanced designs (unlike in P1C2 and P1C2+).

Ryu (2015) also considers MSEM for estimating lower-level interactions in multilevel data, but in contrast to Preacher et al. (2016), Ryu relies on an earlier MSEM approach (Muthén, 1990). The latter decomposes the observed data into between- and pooled within-covariances, whilst fitting separate within- and between-models using the multi-group techniques of SEM. This multi-group approach does not allow for missing data or unbalanced cluster sizes, but more importantly, it cannot account for cross-level interactions. Ryu (2015) considers three types of centering: no centering (UN), grand-mean centering (CGM), and centering within clusters (CWC). First of all, MSEM with uncentred lower-level variables (UN) employs latent cluster means to define the various upper- and lower-level variables. This UN approach therefore corresponds to model (3.10), where the observed means are replaced by their latent counterparts (denoted with a tilde):

$$\begin{aligned} E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) = &\gamma_0 + \gamma_1 (X_{ij} - \tilde{X}_j) + \gamma_2 (Z_{ij} - \tilde{Z}_j) + \gamma_3 (X_{ij} Z_{ij} - \widetilde{XZ}_j) \\ &+ \gamma_4 \tilde{X}_j + \gamma_5 \tilde{Z}_j + \gamma_6 \widetilde{XZ}_j + u_j \end{aligned} \tag{3.13}$$

Here, X_{ij} and Z_{ij} are not grand mean centred. Second, Ryu (2015)'s CGM approach only differs from the UN approach in that X_{ij} and Z_{ij} are first grand mean centred. Third, the CWC-approach described by Ryu (2015) uses the observed cluster means as level 2 covariates. As such, the corresponding estimation model can be written as:

$$\begin{aligned} E(Y_{ij} \mid X_{ij}, Z_{ij}, u_j) = \quad &\gamma_0 + \gamma_1 X^c_{ij} + \gamma_2 Z^c_{ij} + \gamma_3 X^c_{ij} Z^c_{ij} \\ &+ \gamma_4 \overline{X}_j + \gamma_5 \overline{Z}_j + \gamma_6 \overline{X}_j \overline{Z}_j + u_j \end{aligned} \tag{3.14}$$

and will be referred to as 'C1P2+'. We employ this labelling, since the predictors are centred first and only then multiplied as in Preacher et al. (2016), but unlike model (3.12) it does not include any cross-level interactions.

Let us now illustrate how the P1C2- and C1P2-approaches may lead to different estimates of the moderation effect, by means of our example data. To estimate the moderating effect of masturbation on the effect of intimacy on next day's positive relationship feelings, we consider the five different estimation models (3.9), (3.10),(3.11), (3.12) and (3.14). In these models, X_{ij}, Z_{ij} and Y_{ij} denote the grand mean centred intimacy, the grand mean centred masturbation, and next day's positive relationship feelings, respectively. Estimated parameters with associated standard errors, test statistics, and p-values of all within-cluster effects are summarised in table 3.4, together with estimated random intercepts and residual variances.

From these results, we can deduce several things. First, as already stated in the previous section, P1C2 and P1C2+ yield identical results for all within-effects in balanced designs. In our example the data are not perfectly balanced due to a small amount of missingness, and as a consequence the estimates, standard errors and p-values of P1C2 and P1C2+ differ slightly. The estimated within-effects in the different C1P2-approaches, on the other hand, are much more discrepant. Second, the estimated moderating effect of masturbation is more pronounced in the C1P2-approaches compared to the estimates from P1C2. Even though all approaches point in the same direction (the positive effect of intimacy on next day's positive relationship is diluted if the man masturbated), the moderating effect is inflated by about 25% in the C1P2-approaches compared to P1C2. Third, the standard errors of the estimated interaction effect in the C1P2-approaches are about 25% larger than in P1C2. To gain further insights into the performance of the different estimation models, as well as into the precise quantities the different within-effect estimators are targeting, a simulation study is presented in the next section.

5 Simulation study

We consider five different simulation settings under causal model (3.7), where we assess the (relative) bias of the estimators and standard errors of the within-effects for the five different estimation models ((3.9), (3.10), (3.11), (3.12) and (3.14)), as well as their coverage and power. The bias is evaluated by contrasting the sample mean of the estimates from the

1000 simulated data sets to the true parameter value, through the use of a Wald-test. We report the relative bias of the parameter estimates, which is defined as the averaged difference of the estimated (e.g., $\hat{\beta}$) and true parameter value (e.g., β), divided by the latter. Equivalently, the relative bias of the standard errors is defined as the difference between the mean of the estimated standard errors and the empirical standard error, divided by the latter. A negative relative bias thus implies an underestimation of the true variability. The coverage is defined by the proportion of the 95%-confidence intervals that encompass their true parameter value, while the power is determined by the proportion of the 95%-confidence intervals that do not encompass zero.

Mimicking the two-level structure of our illustrating data, we simulated 1000 data sets which contain 66 clusters and 21 observations within each cluster, for five different settings. The true data generating models for Z_{ij} and Y_{ij} are:

$$Z_{ij} = \alpha_0 + \alpha_1 X_{ij} + v_j^Z + \epsilon_{ij}^Z \tag{3.15}$$

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_3 X_{ij} Z_{ij} + v_j^Y + \epsilon_{ij}^Y \tag{3.16}$$

In these models we generate independent lower-level residuals, ϵ_{ij}^Z and ϵ_{ij}^Y, from standard normal distributions. The upper-level confounders, v_j^Z and v_j^Y, follow a multivariate standard normal distribution with a correlation equal to 0.5. As such, we induce unmeasured confounding of the $Z_{ij} - Y_{ij}$ relationship with an additive effect on the outcome. Additionally, we fix $\alpha_0 = 0$, $\beta_0 = 0$, $\beta_1 = 0.1$, $\beta_2 = 0.15$, and $\beta_3 = -0.1$ in all settings, since these values approximately correspond to those seen in our illustrating example (table 3.4). As will become apparent later this section, some of the estimation models will show bias in the interaction effect estimator. Since this bias depends on the distribution of X and Z, we will therefore vary the distribution of X (see table 3.1). Also, as Preacher et al. (2016) showed that the within- and between-components of the product of two lower-level predictors depends on the covariance of the predictors that form the product, we will additionally vary the value of α_1 (see table 3.1). Consequently, in the scenarios where $\alpha_1 \neq 0$, we see that $\text{Cov}(X_{ij}, Z_{ij}) \neq 0$. Note that when $\alpha_1 \neq 0$, Z linearly depends on X and may be viewed as a mediator in the relation between X and Y.

For the first simulation setting, we generate a standard normally distributed X, whereas X is sampled from a zero-centred Bernoulli distribution with success probability .5 in settings $2 - 4$. In simulation setting five,

Simulation	α_1	Distribution of X	$\text{Cov}(X_{ij}, Z_{ij})$
Sim 1	0.000	$N(0.00, 1.00)$	0.000
Sim 2	0.000	$B(1, 0.500) - 0.500$	0.000
Sim 3	-0.200	$B(1, 0.500) - 0.500$	-0.050
Sim 4	-1.500	$B(1, 0.500) - 0.500$	-0.375
Sim 5	-1.500	$B(1, \Phi(v_j^X)) - 0.500$	-0.235

Table 3.1 A summary of the five different simulation settings. Each setting considers a different combination of a value for α_1 and a distribution for X (e.g. $B(1, \Phi(v_j^X)) - 0.5$ reflects a mean centred Bernoulli variable with success probability $\Phi(v_j^X)$, with Φ representing the cumulative normal distribution an v_j^X a standard normally distributed random effect). When $\alpha_1 \neq 0$, or when a random intercept for X is introduced, which is correlated with the random intercepts for Z and Y (as in in Sim 5), the covariance between X_{ij} and Z_{ij}, $\text{Cov}(X_{ij}, Z_{ij})$, will differ from zero.

the true data generating model for X_{ij} is:

$$Probit(X_{ij} = 1) = v_j^X$$

with v_j^X following a standard normal distribution. Furthermore, v_j^X is correlated with v_j^Z and v_j^Y, with a correlation equal to 0.5. The latter implies the existence of an unmeasured upper-level confounder of X, Y and Z, inducing an additional covariation between X_{ij} and Z_{ij}.

In the first two settings $\alpha_1 = 0$, while $\alpha_1 = -0.2$ in the third, and $\alpha_1 = -1.5$ in the fourth and fifth setting. Although setting $\alpha_0 = 0$ and $\beta_0 = 0$ implies that both X and Z already exhibit mean zero at the population level in all settings, we additionally grand mean centre all variables in the samples prior to analysis. Additionally, all estimation models were fitted using the *lmer*-function from the *lme4* R-package. The R-code used to generate the simulated data is available in the supplementary material.

The means of the 1000 parameter estimates (with the relative bias), the mean of the standard errors (with the relative bias), the coverage and power of the estimators are summarised in table 3.2. Estimators that show significant bias are displayed in boldface. Note that we only displayed the results for three of the five approaches, since P1C2+ and C1P2+ yield results identical to P1C2 and C1P2, respectively (for our balanced simulation data). Before we summarise the results, we re-iterate that the fixed effects approach results in the exact same estimates as obtained by the P1C2-approach.

	Estimator	$\hat{\gamma}_1$				$\hat{\gamma}_2$				$\hat{\gamma}_3$			
		Est. ($rel.bias$)	se ($rel.bias$)	Cov.	Power	Est. ($rel.bias$)	se ($rel.bias$)	Cov.	Power	Est. ($rel.bias$)	se ($rel.bias$)	Cov.	Power
Sim 1	P1C2	0.101 (0.006)	0.028 (-0.118)	0.92	0.93	0.150 (0.001)	0.028 (0.011)	0.96	1.00	-0.101 (0.012)	0.020 (0.026)	0.95	1.00
	C1P2	0.101 (0.006)	0.028 (-0.126)	0.92	0.92	0.150 (0.000)	0.028 (0.011)	0.96	1.00	-0.101 (0.008)	0.029 (0.010)	0.95	0.94
	C1P2++	0.101 (0.007)	0.028 (-0.121)	0.93	0.93	0.150 (0.000)	0.028 (0.007)	0.96	1.00	-0.101 (0.009)	0.029 (0.014)	0.95	0.94
Sim 2	P1C2	0.103 (0.030)	0.055 (-0.009)	0.95	0.46	0.150 (-0.001)	0.028 (0.026)	0.96	1.00	-0.101 (0.006)	0.039 (-0.015)	0.95	0.72
	C1P2	0.103 (0.029)	0.055 (-0.007)	0.95	0.47	0.150 (-0.001)	0.028 (0.018)	0.96	1.00	-0.099 (-0.011)	0.058 (-0.003)	0.95	0.41
	C1P2++	0.103 (0.029)	0.055 (-0.009)	0.95	0.46	0.150 (-0.001)	0.028 (0.022)	0.96	1.00	-0.099 (-0.011)	0.058 (-0.003)	0.95	0.41
Sim 3	P1C2	0.103 (0.030)	0.055 (-0.007)	0.95	0.46	0.150 (-0.001)	0.028 (0.026)	0.96	1.00	-0.101 (0.006)	0.039 (-0.015)	0.95	0.72
	C1P2	0.103 (0.029)	0.055 (-0.007)	0.95	0.46	0.150 (-0.001)	0.028 (0.018)	0.96	1.00	-0.099 (-0.013)	0.058 (-0.002)	0.95	0.41
	C1P2++	0.103 (0.029)	0.055 (-0.007)	0.95	0.46	0.150 (-0.001)	0.028 (0.022)	0.96	1.00	-0.099 (-0.011)	0.058 (-0.002)	0.95	0.41
Sim 4	P1C2	0.103 (0.028)	0.069 (0.007)	0.95	0.33	0.150 (-0.001)	0.028 (0.026)	0.96	1.00	-0.101 (0.006)	0.039 (-0.015)	0.95	0.72
	C1P2	0.103 (0.027)	0.069 (0.009)	0.95	0.33	0.150 (-0.001)	0.028 (0.018)	0.96	1.00	**-0.090 (-0.104)**	0.055 (0.009)	0.95	0.37
	C1P2++	0.103 (0.027)	0.069 (0.009)	0.96	0.33	0.150 (-0.001)	0.028 (0.022)	0.96	1.00	-0.099 (-0.011)	0.056 (0.011)	0.96	0.41
Sim 5	P1C2	0.102 (0.020)	0.079 (0.006)	0.95	0.25	0.150 (-0.001)	0.028 (0.000)	0.95	1.00	-0.101 (0.009)	0.044 (0.007)	0.95	0.63
	C1P2	0.102 (0.022)	0.079 (0.005)	0.95	0.25	0.150 (-0.001)	0.028 (-0.004)	0.95	1.00	**-0.083 (-0.166)**	0.058 (0.002)	0.95	0.30
	C1P2++	0.102 (0.021)	0.079 (0.005)	0.95	0.24	0.150 (-0.001)	0.028 (0.000)	0.95	1.00	-0.101 (-0.011)	0.060 (0.050)	0.95	0.39

Table 3.2 Simulation results in the balanced setting. The means of the estimates (*Est.*) over the 1000 simulations (with the relative bias, *rel.bias*), the mean standard errors (*se*) over the 1000 simulations (with the relative bias), the coverage (Cov.), and power for each parameter. Not all approaches are displayed since P1C2+ and C1P2+ provide the exact same results as P1C2 and C1P2, respectively. The true values of β_1, β_2 and β_3 are 0.1, 0.15 and -0.1; when there is significant bias, the mean of the estimates is depicted in boldface.

Let us first focus on the bias. For the first and second simulation setting, where X and Z are independent, we do not observe bias (or more precisely, the relative bias is smaller than 5%, and not significant) for all within-effects under all approaches. In the third to fifth setting, Z depends on X in a linear fashion; when the absolute value of the effect of X on Z is increased (i.e. comparing simulation 4 to simulation 3), we observe bias (i.e. the relative bias is larger than 10% in absolute value, and found to be significant) in the estimator for the interaction effect in C1P2 and C1P2+. When X and Z are zero-mean centred symmetric distributions and Z is linear in X (as is the case in the third, fourth and fifth setting), we see for the OLS-estimator of γ_3 under C1P2:

$$E(\hat{\gamma}_3) = \beta_3 \frac{\text{cov}[X_{ij}Z_{ij}, X_{ij}^c Z_{ij}^c]}{\text{var}[X_{ij}^c Z_{ij}^c]}$$

As pointed out by Croissant and Millo (2008), the OLS-estimators are equivalent to the maximum likelihood estimators (as obtained through the *lmer* function) in our simulations, since we are assuming normality, homoscedasticity and no serial correlation of the errors. The derivation of the above expression can be found in the appendix. Notably, the bias depends on the distribution of X, as well as on the absolute value of α_1. We can see that $\text{cov}[X_{ij}Z_{ij}, X_{ij}^c Z_{ij}^c]$ can be written as the sum of $\text{var}[X_{ij}^c Z_{ij}^c]$ and some other terms that depend on $\alpha_1^2\text{cov}[\overline{X}_j, (X_{ij} - \overline{X}_j)^2]$ and $\alpha_1^2\text{cov}[\overline{X}_j^2, (X_{ij} - \overline{X}_j)^2]$. While the latter two covariances are zero when the distribution of X_{ij} is Gaussian, these covariances no longer equal zero when the distribution of X_{ij} becomes Bernoulli (Dodge and Rousson, 2012). Interestingly, when all cross-level interactions are included in C1P2++, this bias for the interaction effect in C1P2 and C1P2+ disappears. In sum, we find that the estimators of the P1C2, P1C2+, and C1P2++ approaches target the exact same population parameters under the assumed data-generating model.

Next, we take a look at the precision and power. The mean standard error for the estimator of the interaction effect is substantially lower in the P1C2 approaches, compared to C1P2++ in all simulation settings. The mean standard errors of the main effect estimators, on the other hand, are similar across all approaches. Furthermore, from the relative bias of the estimated standard errors we can see that the empirical standard deviation and the mean of the estimated standard error closely correspond under all approaches, for the main and interaction effects (except for the main

effect of X under the first simulation setting). As a consequence, we also observe appropriate coverages for these estimators. We also ran simulations with zero values for all lower-level effects (i.e. all β's equal to zero), and found appropriates type-I errors for all methods (results not shown), in line with the coverages reported. Importantly, given the higher precision of the estimated interaction effect under P1C2, we also observe the highest power for detecting the interaction under this approach. However, it should be noted that the simulation results describe average performances and, in practice, data may be encountered where the P1C2 approach yields a larger p-value for the interaction effect, compared to C1P2++.

So far, our simulation study only considered balanced data. Since our diary study was not always complete over the course of the 21 days, we repeated the above five simulation settings with a missingness pattern similar to the example data. More specifically, we introduced varying cluster sizes by sampling them as rounded values from a shifted beta-distribution, such that cluster sizes varied between 1-21 (with its mode around 18). The substantive findings from this unbalanced setting are essentially the same as in the balanced case (see table 3.3). Note, however, that due to the unbalanced nature of the simulations, the estimators of the P1C2 and P1C2+ approaches, and of the C1P2 and C1P2+ approaches, are no longer identical.

We limit the results of our simulation studies to the settings presented here for two reasons. First, the specific settings we considered allow us to derive analytical expressions for the observed biases. Second, further simulation studies with different choices (e.g., non-symmetric distributions for X and Z, non-linear associations between X and Z, ...) lead to similar conclusions: (1) both P1C2 and C1P2++ yield unbiased estimators for the interaction effect, (2) both exhibit an appropriate coverage of their 95% confidence intervals, but (2) P1C2 is always more precise. This conclusion can also be drawn from our illustrating example: we observe more precise estimators for the interaction effect in P1C2, compared to the C1P2 approaches.

What are the practical implications of these findings in terms of interpretation? First of all, we found that the parameters γ_1, γ_2 and γ_3 in estimation models (3.9) and (3.12) (i.e. the P1C2 and C1P2++ approaches) target the exact same population parameters and can hence be given the same interpretation. Considering the estimates of the P1C2 approach (and C1P2++, respectively) in our illustrating example, we see that masturbation dilutes the positive effect of intimacy on next day's pos-

	Estimator	$\hat{\gamma}_1$				$\hat{\gamma}_2$				$\hat{\gamma}_3$			
		Est. (*rel.bias*)	*se* (*rel.bias*)	Coverage	Power	Est. (*rel.bias*)	*se* (*rel.bias*)	Coverage	Power	Est. (*rel.bias*)	*se* (*rel.bias*)	Coverage	Power
Sim 1	P1C2	0.101 (0.013)	0.032 (-0.061)	0.93	0.87	0.152 (0.011)	0.032 (0.009)	0.95	1.00	-0.101 (0.013)	0.023 (-0.017)	0.96	0.99
	P1C2+	0.101 (0.013)	0.032 (-0.061)	0.93	0.87	0.152 (0.011)	0.032 (0.009)	0.95	1.00	-0.101 (0.013)	0.023 (-0.017)	0.96	0.99
	C1P2	0.101 (0.012)	0.032 (-0.061)	0.94	0.86	0.152 (0.011)	0.032 (0.016)	0.95	1.00	-0.101 (0.010)	0.035 (0.000)	0.94	0.83
	C1P2+	0.101 (0.012)	0.032 (-0.061)	0.94	0.86	0.152 (0.011)	0.032 (0.016)	0.95	1.00	-0.101 (0.011)	0.035 (0.000)	0.94	0.83
	C1P2++	0.101 (0.012)	0.032 (-0.061)	0.93	0.86	0.152 (0.011)	0.032 (0.006)	0.95	1.00	-0.101 (0.010)	0.035 (0.003)	0.94	0.84
Sim 2	P1C2	0.100 (0.004)	0.064 (0.019)	0.96	0.35	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.101 (0.008)	0.046 (0.011)	0.95	0.59
	P1C2+	0.100 (0.004)	0.064 (0.019)	0.96	0.35	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.101 (0.008)	0.046 (0.011)	0.95	0.59
	C1P2	0.101 (0.005)	0.064 (0.019)	0.95	0.35	0.149 (-0.008)	0.032 (0.006)	0.95	1.00	-0.100 (0.001)	0.068 (-0.010)	0.95	0.31
	C1P2+	0.101 (0.005)	0.064 (0.019)	0.95	0.35	0.149 (-0.008)	0.032 (0.006)	0.95	1.00	-0.100 (-0.001)	0.068 (-0.007)	0.95	0.31
	C1P2++	0.100 (0.004)	0.064 (0.017)	0.96	0.35	0.149 (-0.008)	0.032 (0.009)	0.95	1.00	-0.100 (0.000)	0.068 (-0.009)	0.95	0.31
Sim 3	P1C2	0.100 (0.002)	0.065 (0.019)	0.95	0.34	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.101 (0.008)	0.046 (0.011)	0.95	0.59
	P1C2+	0.100 (0.002)	0.065 (0.019)	0.95	0.34	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.101 (0.008)	0.046 (0.011)	0.95	0.59
	C1P2	0.100 (0.003)	0.065 (0.017)	0.95	0.35	0.149 (-0.008)	0.032 (0.006)	0.95	1.00	-0.100 (-0.001)	0.069 (-0.010)	0.95	0.31
	C1P2+	0.100 (0.003)	0.065 (0.017)	0.95	0.35	0.149 (-0.008)	0.032 (0.006)	0.95	1.00	-0.100 (-0.003)	0.069 (-0.007)	0.95	0.31
	C1P2++	0.100 (0.001)	0.065 (0.017)	0.95	0.35	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.100 (0.000)	0.069 (-0.009)	0.95	0.31
Sim 4	P1C2	0.099 (-0.015)	0.080 (0.016)	0.96	0.25	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.101 (0.008)	0.046 (0.011)	0.95	0.59
	P1C2	0.099 (-0.015)	0.080 (0.016)	0.96	0.25	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.101 (0.008)	0.046 (0.011)	0.95	0.59
	C1P2	0.099 (-0.013)	0.081 (0.013)	0.96	0.25	0.149 (-0.008)	0.032 (0.006)	0.95	1.00	**-0.088 (-0.119)**	0.065 (-0.008)	0.95	0.28
	C1P2+	0.099 (-0.013)	0.081 (0.013)	0.96	0.25	0.149 (-0.008)	0.032 (0.006)	0.95	1.00	**-0.088 (-0.120)**	0.065 (-0.006)	0.95	0.29
	C1P2++	0.098 (-0.016)	0.081 (0.014)	0.96	0.25	0.149 (-0.009)	0.032 (0.009)	0.95	1.00	-0.100 (-0.002)	0.066 (-0.009)-	0.95	0.34
Sim 5	P1C2	0.099 (-0.014)	0.093 (-0.019)	0.94	0.20	0.148 (-0.012)	0.032 (-0.019)	0.95	1.00	-0.102 (0.016)	0.052 (0.002)	0.96	0.50
	P1C2	0.099 (-0.014)	0.093 (-0.019)	0.94	0.20	0.148 (-0.012)	0.032 (-0.019)	0.95	1.00	-0.102 (0.016)	0.052 (0.002)	0.96	0.50
	C1P2	0.099 (-0.010)	0.093 (-0.021)	0.94	0.20	0.148 (-0.012)	0.032 (0.003)	0.95	1.00	**-0.079 (-0.209)**	0.069 (-0.024)	0.94	0.22
	C1P2+	0.099 (-0.010)	0.093 (-0.021)	0.94	0.20	0.148 (-0.011)	0.032 (0.003)	0.95	1.00	**-0.079 (-0.211)**	0.069 (-0.025)	0.94	0.23
	C1P2++	0.099 (-0.015)	0.093 (-0.019)	0.94	0.19	0.148 (-0.012)	0.032 (0.000)	0.95	1.00	-0.099 (-0.015)	0.071 (-0.019)-	0.95	0.30

Table 3.3 Simulation results in the unbalanced setting. The means of the estimates (*Est.*) (with relative bias), the mean standard errors (*se*) (with relative bias), the coverage, and power for each parameter and simulation setting. All approaches are displayed, as in unbalanced designs P1C2+ and C1P2+ no longer provide the same results as P1C2 and C1P2, respectively. The true values of β_1, β_2 and β_3 are 0.1, 0.15 and -0.1; when there is significant bias, the mean of the estimates is depicted in boldface.

itive relationship feelings with on average 0.075 units (0.096, respectively), for every unit increase in intimacy (table 3.4). Averaging over days with and without masturbation, a one-unit increase in intimacy within a male individual will result in an 0.079 (0.080, respectively) increase in his next day's positive relationship feelings (table 3.4). Equivalently, at average levels of intimacy, masturbation reduces next day's positive relationship feelings with on average 0.151 points (0.167 respectively) (table 3.4).

6 Discussion

This paper compared two alternative approaches for the centering of lower-level interactions. In our simulation study, the P1C2-approaches outperformed the C1P2-approaches in estimating such interactions: (1) P1C2 results in more precise estimates of the interaction effect, compared to the three C1P2-approaches; (2) P1C2 is not affected by misspecification or omission of upper-level effects, in contrast to C1P2 (unless all cross-level interactions are included).

It can be argued that the data-generating models considered here are somewhat restrictive. However, it is important to note that the performance of the two prevailing approaches for centering interactions was explored in settings where CWC is usually considered a good remedy. That is, we studied settings with additive effects for unmeasured upper-level confounders, because such effects can be effectively eliminated by CWC.

A first important assumption underlying data generating model (3.7) constitutes homogeneous effects amongst subjects. In the presence of heterogeneous subject-effects, random slopes for X, Z, as well as for their interaction can be added to the estimation models. Fortunately, relying on estimation through a simple random intercept model such as (3.9) (which ignores any heterogeneity) will not introduce bias in the effect estimates, provided that the random slopes are independent of the predictors (Baird and Maxwell, 2016). In contrast, if the random slopes were to be correlated with the predictors, CWC would no longer effectively eliminate unmeasured upper-level heterogeneity; alternative approaches such as fixed-effect estimation or per-cluster analysis would then be required (Bates et al., 2014).

A second important assumption underlying data generating model (3.7) entails the absence of unmeasured lower-level confounding. If for example daily intimacy, masturbation, and positive relational feelings were associated with an (unmeasured) daily positive mood (given unmeasured

Parameter	*Intercept*		*Intimacy*		*Masturbation*		*Interaction*		*Variances*	
	Estimate (s.e.) t-value (df)	p	Estimate (s.e.) t-value (df)	p	Estimate (s.e.) t-value (df)	p	Estimate (s.e.)	p	Rand. int.	Res.
P1C2	5.183 (0.102) 50.933 (65)	$< .001$	0.079 (0.015) 5.443 (1015)	$< .001$	-0.151 (0.079) -1.907 (1015)	.057	-0.075 (0.039) -1.930 (1015)	.054	0.6373	0.5576
P1C2+	5.180 (0.080) 64.667 (62)	$< .001$	0.079 (0.015) 5.444 (1015)	$< .001$	-0.150 (0.079) -1.890 (1015)	.059	-0.075 (0.039) -1.928 (1015)	.054	0.3806	0.5576
C1P2	5.179 (0.102) 50.902 (65)	$< .001$	0.080 (0.015) 5.523 (1015)	$< .001$	-0.163 (0.080) -2.041 (1016)	.042	-0.102 (0.050) -2.041 (1030)	.042	0.6366	0.5574
C1P2+	5.197 (0.084) 62.232 (61)	$< .001$	0.080 (0.015) 5.521 (1015)	$< .001$	-0.160 (0.080) -2.012 (1016)	.045	-0.098 (0.045) -1.968 (1038)	.049	0.3774	0.5574
C1P2++	5.197 (0.084) 62.236 (61)	$< .001$	0.080 (0.015) 5.520 (1013)	$< .001$	-0.167 (0.080) -2.091 (1014)	.037	-0.096 (0.050) -1.917 (1036)	.056	0.3773	0.5573

Table 3.4 The parameter estimates, standard errors (s.e.) with associated t-statistics, degrees of freedom, and p-values, for the intercept, the within-subject main effects of intimacy and masturbation, as well as their within-subject interaction effect on next day's positive relationship feelings. Additionally, the estimated random intercept (Rand. int.) variances and residual error variances (Res.) are provided. Five different estimation approaches are considered: P1C2, P1C2+,C1P2, C1P2+ and C1P2++.

subject-specific confounders), this assumption would be violated. Since CWC only eliminates time-invariant confounding, we would expect biased effect estimators under unmeasured lower-level confounding. However, as recently pointed out by Loeys et al. (2016), the assessment of interaction effects in linear models often requires weaker 'no-unmeasured-confounding' assumptions, compared to main effects. Hence, unbiased effect estimators for the interaction may still be found under relatively lenient assumptions.

Third, we limited our discussion to linear settings. As shown by Goetgeluk and Vansteelandt (2008), separating a within- from a between-effect in a random intercept model only yields a consistent estimator of the within-effect in the presence of upper-level confounding when the model is linear. For nonlinear models, it is possible to encounter an inconsistent estimator, though in practice this bias will often be small.

To summarise, when dealing with multilevel data, we recommend that careful consideration be given to the assumptions under which separating within- from between-effects yield valid results. When those assumptions are deemed plausible, CWC can be applied to unbiasedly estimate within-cluster effects. For the estimation of interaction effects we advocate the P1C2-approach rather than the C1P2-approach, as the former is much more efficient. If researchers want to use the C1P2-approach (e.g. because of implementations in software packages for SEM), we recommend not to drop any cross-level or upper-level terms, even when they are not of interest.

Bibliography

Allison, P. (2015). Don't put lagged dependent variables in mixed models. `https://statisticalhorizons.com/lagged-dependent-variables`. Accessed: 2017-12-10.

Baird, M. D. and Maxwell, S. E. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological Methods*, 21(2):175–188.

Bates, M. D., Castellano, K. E., Rabe-Hesketh, S., and Skrondal, A. (2014). Handling Correlations Between Covariates and Random Slopes in Multilevel Models. *Journal of Educational and Behavioral Statistics*, 39(6):524–549.

Brumback, B. A., Dailey, A. B., Brumback, L. C., Livingston, M. D., and

He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters*, 80(21-22):1650–1654.

Castellano, K. E., Rabe-Hesketh, S., and Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics*, 39(5):333–367.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R : The plm package. *Journal of Statistical Software*, 27(2):1–52.

Curran, P. J. and Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual review of psychology*, 62:583–619.

Dewitte, M., Van Lankveld, J., Vandenberghe, S., and Loeys, T. (2015). Sex in its daily relational context. *Journal of Sexual Medicine*, 12(12):2436–2450.

Dodge, Y. and Rousson, V. (2012). The complications of the fourth central moment. *The American Statistician*, 1305(May):2–5.

Enders, C. K. and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2):121–138.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, 21:389–395.

Hofmann, D. A. and Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24(5):623–641.

Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., and Loeys, T. (2015). Within-subject mediation analysis in AB/BA crossover designs. *The International Journal of Biostatistics*, 11(1):1–22.

Kreft, I. G. G., de Leeuw, J., and Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, 30(30):1–21.

Loeys, T., Talloen, W., Goubert, L., Moerkerke, B., and Vansteelandt, S. (2016). Assessing moderated mediation in linear models requires fewer confounding assumptions than assessing mediation. *British Journal of Mathematical and Statistical Psychology*, 69(3):352–374.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., and Muthén, B. (2008). The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychological methods*, 13(3):203–229.

McNeish, D., Stapleton, L. M., and Silverman, R. D. (2016). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, 22(1):114–140.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology - This time forever. *Measurement: Interdisciplinary Research & Perspective*, 2(4):201–218.

Molenaar, P. C. M. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, 18:112–117.

Mundlak, Y. (1978). Pooling of time series and cross-section data. *Econometrica*, 46:69–86.

Muthén, B. O. (1990). Mean and covariance structure analysis of hierarchical data. *UCLA Statistics Series*, 62.

Nesselroade, J. R. and Molenaar, P. C. M. (2016). Some behaviorial science measurement concerns and proposals. *Multivariate Behavioral Research*, 51(2-3):396–412.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

Preacher, K. J., Zhang, Z., and Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, 21(2):189–205.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and data analysis methods.* Sage, Thousand Oaks, CA, second edition.

Ryu, E. (2015). The role of centering for interaction of level 1 variables in multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4):617–630.

Wang, L. P. and Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, 20(1):63–83.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. The MIT Press, Cambridge, MA.

B Appendix

B.1 Bias of the interaction effect estimator under the P1C2 approach

Assume that the true models for Z and Y are:

$$Z_{ij} = \alpha_0 + \alpha_1 X_{ij} + v_j^Z + \epsilon_{ij}^Z \tag{3.17}$$
$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_3 X_{ij} Z_{ij} + v_j^Y + \epsilon_{ij}^Y \tag{3.18}$$

with ϵ_{ij}^Z and ϵ_{ij}^Y i.i.d. with mean zero and variance σ_Z^2 and σ_Y^2, respectively. Consider the estimation model:

$$E[Y_{ij} \mid X_{ij}, Z_{ij}, u_j] = \gamma_0 + \gamma_1 X_{ij}^c + \gamma_2 Z_{ij}^c + \gamma_3 X_{ij}^c Z_{ij}^c + u_j, \tag{3.19}$$

where $X_{ij}^c = (X_{ij} - \overline{X}_j)$ and $Z_{ij}^c = (Z_{ij} - \overline{Z}_j)$.

The OLS-estimators for the parameters of model (3.19), under models (3.17) and (3.18) are given by $\Sigma^{-1}\Sigma_{VY}$ with $V_{ij} = (1 \; X_{ij}^c \; Z_{ij}^c \; X_{ij}^c Z_{ij}^c)'$, $\Sigma = E[V_{ij}V_{ij}']$ and $\Sigma_{VY} = \big(E[Y_{ij}] \; E[X_{ij}^c Y_{ij}] \; E[Z_{ij}^c Y_{ij}] \; E[X_{ij}^c Z_{ij}^c Y_{ij}]\big)'$.

Now, we have that:

$$V_{ij}V_{ij}' = \begin{pmatrix} 1 & X_{ij}^c & Z_{ij}^c & X_{ij}^c Z_{ij}^c \\ X_{ij}^c & {X_{ij}^c}^2 & X_{ij}^c Z_{ij}^c & {X_{ij}^c}^2 Z_{ij}^c \\ Z_{ij}^c & X_{ij}^c Z_{ij}^c & {Z_{ij}^c}^2 & X_{ij}^c {Z_{ij}^c}^2 \\ X_{ij}^c Z_{ij}^c & {X_{ij}^c}^2 Z_{ij}^c & X_{ij}^c {Z_{ij}^c}^2 & {X_{ij}^c}^2 {Z_{ij}^c}^2 \end{pmatrix}$$

Assuming that Z is linear in X, $E(X_{ij}) = E(Z_{ij}) = 0$, while also assuming a symmetric distribution for X, the expectation of $V_{ij}V_{ij}'$ simplifies to

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & \alpha_1 \mathrm{var}[X_{ij}^c] \\ 0 & \mathrm{var}[X_{ij}^c] & \alpha_1 \mathrm{var}[X_{ij}^c] & 0 \\ 0 & \alpha_1 \mathrm{var}[X_{ij}^c] & \alpha_1^2 E[{X_{ij}^c}^2] + \mathrm{var}[\epsilon_{ij}^{Zc}] & 0 \\ \alpha_1 \mathrm{var}[X_{ij}^c] & 0 & 0 & \alpha_1^2 E[{X_{ij}^c}^4] \\ & & & +\mathrm{var}[\epsilon_{ij}^{Zc}]\mathrm{var}[X_{ij}^c] \end{pmatrix},$$

with $\epsilon_{ij}^{Zc} = (\epsilon_{ij}^Z - \overline{\epsilon^Z}_j)$. In order to obtain the elements c_{kl} of Σ^{-1}, we need its determinant. After some tedious calculations, we find that: $|\Sigma| = \mathrm{var}[\epsilon_{ij}^{Zc}]\mathrm{var}[X_{ij}^c]\Big(\alpha_1^2 \mathrm{var}[{X_{ij}^c}^2] + \mathrm{var}[\epsilon_{ij}^{Zc}]\mathrm{var}[X_{ij}^c]\Big)$, $c_{21} = c_{24} = c_{31} = c_{34} = c_{42} = c_{43} = 0$, $c_{22} = \frac{\alpha_1^2}{\mathrm{var}[\epsilon_{ij}^{Zc}]} + \frac{1}{\mathrm{var}[X_{ij}^c]}$, $c_{23} = c_{32} = -\frac{\alpha_1}{\mathrm{var}[\epsilon_{ij}^{Zc}]}$, $c_{33} = \frac{1}{\mathrm{var}[\epsilon_{ij}^{Zc}]}$, $c_{41} = -\frac{\alpha_1 \mathrm{var}[X_{ij}^c]}{\alpha_1^2 \mathrm{var}[{X_{ij}^c}^2] + \mathrm{var}[\epsilon_{ij}^{Zc}]\mathrm{var}[X_{ij}^c]}$, and $c_{44} = \frac{1}{\alpha_1^2 \mathrm{var}[{X_{ij}^c}^2] + \mathrm{var}[\epsilon_{ij}^{Zc}]\mathrm{var}[X_{ij}^c]}$.

As such, we can show that there is no bias in the OLS-estimator $\hat{\gamma}_1$ for β_1:

$$\begin{aligned} E[\hat{\gamma}_1] =& c_{22}E[X_{ij}^c Y_{ij}] + c_{23}E[Z_{ij}^c Y_{ij}] \\ =& (c_{22} + \alpha_1 c_{23})(\beta_1 + \beta_2\alpha_1)\text{var}[X_{ij}^c] + c_{23}\beta_2\text{var}[\epsilon_{ij}^{Zc}] \\ =& \beta_1 \end{aligned}$$

Similarly, we find no bias in the OLS-estimator $\hat{\gamma}_2$ for β_2:

$$\begin{aligned} E(\hat{\gamma}_2) =& c_{32}E[X_{ij}^c Y_{ij}] + c_{33}E[Z_{ij}^c Y_{ij}] \\ =& c_{32}(\beta_1 + \beta_2\alpha_1)\text{var}[X_{ij}^c] + c_{33}\alpha_1(\beta_1 + \beta_2\alpha_1)\text{var}[X_{ij}^c] + c_{33}\beta_2\text{var}[\epsilon_{ij}^{Zc}] \\ =& \beta_2 \end{aligned}$$

And finally, we find for the OLS-estimator $\hat{\gamma}_3$ for β_3 that:

$$\begin{aligned} E(\hat{\gamma}_3) =& c_{41}E[Y_{ij}] + c_{44}E[X_{ij}^c Z_{ij}^c Y_{ij}] \\ =& \beta_3 \frac{\alpha_1^2\text{cov}[X_{ij}^2, {X_{ij}^c}^2]}{\alpha_1^2\text{var}[{X_{ij}^c}^2] + \text{var}[\epsilon_{ij}^{Zc}]\text{var}[X_{ij}^c]} \end{aligned}$$

with $E[{X_{ij}^c}^2 X_{ij}] = E[{X_{ij}^c}^2(X_{ij}^c + \overline{X}_j)] = E[{X_{ij}^c}^3] + \text{cov}[{X_{ij}^c}^2, \overline{X}_j] = 0$ for a symmetric X.

As $\text{var}[X_{ij}^c Z_{ij}^c]$ can be re-expressed as:

$$\begin{aligned} \text{var}[X_{ij}^c Z_{ij}^c] &= \text{var}[X_{ij}^c(\alpha_1 X_{ij}^c + \epsilon_{ij}^{Zc}] \\ &= \alpha_1^2\text{var}[{X_{ij}^c}^2] + \text{var}[X_{ij}^c \epsilon_{ij}^{Zc}] \end{aligned}$$

and $\text{cov}[X_{ij}Z_{ij}, X_{ij}^c Z_{ij}^c]$ as

$$\begin{aligned} \text{cov}[X_{ij}Z_{ij}, X_{ij}^c Z_{ij}^c] &= \text{cov}\big[X_{ij}(\alpha_1 X_{ij} + \epsilon_{ij}^Z), X_{ij}^c(\alpha_1 X_{ij}^c + \epsilon_{ij}^{Zc})\big] \\ &= \alpha_1^2\text{cov}[X_{ij}^2, {X_{ij}^c}^2] \end{aligned}$$

we find that the bias factor for $\hat{\gamma}_3$ can be rewritten as:

$$\frac{\text{cov}[X_{ij}Z_{ij}, X_{ij}^c Z_{ij}^c]}{\text{var}[X_{ij}^c Z_{ij}^c]},$$

which will equal one when the distribution of X is normal, but will be smaller than one when X is Bernoulli distributed.

4 A review of R-packages for random-intercept probit regression in small clusters

Abstract. Generalised Linear Mixed Models (GLMMs) are widely used to model clustered categorical outcomes. To tackle the intractable integration over the random effects distributions, several approximation approaches have been developed for likelihood-based inference. As these seldom yield satisfactory results when analysing binary outcomes from small clusters, estimation within the Structural Equation Modelling (SEM) framework is proposed as an alternative. We compare the performance of R-packages for random-intercept probit regression relying on: the Laplace approximation, adaptive Gaussian quadrature (AGQ), penalised quasi-likelihood, an MCMC-implementation, and integrated nested Laplace approximation within the GLMM-framework, and a robust diagonally weighted least squares estimation within the SEM-framework. In terms of bias for the fixed and random effect estimators, SEM usually performs best for cluster size two, while AGQ prevails in terms of precision (mainly because of SEM's robust standard errors). As the cluster size increases, however, AGQ becomes the best choice for both bias and precision.

This chapter is based on Josephy, H., Loeys, T., & Rosseel, Y. (2016). A review of R-packages for random-intercept probit regression in small clusters. *Frontiers in Applied Mathematics and Statistics*, 2 (18): 1-13.

1 Introduction

In behavioural and social sciences, researchers are frequently confronted with clustered or correlated data structures. Such hierarchical data sets for example arise from educational studies, in which students are measured within classrooms, or from longitudinal studies, in which measurements are repeatedly taken within individuals. In these examples, two levels can be distinguished within the data: measurements or level-1 units (e.g. students or time points), and clusters or level-2 units (e.g. classes or individuals). These lower-level units are correlated, as outcome measures arising from students with the same teacher, or measurements within an individual, will be more alike than data arising from students with different teachers, or measurements from different individuals. As such, an analysis that ignores these dependencies may yield underestimated standard errors, while inappropriate aggregation across levels may result in biased coefficients (Snijders and Bosker, 1999; Raudenbush and Bryk, 2002).

Over the course of decades, several frameworks that can deal with such lower-level correlation have been developed. One such framework entails mixed effect models, which model both the ordinary regression parameters common to all clusters (i.e. the fixed effects), as well as any cluster-specific parameters (i.e. the random effects). Using a parametric approach, two different types can be distinguished: Linear Mixed Models (LMMs) when the outcome is normally distributed, and Generalised Linear Mixed Models (GLMMs) when it is not. A second framework that allows the analysis of multilevel outcomes consists of Structural Equation Models (SEM). Structural Equation Models can be split up into two main classes: 'classic' SEM, which is restricted to balanced data, and multilevel SEM, which is able to deal with unbalanced data structures by relying on likelihood-based or Bayesian approaches. Generally, SEM supersects its GLMM counterpart, as the former is able to additionally include latent measures (and measurement error) and assess mediation, in one big model. Discounting these two assets, however, recent literature proves that SEM is completely equivalent to its GLMM counterpart when considering balanced data (e.g. when considering equal cluster sizes in a random intercept model) (Rovine and Molenaar, 2000; Curran, 2003; Bauer, 2003).

As clustered Gaussian outcomes have already been discussed thoroughly in the LMM and SEM literature (Airy, 1861; Scheffé, 1959; Harville, 1977; Laird and Ware, 1982; Goldstein, 1979; Bauer, 2003; Curran, 2003), we

will focus on GLMM- and SEM-methods for non-normal outcome data. More specifically, we will target binary data from small clusters, with a particular focus on clusters of size two, as such settings have proven difficult for the available GLMM methodologies (Breslow and Clayton, 1993; Rodriguez and Goldman, 1995). Clusters of size two are frequently encountered in practice, e.g. when studying dyads (McMahon et al., 2003), in ophthalmology data (Glynn and Rosner, 2013), in twin studies (Ortqvist et al., 2009), or when analysing measurements from a 2-period - 2-treatment crossover study (Senn, 2002).

Focusing on the two aforementioned frameworks, current literature on the analysis of clustered binary outcomes reveals two major limitations: clusters of size two were either not considered (Rabe-hesketh and Pickles, 2002; Browne and Draper, 2006; Zhang et al., 2009; Capanu et al., 2013), or they were, but limited to only one of both frameworks (Ten Have and Localio, 1999; Sutradhar and Mukerjee, 2005; Broström and Holmberg, 2011; Xu et al., 2014). Here, we compare several estimation procedures within both GLMM- and SEM-frameworks for modelling this type of data, by considering the performance of relevant R-packages. By limiting our comparison to implementations from the statistical environment `R` (version 3.2.3., R Core Team (2013)), we rely on estimation techniques that are easily accessible to all practitioners (this software is freely available, while at the same time enjoying a wide range of open-source packages). Additionally, we choose to only focus on R-packages which stand on themselves and are not dependent on external software. We do, however, check several of the `R`-based implementations against others such as implementations in `SAS`® software (version 9.4, SAS Institute Inc (2015))[1], the `MPLUS`® program (version 7.4, Muthén and Muthén (2010)) or the `JAGS` implementation (version 4.1.0. Plummer (2003)), as to verify the independence of conclusions on the software used.

In the following sections, we first introduce a motivating example. After this we elaborate on the GLMM and SEM frameworks in general, so that the various estimation methods capable of analysing the example can be enumerated. Next, we illustrate these methods on our example data. To facilitate the practitioner's decision on which method is most appropriate in which setting, we subsequently conduct a simulation study. Based on our findings we provide recommendations, and end with a discussion.

[1]SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

2 An example

As a motivating example, we consider data from a randomised study executed by Vandeweghe et al. (2018) in two Flemish nursery schools. As healthy eating habits are important to achieve healthy growth and development in young children, Vandeweghe et al. (2018) focus on strategies to improve the liking of vegetables in preschool children: a child given a tangible or non-tangible reward after tasting should be motivated to taste again. To this end, Vandeweghe et al. (2018) incorporated four possible intervention plans: encouragement towards eating chicory, an active reward after consumption, repeated exposure of the vegetable, and a control group. The binary variable 'vegetable liking' (like/ok versus dislike) was measured during three phases: once during a pretest (to test their inherent liking of chicory), once during a post-test, and once during a follow-up test. When we only consider the pre- and post-test, we end up with two measurements for each child, while additionally including the follow-up measurements will increase this number to three. So irrespective of whether or not the follow-up measurement is included, the authors end up with a small cluster size.

For illustrative purposes, we will only consider the results from a single school, so that the data structure simplifies to a simple two-level setting where a binary outcome is assessed repeatedly within each child. Additionally, we will only contrast the 'encouragement' versus the 'control' group, as to simplify interpretation and results. The sample size of this reduced data set consists of 37 children (only retaining the complete cases), of which 21 were assigned to the control group and 16 to the encouragement group.

To test whether encouragement increases the liking of chicory, we consider the following random-intercept *probit*-regression model:

$$P(y_{ij} = 1 \mid x_{ij}, b_j) = \Phi(\beta_0 + \beta_1 x_{ij} + b_j) \tag{4.1}$$

with index i referring to the measurement moment ($i = 0$, 1 or 2 for pre-, post- and follow-up test, respectively), index j to the individual ($j = 1, \ldots, 37$), and with Φ representing the cumulative normal distribution. Additionally, a random intercept b_j, which is assumed to follow a normal distribution, is included in model (2) to capture the correlation between measures taken from the same toddlers. In this model, the outcome variable Y_{ij} represents *Liking* (*Liking* equals zero when child j dislikes the vegetable at time i, and one when it is liked/tolerated), while the

predictor x_{ij} represents *Encouragement* (x_{ij} equals one when child j is encouraged at time i, and zero when it is not). To capture the effect of *Encouragement* within a single parameter, we have opted to model the intervention as a time-dependent covariate, rather than a between-subject effect interacting with time. This assumption is reasonable here, given the absence of group differences at the pretest, the nonexistence of a time effect in the control group, and a similar effect of *Encouragement* during the post-test and follow-up (see figure 4.1).

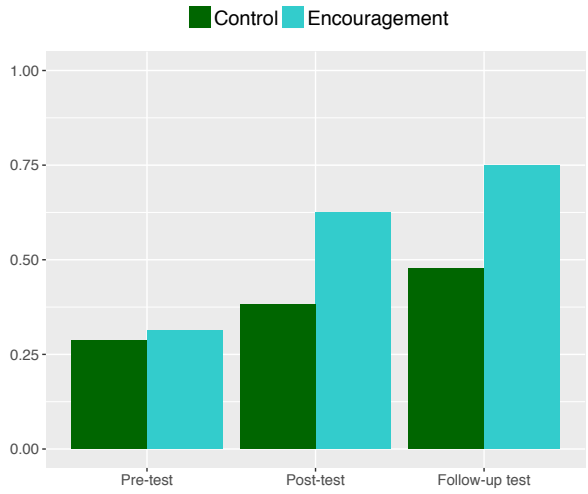

Figure 4.1 Percentages of vegetable liking in 37 preschool children, for the tree measurement moments (pretest, post-test and follow-up) and two reward systems (control versus encouragement).

With model (2) defined, the research question of whether or not a reward system will increase the liking of chicory will amount to testing the null hypothesis $H_0 : \beta_1 = 0$. When this null hypothesis is rejected, we will conclude that the reward system significantly increases (when $\beta > 0$) the probability of liking the vegetable. But how do we estimate and test the fixed effects and random intercept variance? Since there are myriad options and recommendations in current literature, and some of these may not yield satisfactory results for binary outcomes in such small clusters, we will introduce and compare several possibilities. As mentioned in the introduction, these estimation methods stem from both the GLMM- and SEM-frameworks; to this end, the next section provides an introduction of both frameworks, a short note on their equivalence, and an explanation

of the difficulties that accompany marginalising the GLMM-likelihood function over the random effects distribution.

3 Methods

3.1 Generalised Linear Mixed Models

Generalised linear mixed models (GLMMs) are basically extensions of Generalised Linear Models (GLMs) (Nelder and Wedderburn, 1972), which allow for correlated observations through the inclusion of random effects. Such effects can be interpreted as unobserved heterogeneity at the upper level, consequently inducing dependence among lower-level units from the same cluster.

Let x_{ij} and y_{ij} denote the ith measurement from cluster j, for the predictor and the binary outcome respectively (where $i = 1, .., I$ and $j = 1, ..., J$). Note that since we primarily focus on clusters of size two, we will set I to 2. Moreover, as $I = 2$ limits the identification of random effects, we will consider GLMMs with a random intercept only. In a fully parametric framework, this particular GLMM is typically formulated as:

$$E(Y_{ij}|x_{ij}, b_j) = g^{-1}(\beta_0 + \beta_1 x_{ij} + b_j) \qquad \text{with} \quad b_j \sim N(0, \tau) \qquad (4.2)$$

where $g^{-1}(\cdot)$ represents a known inverse link function, β_0 represents the intercept, β_1 the effect of the predictor x_{ij}, and b_j the cluster-specific random intercept. In this paper, we only consider *probit* regression models, where the standard normal cumulative distribution $\Phi(\cdot)$ is defined as the inverse link function $g^{-1}(\cdot)$ (or equivalently the link function $g(\cdot)$ is defined as $probit(\cdot)$). Our reasoning behind this is that *probit*-regression applies to all estimation procedures we investigate, in contrast to the *logit* link. Converting equation (4.2) to a random intercept *probit*-regression model yields us:

$$P(y_{ij} = 1 \mid x_{ij}, b_j) = \Phi(\beta_0 + \beta_1 x_{ij} + b_j) \qquad (4.3)$$

In order to obtain estimates for β_0, β_1 and τ, the marginal likelihood function is typically maximised. For a random-intercept GLMM, this function is obtained by integrating out the cluster-specific random effect,

and can be written as:

$$l(\boldsymbol{\beta}, \tau | y_{ij}) = \prod_{j=1}^{J} \int_{-\infty}^{+\infty} \prod_{i=1}^{I} f(y_{ij} | \boldsymbol{\beta}, b_j) \phi(b_j | \tau) db_j \tag{4.4}$$

where f denotes the density function of the outcomes and ϕ the density of the random intercept (which is assumed to be normal here).

Unfortunately, statistical inference based on maximising (4.4) is hampered, because integrating out the random effects from the joint density of responses and random effects is, except for a few cases, analytically intractable. To tackle this, several techniques have been proposed, which can be divided into two main classes: likelihood-based methods and Bayesian approaches.

3.1.1 Estimation through likelihood-based approximation methods

One way to tackle the intractability of integrating out the random effects of the GLMM likelihood function, is to either approximate the integrand or to approximate the integral itself. We briefly introduce three such methods below, and refer the interested reader to Tuerlinckx et al. (2006) for more details.

Technically speaking, the **Laplace** approximation (Tierney and Kadane, 1986) approximates the integrand by a quadratic Taylor expansion. This results in a closed-form expression of the marginal likelihood, which can be maximised to obtain the maximum likelihood estimates of the fixed effects and random effect variances. In `R`, the implementation based on this approximation is available within the function `glmer`, from the package `lme4` (Bates et al., 2015).

The Penalised Quasi-Likelihood method (**PQL**) (Breslow and Clayton, 1993; Schall, 1991; Stiratelli et al., 1984) also approximates the integrand; more intuitively put, PQL approximates the GLMM with a linear mixed model. This is achieved by considering a Taylor expansion of the response function and by subsequently rewriting this expression in terms of an adjusted dependent variable on which estimation procedures for LMM can be implemented. Consequently, the algorithm cycles between parameter estimation by linear mixed modelling, and updating the adjusted dependent variable until convergence. This approach can be implemented using the function `glmmPQL` from the `R`-package `MASS` (Venables and Ripley, 2002).

Finally, a tractable marginal likelihood can also be obtained by approximating the integral itself with a finite sum. In regular Gauss-Hermite

(GH) Quadrature (e.g. Naylor and Smith (1982)), this summation occurs over a fixed set of nodes, while Adaptive Gaussian Quadrature (**AGQ**) (Pinheiro and Bates, 1995) uses a different set of nodes for each cluster. As such, when applying AGQ, fewer nodes are necessary to achieve equal accuracy as compared to the regular GH quadrature. AGQ estimation in `R` is also possible within the `glmer` function from `lme4`.

The detailed `R`-code on how to implement these three likelihood-based methods for a binary multilevel *probit*-model, can be found in Appendix C.2. To check the `R`-implementation of AGQ against other software, we use the `NLMIXED` procedure within `SAS`® (SAS Institute Inc, 2015).

3.1.2 Estimation through Bayesian methods

A second strategy that tackles the intractability of the GLMM likelihood function, pursues a Bayesian approach where Markov Chain Monte Carlo (**MCMC**) methods are used to obtain a posterior distribution of the parameters. MCMC methods simulate the likelihood rather than computing it, by calculating the sample average of independently simulated realisations of the integrand. As such, MCMC is thought to provide a more robust approach to marginalising the random effects (Zhao et al., 2006; Browne and Draper, 2006).

In `R`, the `MCMCglmm` function from the package `MCMCglmm` (Hadfield, 2010) is available for such an approach. Technically, latent variables are updated in block by means of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994), while the fixed parameters are Gibbs sampled within such a single block (Garcia-Cortés and Sorensen, 2001).

MCMC methods are known to be computationally intensive and sometimes have a hard time in reaching convergence. To this end, hybrid models based on an Integrated Nested Laplace Approximation (**INLA**) of the posterior marginals for latent Gaussian models (Rue et al., 2009) were proposed. In short, the INLA approach provides fast Bayesian inference by using accurate Laplace approximations for the marginal posterior density of the hyperparameter τ, and for the full conditional posterior marginal densities of the fixed and random effects. The final posterior marginals of the model parameters can then be computed through numerical integration, where the integration points are defaultly obtained by estimating the curvature of the approximation for the marginal posterior of the hyperparameter density (Rue et al., 2009). Not surprisingly, these hybrids have shown a steep decline in the computational burden of MCMC algorithms,

while at the same time converging more easily. In R, such an approach is implemented in the function `inla` from the package `R-inla`.

The detailed R-code of both implementations, as well as their prior specifications, can be found in Appendix C.3. To check the R -based MCMC-implementation against other software, we rely on the he JAGS program (Plummer, 2003) through the use of the R-package `rjags` (Plummer, 2016). It has been suggested by Betancourt and Girolami (2013) that a non-centred parameterisation of the hierarchal model works best when data are sparse, while a centred parameterisation prevails when the data strongly identifies the parameters. However, we observed quite similar results stemming from the two parameterisations in our settings (results not shown).

3.2 Structural Equation Models

Although at first sight GLMM and SEM may seem like two completely different modelling frameworks, it is now well established that SEM can also be relied on to model balanced multilevel data structures. For an excellent overview of SEM, we refer the interested reader to Skrondal and Rabe-Hesketh (2004). In order to account for clustered observations, SEM lets its latent factors represent the random effects from their respective multilevel models (Willett and Sayer, 1994; MacCallum et al., 1997). This results in a 'conventional' SEM which is analytically equivalent to the corresponding multilevel model, under a broad set of conditions (Curran, 2003); we illustrate this for model (4.2).

SEM consists of two modelling parts: a measurement model and a structural part (Skrondal and Rabe-Hesketh, 2004). The former defines unobserved variables in terms of observed variables measured with error, so that the latent variables can be interpreted as the 'true' underlying variables (which might be correlated). The structural model on the other hand, links the different latent variables together. When focusing on random intercept models (read: with only one latent variable) with an explanatory variable in clusters of size two, both modelling-parts can be written as:

$$\begin{aligned} \boldsymbol{y_j} &= \boldsymbol{\nu} + \boldsymbol{\Lambda}\eta_j + \boldsymbol{K}\boldsymbol{x_j} + \boldsymbol{\epsilon_j} \\ \eta_j &= \zeta_j \end{aligned} \tag{4.5}$$

where $\boldsymbol{y_j}$ represents the responses within cluster j, $\boldsymbol{\nu}$= (ν ν) the vector of intercepts, η_j a latent variable with its matrix of factor loadings $\boldsymbol{\Lambda}$ = (1

$1)^T$, $\boldsymbol{x_j} = (x_{1j}\ x_{2j})^T$ represents the explanatory variable, with $\boldsymbol{K}$ its matrix of regression coefficients, and $\boldsymbol{\epsilon_j} = (\epsilon_{1j}\ \epsilon_{2j})^T$ the vector of normally distributed measurement errors. In the structural part of the model, ζ_j represents a random disturbance term $\sim N(0, \tau)$. Note that in accordance to equation (4.2), we assume the effect of x to be fixed within- as well as between clusters. Because of this, $\boldsymbol{K}$ reduces to $\left(\begin{smallmatrix} k & 0 \\ 0 & k \end{smallmatrix}\right)$. Alternatively, we can write the above equations in reduced form, resulting in:

$$\begin{aligned} \boldsymbol{y_j} &= \boldsymbol{\nu} + \boldsymbol{K}\boldsymbol{x_j} + \boldsymbol{\Lambda}\zeta_j + \boldsymbol{\epsilon_j} \\ &= \boldsymbol{\nu} + \boldsymbol{K}\boldsymbol{x_j} + \boldsymbol{\zeta_j} + \boldsymbol{\epsilon_j} \end{aligned} \tag{4.6}$$

where $\boldsymbol{\zeta_j} = (\zeta_j\ \zeta_j)^T$.

Traditionally, estimation methods in SEM are based on the assumption that the observed responses are measured on a continuous scale. In order to reconcile SEM with binary outcomes, the Latent Response Variable approach was introduced, where a dichotomous Y is considered a crude approximation of an underlying continuous variable Y^*. Y^* is not directly observed (hence a *latent* response variable), and is written in terms of a linear predictor. When we separate the two observations within each cluster to eliminate matrix notations, we obtain:

$$\begin{cases} y^*_{1j} & = \nu + kx_{1j} + \zeta_j + \epsilon_{1j} \\ y^*_{2j} & = \nu + kx_{2j} + \zeta_j + \epsilon_{2j} \end{cases} \tag{4.7}$$

where ϵ_{1j} and ϵ_{2j} are i.d.d. residuals of the latent response variables $\sim N(0, \theta)$. Because Y^* exhibits an arbitrary mean and variance, a link between Y and Y^* needs to be established through variance constraints. Since the variance of Y^* conditional on x_{ij} is $\tau + \theta$, there are two possible ways to constrain this variance (Muthén et al., 2002). First, Generalized Linear Models standardly fix the residual variance θ to one. In contrast to this theta parameterisation, identification can also be achieved by standardising the latent variable Y^* itself: the delta parameterisation fixes the sum of τ and θ to one. This parameterisation is traditionally used in the SEM-literature.

The relationship between the binary and latent continuous variable is then: $Y = 1 \iff Y^* > \kappa$. Fixing the threshold κ at 0 (for model identifiability, either the threshold or the intercept in (4.7) needs to be constrained), and assuming that $Y^*_{ij} \sim N(0, 1)$ (i.e. making use of the delta parametrisation so that $\zeta_j \sim N(0, \tau_\delta)$ and $\epsilon_{ij} \sim N(0, 1 - \tau_\delta)$), it follows that:

$$
\begin{aligned}
E[Y_{ij}|x_{ij},\zeta_j] &= P(\nu + kx_{ij} + \zeta_j + \epsilon_{ij} > 0|x_{ij},\zeta_j) \\
&= P(\epsilon_{ij} < \nu + kx_{ij} + \zeta_j|x_{ij},\zeta_j) \\
&= P(\frac{\epsilon_{ij}}{\sqrt{1-\tau_\delta}} < \frac{\nu + kx_{ij} + \zeta_j}{\sqrt{1-\tau_\delta}}|x_{ij},\zeta_j) \\
&= \Phi(\frac{\nu}{\sqrt{1-\tau_\delta}} + \frac{kx_{ij}}{\sqrt{1-\tau_\delta}} + \frac{\zeta_j}{\sqrt{1-\tau_\delta}}) \qquad (4.8)
\end{aligned}
$$

which reduces to the random intercept *probit*-model from equation (4.3), where $\frac{\nu}{\sqrt{1-\tau_\delta}}$, $\frac{k}{\sqrt{1-\tau_\delta}}$ and $\frac{\zeta_j}{\sqrt{1-\tau_\delta}}$ are equivalent to β_0, β_1 and b_j, respectively.

3.2.1 Estimation in SEM

Within the SEM-framework, there are two common estimation approaches for modelling binary outcomes: maximum likelihood (ML) estimation and weighted least squares (WLS) (Skrondal and Rabe-Hesketh, 2004). In contrast to WLS, ML estimation for binary outcomes is not widely available in SEM software. Being a 'full information' method, ML is more regularly employed in item response theory (Forero and Maydeu-Olivares, 2009). In contrast, as WLS-based methods adopt a multiple-step estimation procedure in which only first- and second-order information from the data is used, they are referred to as a 'limited information' approach (see Finney and DiStefano (2013) for a review). In SEM, WLS is employed to differentially weigh the residuals resulting from the observed versus the model-implied sample statistics by their full asymptotic covariance matrix W.

Since WLS requires extremely large samples for accurate estimation of the weight matrix W, more contemporary approaches were developed to improve small sample performance. One such version entails diagonally weighted least squares (DWLS), which utilises a diagonal weight matrix instead of a full one (Muthén, 1993; Muthén et al., 1997) (note that statistical inference in DWLS still relies on the full weight matrix, even when a diagonal matrix is used during estimation). Following Muthén et al. (1997), who have shown DWLS to be statistically and computationally efficient in large samples, more recent studies have proven that DWLS is also more stable than WLS in small samples (Forero and Maydeu-Olivares, 2009; Mîndrila, 2010; Bandalos, 2014). Note that WLS and DWLS estimation is limited to *probit*-regression models and therefore

exclude *logit*-models from our current review study.

`SEM` relying on DWLS can be implemented through the `sem`-function from the package `lavaan` (Rosseel, 2012). To check the `lavaan` package against other implementations, we will verify our results with DWLS estimation in MPLUS® software (Muthén and Muthén, 2010) through the use of the `R`-package `MplusAutomation` (Hallquist and Wiley, 2014).

4 Analysis of the example

We illustrate the above six approaches by applying them to our example. To assess the impact of cluster size, we consider the fit of model (2) when solely looking at the pre- and post-test (i.e., cluster size two) versus all three time points together (i.e., cluster size three). The estimated parameters for the fixed effects (and their standard errors), alongside the estimated random intercept variance for each of the estimation approaches are summarised in table 4.1.

Parameter	β_0		β_1		τ	
Cluster size	2	3	2	3	2	3
Laplace	-0.51 (0.22)	-0.44 (0.21)	0.87 (0.42)	1.09 (0.38)	0.21	0.48
AGQ	-0.54 (0.24)	-0.44 (0.22)	0.92 (0.44)	1.11 (0.38)	0.43	0.65
PQL	-0.51 (0.20)	-0.42 (0.19)	0.88 (0.36)	1.05 (0.31)	0.47	0.62
MCMC	-0.72 (0.34)	-0.52 (0.30)	1.20 (0.50)	1.36 (0.43)	1.95	1.79
Hybrid	-0.56 (0.24)	-0.45 (0.23)	0.95 (0.43)	1.14 (0.37)	0.07	0.45
SEM	-0.52 (0.30)	-0.41 (0.27)	0.75 (0.53)	0.91 (0.47)	0.45	0.83

Table 4.1 The estimates (and (robust) standard errors) from the six approaches for the intercept β_0, the slope parameter β_1 and the random intercept variance τ. Each estimate is displayed twice: once for the pre-and post-test only (cluster size two), and once including all three measures (cluster size three).

We observe that for both cluster sizes all methods perform rather similar in their estimation of β_0, except for a higher estimate produced by MCMC. The estimates for β_1 show more variation, especially within clusters of size two (again with an outlying MCMC-estimate). For the random intercept variance τ, we see that the MCMC estimate is somewhat larger than the others, while the estimates from the Laplace approximation and the hybrid approach are at the lower end of this spectrum. In terms of computing times, most approaches performed equivalently, with the Laplace approximation providing the fastest analysis, closely followed by AGQ, SEM and PQL. The MCMC approach took about ten times as long as the aforementioned approaches, while the hybrid approach only increased the computing time threefold.

Now the question becomes: which of these estimation methods is most reliable here? In order to find out, we conduct an extensive simulation study in the next section.

5 Simulation study

In our simulation study we compare the performance of the six above-described estimation methods in different settings. For this, random binary outcome variables from small clusters are generated under a random intercept *probit*-regression model. More specifically, we assume an underlying latent variable Y_{ij}^*, such that $Y_{ij} = 1$ if $Y_{ij}^* > 0$:

$$\begin{aligned} P(Y_{ij} = 1|x_{ij}, b_j) &= P(Y_{ij}^* > 0|x_{ij}, b_j) \\ &= P(\beta_0 + \beta_1 x_{ij} + b_j + \epsilon_{ij} > 0) \\ &\text{with } b_j \sim N(0, \tau) \text{ and } \epsilon_{ij} \sim N(0, 1) \end{aligned} \tag{4.9}$$

First of all, we consider different cluster sizes: we will look at clusters of size two, three and five. Second, we also consider a different numbers of clusters. Since Loeys et al. (2013) reported that sample sizes in studies using the Actor-Partner- Interdependence-Model (Kenny and Ledermann, 2012) within dyads typically ranged from 30 to 300 pairs, we consider sample sizes n of 25, 50, 100, and 300. Third, we also examine different intracluster correlations (icc) for the latent response variable. As the latent icc_l is defined as the proportion of between-group versus total variance in Y^* ($icc_l = \frac{Var(b_j)}{Var(Y_{ij}^*)} = \frac{\tau}{\tau+1}$), a latent icc_l of 0.10, 0.30 and 0.50 corresponds to a random intercept variance of 0.11, 0.43 and 1.00, respectively. Fourth, we consider rare as well as more abundant outcomes, with an overall event rate of 10% and 50%, respectively. Since the marginal expected value of the outcome $E(Y)$ equals $\Phi(\frac{\beta_0}{\sqrt{1.25+\tau}})$, an outcome prevalence of 50% implies that β_0 must be set to zero. Equivalently, when fixing β_0 to -1.50, -1.66, and -1.92 for a random intercept variance $\tau = 0.11$, 0.43, and 1, respectively, an outcome prevalence of 10% is obtained[2]. In all

[2]Note that the observed icc_o is dependent on the intercept β_0, the random intercept variance τ, and the latent icc_l through the following formula (Vangeneugden et al., 2014): $icc_o = \frac{\Phi_2(\frac{\beta_0}{\sqrt{(\tau+1)(1+2\tau)}}, \frac{\beta_0}{\sqrt{\tau+1}}, icc_l) - \Phi_1(\frac{\beta_0}{\sqrt{\tau+1}})^2}{\Phi_1(\frac{\beta_0}{\sqrt{\tau+1}})(1-\Phi_1(\frac{\beta_0}{\sqrt{\tau+1}}))}$. In this equation, Φ_1 represents the cumulative standard normal distribution, and Φ_2 the cumulative bivariate standard normal distribution with correlation icc_l. Since the outcome prevalence dictates the value of the intercept, each combination of icc_l and $E(Y)$ provides different icc_o's; for rare outcomes, the observed icc_o are 0.06, 0.25 and 0.51, while for $E(Y) = 0.5$ they are

simulations, β_1 is fixed to 1. Finally, four types of covariates are compared: we consider a predictor that only varies between clusters, versus one that varies within clusters; and a Gaussian distributed predictor $\sim N(0, 0.25)$, versus a zero-centred Bernoulli x with success rate 0.5.

In total, 2000 simulations are generated for the $3 \times 4 \times 3 \times 2 \times 4$ combinations of clusters size (3), sample size (4), intracluster correlation (3), outcome prevalence (2) and type of predictor (4). The above-introduced methods are compared over these 288 settings in terms of convergence, relative bias, mean squared error (MSE) and coverage. The relative bias is defined as the averaged difference between the estimated (e.g. $\hat{\beta}$) and true parameter values (e.g. β), divided by the latter (so that the relative bias $= \frac{\beta - \hat{\beta}}{\beta}$); as such, a relative bias enclosing zero will indicate an accurate estimator. A relative bias measure was chosen over an absolute one, as the accuracy of some procedures tends to depend on the magnitude of the parameter values (Zhang et al., 2011). The MSE is estimated by summing the empirical variance and the squared bias of the estimates, simultaneously assessing bias and efficiency: the lower the MSE, the more accurate and precise the estimator. The coverage is defined as the proportion of the 95%-confidence intervals that encompass their true parameter value, where coverage rates nearing 95% represent nominal coverages of the intervals. For the likelihood-based and SEM approaches, Wald confidence intervals are used, while the Bayesian approaches rely on the quantile-based 95% posterior credible intervals. Note that coverage rates for τ are not provided, as not all estimation procedures provide this interval. Lastly, in order to conclude model convergence, several criteria must be met: first, whenever fixed effect estimates exceed an absolute value of ten, or the random effect estimate exceeds 25, the fit is classified as 'no convergence'. We decided on this as parameters in a *probit*-regression exceeding an absolute value of five are extremely unlikely for the given covariate distribution and effect sizes. Secondly, convergence has also failed when a model fit does not yield estimators or standard errors. In addition, for `MCMCglmm` we specified that both chains must reach convergence as assessed by Geweke diagnostics; only when this statistic is smaller than two, convergence is concluded. To ensure a fair comparison between methods, we only present results for simulation runs in which all six methods converged.

0.06, 0.19 and 0.33 (corresponding to latent icc_l's of 0.10, 0.30 and 0.50, respectively). As such, the observed icc_o's range from small to large, according to Hox (2010)'s recommendations.

6 Results

Below, we discuss the results of the simulation study for clusters of size two with a Gaussian predictor in detail.

6.1 Convergence

Generally, convergence improves as the number of clusters and the outcome prevalence increase, and as the icc_l decreases (see figure 4.2). In contrast, convergence is rather unaffected by the level of the predictor, except for PQL which tends to show more convergence difficulties for a within-cluster x. The Laplace approximation also shows a slight decline in convergence for rare outcomes combined with a within-cluster predictor. Note that for 300 clusters most approaches reach 100% convergence, except for MCMC (as in Ten Have and Localio (1999)) and at times the Laplace approximation. For rare outcomes in small samples ($n = 25$), however, the hybrid approach and SEM (see e.g. Forero and Maydeu-Olivares (2009); Rhemtulla et al. (2012)) often perform worse than MCMC. Overall, AGQ shows least difficulty in reaching convergence.

6.2 Relative bias

First, for the fixed effect estimators we typically observe that the relative bias decreases as the number of clusters increases (see figure 4.3). The Laplace approximation and PQL contradict this, however: for rare outcomes the relative bias tends to *increase* with n. Second, we see that an increase in the icc_l tends to shift the relative bias downwards. This implies an improvement in the performance of MCMC (in contrast to Ten Have and Localio (1999)), but not of most other methods (Breslow and Lin, 1995; McCulloch, 1997; Rabe-hesketh and Skrondal, 2012; Hox, 2013). As such, we observe that MCMC performs worse than most methods, but that this difference attenuates as the icc_l increases. Third, the relative bias is generally smaller for a 0.5 outcome prevalence, compared to rare events; this is most clear for the hybrid approach, but is also visible in AQG (see Rabe-Hesketh et al. (2004)). For an outcome prevalence of 0.5, the bias in the β_0-estimators even becomes negligible for all methods. For β_1, however, the MCMC method actually performs worse in small samples when $E(Y) = 0.5$, compared to 0.1. Fourth, different measurement levels of the predictor do not much sway the bias, except for PQL; this method reveals slightly more bias for low event rates when the predictor

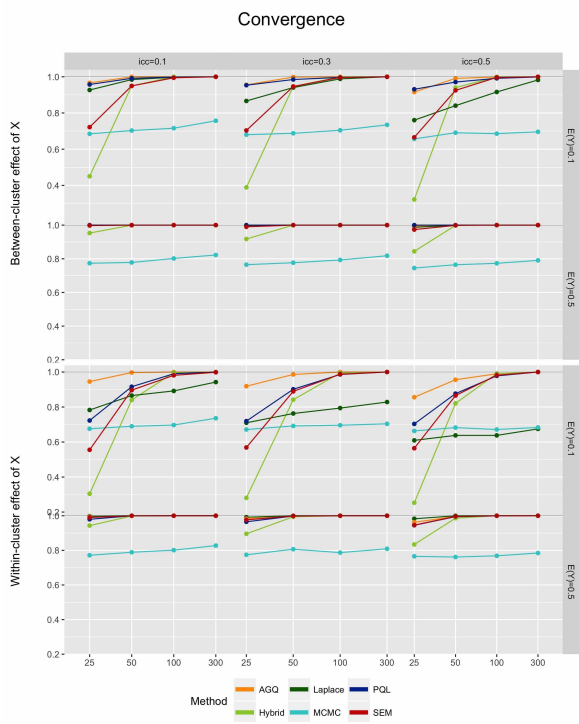

Figure 4.2 Model convergence of the six approaches, for different measurement levels of X (within- or between clusters), outcome prevalence (0.1 and 0.5), icc_l (0.1, 0.3 and 0.5), and sample size (25, 50, 100 and 300).

is measured within- rather than between-clusters. Overall, SEM provides the least biased estimators for the fixed effects, closely followed by AGQ.

For the variance of the random effect, better estimators are typically found in larger samples (see left part of figure 4.4, also see Hox (2013)). Similar to the fixed effect estimators, the Laplace approximation and PQL pose an exception to this rule, by inverting this relation for rare outcomes (see Bauer and Sterba (2011)). As such, the conclusions of Capanu et al. (2013), stating that the hybrid approach outperforms the Laplace approximation by reducing bias in τ, do hold here, but only for large n. We also observe that as the icc_l decreases, bias in the estimates for τ increases in all methods. Finally, a slightly negative bias in the AGQ- and SEM-estimates for τ is observed when the outcome is rare and n small (Raudenbush and Bryk (2002)). This negative bias, however, attenuates as the number of clusters is increased (Bauer and Sterba, 2011). Overall, SEM

yields the least biased estimators for the random intercept variance when the outcome prevalence is rare, while AGQ performs best when $E(Y) = 0.5$.

6.3 MSE

For both β_0 and β_1, the MSE is often higher for rare outcomes, compared to a 0.5 prevalence (see figure 4.5). Additionally, the MSE drops as the sample size grows, and as the icc_l decreases. The Laplace estimator for β_0 again contradicts these trends: for rare events, the MSE *increases* with sample size and icc_l. As before, the measurement level of x does not much alter performance, except in PQL where a within-cluster predictor slightly increases the MSE. For both fixed effects, MCMC often yields the highest MSE when the prevalence equals 0.5, while the hybrid approach regularly performs worst for a prevalence of 0.1. In general, the Laplace approximation yields the lowest MSE when $E(Y) = 0.5$, but performs much worse when the outcome is rare. Overall, AGQ (closely followed by SEM) performs best in terms of MSE.

For the random intercept variance τ, we observe a decrease in MSE as the sample size increases, and as the icc_l decreases (see right part of figure 4.4). The latter conclusion does not hold for MCMC as here the MSE tends to decrease with rising icc_l. Again, PQL performs slightly worse for a within-cluster predictor. In general, the Laplace approximation yields the lowest MSE for 0.5 prevalences, but performs worst when the outcome is rare. Overall, AGQ performs best in terms of MSE, better than SEM, especially in smaller samples.

6.4 Coverage

For both fixed effect estimators, coverage of their 95% confidence intervals is typically better when the outcome prevalence is 0.5 (see figure 4.6). Also, an increasing icc_l usually worsens coverage, except for MCMC (where coverage improves with increasing icc (Ten Have and Localio, 1999)). The impact of the icc_l on coverage has also been observed by Zhang et al. (2011), who found nominal coverages for AGQ and the Laplace approximation for low random intercept variances (i.e. low icc), but more liberal ones as τ increases (i.e. high icc). Generally, SEM and AGQ provide the best coverage rates (Bauer and Sterba, 2011), with SEM taking the upper hand for the coverage of β_0, and AGQ for β_1 with a low to medium icc.

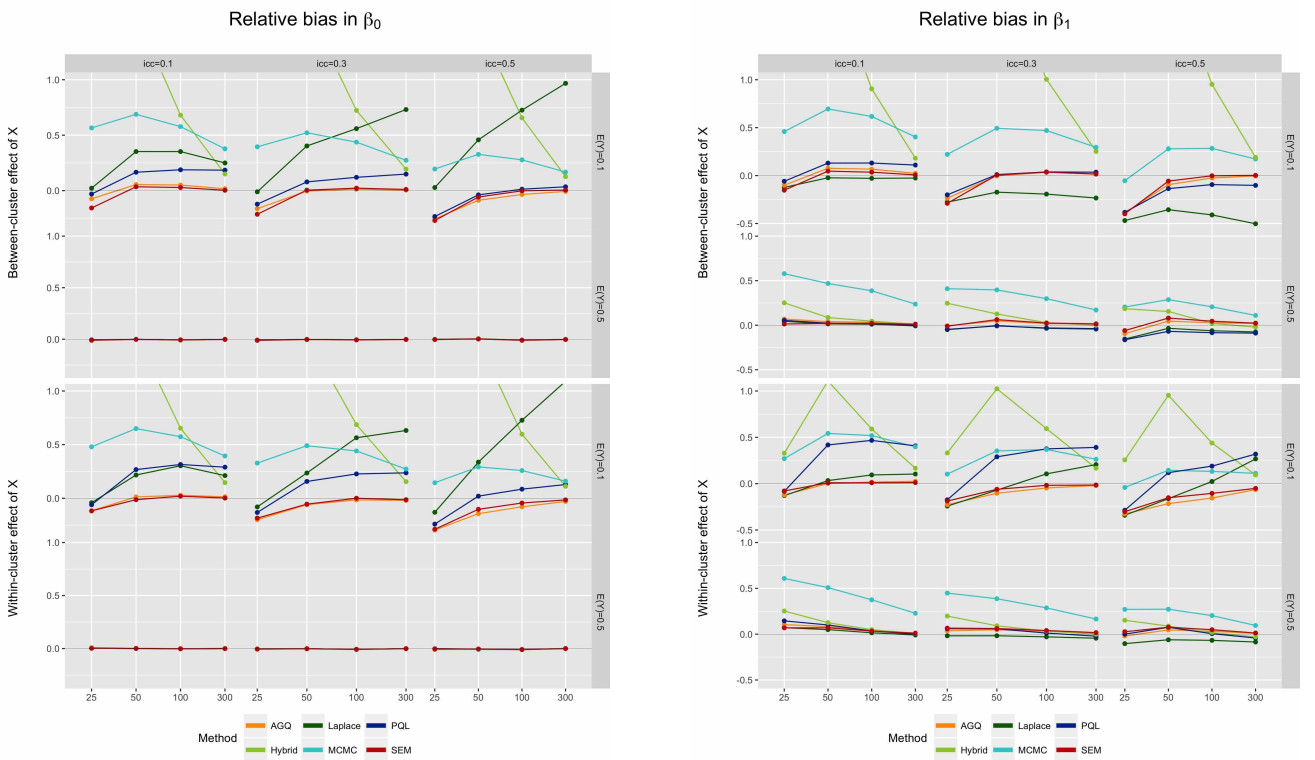

Figure 4.3 Relative bias in β_0 (left) and β_1 (right) for the six approaches, for different measurement levels of X (within- or between clusters), outcome prevalence (0.1 and 0.5), icc_l (0.1, 0.3 and 0.5), and sample size (25, 50, 100 and 300). These results stem from simulation runs where all methods converged.

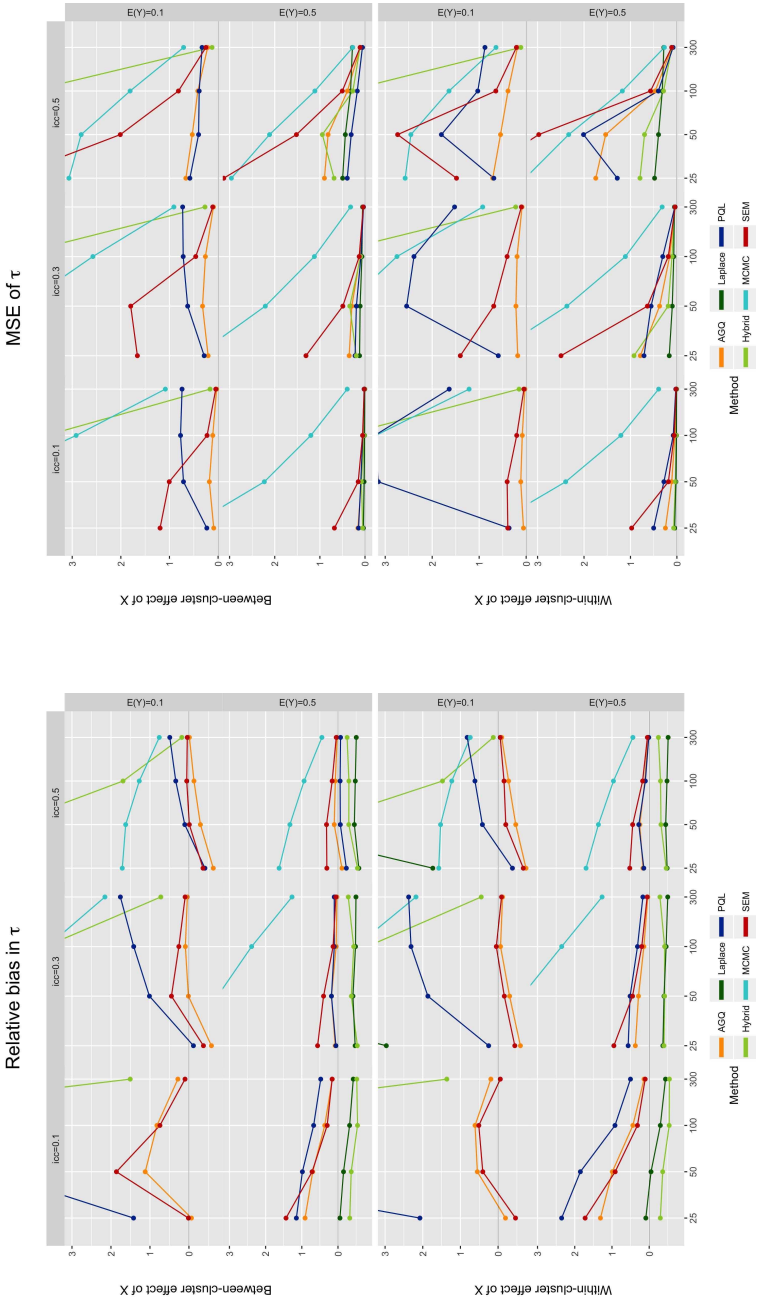

Figure 4.4 Relative bias (left) and MSE (right) of τ for the six approaches, for different measurement levels of X (within- or between clusters), outcome prevalences (0.1 and 0.5), icc_l (0.1, 0.3 and 0.5), and sample sizes (25, 50, 100 and 300). These results stem from simulation runs where all methods converged.

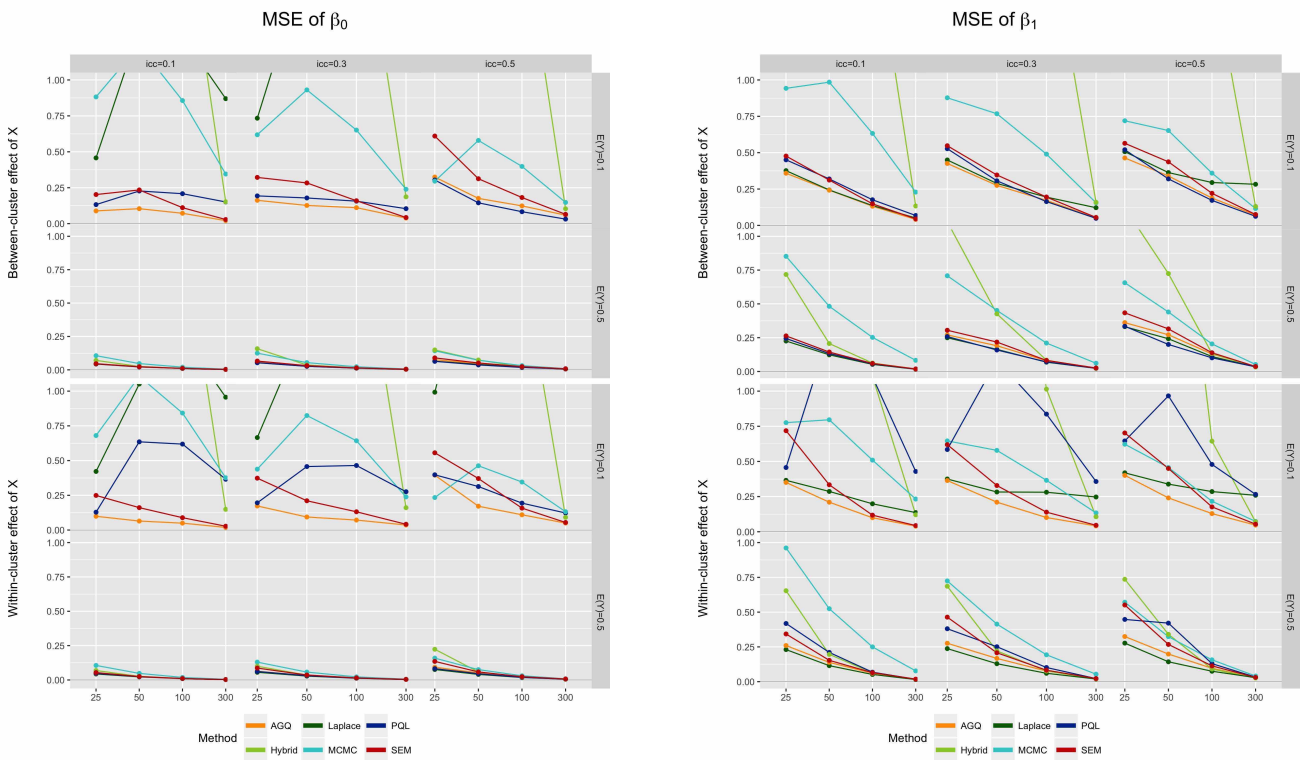

Figure 4.5 MSE of β_0 (left) and β_1 (right) for the six approaches, for different measurement levels of X (within- or between clusters), outcome prevalences (0.1 and 0.5), different icc_l (0.1, 0.3 and 0.5), and sample sizes (25, 50, 100 and 300). These results stem from simulation runs where all methods converged.

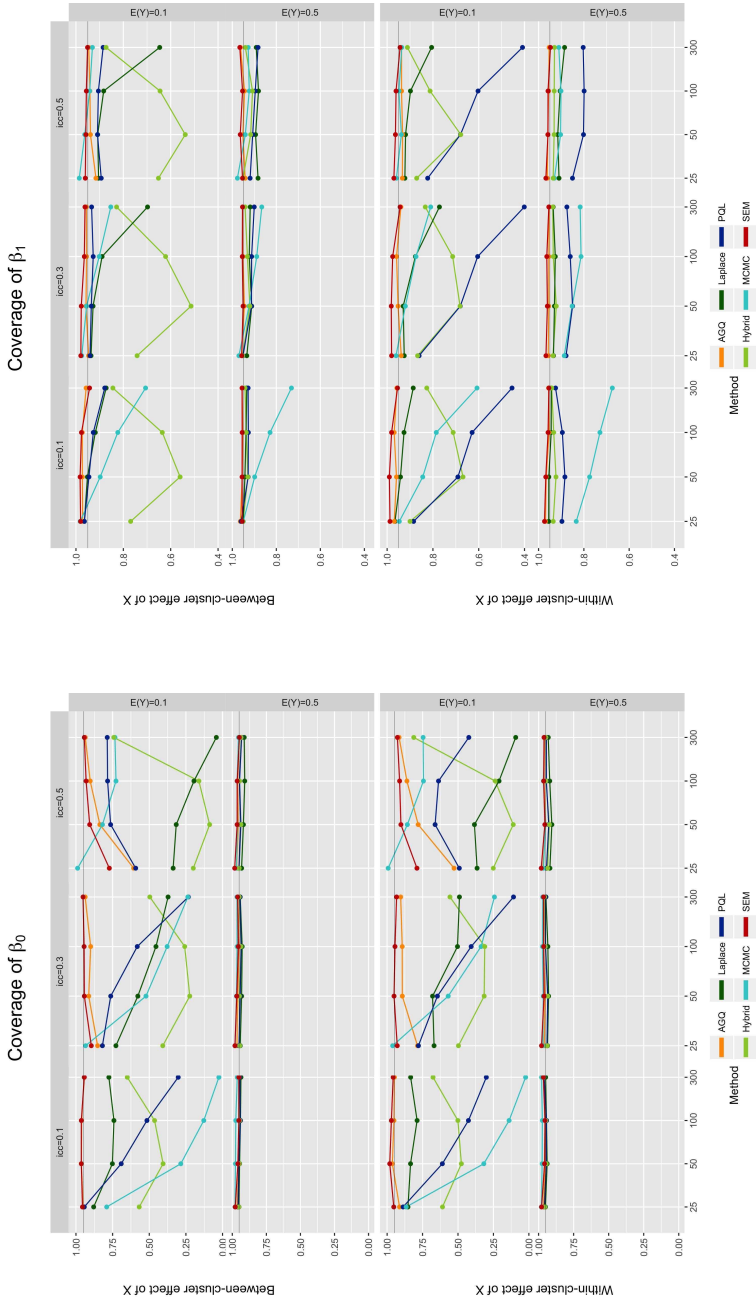

Figure 4.6 Coverage of $\beta 0$ (left) and β_1 (right) for the six approaches, for different measurement levels of X (within- or between clusters), outcome prevalences (0.1 and 0.5), icc_l (0.1, 0.3 and 0.5), and sample sizes (25, 50, 100 and 300). These results stem from simulation runs where all methods converged.

6.5 Summary of the other simulation settings

Until now, we only discussed the results of the simulation study for clusters of size two with a Gaussian predictor. The results for other settings are available online[3] and are briefly discussed in the next paragraphs.

When looking at a binary predictor instead of a Gaussian one, our conclusions remain more or less the same. One exception is that most methods experience a steep decline in convergence for smaller sample sizes, when the predictor is binary compared to continuous. This is most apparent in SEM, where lower convergence rates are due to empty cell combinations of outcome and predictor. In SEM, this produces a warning, which we interpreted as an error (as in `MPLUS`), since such runs yield unreliable results.

As the cluster size increases from two to three or five, we observe a general increase in performance in all methods except SEM. This approach now no longer yields the lowest bias, with AGQ gradually taking over. As such, increasing cluster size favours AGQ in terms of precision, as well as in terms of relative bias.

6.6 MPLUS, JAGS and SAS

`MPLUS` and `lavaan` performed quite similarly throughout our settings, although there were some minor differences (results shown in the online supplementary material). While `MPLUS` version 7 slightly dominates in terms of convergence and coverage, `lavaan` takes the upper hand for the relative bias and the MSE. These differences are trivial, however, and most likely due to `lavaan` incorporating a slightly higher number of iterations in reaching convergence.

When comparing `JAGS` to `MCMCglmm`, we observe some important differences in performance; for most settings, `JAGS` 4.1.0. outperforms `MCMCglmm`, except when a small n is combined with a medium to large icc_l (see supplementary material). Note that although `JAGS` performs slightly better in most settings, its computing times are also significantly higher.

In contrast to Zhang et al. (2011), who found a superior performance of `SAS NLMIXED` compared to `R`'s `glmer`-function, we found that `glmer` performed equally well or even slightly better in terms of convergence rates, relative bias, and coverage (using SAS version 9.4). When the outcome

[3] 'Image 1.pdf' and 'Image 2.pdf' at
https://www.frontiersin.org/articles/10.3389/fams.2016.00018/full

prevalence is 0.5 and for some rare events settings, `glmer` also provided a slightly lower MSE.

7 Discussion

In this paper, we provided an overview of several R-packages based on different estimation techniques, as to fit random-intercept *probit* regression models. More specifically, we focused on techniques capable of modelling binary outcomes in small clusters. Additionally, we presented an extensive simulation study in which we assessed the impact of various data features on a number of performance criteria. In summary, we found that some of our results confirmed findings from previous studies, while others have (to the best of our knowledge) not been observed before:

> Interestingly, both **SEM** and **AGQ** performed considerably well for paired data. Though both approaches disclosed some sensitivity to sample size, they manifested remarkable robustness when varying the *icc*, the event rate, and the measurement level of the predictor. As such, these methods can be considered the most stable over all settings in terms of relative bias, for the fixed effect regression coefficients as well as the random intercept variance. While AGQ performs slightly better than SEM in terms of convergence and MSE, SEM performs slightly better when considering the relative bias. As SEM relies on robust standard errors, it yields higher MSE's, but also provides robustness against model misspecification (which was not investigated here). For the coverage, we observed that SEM performs slightly better for β_0, while AGQ tentatively gains the upper hand for β_1. As the cluster size increases, however, AGQ takes over and becomes most reliable in terms of bias and precision.
>
> Since the **Laplace** approximation is known to be precise only for normally distributed data or for non-normal data in large clusters (Tuerlinckx et al., 2006), we observed an expected poor performance of this approximation in our settings (Broström and Holmberg, 2011). **PQL** also exhibits an inferior performance for low *icc*'s and a low outcome prevalence, while additionally revealing disconcerting performance issues for a within-cluster measured predictor. Finally, the two **Bayesian** approaches performed below par in terms of most criteria considered.

Let us once again consider our motivating example with a within-cluster

measured predictor, a sample size of 37, an outcome prevalence of 0.4, and a medium to large latent *icc*. When we apply our conclusions to these settings, we can state that SEM will yield the most trustworthy estimates when the cluster size is two, while AGQ will take over as a measurement is added. MCMC will yield the most biased estimates in both cases (as can be clearly seen in table 4.1).

Several limitations can still be ascribed to this paper. First, we restricted our comparisons to estimation techniques available in `R`-packages. As such, several improvements regarding the estimation methods discussed, could not be explored. For example, while the `glmmPQL` function employed in this paper is based on Breslow and Clayton (1993)'s PQL version, a second-order Taylor expansion (Goldstein and Rasbash, 1996) might provide a more precise linear approximation (this is referred to as PQL-2, in contrast to the first order version PQL-1). Be that as it may, not all the evidence speaks in favour of PQL-2: even though it yields less bias than PQL-1 when analysing binary outcomes, Rodriguez (2001) found that the estimates for both fixed and random effects were still attenuated for PQL-2. Furthermore, PQL-2 was found to be less efficient and somewhat less likely to converge (Rodriguez, 2001). Second, certain choices were made with respect to several estimation techniques, such as the number of quadrature points used in the AGQ-procedure. However, acting upon the recommendation of 8 nodes for each random effect (Rabe-hesketh and Pickles, 2002), we argue that surpassing the ten quadrature points considered, would carry but little impact in our random intercept model. Also, the repercussions of our choices on prior specification in the Bayesian framework deserves a more thorough examination, as different priors may lead to somewhat different findings. Third, the performance results presented here may not be intrinsic to their respective estimation techniques, but instead due to decisions made during implementation. As we demonstrated for `MCMCglmm` when comparing it to `JAGS`, its disappointing performance is most likely due to a suboptimal implementation, and not an inherent treat of the MCMC estimation procedure. Fourth, some scholars (Skrondal, 2000) have recommended the evaluation of different estimation methods and their dependence on different data features, by applying ANOVA-models rather than graphical summaries. Treating the different settings (sample size, *icc*, level of the predictor, the event rate, and their two-way interactions) as factors, did not provide us much insight since almost all variables (as well as their interactions) were found to be highly significant. Fifth, in our simulation study we only considered complete data; in the

presence of missing data, however, DWLS estimation in SEM will exclude clusters with one or more missing outcomes, resulting in a complete case analysis. This exclusion stands in contrast to maximum likelihood and Bayesian approaches from the GLMM-framework, as they consider all available outcomes when there is missingness present. Consequently, the GLMM-framework will not introduce any (additional) bias under the missing at random assumption, while DWLS-estimation requires the more stringent assumption of data missing completely at random. Sixth, we do not focus on measurement imprecision in this study and assume that all observed variables are measured without error. Of course, as Westfall and Yarkoni (2016) recently pointed out, this rather optimistic view might pose inferential invalidity when this assumption fails. In light of this, it is important to note that SEM can deal with such measurement error, in contrast to GLMM-based approaches.

With the results, as well as the limitations of the current paper in mind, some potential angles for future research might be worth considering. As we explicitly focused on conditional models, we deliberately excluded marginal approaches such as Generalised Estimating Equations (GEE), because such a comparison is impeded by the fact that marginal and conditional effects differ for binary outcomes. Whereas multilevel models allow for the separation of variability at different levels by modelling the cluster-specific expectation in terms of the explanatory variables, GEE only focuses on the respective marginal expectations. Previous research (Loeys and Molenberghs, 2013) has revealed excellent small sample performance of GEE in terms of bias, when analysing binary data in clusters of size two. Also, it might we worth considering a pairwise maximum likelihood (PML) approach, as PML estimators have the desired properties of being normally distributed, asymptotically unbiased and consistent (Varin et al., 2011). This estimation method breaks up the likelihood into little pieces and consequently maximises a composite likelihood of weighted events. PML in `R` is currently unable to cope with predictors, but this will most likely be possible in the near future. And finally, as pointed out by one of the reviewers, Hamiltonian Monte Carlo (used in Stan software, Carpenter et al. (2016)) may be a more efficient sampler compared to a Metropolis-Hastings (i.e. `MCMCglmm`) or a Gibbs sampler (i.e. `JAGS`). To this end, exploring the performance of the Stan software might prove worthwhile when further focusing on Bayesian analysis.

Bibliography

Airy, G. B. (1861). *On the algebraical and numerical theory of errors of observations and the combination of observations*. Macmillan, London.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1):102–116.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2):135–167.

Bauer, D. J. and Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, 16(4):373–390.

Betancourt, M. and Girolami, M. (2013). Hamiltonian Monte Carlo for hierarchical models. *ArXiv e-prints*.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3):261–280.

Breslow, N. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Broström, G. and Holmberg, H. (2011). Generalized linear models with clustered data: Fixed and random effects models. *Computational Statistics & Data Analysis*, 55(12):3123–3134.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Capanu, M., Gönen, M., and Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine*, 32(26):4550–4566.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, (in press).

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38(4):529–569.

Ferkingstad, E. and Rue, H. (2015). Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *ArXiv e-prints*.

Finney, S. J. and DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In Hancock & Mueller, editor, *Structural equation modeling: a second course*, chapter 11, pages 439–492. Information Age Publishing, Charlotte, NC, second edition.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2009). *Longitudinal data analysis*. Chapman & Hall/CRC, Boston, MA.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Forero, C. G. and Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14(3):275–299.

Garcia-Cortés, L. A. and Sorensen, D. (2001). Alternative implementations of Monte Carlo EM algorithms for likelihood inferences. *Genetics Selection Evolution*, 33:443–452.

Glynn, R. J. and Rosner, B. (2013). Regression methods when the eye is the unit of analysis. *Ophthalmic Epidemiology*, 19(3):159–165.

Goldstein, H. (1979). *The design and analysis of longutudinal studies*. Academic Press, London.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A*, 159(3):505–513.

Hadfield, J. D. (2010). MCMC methods for multi-respoinse generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.

Hallquist, M. and Wiley, J. (2014). MplusAutomation: Automating Mplus Model Estimation and Interpretation.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge, New York, NY, second edition.

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In Little, T. D., editor, *The Oxford Handbook of Quantitative Methods. Vol. 2.*, chapter 14, pages 281–294. Oxford University Press, Oxford.

Kenny, D. A. and Ledermann, T. (2012). Bibliography of Actor-Partner interdependence model.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., and Vansteelandt, S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivariate Behavioral Research*, 48(6):871–894.

Loeys, T. and Molenberghs, G. (2013). Modeling actor and partner effects in dyadic data when outcomes are categorical. *Psychological Methods*, 18(2):220–236.

MacCallum, R. C., Cheongtan, K., and Malarkey, W. B. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, 32(3):215–253.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):162–170.

McMahon, J. M., Tortu, S., Torres, L., Pouget, E. R., and Hamid, R. (2003). Recruitment of heterosexual couples in public health research: a study protocol. *BMC medical research methodology*, 3:24.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Mîndrila, D. (2010). Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: a comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, 1(1):60–66.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In Bollen, K. and Long, J., editors, *Testing Structural Equation Models*, chapter 9, pages 205–243. Sage, Newbury Park, CA.

Muthén, B., du Toit, S. H. C., and Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Conditionallly accepted for publication in Psychometrika.*

Muthén, B., Muthén, B., and Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus Web Notes*, 4(5):0–22.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus User's Guide.* Muthén & Muthén, Los Angeles, CA, sixth edition.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a method for the efficients computation of posterior distributions. *Applied Statistics*, 31(3):214–225.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135(3):370–384.

Ortqvist, A. K., Lundholm, C., Carlström, E., Lichtenstein, P., Cnattingius, S., and Almqvist, C. (2009). Familial factors do not confound the association between birth weight and childhood asthma. *Pediatrics*, 124(4):e737–43.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Plummer, M. (2003). JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik, K., Leisch, F., and Zeileis, A.,

editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20-22*, pages 1–10, Vienna, Austria.

Plummer, M. (2016). rjags: Bayesian Graphical Models using MCMC.

R Core Team (2013). R: A language and environment for statistical computing.

Rabe-hesketh, S. and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

Rabe-hesketh, S. and Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata - Volume II: Categorical responses, counts and survival*. Stata Press, Texas, TX, third edition.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, 69(2):167–190.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and data analysis methods*. Sage, Thousand Oaks, CA, second edition.

Rhemtulla, M., Brosseau-Liard, P. É., and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3):354–373.

Rodriguez, G. (2001). Improved estimation procedures for multilevel models with binary response: a case-study. *Journal of the Royal Statistical Society: Series A*, 164(2):339–355.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A*, 158(1):73–89.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rovine, M. J. and Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, 35(1):55–88.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392.

SAS Institute Inc (2015). Base SAS® 9.4 Procedures Guide, Fifth Edition.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Scheffé, H. (1959). *The analysis of variance*. John Wiley & Sons, New York, NY.

Senn, S. (2002). *Cross-over trials in clinical research*. John Wiley & Sons, Chichester.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, 35(2):137–167.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, New York.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage, Thousand Oaks, CA.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971.

Sutradhar, B. C. and Mukerjee, R. (2005). On likelihood inference in binary mixed model with an application to COPD data. *Computational Statistics & Data Analysis*, 48(2):345–361.

Ten Have, T. R. and Localio, A. R. (1999). Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics*, 55:1022–1029.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.

Tierney, L. and Kadane, J. B. (1986). Acccurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *The British Journal of Mathematical and Statistical Psychology*, 59(2):225–255.

Vandeweghe, L., Verbeken, S., Braet, C., Loeys, T., De Henauw, S., and Moens, E. (2018). Strategies to increase preschoolers' vegetable liking and consumption: The role of reward sensitivity. *Food Quality and Preference*, 66:153–159.

Vangeneugden, T., Molenberghs, G., Verbeke, G., and Demétrio, C. G. (2014). Marginal correlation from logit- and probit-beta-normal models for hierarchical binary data. *Communications in Statistics - Theory and Methods*, 43(19):4164–4178.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Venables, W. N. and Ripley, B. D. (2002). Modern applied statistics with S.

Westfall, J. and Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11(3):1–22.

Willett, J. B. and Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, 116(2):363–381.

Xu, Y., Lee, C. F., and Cheung, Y. B. (2014). Analyzing binary outcome data with small clusters: A simulation study. *Communications in Statistics - Simulation and Computation*, 43(7):1771–1782.

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L., and Tu, X. M. (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, 30:2562–2572.

Zhang, Z., Zyphur, M. J., and Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models. *Organizational Research Methods*, 12(4):695–719.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science: a review journal*, 21(1):35–51.

C Appendix

This appendix contains the R-code for the data generating mechanism and the implementation of the six methods discussed in this paper. Note that in these scripts, `y` represents the binary outcome, `x` the predictor and `id` the cluster identifier.

C.1 Data generating mechanism

The following R-code allows the generation of data with clusters of size two, a sample size 'n', a latent intracluster correlation 'icc', and an outcome prevalence 'prev'. In this script, the normally distributed exposure is generated to vary within clusters.

```
#Generate 2000 data sets for the current n, icc and prev:
for (i in 1:2000){
  print(i)
  set.seed(123456+i)
  #Population parameters:
  tau<-icc/(1-icc)
  beta0<-qnorm(prev)*sqrt(1.25+tau)
  beta1<-1

  #Cluster identifier:
  ind<-seq(1,n)
  #Random intercept for each cluster:
  ri<-rnorm(n,0,sqrt(tau))

  #Two normally distributed within-cluster exposures:
  x0<-rnorm(n,0,0.5)
  x1<-rnorm(n,0,0.5)
  #Two binary outcomes y0 and y1:
  y0<-rbinom(n,1,pnorm(beta0+beta1*x0+ri))
  y1<-rbinom(n,1,pnorm(beta0+beta1*x1+ri))

  #Convert variables to long format for GLMM-analyses:
  y<-c(y0,y1)
  x<-c(x0,x1)
  id<-rep(ind,2)}
```

C.2 Likelihood-based methods

The **Laplace** approximation can be applied by relying on the function `glmer`, from the package `lme4` (Bates et al., 2015). The `R`-syntax that corresponds to model (2) is:

```
glmer(y∼x+(1|id),family=binomial(link="probit"))
```

PQL can be implemented using the function `glmmPQL` from the `R`-package `MASS` (Venables and Ripley, 2002), with the following syntax for model (2):

```
glmmPQL(y∼x,random=∼ 1|id,family = binomial(link=
        "probit"))
```

AGQ estimation in `R` is also possible within the `glmer` function from `lme4`, by additionally specifying the number of quadrature nodes (note that when the number of nodes is set to 1 (the default option), AGQ reduces to the Laplace approximation). As more quadrature points usually improve estimation but also increase computational time (Pinheiro and Bates, 1995; Fitzmaurice et al., 2009; Bauer and Sterba, 2011) and we only consider one random effect, we fix this number to ten (Bock et al., 1988):

```
glmer(y∼x+(1|id),family=binomial(link="probit"),nAGQ=10)
```

C.3 Bayesian methods

MCMC-modelling can be achieved by the `MCMCglmm` function from the package `MCMCglmm` (Hadfield, 2010). We consider two chains to assess convergence, each with a burnin period of 3000, a thinning interval of 10, and 1000 random posterior draws (providing us with 2000 posterior estimates in total, of which we reported the posterior median). In order to fit the random-intercept *probit*-regression model in equation (2), the option '`family = "ordinal"`' needs to be specified, resulting in:

```
MCMCglmm(y∼x, random = id,family = "ordinal",verbose=
FALSE, prior=priors,nitt=13000,thin=10,burnin=3000)
```

In a Bayesian framework priors need to be specified. Since *probit*-regression coefficients larger than five are unlikely given our data, the specification of more informative priors will prevent estimates from becoming excessively large (Bauer and Sterba, 2011). With this range in mind, we define a multivariate normal prior for the fixed effects `B` with mean zero and a covariance matrix `V` with 5^2 on the diagonal and zero

elsewhere. An inverse gamma distribution `G` for the random effects is typically parameterised in terms of a shape parameter α and a scale parameter β, but `MCMCglmm` relies on $\nu = 2\alpha$ and $V = \frac{\beta}{\alpha}$. As such, expressions of the mean ($\frac{\beta}{\alpha-1}$) and variance ($\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$) yield the following derivations for ν and V: $\nu = 2\frac{\mu_\tau^2}{\sigma_\tau^2} + 4$ and $V = 2\frac{\mu_\tau}{\nu}(\frac{\nu}{2} - 1)$. Consequently, values for ν and V can be deduced by simulating an appropriate distribution for τ and by extracting its mean and variance. Since we likewise argue that values exceeding five are equally unlikely for the standard deviation of the random intercept (and as such, values exceeding 25 for τ), we simulated a uniform distribution for $\sqrt{\tau} \sim U(0, 5)$. Last of all, the residual variance `R` is fixed at one. The resulting `R`-code is:

```
priors=list(B=list(mu=c(0,0),V=diag(2)*5**2),
R=list(V=1,fix=1), G=list(G1=list(V=V, n=nu)))
```

Integrated Nested Laplace Approximation, can be implemented by the function `inla` from the package `R-inla` (Rue et al., 2009). In this package a prior for the logarithm of the random intercept precision $log(\frac{1}{\tau})$, which follows a logGamma distribution, needs to be defined. We follow the same reasoning as before: a uniform distribution between zero and five is defined for the standard deviation, which is subsequently squared and inverted to achieve a distribution for the precision. The mean ($\mu_{1/\tau}$) and variance ($\sigma^2_{1/\tau}$) of this distribution help define the shape parameter $\alpha = \frac{\mu^2_{1/\tau}}{\sigma^2_{1/\tau}}$, and the scale parameter $\beta = \frac{\mu_{1/\tau}}{\sigma^2_{1/\tau}}$. Since Fong et al. (2010) showed that binary data prove particularly problematic for this package, the authors of `R-inla` have suggested an improvement by constructing better approximations to the posterior marginals, without any additional computational costs (Ferkingstad and Rue, 2015). We too have included this improvement in our analysis, through the following R-syntax to fit model (2):

```
inla(y x+f(id,model="iid",param=c(a,b)),family="binomial",
control.family=list(link = "probit"),Ntrials=1,
control.inla=list(correct=TRUE,correct.factor = 10))
```

C.4 SEM methods

SEM can be applied to the data by use of the function `sem` from the package `lavaan` (Rosseel, 2012). This `R`-function allows both the theta- and delta-parametrisation (see section 3.2) but since these are practically

equivalent, we only focussed on the latter. As the delta-parameterisation and the DWLS estimator with robust standard errors are executed by default, we do not need to specify any additional options for this function. Note that the data is now in wide format, with the following `model`-specification for a within-cluster predictor in clusters of size two:

```
Data <- data.frame(y0 = y0, y1 = y1, x1 = x1, x0 = x0)
model <- '        int=~1*y0+1*y1
                  y0~a1*x0;  y0|a0*t1;  y0 ~~ v1*y0
                  y1~a1*x1;  y1|a0*t1;  y1 ~~ v1*y1 '
fit <- sem(model,ordered=c("y0","y1"),data=Data)
summary(fit)
```

In this code, *int* represents the random intercept, influencing both outcomes from each cluster ($y0$ and $y1$). $y0$ and $y1$ are regressed on $x0$ and $x1$, respectively, with a common intercept $(-1) \cdot a0$ (with $a0$ the threshold value), and regression parameter $a1$. We also define identical residual variances $v1$ for both outcomes. For a between-cluster predictor, $x0$ and $x1$ are substituted by one x value for both measurement moments. Note that the estimates obtained by this parameterisation are on a different scale than the parameters from model (2); in order to adjust them, they need to be divided by the square root of the residual variance $v1$ (see section 3.2).

5 Lower-level mediation with a binary outcome

Abstract. In recent literature, researchers have put a lot of time and effort in expanding mediation to multilevel settings. Unfortunately, such extensions are often limited to a continuous outcome, whereas research concerning multilevel mediation within binary settings remains rather sparse. Additionally, in lower-level mediation, the effect of the lower-level mediator on the outcome may oftentimes be confounded by an (un)measured upper-level variable. When such confounding is left unaddressed, the effect of the mediator, as well as the causal mediation effects, will be estimated with bias. In linear settings, bias due to unmeasured additive upper-level confounding is often remedied by separating the effect of the mediator into a within- and between-cluster component. However, this solution is no longer valid when considering binary outcome measures. To assess the severity of this transgression, we aim to tackle lower-level mediation with a binary outcome and a binary randomised exposure from a counterfactual point of view, with a special focus on small clusters. We do this by 1) providing non-parametrical identification assumptions of the direct and indirect effect, 2) parametrically identifying these effects based on appropriate modelling equations, 3) considering estimation models for the mediator and the outcome, and 4) estimating the causal effects through an imputation algorithm that samples counterfactuals. Since steps three and four can be completed in various ways, we compare the performance of three different estimation models (an uncentered and centred separate modelling method, and a joint approach), and two different ways of predicting random effects (marginally versus conditionally). Employing simulations, we observe that the joint modelling approach combined with a marginal generation of the random effects performs best.

This chapter is joint work with Tom Loeys and Sara Kindt.

1 Introduction

We must acknowledge that clustered or multilevel data have become protagonists in numerous research fields, either through the application of family- or twin studies, during multicenter research, or in longitudinal designs. In this type of studies, we always encounter a specific kind of hierarchy within our data where usually, two levels can be distinguished: lower-level measurements are nested within clusters or upper-level units. Examples of such hierarchically nested entities, consist of relatives nested within a family, students within classrooms, or measurement moments within an individual. These lower-level measures show dependencies amongst each other, as measures arising from within a family, a classroom, or an individual, will be more alike than data arising from two random units. Such correlated data structures need special care, as analyses that either ignore these dependencies or inappropriately aggregate the data across levels, will often lead to invalid inferences (Snijders and Bosker, 1999; Raudenbush and Bryk, 2002). Over the course of decades, two major frameworks have been put forward that are able to deal with such correlations: Mixed-effect Models (MM) and Structural Equation Models (SEM). Although SEM holds several advantages over its MM counterpart, both frameworks turn out to be entirely equivalent when considering balanced multilevel data within a random intercept model (Rovine and Molenaar, 2000; Curran, 2003; Bauer, 2003).

Taking the extreme usefulness of multilevel designs into account, expanding mediation to multilevel settings has become an increasingly popular topic (Bauer et al., 2006; VanderWeele and Vansteelandt, 2009; Zhang et al., 2009; Preacher et al., 2010; Preacher, 2015; Tofighi and Kelley, 2016). When looking at the effect of a randomised binary exposure that varies within clusters, researchers usually consider a design where the mediator and the outcome are also measured at the lower-level. This type of mediation is appropriately termed lower-level or 1-1-1 mediation (i.e., the exposure, mediator, and outcome are all measured at level-1). Despite becoming a quite established subject, the lower-level mediation literature has almost exclusively relied on extending the product-of-coefficients approach to multilevel settings (Judd et al., 2001; Kenny et al., 2003; Bauer et al., 2006; Preacher et al., 2010). Unfortunately, this procedure does not offer a general definition of the causal effects that is applicable beyond the few (linear) statistical models considered. Also, these extensions to multilevel settings have mostly been executed without due attention to

the interpretation of the effects as causal parameters, nor to the underlying assumptions needed to identify these. Some researchers have tried to surmount these shortcomings by tackling multilevel mediation from a counterfactual perspective (Imai et al., 2010a; VanderWeele, 2010b,a; Josephy et al., 2015). This has proven very fruitful, as this framework is able to explicate the assumptions underlying multilevel mediation, put forward a general non-parametrical definition of the causal effects, as well as identify these effects based on appropriate parametrical models (Pearl, 2001; VanderWeele and Vansteelandt, 2009; Imai et al., 2010a; VanderWeele, 2010b; Pearl, 2012).

1.1 Estimation of the causal mediation effects in four steps

When resorting to the counterfactual framework, four steps need to be considered if we want to unbiasedly estimate the causal mediation effects.

1.1.1 A first step - Nonparametric definition & identification of the causal effects

First, we define non-parametrical expressions for the direct and indirect effect. For a continuous outcome, this is usually achieved on a linear scale (VanderWeele, 2010b; Josephy et al., 2015), while a linear-, risk ratio- (RR), and odds ratio (OR)-type definition have been used for categorical outcomes (Imai et al., 2010b; VanderWeele, 2013; Loeys et al., 2013; Bind et al., 2016). We would like to focus on deriving these expressions on a linear scale, as to provide a counterfactual definition of the causal effects in terms of differences.

As a second part of the first step, we also recite the assumptions needed to identify the above-mentioned effects. One very important assumption in lower-level mediation studies entails the absence of unmeasured upper-level confounders of the mediator-outcome relationship, alternatively referred to in econometrics as upper-level endogeneity of the mediator and the outcome (Wooldridge, 2010). This type of confounding is very common in many contexts, and may lead to serious bias in the estimation of the intervening effect if not appropriately dealt with. However, although the absence of such endogeneity is often claimed as a necessary prerequisite for unbiased estimation, specific conditions allow researchers to relax this assumption (e.g., when the confounder has a linear and additive effect on mediator and outcome). As this assumption portrays such an important, complex,

and recurrent issue in lower-level mediation settings, this manuscript intends to emphasise the consequences of upper-level endogeneity of the mediator-outcome relation.

1.1.2 A second step - Parametric identification of the causal effects

Next, we identify parametrical expressions for the causal mediation effects based on modelling equations that satisfy the assumptions explicated in the previous step. Traditionally, most such attempts were made with a continuous scaled mediator and outcome in mind (VanderWeele, 2010b; Josephy et al., 2015). Social and behavioural sciences, however, have developed a natural interest in hierarchical models for dichotomous data, and hence, the corresponding mediation analyses that may ensue. In practice, binary variables often arise through the occurrence of discrete events (e.g., disease vs. no disease, pass vs. fail, ...), or through an artificial classification of a continuous variable based (high vs. low rumination, small vs. large families, ...). Because research on multilevel mediation with a binary outcome is relatively sparse, we will focus on this setting in particular.

When considering modelling equations for the outcome, researchers have mostly focussed on the *logit*-link for binary multilevel models (Robins et al., 2000; Neuhaus and McCulloch, 2006; Bind et al., 2016). In practice, however, dichotomous outcomes are also often predicted through the use of *probit*-regression models. As such, we will consider both *logit*- and- *probit* link-functions when deriving and evaluating parametrical expressions for the direct and indirect effect.

1.1.3 A third step - Estimation models for the mediator and outcome

In a third step, we require unbiased and efficient estimation of the regression coefficients of the mediator and outcome models. This unbiasedness of course depends upon the assumptions mentioned during the first step of this process; if, for example, the assumption of 'no upper-level endogeneity of the mediator-outcome relation' is not met, a traditional multilevel model for the outcome (with the mediator as a predictor) will estimate its regression coefficients with bias (Zhang et al., 2009; Josephy et al., 2015). In two-level linear settings, Centering Within-Clusters, or CWC (i.e., the subtraction of a cluster-specific mean), was proposed to solve potential confounding issues when estimating the effect of a predictor on an outcome (Neuhaus and Kalbfleisch, 1998). Unfortunately, when the

outcome is binary, CWC will no longer yield proper parameter estimates, although in practice the resulting bias may often be small (Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010).

Alternatively, the mediator and outcome can also be modelled jointly under a slightly more stringent set of conditions (Bauer et al., 2006; Josephy et al., 2015). Such a joint modelling approach allows for unmeasured cluster-specific common causes of the mediator and the outcome, by estimating a covariance term between the two random intercepts (Bauer et al., 2006; Skrondal and Rabe-Hesketh, 2014). Bind et al. (2016) incorporate such a joint strategy (even though their applications are limited to linear settings), hereby allowing for upper-level confounding (Skrondal and Rabe-Hesketh, 2014). As CWC no longer provides unbiased parameter estimates in binary settings, we too, aim to focus on joint modelling in order to confront and solve upper-level endogeneity of the mediator-outcome relation.

1.1.4 A fourth step - Estimation of the causal effects through Monte Carlo potential outcome generation

Finally, a fourth step aims to estimate the causal mediation effects themselves. Typically, the expressions derived during the second step are conditional on the cluster-specific random effects. In linear settings, such effects are effectively eliminated from the parametrical expressions of the causal effects (when these are defined on a difference-scale), but unfortunately, this is no longer the case when the outcome is binary. If we want to obtain expressions for the indirect and direct effect marginalised over the random effects, we need to sample the random effects from their assumed distribution and average them out. From a counterfactual point of view, this can be achieved by an imputation algorithm that sequentially 1) simulates 'potential values' for the mediator, conditional on the random effects, 2) simulates 'potential values' for the outcome, given the sampled values of the mediator and conditional on the random effects, 3) computes the causal mediation effects for each simulated draw, and 4) calculates the summary statistics of these effects, over the draws (Imai et al., 2010a). This sampling algorithm may base itself upon the empirical Bayes predictor for the random effects (Skrondal and Rabe-Hesketh, 2004), which can rely on one of two possible mechanisms during the sampling process of the mediator and outcome models. A first possibility draws the random effect from a marginal zero-centred distribution (i.e., the marginal sampling (co)variances), while a second relies on a distribution that is conditional on the cluster identifier (i.e., the posterior (co)variances) (Tingley et al.,

2014). It has been stated that the latter might underestimate the variance of the random effects distribution (Skrondal and Rabe-Hesketh, 2004), which is why we aim to quantify and compare the performance of both methods.

In their manuscript, Bind et al. (2016) parametrically identify the direct and indirect effect in binary settings and subsequently rely on these expressions to estimate the causal effects. In doing so, the authors are able to circumvent the above-described algorithm, through the explication of additional assumptions that enable them to remove the random effects from the parametrical expressions of the causal effects (i.e., the assumption of a rare binary mediator and/or outcome, and small random slopes). In contrast, the algorithm we propose in the fourth step does not require such additional assumptions, since the imputation algorithm allows us to marginalise over the random effects rather than remove them. In this, the algorithm broadens the applicability of causal effect estimation in multilevel mediation models.

1.2 Our work

In summary, we aim to investigate which multilevel estimation models are able to effectively eliminate unmeasured upper-level confounding of the mediator and the outcome, when the latter is binary. In addition, we will focus on a randomised binary exposure that varies within small clusters, as such group sizes have proven difficult for the available estimation techniques (Breslow and Clayton, 1993; Rodriguez and Goldman, 1995). These settings are often encountered in practice, e.g. when studying dyads (McMahon et al., 2003), twins (Ortqvist et al., 2009), or few repeated measures within each individual (Senn, 2002). On top of this, we want to evaluate if, and how, the link-function and/or a conditional versus a marginal sampling of the upper-level residual distribution (within the imputation algorithm), may affect the estimation of the mediation effects. In an attempt to answer these questions, we aim to conduct a large simulation study in which we compare three estimation models for mediator and outcome (an uncentered separate modelling approach, a separate approach that relies on CWC, and a joint method), two link-functions (*logit* and *probit*), and two ways in which to generate the random effects (marginally vs. conditionally). In this respect, our work distinguishes itself from other papers on lower-lever mediation, as most of these either 1) do not offer a generalisable approach to lower-level mediation from a counterfactual point-of-view, 2) do not

investigate the case of a binary outcome (and/or mediator), or 3) do not evaluate performance measures based on an extensive simulation study (Judd et al., 2001; Kenny et al., 2003; Bauer et al., 2006; Raykov and Mels, 2007; Montoya and Hayes, 2017; Vuorre and Bolger, 2017).

In the following sections, we go over the four different steps one at a time. First though, we introduce a motivating example with a small clusters size, where the outcome is measured on a binary scale. Next, we go over the first step in the process of estimating the causal mediation effects: we start by defining counterfactual outcomes in lower-level mediation settings, introduce non-parametric expressions for the direct and indirect effect, and discuss the assumptions needed for their identification. Then, in a second step, we derive parametrical expressions for these effects under a set of equations for the mediator and outcome that satisfy these assumptions. In the third part, we elaborate on three possible estimation methods that enable us to estimate the regression coefficients of random intercept models for binary measures. During the fourth part, we discuss the mechanism through which the causal mediation effects are estimated, as well as the two possible ways through which the random effects can be generated. Next, we illustrate these methods on our example data. To facilitate the practitioner's decision on which method is most appropriate in which setting, we subsequently conduct a simulation study where we compare the relative performance of the different estimation techniques and random effect generating mechanisms. Based on our findings we provide recommendations, and recap with a discussion.

2 Illustrating example

We consider data from a crossover study that aims to assess the impact of experimentally induced goal conflict on the helping behaviour of partners of individuals with chronic pain (ICP) (Kindt et al., 2018). During this study, couples (with at least one person having chronic pain) were asked to perform a series of household activities, while the presence of goal conflict in partners was randomly manipulated in a counterbalanced way. Partners were asked to stay available for help, while simultaneously working on a puzzle task (i.e., the goal conflict condition) or simply asked to be available (i.e., the control condition). After each series of chores, couples reported on several intra- and interpersonal outcomes, as well as the partners' quantity and quality of help. We will focus on the effect of goal conflict (a binary exposure) on the amount of help provided by the ICP's partner (the

binary outcome, high vs. low amount of help). As the amount of help is encoded within 10-second time frames (absence vs. presence of help), we regarded the amount of help as 'high' when, on average, help was more present than absent, and regarded it as 'low' otherwise. Additionally, we wanted to check whether or not this relation is mediated by the partner's amount of autonomous helping motivation, as perceived by the ICP (a continuous mediator, based on eight items on a 7-point scale). For this research question, we focus on data from 56 out of the original 68 couples, where no missingness is observed in the mediator or the outcome for either experimental condition[1]. As all three variables are measured within clusters (i.e., the couple) and each couple is exposed to two experimental conditions, we end up with a lower-level mediation design where two measurements are taken within each cluster.

3 Step 1 - Nonparametric definition & identification of the causal effects

Traditionally, mediation analysis has been formulated, understood, and implemented within a framework of linear regression models. This has proven problematic, since this line of thinking cannot offer general definitions of the causal effects beyond a few specific models. On top of this, these conclusions cannot be generalised to nonlinear models for discrete mediators and outcomes. In response, researchers have proposed and relied on the counterfactual framework to include the definition, identification, and estimation of causal mediation effects, without any reference to one specific statistical model (VanderWeele and Vansteelandt, 2009; Pearl, 2010; Imai et al., 2010a; Pearl, 2012).

3.1 The counterfactual framework

Before we introduce a nonparametric definition for the causal effects, let us explain the concept of 'counterfactual outcomes' in settings where all variables are measured at the lower-level. A 'counterfactual' or 'potential' outcome $Y_{ij}(x)$ represents the outcome that we would, possibly contrary to fact, have observed for measurement j within cluster i, had the exposure X_{ij} been manipulated to a value x (Rubin, 1978). When considering a dichotomous exposure (with value 0 for baseline/no exposure, and

[1] We run a complete case analysis, as missingness proves problematic for the joint approach implemented in this manuscript.

1 otherwise), we can define two possible potential outcomes for each measurement within a cluster: $Y_{ij}(0)$ and $Y_{ij}(1)$. Keeping this in mind, the measure- and cluster-specific total effect of X on Y is defined as the difference between both counterfactuals: $Y_{ij}(1) - Y_{ij}(0)$. Unfortunately, since only one of these counterfactuals is observed for each measurement, this effect cannot be estimated. The population average of the total causal effect $E[Y_{ij}(1) - Y_{ij}(0)]$, on the other hand, can be identified under specific assumptions (cfr. next section).

Similarly, counterfactuals for the mediator, $M_{ij}(0)$ and $M_{ij}(1)$, and nested counterfactuals for the outcome, $Y_{ij}(x, M_{ij}(x^*))$, can be devised (Robins and Greenland, 1992; Pearl, 2001). The latter counterfactual represents the value for the outcome Y_{ij}, when X_{ij} is set to x and M_{ij} is fixed at the value it would obtain when $X_{ij} = x^*$. Nested counterfactuals allow us to rephrase the average total effect of X on Y, to include a mediator: $E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))] = E[Y_{ij}(1) - Y_{ij}(0)]$, enabling us to partition the total causal effect into a total natural indirect and a pure natural direct effect (Hafeman and Schwartz, 2009; VanderWeele, 2013):

$$\begin{aligned} TCE =& E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))] \\ =& E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0)) + Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))] \\ =& TNIE + PNDE \end{aligned} \tag{5.1}$$

Note that we define non-parametrical expressions for the direct and indirect effect on a linear scale, in contrast to e.g. Imai et al. (2010b); VanderWeele (2013); Loeys et al. (2013); Bind et al. (2016), where these effects are often defined in terms of risk- or odds ratios for binary outcomes.

3.2 Causal and modelling assumptions to identify the causal mediation effects

In order to identify the above-defined non-parametrical effects in lower-level mediation settings with a randomised exposure, we need to postulate the following set of assumptions (VanderWeele, 2010b; Josephy et al., 2015):

(i) There are no unmeasured upper- or lower-level confounders of the association between mediator and outcome.

(ii) There are no confounders of the association between mediator and outcome, caused by exposure (i.e. no intermediate confounding).

(iii) There is no carry-over effect when lower-level measures represent time points.

For lower-level mediation in clusters of size two, assumptions *(i)-(iii)* can be summarised by the (lack of) arrows within the diagram in figure 5.1 (Robins and Richardson, 2010).

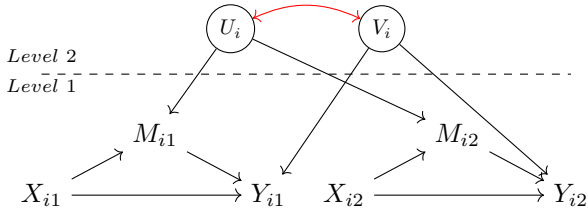

Figure 5.1 This causal diagram graphically represents assumptions *(i)-(iii)*, which are needed to identify the causal effects in a randomised lower-level mediation setting with clusters of size two. X_{i1}, M_{i1} and Y_{i1} represent the respective values of the exposure, mediator, and outcome for the first measure within cluster i, while X_{i2}, M_{i2} and Y_{i2} reflect these variables for the second measurement. Absence of a unidirectional arrow between two variables indicates the absence of a direct causal effect, while a bidirectional arrow captures an unmeasured common cause.

Note that including the red arrow in figure 5.1, allows for the unmeasured cluster-specific common causes of the outcome (V) and those of the mediator (U), to correlate. As such, V can be expressed as a function of U (i.e. $h(U)$) without a loss of generality, rendering the unmeasured upper-level confounder of the M-Y relationship more explicit. Consequently, this arrow directly violates assumption *(i)*: there *are* unmeasured *upper-level* confounders of the mediator-outcome relation. Josephy et al. (2015) showed that in linear lower-level mediation settings, this assumption is not necessary for the identification of the causal mediation effects; they demonstrate that researchers can estimate the direct and indirect effects without bias, even in the presence of such upper-level confounding. In this manuscript, we wish to additionally demonstrate the redundancy of this assumption in lower-level mediation settings with a binary outcome.

In addition to these three causal assumptions, we will consider the following modelling assumptions throughout the paper:

(iv) Unmeasured upper-level confounders of the mediator and outcome exert an additive effect on both the mediator and the outcome[2].

[2]This assumption is made on the scale of the parametrical models for the mediator and outcome

(*v*) There is no unmeasured heterogeneity among clusters in the effect of exposure on mediator, nor in the effect of exposure and mediator on the outcome.

Unlike the previous three assumptions, assumptions *(iv)* and *(v)* cannot be represented on a causal diagram; hence, they are not depicted in figure 5.1.

4 Step 2 - Parametric identification of the causal effects

Now that we possess non-parametric definitions of the causal effects, we can pursue their identification based on parametrical statistical models for the mediator and binary outcome. Let us consider the following multilevel models, with i the cluster, and j a within-cluster observation:

$$\begin{aligned} E[M_{ij}|X_{ij}, U_i] &= g_M^{-1}\big(\delta_M + \alpha X_{ij} + \eta_i\big) \\ E[Y_{ij}|X_{ij}, M_{ij}, U_i] &= g_Y^{-1}\big(\delta_Y + \zeta' X_{ij} + \beta M_{ij} + \phi X_{ij} M_{ij} + h(\eta_i)\big) \end{aligned} \tag{5.2}$$

where g_M^{-1} and g_Y^{-1} represent known inverse link functions for M and Y, respectively. In these equations, δ_M and δ_Y represent the intercepts for the mediator and the outcome, while α, β, ζ' , and ϕ represent the effects of exposure on mediator, mediator on outcome, exposure on outcome, and the interaction between exposure and mediator on the outcome, respectively. Since the unmeasured upper-level confounders of the mediator, η_i, and of the outcome, $\nu_i = h(\eta_i)$, are allowed to correlate, this induces unmeasured cluster-specific confounding of the M-Y relationship (see red arrow in figure 5.1). Note that we additionally assume that these effects are homogeneous across clusters, in accordance with assumption *(v)*. Under this data-generating mechanism, the assumptions introduced in section 3.2 are met (except for the upper-level confounders of assumption *(i)*, of which we aim to prove its redundancy under a lenient set of modelling assumptions). This enables us to operate Pearl's mediation formula (Pearl, 2001, 2010) to derive the total, pure natural direct, and total natural indirect effect for each measurement j within cluster i.

For example, when $g_M = g_Y = probit$ (and hence with Φ representing the standard normal cumulative distribution), we find a "ij-th"-specific *total natural indirect effect* of:

$$
\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))|U_i, t)] \\
&\quad = \Big(\Phi(\delta_M + U_i) - \Phi(\delta_M + \alpha + U_i)\Big)\Big(\Phi(\delta_Y + \zeta' + h(U_i)) \\
&\qquad - \Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\Big) \qquad (5.3)
\end{aligned}
$$

The parametrical derivations and expressions for the causal effects can be found in the appendix, for both $g_M = g_Y = logit$ and $g_M = g_Y = probit$. We do not offer any derivations for these effects when g_M represents the *identity*-link and g_Y either the *probit*- or *logit*-link, as this case does not provide any closed-form expressions.

5 Step 3 - Estimation models for the mediator and outcome

Now that we know how to identify the causal mediation effects, a next logical step considers their estimation. Before we can achieve this, however, we first need to estimate the regression parameters for the mediator and outcome models (5.2), with the aid of appropriate estimation models. To this end, the following sections summarise three potential approaches. Note that the next few equations represent estimation models, in contrast to the causal model from the previous section (i.e., equation (5.2)).

5.1 Separate modelling of a binary mediator and outcome

One such approach fits the mediator and outcome measures by use of two separate multilevel models, with i the cluster ($i = 1...I$) and j the measurement within a cluster ($j = 1...J$):

$$
\begin{aligned}
E(M_{ij}|X_{ij}, u_i) &= g_M^{-1}\big(d_M + aX_{ij} + u_i\big) \\
E(Y_{ij}|X_{ij}, M_{ij}, v_i) &= g_Y^{-1}\big(d_Y + c'X_{ij} + bM_{ij} + fX_{ij}M_{ij} + v_i\big) \\
&\qquad \text{with } u_i \amalg X_{ij} \text{ and } v_i \amalg X_{ij}, M_{ij}, X_{ij}M_{ij} \qquad (5.4)
\end{aligned}
$$

Here, g_M^{-1} and g_Y^{-1} again represent known inverse link functions and u_i and v_i the random intercepts for M and Y, respectively. These upper-level residuals are assumed to be normally distributed with mean zero and variance σ_M^2 for u_i and σ_Y^2 for v_i. Note that this uncentred (UN) separate modelling approach assumes that the upper-level residuals are independent of the predictors. If, however, there is upper-level confounding of the M-Y relation, both random intercepts will be correlated and, since u_i predicts

M_{ij}, M_{ij} and v_i will be correlated as well. This is in direct violation of the assumption $v_i \amalg X_{ij}, M_{ij}$ and, as a result, the above-described multilevel model for the outcome will estimate the regression coefficients of model (5.2) with bias.

In linear multilevel settings (i.e., when g_M and g_Y both represent the identity-link), many scholars have attempted to solve this confounding issue by separating within- from between-cluster effects (Louis, 1988; Neuhaus and Kalbfleisch, 1998; Begg and Parides, 2003; Zhang et al., 2009; Kenward and Roger, 2010; Preacher et al., 2010; Pituch and Stapleton, 2012). Such centering within-clusters (CWC) can be achieved by regressing a continuous dependent variable on the cluster-mean centred values of the predictors: $(X_{ij} - \overline{X}_i)$ and $(M_{ij} - \overline{M}_i)$. In these expressions, $\overline{X}_i$ and $\overline{M}_i$ denote the cluster-specific averages of the exposure- and mediator-scores for cluster i across its measurements (MacKinnon, 2008). Subtracting these means from the raw scores will remove any cluster-specific effects that may influence the predictors and hence, any possible impact of unmeasured upper-level confounders. As such, the upper-level residuals will be uncorrelated with these within-cluster deviations, implying that the parameter coefficients of model (5.2) can be estimated without bias in the presence of upper-level endogeneity.

A similar approach is possible for a binary outcome, through the following set of multilevel models:

$$\begin{aligned} E(M_{ij}|X_{ij}, u_i) &= g_M^{-1}\big(d_M + a(X_{ij} - \overline{X}_i) + u_i\big) \\ E(Y_{ij}|X_{ij}, M_{ij}, v_i) &= g_Y^{-1}\big(d_Y + c'(X_{ij} - \overline{X}_i) + b(M_{ij} - \overline{M}_i) \\ &\quad + f(X_{ij}M_{ij} - \overline{XM}_i) + v_i\big) \\ &\quad \text{with } u_i \amalg (X_{ij} - \overline{X}_i) \\ &\quad \text{and } v_i \amalg (X_{ij} - \overline{X}_i), (M_{ij} - \overline{M}_i), (X_{ij}M_{ij} - \overline{XM}_i) \end{aligned} \quad (5.5)$$

Again, both upper-level residuals are assumed to be independently and normally distributed with mean zero and a fixed variance. Unfortunately, when the outcome is measured on a binary scale, CWC no longer yields proper parameter estimates, although in practice, the bias may often be small (Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010).

Similar to Greenland (2002), Goetgeluk and Vansteelandt (2008), and Brumback et al. (2010), we argue that model (5.5) cannot be considered a valid causal model. In the longitudinal setting considered here, this model would imply that the future causes the past (e.g., future X_{ij} would cause past Y_{ij^*} for $j > j^*$, since X_{ij} is contained within $\overline{X}_i$). This remark does not

downgrade the usefulness of this estimation model, but rather emphasises that model (5.5) should not be attributed a causal interpretation.

5.2 Joint modelling of a binary mediator and outcome

A second approach consists of jointly modelling the mediator and the outcome. This can be achieved by either relying on multivariate techniques, or by tricking univariate software into modelling the mediator and outcome in a multivariate way (Bauer et al., 2006). The set of equations resembles (5.4) except that now, u_i and v_i are allowed to covary:

$$\begin{aligned} E(M_{ij}|X_{ij}, u_i) &= g_M^{-1}\left(d_M + aX_{ij} + u_i\right) \\ E(Y_{ij}|X_{ij}, M_{ij}, v_i) &= g_Y^{-1}\left(d_Y + c'X_{ij} + bM_{ij} + fX_{ij}M_{ij} + v_i\right) \\ &\text{with } (u_i, v_i) \sim N(\mathbf{0}, \mathbf{\Sigma}) \end{aligned} \tag{5.6}$$

Here, the upper-level residuals are assumed to be multivariate normally distributed, with zero mean and covariance matrix Σ. This matrix is defined by the variances of u_i and v_i on its diagonal (σ_M^2 and σ_Y^2, respectively), and by the covariance between both upper-level residuals (σ_{MY}) elsewhere. Since this set of models allow both random intercepts to covary, unmeasured upper-level M-Y confounding may be accounted for through the modelling of this correlation. As estimation model (5.6) equals the true data-generating model (5.2) from section 4, we expect unbiased estimators for the regression coefficients in the outcome model.

6 Step 4 - Estimation of the causal effects through Monte Carlo potential outcome generation

After estimating the regression coefficients in the models for the mediator and outcome, a final step aims to estimate the mediation effects themselves. This can be achieved by sampling potential outcomes from the estimated mediator and outcome distributions with the aid of an imputation algorithm. Recall that, for a randomised binary exposure X, we observe $Y_{ij}(X_{ij}, M_{ij}(X_{ij}))$ for each within-cluster measure. However, in order to estimate the population averaged indirect effect, we additionally require the counterfactual outcome $Y_{ij}(X_{ij}, M_{ij}(1 - X_{ij}))$ for every measurement. In an algorithm proposed by Imai et al. (2010a), we can obtain a Monte Carlo draw from the potential outcome $Y_{ij}(x, M_{ij}(x^*))$ by using model predictions. This transpires through a parametrical or quasi-Bayesian Monte

Carlo approximation in which the posterior distribution of the quantities of interest is approximated by their sampling distribution:

1. Fit models for the observed mediator and outcome variables.
2. Simulate estimated model parameters from their sampling distributions (e.g., 1000 draws).
3. Repeat the following three processes within a single draw from the previous step: (a) predict both potential values of the mediator ($M_{ij}(0)$ and $M_{ij}(1)$) for each measure within each cluster, (b) predict the potential outcomes for each within-cluster measurement, given the predicted values of the mediator ($Y_{ij}(0, M_{ij}(0))$, $Y_{ij}(1, M_{ij}(0))$, $Y_{ij}(0, M_{ij}(1))$, and $Y_{ij}(1, M_{ij}(1))$), (c) compute the causal mediation effects, averaged over clusters and measurements within clusters.
4. Compute the summary statistics, such as point estimates and confidence intervals, over all simulated draws.

7 Estimation techniques and software implementations

Up until now, we merely focussed on the models for the mediator and the outcome and the estimation of the causal mediation effects. Of course, there are a lot of possible combinations of estimation techniques and software implementations that allow us to fit the above-mentioned statistical models and to generate potential outcomes for the estimation of causal mediation effects. We discuss several such options next.

7.1 Step 3 - Estimation of the regression parameters

Let us first tackle estimation techniques for the uncentered and centred approaches that model the mediator and the outcome separately. With the aid of simulation studies, Josephy et al. (2016) concluded that generalised linear mixed models (GLMMs) that rely on Maximum Likelihood (ML) estimation through Adaptive Gaussian Quadrature (AGQ) provided the most reliable estimates when analysing binary *probit*-regression models within small clusters. For dyadic cluster sizes, AGQ operated on par with Diagonally Weighted Least Squares (DWLS) estimation within the SEM framework, but took the upper hand as the cluster size increased. Note

that using conditional logistic regression within the third step is not really an option, although this approach is perfectly capable of dealing with unmeasured upper-level confounding of mediator and outcome. The reason for this is that conditional likelihood approaches do not estimate any intercepts, nor do they provide estimates for the random effect variances, making the prediction of potential values for the mediator and the outcome within the fourth step impossible. On top of this, conditional logistic regression cannot be implemented for any link-functions other than the *logit*-link.

Next, let us look at possible implementations that allow us to jointly model the mediator and outcome. Bauer et al. (2006) first introduced a joint modelling approach for linear mixed models (LMMs) in `SAS`®, by fitting a multivariate model using univariate multilevel software (i.e., the Proc Mixed procedure). A next logical step extends this line of thinking to GLMMs, but regrettably proves unattainable within the current `SAS`-software, as the method requires the specification of random effects in combination with a residual covariance structure that differentiates between mediator and outcome. Unfortunately, Proc Glimmix cannot integrate marginal covariances within AGQ, while Proc NLmixed is unable to model residual covariances in the first place.

A second possible candidate consists of a Bayesian approach through Komárek and Komárková (2014)'s *mixAK*-package within the statistical environment `R` (R Core Team, 2013). However, while exploring the numerous possibilities of this package, we experienced several difficulties in estimating the covariance between the random intercepts of M and Y; this random term did not attain stable convergence measures, even when considering high burn-in and thinning values. On top of this, the package is restrained to Bayesian estimation through the *logit*-link, disregarding its *probit*-alternative entirely.

Structural Equation Models (SEM), where a categorical outcome is considered a crude approximation of an underlying continuous variable, offer a third possibility. As we mentioned at the beginning of this section, DWLS has proven very auspicious when estimating models for binary measures within small clusters. Since SEM naturally considers data in a multivariate way, it allows the joint modelling of mediator and outcome. However, within (D)WLS, a binary measure that simultaneously acts as both a dependent and independent variable (i.e., a so-called endogenous variable), is treated as its underlying continuous measure during the entire estimation process. As such, (D)WLS encounters problems when estimat-

ing the parameter coefficient of a binary endogenous mediator within the outcome model. Fortunately, ML-estimation can treat the mediator as its underlying measure when it serves as a dependent variable, while considering its observed values when the mediator serves as a predictor. In Rosseel (2012)'s *lavaan*-package within R, only (D)WLS-estimation is currently able to deal with endogenous categorical variables. In contrast, ML-estimation through AGQ in MPLUS®-software is able to model endogenous binary mediators (Muthén and Muthén, 2010).

With these considerations and limitations in mind, we will consider ML-estimation through AGQ for the joint and both separate modelling methods in the third step. The uncentred and centred separate modelling approaches will be fitted with the aid of the *lme4*-package (version 1.1-17) within R version 3.5.0 (Bates et al., 2015), while the joint approach will take place within the MPLUS-software (version 7.4).

7.2 Step 4 - Estimation of the causal mediation effects

Conveniently, Tingley et al. (2014) developed their R-package *mediation* in which separate models for the mediator and the outcome can be inserted, to subsequently generate estimates for the causal mediation effects through the implementation of the algorithm described in section 6. We do, however, have a few concerns regarding the implementation as described in Imai et al. (2010a).

For one, as this package can only model the mediator and the outcome separately, it cannot quantify any unmeasured upper-level confounding of M and Y through a covariance term between both random intercepts. As a consequence, the random effects will be generated from independent normal distributions rather than from a less stringent multivariate one, and hence, will not be able to appropriately deal with upper-level endogeneity of a mediator and a binary outcome. This package's documentation consequently assumes the (somewhat improbable) absence of upper-level endogeneity of mediator and outcome.

Two, the authors do not provide any recommendations concerning which estimation techniques ought to be used in which settings. Tingley et al. (2014) do not explicitly recommend the use of AGQ when the fitted models for the mediator and/or outcome constitute GLMMs. As such, uninformed researchers might not be aware that they are relying upon the Laplace approximation by default and consequently, shoulder the approach's shortcomings when dealing with non-normal data within small

clusters (Tuerlinckx et al., 2006; Josephy et al., 2016).

Three, in the third step of their algorithm (see section 6), the authors rely on a conditional approach for generating the random effects. During this process, these are sampled conditionally on the cluster identifier: the upper-level residuals are assumed to follow a normal distribution, conditional on the estimated random effect within that cluster, as well as the estimated conditional variance. It has been pointed out that this method for generating the random effects may not lead to a realistic sampling distribution (Skrondal and Rabe-Hesketh, 2004). Rather, a marginal sampling process, in which all random effects are drawn from a normal distribution with a zero mean and a standard deviation based on the estimated variance component, may lead to better estimates of the causal mediation effects.

As such, we will compare the performance of two possible random effect generating mechanisms within the fourth estimation step: a conditional versus a marginal procedure. As we were unable to extract the conditional variances from `MPLUS`, we only compare both approaches within the two separate modelling techniques.

8 Simulation study

In summary, we compare five different approaches for estimating the causal mediation effects: (1) an uncentered separate, (2) a centred separate, and (3) a joint modelling procedure with marginally generated random effects, alongside (4) an uncentered and (5) a centred separate modelling approach with conditionally generated random effects. The detailed code on the software implementation can be found in the appendix.

To gain insight into the performance of these procedures, we compare them through simulations under a variety of settings. For this, we generated random binary mediator and outcome values within small clusters, according to random intercept *probit*- or *logit*-models. For simplicity, our data generating mechanism omits an interaction between the exposure and mediator in the outcome model:

$$\begin{aligned} P(M_{ij} = 1|X_{ij}, u_i) &= P(M^*_{ij} > 0|X_{ij}, u_i) \\ &= P(\delta_M + \alpha X_{ij} + u_i + \epsilon^M_{ij} > 0) \\ P(Y_{ij} = 1|X_{ij}, M_{ij}, v_i) &= P(Y^*_{ij} > 0|X_{ij}, M_{ij}, v_i) \\ &= P(\delta_Y + \zeta' X_{ij} + \beta M_{ij} + v_i + \epsilon^Y_{ij} > 0) \\ &\text{with } (u_i, v_i) \sim N(\mathbf{0}, \mathbf{\Sigma}) \end{aligned} \tag{5.7}$$

Here, M_{ij}^* and Y_{ij}^* represent the underlying latent variables of the binary mediator and outcome, respectively, such that $M_{ij} = 1$ if $M_{ij}^* > 0$, and $Y_{ij} = 1$ if $Y_{ij}^* > 0$. In these equations, the lower-level residuals of the latent variables, ϵ_{ij}^M and ϵ_{ij}^M, are both i.i.d. drawn from a normal distribution with mean zero and a variance, σ^2, that changes according to the link function. For the *probit*-link, $\sigma^2 = 1$, while for the *logit*-link $\sigma^2 = \frac{\pi^2}{3}$. The random intercepts are sampled from a multivariate normal distribution with zero means and variance-covariance matrix Σ:

$$\Sigma = \begin{pmatrix} \tau_M & \rho\sqrt{\tau_M \tau_Y} \\ \rho\sqrt{\tau_M \tau_Y} & \tau_Y \end{pmatrix}$$

For the different simulation settings, we vary several parameters. First of all, we consider different clusters sizes: we will look at clusters of size two and five. Second, we also regard a different number of clusters: we consider sample sizes n of 50, 100, and 300. Three, as we wish to study whether or not the link function impacts our conclusions, we consider both the *probit*- and the *logit*-link in generating the mediator and outcome measures. Note that the coefficients of the *logit*-link are about 1.7 times larger compared to those defined for the *probit*-link (see table 5.1).

	Link-function	
Parameter	*Probit*-link	*Logit*-link
δ_M	0.00	0.00
α	1.00	1.70
δ_Y	–0.70	-1.20
ζ'	0.50	0.85
β	0.80	1.35

Table 5.1 A summary of the parameter values for the data-generating mechanism, according to one of two possible link functions.

Four, we also wish to examine the impact of different intracluster correlations (*icc*'s) for the latent response variables of the mediator and outcome. As the latent icc_l is defined as the proportion of between-cluster versus total variance in the latent responses (e.g., for the mediator, $icc_l = \frac{Var(u_i)}{Var(M_{ij}^*)} = \frac{\tau_M}{\tau_M + \sigma^2}$), this value depends upon the variance of the lower-level residuals and hence, on the link function. As such, a latent icc_l of 0.10, 0.30, and 0.50 corresponds to a respective random intercept variance of 0.11, 0.43, and 1.00 for the *probit*-link, and 0.36, 1.41, and 3.29 for the *logit*-link (with $icc_l^M = icc_l^Y$). Finally, we also look at the impact of unmeasured

upper-level confounding of mediator and outcome by varying the value of ρ in the covariance matrix Σ: we consider a correlation $\rho = 0$ (i.e., no unmeasured upper-level confounding) and $\rho = 0.50$.

In total, 1000 simulations are generated for different combinations of cluster size (2), sample size (3), link-function (2), *icc* (3), and random intercept correlation (2). The five above-introduced methods are compared over these settings in terms of convergence, relative bias, mean squared error (MSE), and coverage. The relative bias is defined as the averaged difference between the estimated (e.g. $\hat{\beta}$) and true parameter values (e.g. β), divided by the latter (so that the relative bias equals $\frac{\beta-\hat{\beta}}{\beta}$); as such, a relative bias enclosing zero will indicate an unbiased estimator. The MSE is estimated by summing the empirical variance and the squared bias of the estimates, simultaneously assessing bias and precision: the lower the MSE, the more accurate and precise the estimator. The coverage is defined as the proportion of the 95% Wald-confidence intervals that encompass their true parameter value; coverage rates nearing 95% represent nominal coverages of the intervals. Lastly, in order to conclude model convergence, a model fit must yield both estimates and standard errors. To ensure a fair comparison between methods, we only present results for simulation runs in which all five methods converged.

9 Results

Below, we discuss the results of the simulation study for the *probit*-link in detail, comparing clusters of size two and five. In addition, we report the results comparing the *logit*- and *probit*-link for clusters of size two. In the following sections, we refer to the five approaches as: UN (separate modelling, uncentred) and CWC (separate modelling, centred within-clusters), both with a marginal ("-Marg") and a conditional ("-Cond") approach to modelling the random effects, and to the Joint approach with a marginal random effects generation as "Joint-Marg".

9.1 Convergence

For the convergence, we can but observe the results of three rather than five approaches, since both ways of generating the random intercepts overlap up until the fourth step of our estimation process; hence, their convergence will be identical. Generally, convergence improves as the number of clusters (i.e. the sample size n) and the number of measurements within a cluster

increase (i.e. from two to five) (see left part of figure 5.2). Note that for 300 clusters most approaches reach 100% convergence, except for the joint approach when the icc_l is low. In contrast to changes in sample and cluster sizes, convergence seems more or less unaffected by the presence of unmeasured upper-level confounding of the mediator-outcome relation. Moreover, convergence is seemingly unaffected by the latent intracluster correlation for UN and CWC, whereas it seems to improve for the joint approach with increasing icc_l. Lastly, it appears that the convergence fares slightly better for all approaches when the *logit*-link, rather than the *probit*-link is used (see right part of figure 5.2). Overall, the joint approach shows the most difficulty in reaching convergence.

9.2 Relative bias

First of all, for the direct and indirect effect estimators we typically observe that the relative bias decreases as the number of clusters and the number of measurements within each cluster increases (see figure 5.3). Only when there is upper-level endogeneity of mediator and outcome, does the relative bias of the indirect effect increase instead of decrease with larger samples, for both uncentred approaches. Second, both causal mediation effects are not influenced by an increase in the icc_l for the joint approach (with 'CWC-Cond' a close second), while it does impact others, especially when $\rho \neq 0$: in this case, their relative bias increases with rising *icc*. Third, when comparing link functions, we see no obvious changes in the performance of the joint approach, nor for both conditional approaches to generating the random effects (see figure 5.6 in the Appendix). Both marginal approaches, however, exhibit a strong increase in relative bias when relying on the *logit*-, compared to the *probit*-link. Overall, we observe that the joint approach provides the least biased estimates.

9.3 MSE

Generally, the mean squared error declines with increasing sample size and number of within-cluster measures, as well as with a rising icc_l (see figure 5.4). Furthermore, we do not observe any differences in MSE when comparing settings with and without unmeasured upper-level confounding of the M-Y relation, nor when comparing link functions. The only deviation from this consists of a slightly increased MSE for the direct effect estimator of both CWC approaches when comparing *logit*- to *probit* regression (see

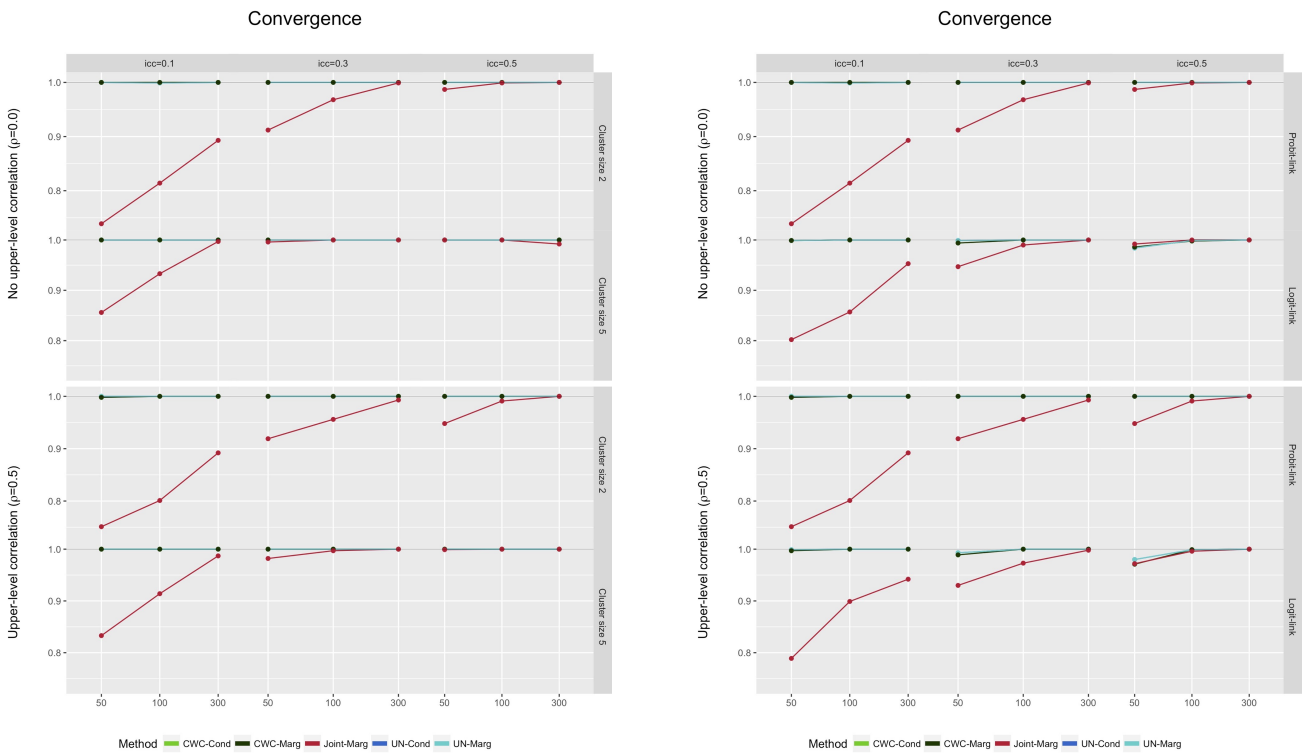

Figure 5.2 Model convergence of the five approaches comparing cluster sizes for the *probit*-link (left), and comparing link-functions for cluster size two (right), with different icc_l's (0.1, 0.3 and 0.5) and sample sizes (50, 100 and 300).

Figure 5.3 The relative bias of the direct (left) and indirect (right) for the five approaches modelled with a *probit*-link, for different upper-level correlations between the random intercepts (zero or 0.5), cluster sizes (2 and 5), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

figure 5.7 in the Appendix). Over all settings considered, the MSE is generally lowest for the joint modelling approach and CWC.

9.4 Coverage

For both causal mediation effects, the coverage of their 95% confidence intervals is typically better when the cluster size equals two rather than five, and when the latent intracluster correlation is low (see figure 5.5). This observation holds for all methods, although the joint approach and 'CWC-Cond' seem least influenced by changes in these measures. Additionally, when the icc_l is high, we often observe a decrease in the coverage rate as the upper-level sample size increases, for both UN-approaches and 'CWC-Marg'. Also, the presence of unmeasured upper-level confounding of the mediator and outcome does not seem to impact the joint approach or 'CWC-Cond' that much, whereas the other approaches show a steep decrease in coverage, especially when samples sizes are large. Again, the link function does not seem to impact the coverage of the joint and both conditional approaches, whereas the it tends to decrease for both marginal approaches when comparing the *logit*- to the *probit*-link (see figure 5.8 in the Appendix). Generally, the joint approach and 'CWC-Cond' provide the best coverage.

9.5 Analysis of the example

Next, we illustrate the above five approaches by applying them to our example data, where we assess whether or not the effect of goal conflict (i.e., the binary exposure) on the observed amount of help (i.e., the binary outcome) is mediated by the partner's amount of autonomous helping motivation, as perceived by the patient (i.e., the continuous mediator). Within the third step of our estimation procedure, we modelled the mediator and the outcome according to estimation models (5.4)-(5.6), where g_M represents the identity-link, g_Y the *probit*-link, and without an interaction (i.e., $f = 0$). The estimated regression coefficients and random intercept variances (alongside the estimated standard errors and p-values) based on the joint modelling approach (i.e., estimation model (5.6)), are summarised in table 5.2.

In doing so, we observe a significant effect of goal conflict on the observed amount of help (i.e., the c-path, $p < 0.001$) and on the perceived amount of autonomous helping motivation (i.e., the a-path, $p = 0.032$). Additionally, we also discern a significant effect of goal conflict on the

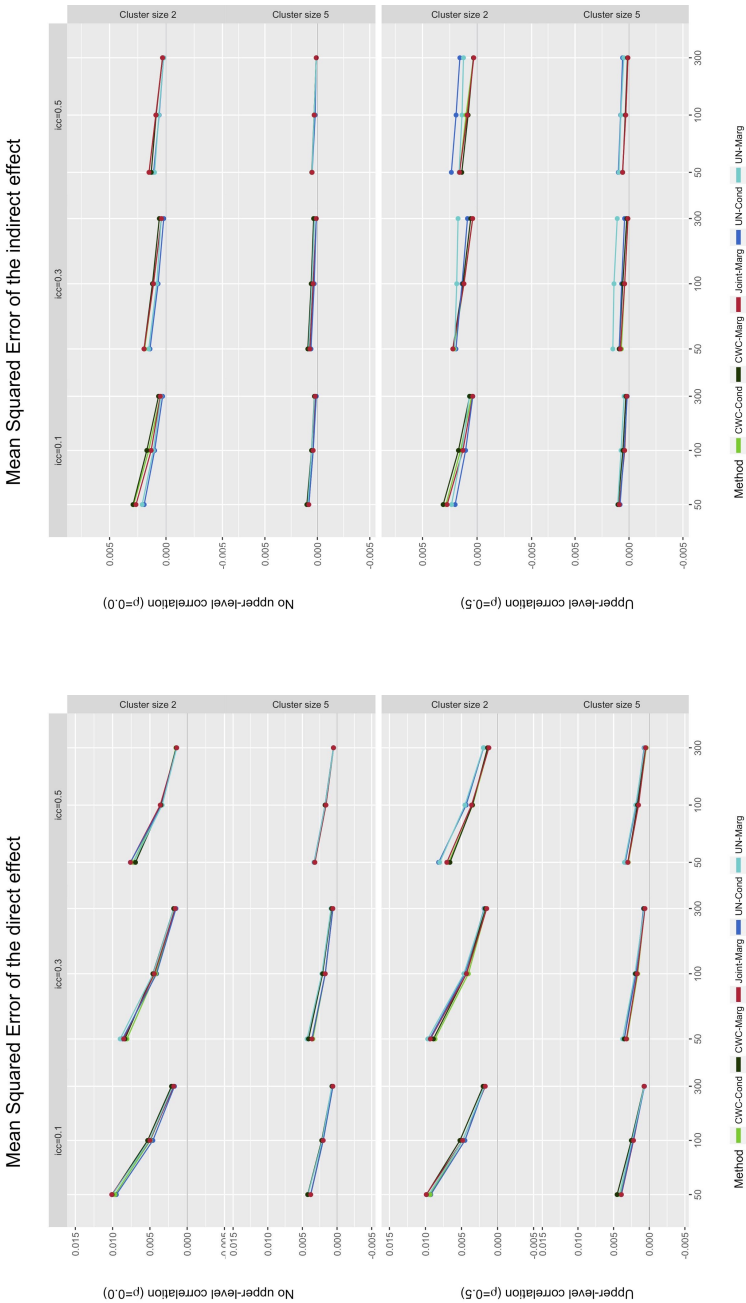

Figure 5.4 The MSE of the direct (left) and indirect (right) for the five approaches modelled with a *probit*-link, for different upper-level correlations between the random intercepts (zero or 0.5), cluster sizes (2 and 5), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

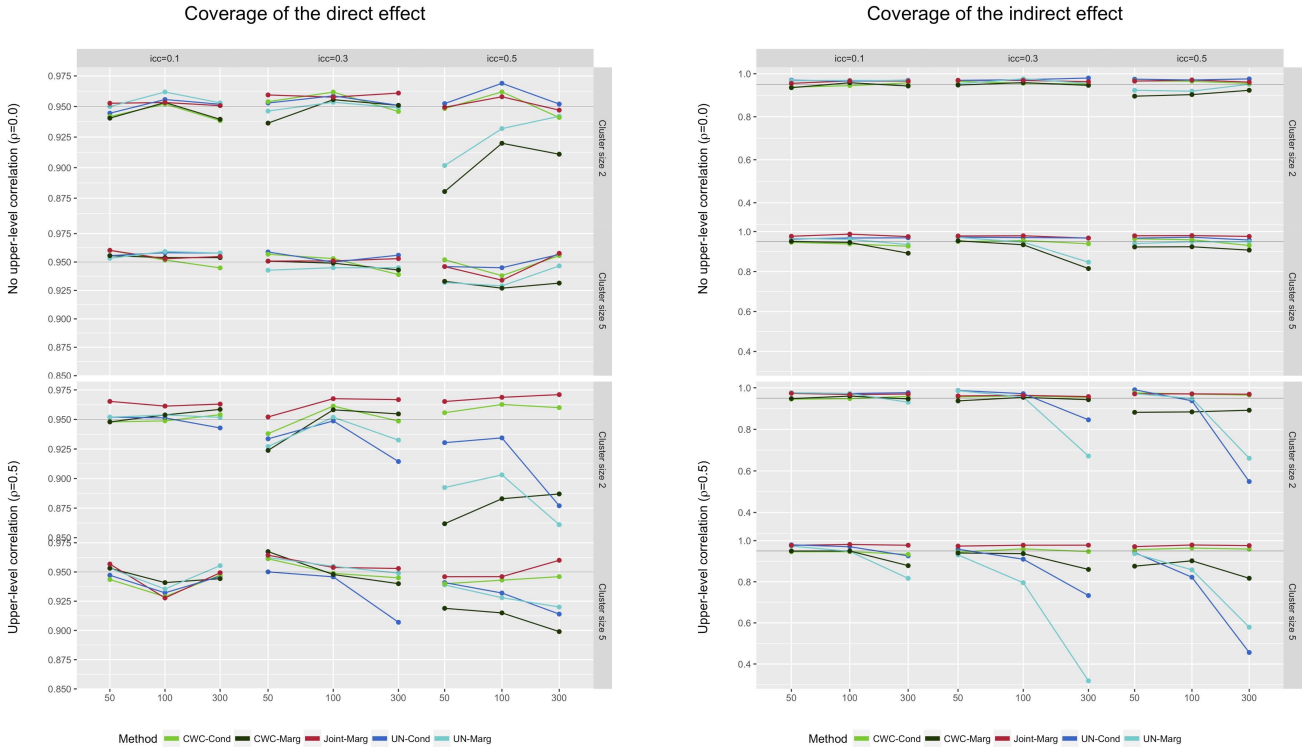

Figure 5.5 The coverage of the direct (left) and indirect (right) for the five approaches modelled with a *probit*-link, for different upper-level correlations between the random intercepts (zero or 0.5), cluster sizes (2 and 5), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

	Estimate (se)	p-value
$\hat{d_M}$	3.896 (0.200)	< 0.001
$\hat{a}$	-0.356 (0.160)	0.026
$\hat{d_Y}$	2.678 (1.350)	0.047
$\hat{c'}$	-1.453 (0.446)	0.001
$\hat{b}$	0.699 (0.342)	0.041
$\hat{\sigma^2_M}$	1.520 (0.361)	< 0.001
$\hat{\sigma^2_Y}$	0.765 (0.930)	0.411
$\hat{\sigma_{MY}}$	-0.704 (0.610)	0.249

Table 5.2 The estimated regression coefficients of the models for the effect of goal conflict on the partner's amount of autonomous helping motivation, and of goal conflict and the partner's amount of autonomous helping motivation on the observed amount of help. The estimates (with estimated standard errors, se) and p-values are provided for the joint modelling procedure.

amount of help, when controlling for the amount of autonomous helping behaviour (i.e., the c'-path, $p = 0.001$), and a significant effect of the amount of autonomous helping behaviour, when controlling for the exposure (the b-path, $p = 0.041$).

Although all these pathways appear significant, we observe a significant direct, but no indirect effect, for either of the five estimation procedures during the fourth step of our estimation process (see table 5.3). This is most likely due to the rather modest upper-level sample size, which might provide a very low power for detecting a small indirect effect. For the intervening effect, we see that the uncentred estimates are larger than those of both CWC-approaches, whereas the estimates of the Joint approach are somewhere in between (although closer to those of CWC). In contrast, the estimates for the direct effect are smallest for the uncentred approaches and largest for the CWC-approaches, with the Joint approach again providing estimates between both. Additionally, we also observe the smallest empirical standard errors for the uncentred approaches, the largest for CWC, with the joint approach again taking the middle ground.

Our motivating example manifests 56 clusters of size two and, according to $icc_l = \frac{\tau}{\tau+\epsilon}$, a latent intracluster correlation of 0.68 for the mediator and 0.43 for the outcome (according to the joint approach). Additionally, the upper-level correlation between the random intercepts of mediator and outcome is estimated at -0.65 (with $p = 0.25$). An example of a couple-level confounder that is negatively associated with the amount of autonomous motivation and positively associated with the observed amount of help

	Indirect Effect		Direct Effect		Total Effect	
	Estimate (se)	95%-CI	Estimate (se)	95%-CI	Estimate (se)	95%-CI
UN-Cond	-0.020 (0.016)	(-0.058; 0.004)	-0.370 (0.086)	(-0.523; -0.202)	-0.390 (0.086)	(-0.544; -0.214)
UN-Marg	-0.020 (0.016)	(-0.056; 0.005)	-0.374 (0.087)	(-0.531; -0.202)	-0.394 (0.087)	(-0.547; -0.226)
CWC-Cond	-0.042 (0.035)	(-0.115; 0.007)	-0.275 (0.112)	(-0.481; -0.067)	-0.317 (0.116)	(-0.525; -0.077)
CWC-Marg	-0.042 (0.035)	(-0.120; 0.005)	-0.281 (0.114)	(-0.491; -0.062)	-0.323 (0.118)	(-0.527; -0.074)
Joint-Marg	-0.037 (0.031)	(-0.109; 0.012)	-0.339 (0.093)	(-0.514; -0.151)	-0.376 (0.092)	(-0.539; -0.187)

Table 5.3 The estimates of the indirect, direct, and total effect of goal conflict on the observed amount of help, mediated by the patients' perceived amount of autonomous helping motivation. The estimates (and empirical standard errors, se) are provided for the five different estimation procedures, alongside the percentile-based 95% confidence intervals.

might be found in a worried/solicitous partner. Partners who are naturally worrisome will often offer a considerable amount of help to regulate their own distress and emotions, irregardless of the needs of the ICP (Vervoort and Trost, 2017). As such, solicitous partners may help the ICP out of an internal pressure (i.e., controlled motivation) to temper his or her own guilt. If the patient picks up on this, the perceived amount of autonomous helping behaviour may be low (i.e., the partner helps out of obligation), while at the same time, the observed amount of offered help is high.

In summary, we are confronted with a small upper-level sample size with only two observations within each cluster, medium to large latent *icc*'s, and a high (nonsignificant) amount of unmeasured upper-level M-Y confounding. In these settings, our simulation studies suggest that the joint approach is most likely to provide unbiased estimates for both causal mediation effects (as can be seen in figure 5.3). In the absence of unmeasured upper-level confounding of mediator and outcome, our preference would slightly gravitate towards the estimates of the UN-approaches, but in the end, we would still prefer the estimates of the joint approach as they have proven much more reliable.

In section 3.2 we saw that assumption *(i)* can be relaxed to 'no unmeasured lower-level confounders of mediator and outcome' for some estimation models, but not for others. The other assumptions, however, still need to hold in order for our inferences to be valid. Unfortunately, there is no way of checking the plausibility of assumptions *(i)*), *(ii)*, and *(iv)*. Concerning assumption *(iii)*, we cannot exclude the possibility of a carry-over effect, as there was no notable wash-out period in this study. Unfortunately, there is also no way of checking assumption *(v)*, as random slopes are unidentifiable in designs with a mere two measurements within each cluster. These shortcomings should be kept in mind when interpreting this lower-level mediation analysis.

10 Discussion

In this paper, we provided an overview of several possible estimation techniques that allow us to evaluate lower-level mediation where the outcome is binary (with a specific focus on small cluster sizes). Additionally, we presented an extensive simulation study in which we assessed the impact of several data features on the convergence, relative bias, mean squared error, and coverage of the estimates of the various methods. Overall, we found that jointly modelling the mediator and the outcome provided the best performance measures (combined with a marginal approach to simulating the random effects), especially in the presence of unmeasured upper-level confounding of mediator and outcome. A separate modelling approach that centres the lower-level variables within-clusters and draws the random effects in a conditional way, comes in as a close second performance-wise, confirming the reports of Goetgeluk and Vansteelandt (2008); Brumback et al. (2010), who stated that although CWC no longer yields proper parameter estimates when the outcome is binary, the resulting bias is often small. Unsurprisingly, not centering the lower-level predictors provided very biased estimates in the presence of upper-level mediator-outcome endogeneity (irrespective of the assumed random effects distribution).

With these conclusions in mind, we must acknowledge several limitations to this manuscript. For one, we restricted our simulations to settings where the intracluster correlations for the mediator and outcome are identical, because allowing them to vary independently of each other would have incremented the duration of our simulation study and its computational demands by sixfold. Of course, as witnessed in our example data, unequal *icc*'s are often encountered in practice.

Two, there also exists a non-parametrical implementation of the algorithm that we described in the fourth step of the estimation process. This alternative assesses mediation based on bootstrapping mediator and outcome values, rather than on simulated draws from the estimated parameter distributions. However, this non-parametrical approach has not (yet) been implemented into Imai et al. (2010a)'s *mediation*-package for multilevel data structures. We also did not incorporate this procedure within our study, as the bootstrapping process takes up an enormous amount of time in multilevel samples. However, our (and Imai et al. (2010a)'s) sole reliance on a parametrical approach might provide suboptimal estimates for the causal mediation effects, especially when estimation procedures are used that are known to produce biased estimates (e.g., the uncentred approaches

when there is upper-level endogeneity of mediator and outcome).

Three, we only considered complete data in our simulation study, as well as in our example data set, as missingness in either the mediator or the outcome will cause `Mplus` to produce error messages when the method of integration is specified as 'Gaussian' (i.e., as in Gaussian Adaptive Quadrature). This stands in contrast to the two separate modelling approaches in `R`, where all available outcomes are considered even when there is missingness.

With the current results and limitations in mind, future research might consider an important potential addition to Imai et al. (2010a)'s *mediation*-package. Unfortunately, to this day, we lack an easy-to-use software implementation that allows us to estimate the causal mediation effects in the presence of upper-level confounding of mediator and outcome. To this end, it would well be worthwhile investigating whether a joint approach can be implemented to achieve this, in linear as well as in binary settings.

Bibliography

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2):135–167.

Bauer, D. J., Preacher, K. J., and Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, 11(2):142–163.

Begg, M. D. and Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, 22(16):2591–2602.

Bind, M. A. C., Vanderweele, T. J., Coull, B. A., and Schwartz, J. D. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics*, 17(1):122–134.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brumback, B. A., Dailey, A. B., Brumback, L. C., Livingston, M. D., and He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters*, 80(21-22):1650–1654.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38(4):529–569.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780.

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, 21:389–395.

Hafeman, D. M. and Schwartz, S. (2009). Opening the Black Box: A motivation for the assessment of mediation. *International Journal of Epidemiology*, 38(3):838–845.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Josephy, H., Loeys, T., and Rosseel, Y. (2016). A Review of R-packages for random-intercept probit regression in small clusters. *Frontiers in Applied Mathematics and Statistics*, 2(18):1–13.

Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., and Loeys, T. (2015). Within-subject mediation analysis in AB/BA crossover designs. *The International Journal of Biostatistics*, 11(1):1–22.

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115–134.

Kenny, D. A., Korchmaros, J. D., and Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, 8(2):115–128.

Kenward, M. G. and Roger, J. H. (2010). The use of baseline covariates in crossover studies. *Biostatistics*, 11(1):1–17.

Kindt, S., Vansteenkiste, M., De Ruddere, L., Cano, A., and Goubert, L. (2018). "What should I do first?" The effect of manipulated goal conflict on affect, motivation and helping behavior in chronic pain. *Under review.*

Komárek, A. and Komárková, L. (2014). Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, 59(12):1–38.

Loeys, T., Moerkerke, B., De Smet, O., Buysse, A., Steen, J., and Vansteelandt, S. (2013). Flexible mediation analysis in the presence of nonlinear relations: Beyond the mediation formula. *Multivariate Behavioral Research*, 48(6):871–894.

Louis, T. A. (1988). General methods for analyzing repeated measures. *Statistics in Medicine*, 7(1-2):29–45.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Taylor & Francis Group, LLC, New York.

McMahon, J. M., Tortu, S., Torres, L., Pouget, E. R., and Hamid, R. (2003). Recruitment of heterosexual couples in public health research: a study protocol. *BMC medical research methodology*, 3:24.

Montoya, A. K. and Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, 22(1):6–27.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus User's Guide.* Muthén & Muthén, Los Angeles, CA, sixth edition.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):638–645.

Neuhaus, J. M. and McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(5):859–872.

Ortqvist, A. K., Lundholm, C., Carlström, E., Lichtenstein, P., Cnattingius, S., and Almqvist, C. (2009). Familial factors do not confound the

association between birth weight and childhood asthma. *Pediatrics*, 124(4):e737–43.

Pan, W. (2002). A note on the use of marginal likelihood and conditional likelihood in analyzing clustered data. *The American Statistician*, 56(3):171–174.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):Article 7.

Pearl, J. (2012). The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–36.

Pituch, K. a. and Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research*, 41(4):630–670.

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(4):1–28.

Preacher, K. J., Zyphur, M. J., and Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3):209–233.

R Core Team (2013). R: A language and environment for statistical computing.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models. Applications and data analysis methods*. Sage, Thousand Oaks, CA, second edition.

Raykov, T. and Mels, G. (2007). Lower level mediation effect analysis in two-level studies: A note on a multilevel structural equation modeling approach. *Structural Equation Modeling*, 14(4):636–648.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):153–155.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M. and Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A*, 158(1):73–89.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rovine, M. J. and Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, 35(1):55–88.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58.

Senn, S. (2002). *Cross-over trials in clinical research*. John Wiley & Sons, Chichester.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, New York.

Skrondal, A. and Rabe-Hesketh, S. (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 63(2):211–237.

Snijders, T. and Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage, Thousand Oaks, CA.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Tofighi, D. and Kelley, K. (2016). Assessing omitted confounder bias in multilevel mediation models. *Multivariate Behavioral Research*, 51(1):86–105.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *The British Journal of Mathematical and Statistical Psychology*, 59(2):225–255.

VanderWeele, T. J. (2010a). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21(4):540–551.

VanderWeele, T. J. (2010b). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, 38(4):515–544.

VanderWeele, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

Vervoort, T. and Trost, Z. (2017). Examining affective-motivational dynamics and behavioural implications within the interpersonal context of pain. *The Journal of Pain*, 18(10):1174–1183.

Vuorre, M. and Bolger, N. (2017). Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behavior Research Methods*, pages 1–19.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. The MIT Press, Cambridge, MA.

Zhang, Z., Zyphur, M. J., and Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models. *Organizational Research Methods*, 12(4):695–719.

D Appendix

D.1 Identification of the causal effects in general settings

In equation (5.2) we worked with unspecified, known inverse link functions, g_M and g_Y. In the following sections, we derive the parametrical expressions for the causal effects when $g_M = g_Y = probit$, and when $g_M = g_Y = logit$.

D.1.1 *Probit*-regression models

Consider the following *probit*-models for both a binary mediator and outcome (with i the cluster, and j the measurement within a cluster):

$$E[M_{ij}|X_{ij}, U_i] = \Phi\big(\delta_M + \alpha X_{ij} + U_i\big)$$
$$E[Y_{ij}|X_{ij}, M_{ij}, V_i] = \Phi\big(\delta_Y + \zeta' X_{ij} + \beta M_{ij} + \phi X_{ij} M_{ij} + h(U_i)\big)$$

with Φ representing the standard normal cumulative distribution. Based on this data generating mechanism, the "ij-th"-specific *total natural indirect effect* can be identified, when the assumptions *(i)-(v)* from section 3.2 are satisfied:

$$\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))|U_i, t)]) \\
&= \sum_m \{E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]P(M_{ij} = m|X_{ij} = 1, U_i) \\
&\quad - E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]P(M_{ij} = m|X_{ij} = 0, U_i)\} \\
&= P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)(1 - P(M_{ij} = 1|X_{ij} = 1, U_i)) \\
&\quad - P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)(1 - P(M_{ij} = 1|X_{ij} = 0, U_i)) \\
&\quad + P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i)(P(M_{ij} = 1|X_{ij} = 1, U_i) \\
&\quad - P(M_{ij} = 1|X_{ij} = 0, U_i)) \\
&= P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)(P(M_{ij} = 1|X_{ij} = 0, U_i) \\
&\quad - P(M_{ij} = 1|X_{ij} = 1, U_i)) + P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i) \\
&\quad (P(M_{ij} = 1|X_{ij} = 1, U_i) - P(M_{ij} = 1|X_{ij} = 0, U_i)) \\
&= (P(M_{ij} = 1|X_{ij} = 0, U_i) - P(M_{ij} = 1|X_{ij} = 1, U_i)) \\
&\quad (P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i) - P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i)) \\
&= \big(\Phi(\delta_M + U_i) - \Phi(\delta_M + \alpha + U_i)\big)\big(\Phi(\delta_Y + \zeta' + h(U_i)) \\
&\quad - \Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\big)
\end{aligned}$$

Similarly, the "ij-th"-specific *pure natural direct effect* can be identified:

$$\begin{aligned}
&E[Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&= \sum_m \{E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]P(M_{ij} = m|X_{ij} = 0, U_i) \\
&\quad - E[Y_{ij}|X_{ij} = 0, M_{ij} = m, U_i]P(M_{ij} = m|X_{ij} = 0, U_i)\} \\
&= (P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i) - P(Y_{ij} = 1|X_{ij} = 0, M_{ij} = 0, U_i)) \\
&\quad P(M_{ij} = 0|X_{ij} = 0, U_i) + (P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i) \\
&\quad - P(Y_{ij} = 1|X_{ij} = 0, M_{ij} = 1, U_i))P(M_{ij} = 1|X_{ij} = 0, U_i) \\
&= \big(1 - \Phi(\delta_M + U_i)\big)\big(\Phi(\delta_Y + \zeta' + h(U_i)) - \Phi(\delta_Y + h(U_i))\big) \\
&\quad + \Phi(\delta_M + U_i)\big(\Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i)) - \Phi(\delta_Y + \beta + h(U_i))\big)
\end{aligned}$$

Finally, the "ij-th"-specific *total causal effect* can be identified as well:

$$\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&= \sum_m \{E[Y_{ij}|X_{ij} = 1, M_{ij} = m, U_i]P(M_{ij} = m|X_{ij} = 1, U_i) \\
&\quad - E[Y_{ij}|X_{ij} = 0, M_{ij} = m, U_i]P(M_{ij} = m|X_{ij} = 0, U_i)\} \\
&= P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 0, U_i)P(M_{ij} = 0|X_{ij} = 1, U_i) \\
&\quad - P(Y_{ij} = 1|X_{ij} = 0, M_{ij} = 0, U_i)(P(M_{ij} = 0|X_{ij} = 0, U_i)) \\
&\quad + P(Y_{ij} = 1|X_{ij} = 1, M_{ij} = 1, U_i)(P(M_{ij} = 1|X_{ij} = 1, U_i)) \\
&\quad - P(Y_{ij} = 1|X_{ij} = 0, M_{ij} = 1, U_i)(P(M_{ij} = 1|X_{ij} = 0, U_i)) \\
&= \Phi(\delta_Y + \zeta' + h(U_i))\big(1 - \Phi(\delta_M + \alpha + U_i)\big) - \Phi(\delta_Y + h(U_i))\big(1 - \Phi(\delta_M + U_i)\big) \\
&\quad + \Phi(\delta_Y + \zeta' + \beta + \phi + h(U_i))\Phi(\delta_M + \alpha + U_i) - \Phi(\delta_Y + \beta + h(U_i))\Phi(\delta_M + U_i)
\end{aligned}$$

D.1.2 *Logit*-regression models

Consider the following *logit*-models for both a binary mediator and outcome:

$$\begin{aligned}
E[M_{ij}|X_{ij}, U_i] &= \frac{1}{1 + e^{-\delta_M - \alpha X_{ij} - U_i}} \\
E[Y_{ij}|X_{ij}, M_{ij}, U_i] &= \frac{1}{1 + e^{-\delta_Y - \zeta' X_{ij} - \beta M_{ij} - \phi X_{ij} M_{ij} - h(U_i)}}
\end{aligned}$$

Based on this data generating mechanism, the "ij-th"-specific *total natural indirect effect* can be identified, when the assumptions *(i)-(v)* from

section 3.2 are satisfied:

$$\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(1, M_{ij}(0))|U_i, t)] \\
&\quad = \left(\frac{1}{1+e^{-\delta_M - U_i}} - \frac{1}{1+e^{-\delta_M - \alpha - U_i}}\right)\left(\frac{1}{1+e^{-\delta_Y - \zeta' - h(U_i)}}\right. \\
&\quad \left. - \frac{1}{1+e^{-\delta_Y - \zeta' - \beta - \phi - h(U_i)}}\right)
\end{aligned}$$

Similarly, the "ij-th"-specific *pure natural direct effect* can be identified:

$$\begin{aligned}
&E[Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&\quad = \frac{e^{-\delta_M - U_i}}{1+e^{-\delta_M - U_i}}\left(\frac{1}{1+e^{-\delta_Y - \zeta' - h(U_i)}} - \frac{1}{1+e^{-\delta_Y - h(U_i)}}\right) \\
&\qquad + \frac{1}{1+e^{-\delta_M - U_i}}\left(\frac{1}{1+e^{-\delta_Y - \zeta' - \beta - \phi - h(U_i)}} - \frac{1}{1+e^{-\delta_Y - \beta - h(U_i)}}\right)
\end{aligned}$$

Finally, the "ij-th"-specific *total causal effect* can be identified as well:

$$\begin{aligned}
&E[Y_{ij}(1, M_{ij}(1)) - Y_{ij}(0, M_{ij}(0))|U_i, t)] \\
&\quad = \frac{1}{1+e^{-\delta_Y - \zeta' - h(U_i)}}\frac{e^{-\delta_M - \alpha - U_i}}{1+e^{-\delta_M - \alpha - U_i}} - \frac{1}{1+e^{-\delta_Y - h(U_i)}}\frac{e^{-\delta_M - U_i}}{1+e^{-\delta_M - U_i}} \\
&\qquad + \frac{1}{1+e^{-\delta_Y - \zeta' - \beta - \phi - h(U_i)}}\frac{1}{1+e^{-\delta_M - \alpha - U_i}} - \frac{1}{1+e^{-\delta_Y - \beta - h(U_i)}}\frac{1}{1+e^{-\delta_M - U_i}}
\end{aligned}$$

D.2 Identification of the causal effects in general settings

This appendix contains the R-code for the data generating mechanism and the implementation of the different methods discussed in this paper. Note that in these scripts, `y` represents the binary outcome, `x` the exposure, and `m` the mediator.

D.2.1 Data generating mechanism for *probit*- regression

The following R-code allows the generation of data with clusters of size two, a sample size 'n', a latent intracluster correlation for M and Y of 'icc', a correlation between the random intercepts of mediator and outcome of rho', and generated through the *probit*-link.

```
#Generate 1000 data sets for the current n, icc, and prev:
for (i in 1:1000){
  print(i);  set.seed(123456+i)
  #Population parameters:
  iM<-0; ia<-1; iY<--0.7; ic<-0.5; ib<-0.8
  #Random intercept covariance matrix (with tau<-icc/(1-icc)):
```

```
  sig<-matrix(c(tau,sqrt(tau)*sqrt(tau)*rho,
          sqrt(tau)*sqrt(tau)*rho,tau),byrow=T,nrow=2)
  #Random intercepts for M and Y within each cluster:
  ri<-mvrnorm(n,c(0,0),sig)

  #Generate data for binary X, M and Y:
  x0<-rbinom(n,1,0.5);  x1<-1-x0
  m0<-rbinom(n,1,pnorm(iM+ia*x0+ri[,1]))
  m1<-rbinom(n,1,pnorm(iM+ia*x1+ri[,1]))
  y0<-rbinom(n,1,pnorm(iY+ic*x0+ib*m0+ri[,2]))
  y1<-rbinom(n,1,pnorm(iY+ic*x1+ib*m1+ri[,2]))

  #Centring of X and M within-clusters:
  xmean<-colMeans(rbind(x0,x1))
  xx0<-x0-xmean;  xx1<-x1-xmean
  mmean<-colMeans(rbind(m0,m1))
  mm0<-m0-mmean;  mm1<-m1-mmean

  #Convert the variables to long format:
  x<-c(x0,x1);  m<-c(m0,m1)
  xx<-c(xx0,xx1);  mm<-c(mm0,mm1)
  y<-c(y0,y1)
  #Cluster identifier:
  ind<-rep(seq(1,n),2)
  #Create dataset:
  data<-as.data.frame(cbind(ind,x,m,xx,mm,y)) }
```

D.2.2 Estimation models for mediator and outcome

For the uncentred separate modelling approach, by use of the `lme4`-package in `R` (Bates et al., 2015):

```
med.UN<-glmer(m~x+(1|ind),family=binomial(link="probit"),
         data=data,nAGQ=15)
out.UN<-glmer(y~x+m+(1|ind),family=binomial(link="probit"),
         data=data,nAGQ=15)
```

For the separate modelling approach centred within clusters, by use of the `lme4`-package in `R` (Bates et al., 2015):

```
med.CWC<-glmer(m~xx+(1|ind),family=binomial(link="probit"),
         data=data,nAGQ=15)
out.CWC<-glmer(y~xx+mm+(1|ind),family=binomial(link="probit"),
         data=data,nAGQ=15)
```

For the joint modelling approach in `Mplus` (Muthén and Muthén, 2010):

```
DATA:
  file = mplus.raw;  type = individual;
VARIABLE:
  names = x0 x1 m0 m1 y0 y1;  usevariables = x0 x1 m0 m1 y0 y1;
  missing = .;  categorical = m0 m1 y0 y1;
ANALYSIS:
  type = general;  estimator = ML;  integration= GAUSS;
  adaptive = on;  link = probit;
MODEL:
  i0 BY m0@1 m1@1;  i1 BY y0@1 y1@1;  i0 (Mvar);  i1 (Yvar);
  m0 ON x0 (a);  m1 ON x1 (a);
  y0 ON x0 (c);  y1 ON x1 (c);  y0 ON m0 (b);  y1 ON m1 (b);
  [m0$1] (iM);  [m1$1] (iM);  [y0$1] (iY);  [y1$1] (iY);
OUTPUT:
  sampstat cinterval tech3;
```

D.2.3 Generation of the random effects

For a marginal generation of the random effects, based on the uncentred separate modelling approach and the *probit*-link:

```
#Extract the estimates and estimated covariance matrix:
b_est<-c(fixef(med.UN),out.UN)
b_vcov[c(1:2),c(1:2)]<-as.matrix(vcov(med.UN))
b_vcov[c(3:5),c(3:5)]<-as.matrix(vcov(out.UN))
#Extract the estimated random intercept variances:
ri_varM<-med.UN@theta**2, ri_varY<-out.UN@theta**2

#Simulate draws from the sampling distribution:
b_sim<-mvrnorm(1000,b_est,b_vcov)
#Simulated draws:
for (t in 1:1000){
 set.seed(12345+t)
 riMm<-rep(rnorm(n,mean=0,sd=ri_varM),each=2)
 riYm<-rep(rnorm(n,mean=0,sd=ri_varY),each=2)
 m_0m<-rbinom(2*n,1,pnorm(b_sim[t,1]+riMm))
 m_1m<-rbinom(2*n,1,pnorm(b_sim[t,1]+riMm+b_sim[t,2]))
 y00m<-pnorm(b_sim[t,3]+riYm+b_sim[t,5]*m_0m)
 y11m<-pnorm(b_sim[t,3]+riYm+b_sim[t,4]+b_sim[t,5]*m_1m)
 y10m<-pnorm(b_sim[t,3]+riYm+b_sim[t,4]+b_sim[t,5]*m_0m)
 ie_m[t]<-mean(y11m-y10m)
 de_m[t]<-mean(y10m-y00m)
```

```
 te_m[t]<-mean(y11m-y00m)  }

#Marginal Causal effects:
ie[i,2]<-mean(ie_m); de[i,2]<-mean(de_m); te[i,2]<-mean(te_m)
```

For a conditional generation of the random effects, based on the uncentred separate modelling approach and the *probit*-link:

```
#Extract the parameter estimates and estimated covariance matrix:
b_est<-c(fixef(med.UN),fixef(out.UN))
b_vcov[c(1:2),c(1:2)]<-as.matrix(vcov(med.UN))
b_vcov[c(3:5),c(3:5)]<-as.matrix(vcov(out.UN))
#Extract the estimated conditional means and random intercept vars:
ri_meanM<-ranef(med.UN)[[1]][,1]
ri_meanY<-ranef(out.UN)[[1]][,1]
ri_cond_varM<-cond.se(med.UN)[[1]][,1]
ri_cond_varY<-cond.se(out.UN)[[1]][,1]

#With the function to extract the conditional standard errors:
cond.se<-function(object){
  se.bygroup<-ranef(object,condVar=T)
  vars<-attr(se.bygroup[[1]],"postVar")
  se.by.clust[[1]]<-array(NA,c(n,1))
  for (j in 1:n){
    se.by.clust[[1]][j,]<-sqrt(diag(as.matrix(vars[,,j])))  }
  return(se.by.clust)}

#Simulate draws from the sampling distribution:
b_sim<-mvrnorm(1000,b_est,b_vcov)
#Simulated draws:
for (t in 1:1000){
 set.seed(12345+t)
 riMc<-rep(rnorm(n,mean=ri_meanM,sd=ri_cond_varM),each=2)
 riYc<-rep(rnorm(n,mean=ri_meanY,sd=ri_cond_varY),each=2)
 m_0c<-rbinom(2*n,1,pnorm(b_sim[t,1]+riMc))
 m_1c<-rbinom(2*n,1,pnorm(b_sim[t,1]+riMc+b_sim[t,2]))
 y00c<-pnorm(b_sim[t,3]+riYc+b_sim[t,5]*m_0c)
 y11c<-pnorm(b_sim[t,3]+riYc+b_sim[t,4]+b_sim[t,5]*m_1c)
 y10c<-pnorm(b_sim[t,3]+riYc+b_sim[t,4]+b_sim[t,5]*m_0c)
 ie_c[t]<-mean(y11c-y10c)
 de_c[t]<-mean(y10c-y00c)
 te_c[t]<-mean(y11c-y00c)  }
```

```
#Conditional causal effects:
ie[i,1]<-mean(ie_c); de[i,1]<-mean(de_c); te[i,1]<-mean(te_c)
```

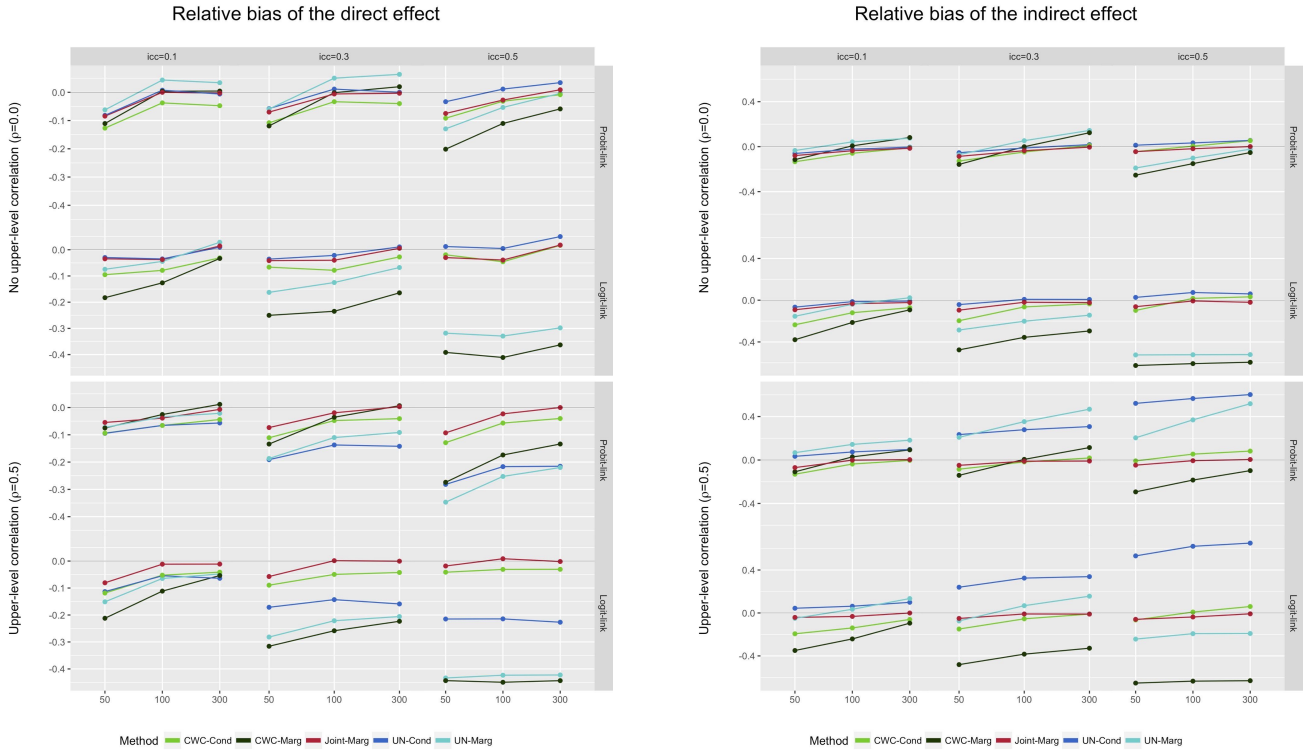

Figure 5.6 The relative bias of the direct (left) and indirect (right) for the five approaches modelled within clusters of size two, for different upper-level correlations between the random intercepts (zero or 0.5), link-functions (*probit* and *logit*), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

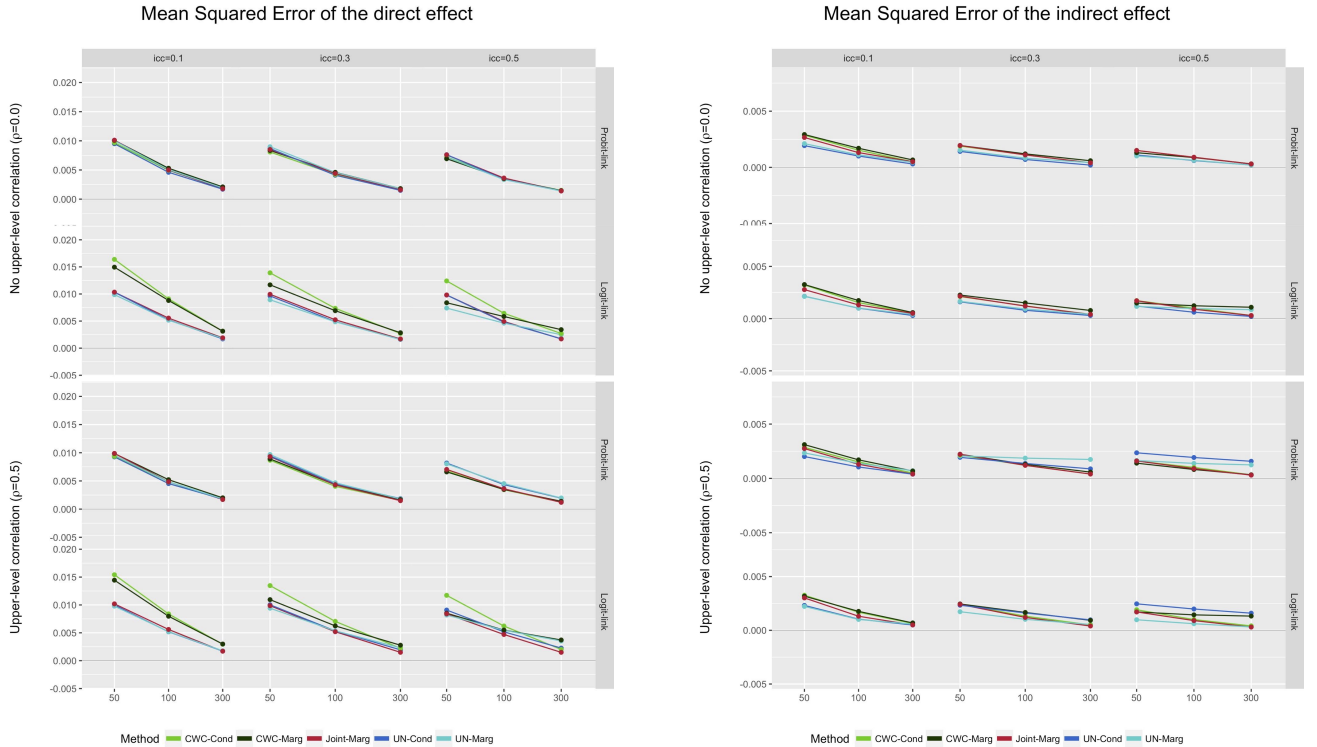

Figure 5.7 The MSE of the direct (left) and indirect (right) for the five approaches modelled within clusters of size two, for different upper-level correlations between the random intercepts (zero or 0.5), link-functions (*probit* and *logit*), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

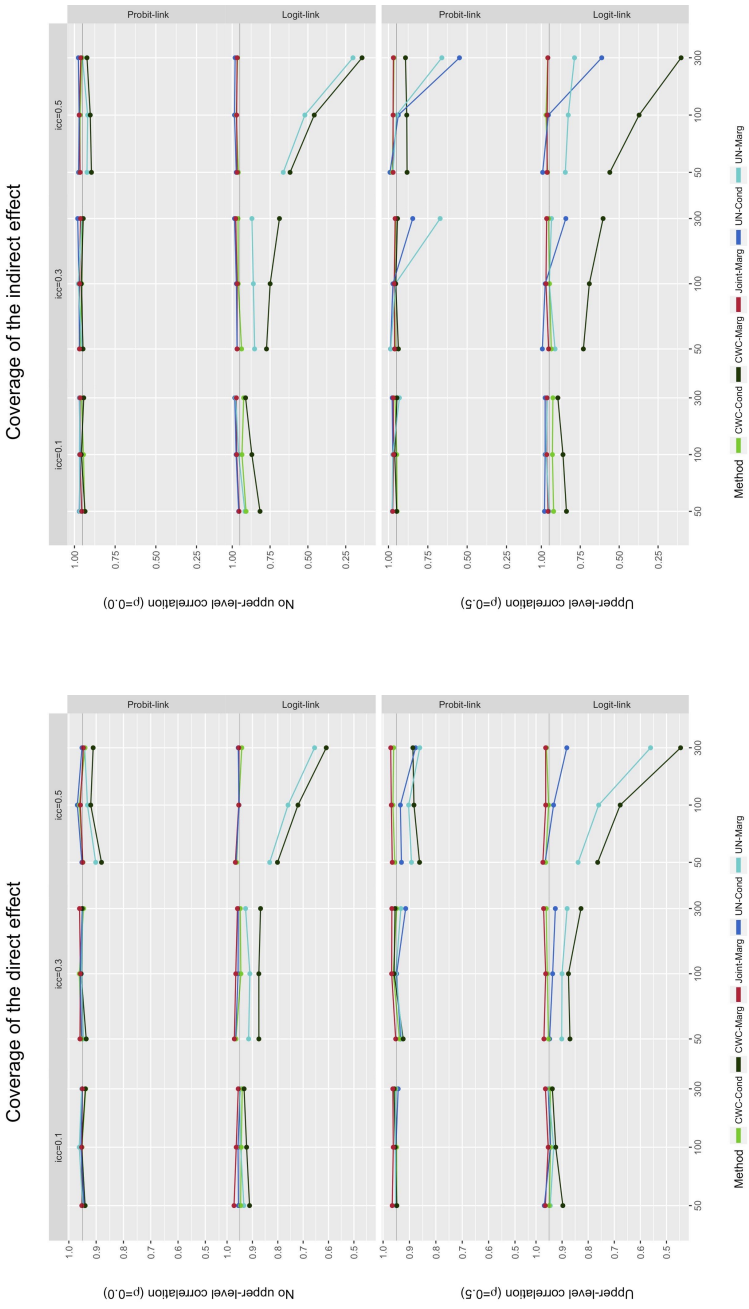

Figure 5.8 The coverage of the direct (left) and indirect (right) for the five approaches modelled within clusters of size two, for different upper-level correlations between the random intercepts (zero or 0.5), link-functions (*probit* and *logit*), different icc_l for the mediator and outcome (0.1, 0.3 and 0.5), and sample sizes (50, 100 and 300). These results stem from simulation runs where all methods converged.

6 General discussion

1 General Overview

With this thesis, we aim to provide applied researchers with a tangible set of guidelines on how to assess multilevel mediation in within-subject designs, when confronted with one of several issues. These items include (1) dealing with unmeasured upper-level confounding of the mediator-outcome relation, (2) appropriate inclusion and assessment of lower-level mediation in the presence of interaction terms, (3) assessing mediation in multilevel settings with binary measures, and (4) exploring which estimation techniques provide the best overall performance (i.e., in terms of bias and efficiency) under a broad variety of settings (with a special focus on small cluster sizes).

In chapter 2, we discussed multilevel mediation in *linear settings* within crossover designs, where a mere two observations within each cluster are observed. More specifically, we tried to address the first and fourth above-raised issues, by deriving expressions for the direct and indirect effect based on the counterfactual framework. In doing so, we were able to demonstrate that, in the presence of upper-level confounding of the mediator-outcome relationship, the intervening effect can be identified in some statistical models, but *not* in others. When multilevel mediation was considered within linear settings (i.e., with a continuous mediator and outcome) we revised three possible ways in which researchers can model the mediator and the outcome. A first possibility consisted of separately modelling the mediator and the outcome through the use of multilevel models. Unfortunately, as soon as upper-level endogeneity of the mediator and the outcome is present, this option may result in biased estimates for the parameter coefficients in the model for the outcome. A second option was found in within-cluster centering the lower-level predictors in the outcome equation, or equivalently, through the difference-approach. Such within-cluster centering can be achieved by separating both the exposure and the mediator into a within- and a between-cluster part. Since these within-cluster parts no longer contain any upper-level variation, they will be uncorrelated with *all* upper-level variables: measured *and* unmeasured. As such, the within-cluster regression coefficients of exposure and mediator in the model for the outcome will be estimated without bias, even in the presence of upper-level endogeneity of mediator and outcome. A third possibility was found in an approach that models the mediator and the outcome jointly. Doing so, allowed us to estimate a covariance term between the random intercepts of both mediator and outcome, which in

turn indirectly captured any unmeasured cluster-level common causes of both variables. In chapter 2, we saw that the first approach was unequipped to deal with upper-level endogeneity of mediator and outcome. In contrast, the latter two approaches were able to unbiasedly estimate the causal mediation effects in the presence of such endogeneity, although the joint approach required a slightly more stringent set of assumptions. For the latter method, we find bias in its parameter estimates when either (1) the mediator values are non-normally distributed, (2) when the random intercept of the outcome is non-normal, or when (3) the random intercept for Y interacts with the mediator in the model for the outcome. Because of the stronger assumptions required for the joint modelling approach, we recommend the use of within-cluster centring of lower-level predictors to deal with upper-level endogeneity of the mediator and outcome in linear settings.

In chapter 3, we continued exploring the first and fourth above-raised issues within *linear within-subject mediation settings*, but now additionally focussed on the inclusion of *lower-level interaction terms* (i.e., the second issue mentioned above). In the previous paragraph, we concluded that separating the lower-level predictors into a within- and a between-cluster part constitutes the least restrictive and most appropriate way of modelling (linear) lower-level mediation in the presence of upper-level endogeneity of mediator and outcome. However, including an interaction between the exposure and mediator within the model for the outcome, introduces some questions concerning the correct way of centering this lower-level product term. A first option considers a model where both the exposure and mediator are first centred within-clusters, after which the interaction term is defined by multiplying these two centred variables (C1P2, first *c*entre and take the *p*roduct term next). A second possibility first multiplies both the exposure and the mediator, and only afterwards centres this product term within-clusters (P1C2, take *p*roduct first and *c*entre second). As such, chapter 3 focussed on the fourth issue by comparing the performance of estimation techniques that differ in their centring of lower-level interactions. We observed that unmeasured upper-level endogeneity of the mediator-outcome relation may lead to biased parameter estimates for the interaction, when the lower-level variables are centred according to C1P2 (unless *all* cross-level interaction are included). In addition, P1C2 also provided more precise estimates of the interaction effect, compared to C1P2. Consequently, when dealing with a within-subject moderated mediation model in linear settings, we advocate to first multiply the values of lower-level predictors,

and to only then apply within-cluster centering.

The third issue we mentioned at the beginning of this section, considers the assessment of lower-level mediation when switching from continuous to *binary settings*. To address this, however, we first needed to figure out which estimation techniques (and implementation procedures) provide unbiased and efficient estimates when modelling dichotomous outcomes in small clusters. As such, in order to address issue number four in binary settings, chapter 4 temporarily digressed from multilevel mediation in preparation for chapter 5. In the introduction, we saw that binary responses are typically modelled through the aid of GLMMs, but unfortunately, the marginal likelihood functions of these models prove analytically intractable. To tackle this, researchers can either resort to a likelihood-based method by approximating the integrand (e.g. the Laplace approximation, Penalised Quasi-Likelihood), to approximate the the integral itself (e.g., Adaptive Gaussian Quadrature or AGQ), or through a Bayesian approach (e.g., Markov Chain monte Carlo methods, hybrid models based on integrated nested Laplace approximations). Alternatively, it is also possible to turn form GLMMs in general and resort to diagonally weighted least squares estimation (DWLS) within structural equation models. To evaluate the performance of these different techniques, we provided an overview of several `R`-based packages that are able to fit random intercept *probit*-models and subsequently presented an extensive simulation study. We found that AGQ and DWLS-estimation performed best when considering cluster sizes with only two measurements, with AGQ clearly taking the upper hand as the cluster size increases. As AGQ best withstood our performance assessment, it is consequently chosen to lay the foundations of chapter 5 .

In chapter 5 we redirected our attention to *multilevel mediation*, with our continued focus on *binary, rather than continuous outcome measures*. More specifically, in chapter 5, we aimed to address issue number one, three, and four by estimating the causal mediation effects in four consecutive steps: (1) providing non-parametrical expressions for the direct and indirect effect in binary settings (alongside the assumptions required for their identification), (2) identifying these effects parametrically, based on appropriate statistical models, (3) considering estimation models for the mediator and the outcome, and (4) estimating the causal mediation effects through an imputation algorithm that samples counterfactual outcomes. From the previous paragraph, we concluded that GLMMs with AGQ provided the best overal performance when modelling binary outcomes in

small clusters. Hence, we decided on comparing several estimation models for the mediator and outcome within the third step, based on this approximation. Similar to linear mediation settings, we assessed three different techniques: a separate modelling approach that does not centre the lower-level predictors, a separate modelling approach than centres these predictors within-clusters, and a method that models the mediator and outcome jointly. During the fourth step in our estimation process we additionally evaluated two different ways of generating random intercepts: a marginal versus a conditional approach. Employing simulation studies allowed us to evaluate the performance of the different methods during steps three and four, under a broad variety of settings. These results suggest that jointly modelling the mediator and binary outcome, combined with a marginal generation of the random effects, generally provided the most reliable results, even in the presence of upper-level endogeneity of mediator and outcome.

Since it is impossible to completely rule out the presence of unmeasured upper-level confounders of the mediator and outcome in within-subject mediation settings, we advise to always act as if such confounding is indeed present. We also caution applied scientists towards careful consideration of estimation techniques and implementations when cluster sizes are small, as such settings have proven straining on the available methodologies. In summary, we conclude this thesis with the following suggestions:

- We strongly *advise against* assessing mediation through the use of a separate modelling approach that does *not centre* the lower-level predictors within-clusters. This approach is, in both linear or non-linear within-subject mediation settings, unequipped to deal with upper-level endogeneity of mediator and outcome.

- In *linear settings*, we recommend researchers to centre all lower-level predictors within-clusters, as to correctly assess mediation in the presence of upper-level endogeneity of mediator and outcome.

- When researchers wish to include *lower-level interactions* into their linear within-subject mediation model, we advise to first multiply the corresponding variables and to only then centre this product term within-clusters.

- In *binary* settings, we encourage the use of Generalised Linear Mixed models, where the integral of the likelihood function is approximated

through Adaptive Gaussian Quadrature. This suggestion becomes increasingly substantial as the lower-level cluster size decreases.

- Moreover, when assessing *binary within-subject mediation*, we advise to jointly model the mediator and outcome, as to correctly assess the intervening effect in the presence of upper-level endogeneity of mediator and outcome. We do not recommend the use of within-cluster centring, as this approach will no longer yield unbiased parameter estimates when the outcome is binary.

2 Limitations and Future Research

Throughout this thesis, we looked at multilevel mediation from a counterfactual point of view, defining the causal mediation effects in terms of so-called 'potential outcomes'. We only considered *natural* direct and indirect effects, glossing over the concept of a *controlled* direct effect (Robins and Greenland, 1992; Pearl, 2001). For a dichotomous exposure in single-level settings, this controlled direct effect of exposure on outcome (controlling for the mediator), can be defined by: $E[Y(1,m) - Y(0,m)]$. It expresses the effect of exposure on outcome, if the mediator would be fixed at level m in the entire population. This controlled direct effect in fact requires a smaller set of assumptions for its identification, compared to the natural effects we considered within our chapters:

(i) A consistency assumption: for measurement moments within subjects with observed exposure level $X_{ij} = x$ and observed mediator $M_{ij} = m$, the observed outcome Y_{ij} equals the potential outcome $Y(x,m)$.

(ii) There are no unmeasured upper- or lower-level confounders of the association between exposure and outcome.

(iii) There are no unmeasured upper- or lower-level confounders of the association between mediator and outcome.

Unfortunately, an indirect effect cannot be defined in a similar *controlled* manner, as it is impossible to keep a set of variables fixed in a way that would exclude the direct effect (Pearl, 2001). On top of this, it might often not be realistic to force the mediator to a specific value for all measurements within all individuals. Because of these limitations, we instead focussed on *natural* direct and indirect effects in this thesis. For example, the (pure) natural direct effect is defined as $E[Y_{ij}(1, M_{ij}(0)) - Y_{ij}(0, M_{ij}(0)]$, where $Y_{ij}(x, M_{ij}(x^*))$ represents the value for the outcome Y_{ij}, when X_{ij} is set to x and M_{ij} is fixed at the value it would obtain when $X_{ij} = x^*$ (see chapter 2). However, the identification of these natural effects requires several additional assumptions:

(iv) A composition assumption: for measurement moments within subjects with observed exposure level $X_{ij} = x$, the observed outcome Y_{ij} equals the potential outcome $Y(x, M_{ij}(x))$.

(v) There are no unmeasured upper or lower-level confounders of the association between exposure and mediator.

(vi) There are no confounders of the association between mediator and outcome, caused by exposure (i.e., no intermediate confounding).

As should be quite clear by now, the upper-level part of assumption (iii) constituted the primary focus of this thesis. With respect to this assumption, our most important conclusion reported that bias induced by unmeasured upper-level confounders of mediator and outcome may be corrected for by some modelling strategies, but not by others. The lower-level chunk of this assumption was partly addressed within chapter 2, where we proposed a sensitivity analysis to asses the impact of possible lower-level confounders of mediator and outcome on the estimated causal mediation effects in linear settings. Equivalently, future work might attempt to evaluate the effect of such lower-level confounders within binary within-subject mediation settings, as introduced in chapter 5. We additionally postulated that the effect of upper-level confounders of mediator and outcome exert an additive effect on both variables. As such, studying the impact of nonlinear effects of unmeasured confounders (such as interactions or quadratic terms) may also prove rewarding as a future research topic. On top of this, we limited our research to random intercept models without any random slopes. The main reason for this was that small clusters sizes often limit the number of random effects that can be identified within multilevel models. Hence, it would be extremely interesting to extend our work towards more general settings that include random slopes for the lower-level predictors (and hence also consider a lower-level sample size that is appropriate 'large').

Similar to our focus on assumption (iii), the evaluation of the plausibility as well as the possible violation of the remaining assumptions may entail an important part of possible future studies. As we limited our research to within-subject designs with a randomised exposure, we hence automatically placated assumptions (ii) and (v). With this in mind, our work could be regarded as a first preparatory step towards more general longitudinal settings within observations studies. In observational research, we may not only be confronted with unmeasured confounding of mediator and outcome, but also with unmeasured confounders concerning the exposure. As some methods have proven able to appropriately deal with upper-

level endogeneity of mediator and outcome, this line of thinking may be extended towards unmeasured upper-level confounders of the exposure and mediator, or of the exposure and outcome. Within longitudinal studies, researchers may often be confronted with the risk of causal transience across time points. The most easy way to counter such lingering effects over time, is to incorporate a sufficiently long washout period between individual measurements within the study design itself. When such a carry-over effect cannot be excluded by design, however, the mediator or outcome values from the first measurement occasion may influence the mediator and/or outcome values of consecutive time points. When this is the case, we are confronted with exposure-dependent confounders (i.e., the mediator or outcome measure at the first measurement occasion) of the mediator-outcome relation, hence directly violating assumption (vi). When this is the case, we will no longer be able to identify the natural direct and indirect effects. Note that the *controlled* direct effect can still be identified in these settings, through the use of e.g. G-estimation or structural equation models. It may therefore prove very interesting to investigate the performance of such methodologies within (binary) lower-level mediation settings, when there is additional unmeasured upper-level confounding of mediator and outcome.

Finally, setting the above-mentioned assumptions aside, we think it might also prove worthwhile to invest in an application that enables researchers to evaluate and estimate causal mediation effects within lower-level settings. This software implementation would have to incorporate approaches that are able to deal with possible unmeasured upper-level confounding of mediator and outcome, in linear as well as in binary settings. Additionally, it might also facilitate unbiased and precise estimation of moderated mediation effects, irrespective of the measurement levels of the exposure, the mediator, or the outcome. Ideally, this program would also include appropriate sensitivity analyses, which ought to enable researchers to evaluate the impact of possible lower-level confounders of mediator and outcome on their inferences concerning the causal mediation effects.

Bibliography

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):153–155.

7 English summary

1 Introduction

When a clinical experiment establishes an effect of an intervention X on an outcome Y, researchers often wonder which underlying processes make up this effect. More specifically, we can ask ourselves whether this effect can be ascribed to an underlying process than runs indirectly from X to Y, through an intermediary measure M. This constitutes the kind of question mediation analyses attempt to answer. If mediation is indeed present, the total effect of intervention on outcome can be partitioned into an indirect effect that runs through the mediator M, and a possible lingering direct effect of X on Y. As science is generally enormously interested in the existence of and search for such explanatory processes, researchers have studied this phenomenon quite extensively during the last couple of decades. In the eighties, Baron and Kenny (1986) came up with a relatively straightforward approach to assess mediation, in a manuscript that makes up one of the most cited social science papers of all time. But even though mediation analysis has come a long way since their groundbreaking work (Pearl, 2001; Imai et al., 2010; Pearl, 2012), a lot of questions still remain.

One such stingy subject entails the extension of mediation to multilevel data structures, mainly because single-level techniques assume independent observations; an assumption which is clearly violated when we deal with clustered or multilevel data. Most often, hierarchical data consist of two levels (often referred to as level-1 and level-2), where level-1 units or lower-level units are nested within level-2 or upper-level units. Examples of such data structures entail students grouped within classes or repeated measurements nested within individuals. When we extend single-level mediation to nested data structures, we are able to discern three different models: 2-2-1, 2-1-1, and 1-1-1 mediation. The first type of multilevel mediation occurs when the effect of an upper-level exposure on a lower-level outcome is explained by an upper-level mediator. The second type of mediation, on the other hand, exhibits an exposure at the upper-level, while both the mediator and the outcome are defined at the lower-level. The last type of multilevel mediation occurs when all three variables, the exposure, the mediator, and the outcome are defined at the lower-level. The latter type is alternatively termed within-subject mediation, when the upper-level constitutes the individual. This model brings along several additional challenges, as longitudinal measurements are not only influenced by other time-dependent measures, but also by properties that remain constant

over time.

Because of the additional complexity provided by 1-1-1 or within-subject mediation models, this type of multilevel mediation will be the major focus of this thesis. We want to provide applied researchers with a concrete set of guidelines on how to assess within-subject mediation, when confronted with one of several issues. These issues include (1) dealing with unmeasured upper-level confounding of the mediator-outcome relation, (2) appropriate inclusion and assessment of multilevel mediation in the presence of lower-level interactions, (3) assessing mediation in multilevel settings with binary measures, and (4) exploring which estimation techniques provide the best overall performance (i.e., in terms of bias and efficiency) under a broad variety of settings (with a special focus on small sample sizes).

2 Chapter 2

In chapter 2 we take a closer look at one specific subcategory of within-subject mediation, namely mediation within crossover designs in linear settings. In this type of studies, each participant is observed exactly twice: once under exposure A and once under exposure B. Since two subsequent observations within the same individual are usually correlated with each other, this design exhibits a multilevel structure where the subject is considered the upper-level and the (repeated) measurements within an individual constitute the lower-level. On top of this, crossover studies administer both treatments in a randomised order, implying that every participant is assigned to one of two possible intervention series: sequence AB or sequence BA. Consequently, this type of designs are often referred to as AB/BA crossover studies.

Unfortunately, the currently available methodologies aimed at assessing mediation in AB/BA studies are limited to the approach suggested by Judd et al. (2001). And even though this method offers a simple and elegant way of testing for within-subject mediation, researchers have come up with a few points of criticism. For one, this method does not provide a definite estimate for the indirect effect; it merely answers the question of whether or not mediation has occurred. Moreover, this question is answered through an adaptation of Baron and Kenny (1986)'s causal steps approach, which is not without criticism itself. Apparently, this approach has been attributed a low power in detecting the indirect effect (Hayes, 2009), while the necessity of some of its constituent steps have been questioned (Collins

et al., 1998; MacKinnon et al., 2000; Preacher et al., 2007; Zhao et al., 2010). Other shortcomings of the method suggested by Judd et al. (2001), entail that it but allows for one type of moderation (namely, an interaction between the intervention X and the mediator M) and that possible period- and carry-over effects cannot be taken into account (Tucker-Drob, 2011). Finally, as is the case for Baron and Kenny (1986)'s causal steps, underlying assumptions concerning measured and unmeasured confounders are not clearly explicated.

To counter the existing limitations in current literature, chapter 2 aimed to investigate mediation analysis in crossover studies from a counterfactual point of view. This enabled us to formulate non-parametrical expressions for the direct and indirect effect in 1-1-1 designs, based on so-called 'potential outcomes'. These formulas in turn allowed us to shed some light on the assumptions needed to identify the causal mediation effects. When we subsequently focussed on a data generating mechanism that satisfied these assumptions, we were able to draw up parametric expressions for the direct and indirect effect, based on Pearl (2001)'s mediation formula. In this way, different data generating mechanisms (ranging from simple to complex settings with different kinds of interactions) enabled us to contrast the performance of different statistical models during a simulation study. This comparison included a 'naive' Linear Mixed Model (LMM) without centring, an LMM with centring within-subjects, and an LMM approach that models the mediator and the outcome jointly. Apart from these three approaches, we also implemented a technique that is based on regressing the difference scores of the outcome on the difference scores of the mediator (i.e., an adaptation of Judd et al. (2001)'s method), while at the same time allowing for period effects and different types of moderation.

Our simulation studies revealed that the 'naive' approach cannot adequately handle unmeasured upper-level confounding of the mediator-outcome relationship. Since the absence of such confounders can never be guaranteed, we strongly advise against the use of this method. The other statistical models were able to unbiasedly estimate both the direct and indirect effect, even in the presence of unmeasured time-independent confounders of mediator and outcome. On top of this, we were able to establish that the joint procedure requires a slightly more stringent set of assumptions, compared to the centring- and difference-score methods (both of which usually yield identical estimators). Finally, we illustrated these conclusions with a neurostimulation experiment, after which we

proposed a sensitivity analysis that attempted to assess the impact of possible lower-level confounders of M and Y, on the direct and indirect effect.

3 Chapter 3

In chapter 3, we continue focussing on the different centring techniques within LMMs. Basically, multilevel data can be centred according to one of three possible techniques: either no centring is employed, data are grand-mean centred, or centring is applied within clusters. Within current multilevel literature, there exists a general consensus that centring within clusters is best suited when researchers' main interests lie with the effects of lower-level predictors. This recommendation is in accordance with the results from the previous chapter, where we saw that an uncentred (or grand-mean centred) approach provided biased estimates in the presence of unmeasured upper-level confounding of M and Y, in contrast to an approach where measures were centred within-subjects.

Unfortunately, discussions within multilevel literature on the role of centering are mostly limited to the assessment of main effects in multilevel models (MLM) and ignore the centering of interactions. An issue of particular importance entails the centering of interactions in a $1 \times (1 \rightarrow 1)$ design, where the first '1' corresponds to the level at which the moderator is measured, the second '1' represents the level of the exposure, while the last '1' defines the level of the outcome (Preacher et al., 2016; Ryu, 2015); we will refer to such moderated effects as 'lower-level interactions'. When cluster-mean centering these interaction terms, the question arises whether the exposure and moderator should be centred first and multiplied next (labeled as 'C1P2', centre-first and product-second), or whether it should be the other way around (labeled hereafter as 'P1C2'). This questions entails an important concern as, in contrast to an interaction between an upper- and a lower-level variable or between two upper-level variables, C1P2 and P1C2 produce diverging results when cluster-mean centering a lower-level product term.

Consequently, chapter 3 investigates whether these approaches can unbiasedly estimate a moderated within-subject effect of exposure on outcome. To better understand the performance of both techniques, we explored why and when those centering approaches perform differently by means of a simulation study. To this end, we looked at different settings where we investigated the relative bias of the estimators and standard

errors, as well as their coverage and power. We were able to determine that the estimators remained unbiased, as long as the predictor and moderator remained independent of each other. However, as soon as the moderator is affected by the exposure (i.e., as soon as it becomes a mediator), we detected bias in the C1P2 estimators of the interaction effect. An analytical determination of this bias demonstrated its dependence on the distribution of the exposure *and* the size of the effect of exposure on mediator: when X is binary, the bias inflates as the effect of exposure on moderator increases. On top of this, we were able to observe smaller standard errors for P1C2, compared to C1P2. Taking both conclusions into account, we advise to always multiply any level-1 predictors first and only afterwards centre their product term within clusters, because: (1) P1C2 results in more precise estimates of the interaction effect, and (2) P1C2 is not affected by misspecification or omission of upper-level effects, in contrast to C1P2 (unless all cross-level interactions are included).

To demonstrate these centring techniques, we illustrated our results on a longitudinal diary study on sexual behaviour in Flanders. More specifically, we focussed on male participants and investigated the effect of intimacy on next day positive relationship feelings, and to what extent this effect was moderated by masturbation.

4 Chapter 4

As chapters 2 and 3 solely focussed on within-subject mediation in linear settings, the next chapters aim to investigate whether 1-1-1 mediation is easily extendable to binary settings. However, before we are able to answer this question, we must attempt to figure out which estimation methods are able to unbiasedly and efficiently estimate a simple (unmediated) effect of an exposure on a binary outcome. This smoothly transitions us to 4, where Generalised Linear Mixed Models (GLMMs) take the center stage, rather than the LMMs from the previous chapters.

Although GLMMs are widely used to model clustered categorical outcomes, their statistical inference is hampered as integrating out the random effects from the likelihood function is, except for a few cases, analytically intractable. To tackle this, several techniques have been proposed, which can be roughly divided into two main classes: likelihood-based methods and Bayesian approaches. One way to tackle the intractability of the GLMM likelihood function, is to either approximate the integrand, as does the Laplace approximation (Tierney and Kadane, 1986) or Penalised

Quasi-Likelihood (PQL, Breslow and Clayton (1993); Schall (1991); Stiratelli et al. (1984)), or to approximate the integral itself by a finite sum, as in Adaptive Gaussian Quadrature (AGQ, Pinheiro and Bates (1995)). Bayesian methods, on the other hand, make use of Markov Chain Monte Carlo (MCMC) implementations, where the likelihood is simulated rather than calculated analytically, as to obtain the posterior distribution of the parameters of interest. As MCMC methods are known to be computationally intensive, hybrid models based on Integrated Nested Laplace Approximations (INLA) of the posterior marginals were proposed (Rue et al., 2009) .

As these approximations rarely yield satisfactory results when analysing binary outcomes within small clusters (Breslow and Clayton, 1993; McMahon et al., 2003), we proposed estimation within the Structural Equation Modelling (SEM) framework as an alternative. Although at first glance SEM and GLMM may seem like two different edifices, recent literature proves that SEM is completely equivalent to its GLMM counterpart in the absence of latent variables, and this under a broad set of conditions (Rovine and Molenaar, 2000; Curran, 2003; Bauer, 2003). Within the SEM-framework, there are two common estimation approaches for modelling binary outcomes: maximum likelihood (ML) estimation and (diagonally) weighted least squares (DWLS) (Skrondal and Rabe-Hesketh, 2004). As ML-estimation is not widely used within traditional SEM literature, but also proves equivalent to ML-estimation within GLMMs, this chapter puts an emphasis on DWLS.

Apparently, we are confronted with a myriad of options to estimate binary clustered outcomes: the Laplace approximation, AGQ, PQL, MCMC, and INLA within the GLMM framework, and robust DWLS estimation within SEM. But which method yields the best and most efficient estimators? To answer this question, we conducted an extensive simulation study in chapter 4, where we assessed the performance of six different `R`-packages for random-intercept probit regression (`R` version 3.2.3, R Core Team (2013)). To compare these methods as thoroughly as possible, we decided on varying a range of settings: we considered a cluster size of 2, 3, and 5, a sample size ranging from 25 to 300, a rare versus an average outcome prevalence, small, medium, and large latent intracluster correlations, as well as different types of predictors (continuous versus binary, and varying within- versus between subjects). Over these $3 \times 4 \times 3 \times 2 \times 4$ possible settings, we assessed the convergence, relative bias, mean squared error, and coverage of the six different estimation methods. To ascertain if the

conclusions hold, irrespective of software implementations, we reran some prominent settings by means of other programs, such as SAS® (version 9.4 (SAS Institute Inc, 2015)), MPLUS® (version 7 (Muthén and Muthén, 2010)), and JAGS (version 4.1.0. (Plummer, 2003)).

For clusters of size two, we were able to conclude that SEM usually performs best in terms of bias for the fixed and random effect estimators, while AGQ prevails in terms of precision (mainly because of SEM's robust standard errors). As the cluster size increases, however, AGQ becomes the best choice for both bias and precision. These results proved independent of the software used, with one notable exception: the MCMC implementation in R appeared to be suboptimal, compared to JAGS software.

Finally, these conclusions were highlighted by means of a dataset on eating habits of toddlers in Flemish nursery schools. In this study, we considered whether or not encouragement towards the eating of chicory (the intervention X) affected the children's disliking of the vegetable (the binary outcome Y).

5 Chapter 5

Although recent literature has devoted a lot of time and attention to expanding mediation to multilevel settings, such extensions were often limited to continuous outcome measures. Hence, in chapter 5, we attempt to address this issue by expanding 1-1-1 mediation to settings with a binary outcome. To this end, we continue and expand the preparatory work from chapter 4, where we demonstrated that GLMMs with AGQ provide the best estimators, when assessing the effect of an exposure on a binary outcome within small clusters.

Once again, we intend to focus on the consequences of upper-level endogeneity of M and Y, as such confounding may generate bias in the estimates of the regression coefficients, as well as those of the direct and indirect effect. As shown in chapters 2 and 3, in linear settings, bias due to unmeasured additive upper-level confounding of mediator and outcome is often remedied by separating lower-level predictors into a within- and a between-cluster component. However, as this solution is no longer valid when considering binary outcome measures (Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010), we need to search for a different solution when confronted with upper-level endogeneity of a mediator and binary outcome.

To assess the severity of this transgression, chapter 5 aims to tackle

lower-level mediation with a binary outcome from a counterfactual point of view, with a special focus on small cluster sizes. We proposed doing this through the evaluation of four consecutive steps. A first step offers non-parametric definitions of the causal mediation effects, as well as the assumptions needed for their identification. Within this step, we focussed on expressions defined on the linear scale, as to provide a counterfactual definition based on differences. A second step identifies the direct and indirect effect based on parametric models for the mediator and outcome. For these models, we considered two link functions that combine the binary outcome and linear predictor term: the probit and the logit link. A third step estimates the regression coefficients of the models for mediator and outcome, by use of three different estimation models: (1) an uncentred method that estimates the mediator and outcome separately, (2) a separate modelling technique that centres the predictors within subjects, and (3) a method that jointly models the mediator and outcome. Finally, a fourth step estimates the causal mediation effects themselves, by predicting potential outcomes for M and Y. This was done through a parametric algorithm, in which the posterior distributions of both variables are approximated by their sampling distribution. This predicting of random effects can be achieved in one of two ways: a marginal versus a conditional approach.

In summary, chapter 5 aimed to check which multilevel estimation models are capable of effectively eliminating unmeasured upper-level confounding of mediator and outcome, by the use of the four above-mentioned steps. In doing so, we focussed on a binary randomised exposure and a binary outcome within smal clusters. To verify this research question, we subsequently presented an extensive simulation study in which we compared three different estimation models (an uncentred, a centred, and a joint modelling approach), two link functions for the outcome (the logit and the probit link), and two ways of generating the random effects (marginally versus conditionally). Ensuring a relevant comparison of these different techniques, we varied a number of factors within our simulated datasets: the cluster size and sample size, the intracluster correlation, as well as the presence or absence of unmeasured upper-level confounding of mediator and outcome.

Overall, we found that jointly modelling the mediator and the outcome provided the best performance measures (combined with a marginal approach to simulating the random effects), especially in the presence of unmeasured upper-level confounding of mediator and outcome. A separate modelling approach that centres the lower-level variables within-clusters

and draws the random effects in a conditional way, comes in as a close second performance-wise. Unsurprisingly, not centering the lower-level predictors provided very biased estimates in the presence of upper-level mediator-outcome endogeneity (irrespective of the assumed random effects distribution).

To illustrate these results, we applied the different methods to a crossover study that assessed the impact of an induced goal conflict situation on the observed helping behaviour in partners of individuals with chronic pain. Additionally, we wanted to assess whether or not this causal effect was mediated by the partner's amount of autonomous helping behaviour, as perceived by the patients.

6 Discussion

With this thesis, we aim to provide applied researchers with a concrete set of guidelines on how to assess within-subject mediation, when confronted with: (1) unmeasured upper-level confounding of the mediator-outcome relation, (2) lower-level interactions, (3) binary outcome measures, and (4) challenging or demanding settings (e.g., small cluster sizes). As researchers can never rule out the presence of unmeasured upper-level confounders of the mediator and outcome in within-subject mediation settings, we advise to always act as if such confounding is indeed present. We also caution applied scientists towards careful consideration of estimation techniques and implementations when cluster sizes are small, as such settings have proven straining on the available methodologies.

In summary, we conclude this thesis with the following suggestions. One, we strongly advise against assessing mediation through the use of a separate modelling approach that does not centre the lower-level predictors within-clusters. This approach is, in both linear and non-linear within-subject mediation settings, unequipped to deal with the presence of possible upper-level endogeneity of mediator and outcome. Two, in linear settings, we recommend researchers to centre all lower-level predictors within-clusters, as to correctly assess mediation in the presence of unmeasured upper-level confounding of mediator and outcome. Three, when researchers wish to include lower-level interactions into their linear within-subject mediation model, we advise to first multiply the corresponding variables and only then centre this product term within-clusters. Four, in binary settings, we encourage the use of Generalised Linear Mixed models where the integral of the likelihood function is approximated through Adaptive

Gaussian Quadrature. This suggestion becomes increasingly substantial as the lower-level cluster size decreases. And finally, when assessing binary within-subject mediation, we advise to jointly model the mediator and binary outcome, as to correctly assess the intervening effect in the presence of upper-level endogeneity of mediator and outcome. We do not recommend the use of within-cluster centring in this setting, as this approach will no longer yield unbiased parameter estimates when the outcome is binary.

Concerning the current limitations of this thesis and possible directions for future research, we referred to the various assumptions that were postulated during the different chapters. Investigating the impact of the violations of each of these in turn, would prove a valuable addition to our work. Moreover, we strongly promote the construction of a software implementation which would enable applied researchers to unbiasedly assess within-subject (moderated) mediation in linear or binary settings, in the presence of upper-level endogeneity of mediator and outcome.

Bibliography

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2):135–167.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brumback, B. A., Dailey, A. B., Brumback, L. C., Livingston, M. D., and He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters*, 80(21-22):1650–1654.

Collins, L. M., Graham, J. J., and Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, 33(2):295–312.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38(4):529–569.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4):408–420.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115–134.

MacKinnon, D. P., Krull, J. L., and Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4):173–181.

McMahon, J. M., Tortu, S., Torres, L., Pouget, E. R., and Hamid, R. (2003). Recruitment of heterosexual couples in public health research: a study protocol. *BMC medical research methodology*, 3:24.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA, sixth edition.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Pearl, J. (2012). The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–36.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Plummer, M. (2003). JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20-22*, pages 1–10, Vienna, Austria.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1):185–227.

Preacher, K. J., Zhang, Z., and Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, 21(2):189–205.

R Core Team (2013). R: A language and environment for statistical computing.

Rovine, M. J. and Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, 35(1):55–88.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392.

Ryu, E. (2015). The role of centering for interaction of level 1 variables in multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4):617–630.

SAS Institute Inc (2015). Base SAS® 9.4 Procedures Guide, Fifth Edition.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, New York.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971.

Tierney, L. and Kadane, J. B. (1986). Acccurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tucker-Drob, E. M. (2011). Individual differences methods for randomized experiments. *Psychological Methods*, 16(3):298–318.

Zhao, X., Lynch Jr., J. G., and Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2):197–206.

8 Nederlandstalige Samenvatting

1 Inleiding

Wanneer een klinisch experiment een effect vertoont van een interventie X op een uitkomst Y, kan men zich afvragen welke onderliggende processen dit effect bepalen. Of specifieker uitgedrukt: kan het effect van X op Y (deels) worden toegeschreven aan een onderliggend effect dat loopt via een intermediaire meting M? Dit is het soort vraag waar mediatie-analyses pogen een antwoord op te geven. Indien er sprake is van mediatie, zal het totale effect van interventie op de uitkomst kunnen worden opgedeeld in een indirect effect dat loopt via de mediator M en een eventueel overblijvend direct effect van X op Y. Aangezien de wetenschap enorm geïnteresseerd is in het bestaan van en de zoektocht naar dergelijk verklarende processen, hebben onderzoekers dit fenomeen de voorbije decennia dan ook uitvoerig bestudeerd. Zo werkten Baron and Kenny (1986) een relatief eenvoudig stappen-plan voor mediatie uit, in een manuscript dat intussen één van de meest geciteerde werken uit de sociaal wetenschappelijke literatuur vormt. Hoewel methodologisch reeds een lange weg werd afgelegd sinds hun baanbrekende werk (Pearl, 2001; Imai et al., 2010; Pearl, 2012), resten er ons nog steeds enkele grote uitdagingen.

Eén zo'n uitdaging omvat de extensie van mediatie naar multilevel data, aangezien traditionele mediatie-technieken onafhankelijke observaties veronderstellen; een assumptie die duidelijk geschonden wordt in het geval van geclusterde of multilevel data. De meest voorkomende hiërarchisch gestructureerde data omvatten twee levels (vaak level-1 en level-2 genoemd), waarbij level-1 of onderste-level units genest zijn binnen level-2 of bovenste-level units. Voorbeelden van dergelijke hiërarchische structuren omvatten studenten gegroepeerd binnen klassen of herhaalde metingen genest binnen individuën. Wanneer men klassieke mediatie-analyse uitbreidt naar geneste data, kan men drie verschillende mediatie-modellen onderscheiden: 2-2-1, 2-1-1 en 1-1-1 mediatie. Het eerste type multilevel mediatie komt voor wanneer het effect van een level-2 interventie op een level-1 uitkomst gemediëerd word door een level-2 mediator. Het tweede type mediatie vertoont dan weer een interventie op het onderste level, terwijl zowel de mediator als de uitkomst gemeten zijn op level-2. Het laatste type mediatie komt dan weer voor wanneer alle drie de variabelen, zowel de interventie, de mediator, als de uitkomst, gedefiniëerd zijn op het onderste-level van de data. Dit laatste type wordt alternatief binnen-subject mediatie gedoopt, wanneer men mediatie beschouwt in longitudinale settings. Dit soort mediatie brengt extra complicaties met zich mee, aangezien longitu-

dinale metingen niet alleen beïnvloed worden door andere tijdsafhankelijke variabelen, maar evenzeer door eigenschappen die constant blijven over tijdspannes heen.

Omdat deze 1-1-1- of binnen-subject mediatie extra uitdagingen met zich meebrengt, zal deze setting de voornaamste focus van dit doctoraat vormen. Zo willen we toegepaste onderzoekers voorzien van een concrete set aan richtlijnen over hoe mediatie correct te evalueren en te schatten, wanneer ze geconfronteerd worden met enkele belangrijke kwesties. Deze uitdagingen omvatten ondermeer (1) ongemeten level-2 confounders van de M-Y relatie, (2) de aanwezigheid van onderste-level interacties, (3) een uitkomst gemeten op een binaire, eerder dan een lineaire schaal, en (4) het op de proef stellen van de betrokken schattingsmethoden (en hun implementaties) door verschillende eigenschappen van de data te laten variëren (met een speciale focus op kleine clustergroottes).

2 Hoofdstuk 2

In hoofdstuk 2 wordt er dieper ingegaan op één specifieke subcategorie van binnen-subject mediatie, namelijk mediatie binnen crossover designs in lineaire settings. In dit type studies worden er exact twee metingen afgenomen bij elk participerend individu: één meting onder interventie A en één onder interventie B. Aangezien twee opeenvolgende observaties binnen hetzelfde individu meestal afhankelijk zijn van elkaar, vertoont dergelijk design een multilevel structuur waarbij level-2 verwijst naar het individu, terwijl level-1 refereert naar de (herhaalde) metingen binnen dat subject. Bovendien worden in crossover studies beide behandelingen in een gerandomiseerde volgorde aangeboden, waardoor alle participanten worden toegewezen aan één van twee mogelijke interventie-armen: de sequentie AB of de sequentie BA. Vandaar dat naar dit soort designs ook vaak wordt verwezen als AB/BA-crossover studies.

Helaas is de methodologie voor mediatie-analyse in AB/BA studies binnen de bestaande literatuur beperkt tot de aanpak voorgesteld door Judd et al. (2001). En hoewel deze methode een eenvoudige en elegante manier aanreikt om binnen-subject mediatie te testen, bestaan er toch enkele belangrijke punten van kritiek. Ten eerste levert deze methode geen concrete schatter van het indirecte effect op; het levert slechts een antwoord op de vraag of er al dan niet mediatie aanwezig is. Daarboven wordt deze vraag beantwoord via een adaptatie van de causale-stappen methode van Baron and Kenny (1986), die zelf ook niet vrij is van kritiek.

Uit verschillende studies blijkt immers dat de causale-stappen methode een lage power heeft (Hayes, 2009) en wordt de noodzaak van sommige onderdelen uit hun stappen-plan in twijfel getrokken (Collins et al., 1998; MacKinnon et al., 2000; Preacher et al., 2007; Zhao et al., 2010). Andere beperkingen van de methode voorgesteld door Judd et al. (2001), houden in dat het slechts één type moderatie toestaat (namelijk, een interactie tussen de interventie X en de mediator M) en dat er geen rekening gehouden wordt met mogelijke periode- of carry-over effecten (Tucker-Drob, 2011). Tenslotte worden -net zoals in Baron and Kenny (1986)- onderliggende assumpties over gemeten en ongemeten confounders niet duidelijk geëxpliciteerd.

Om deze beperkingen in de huidige literatuur tegemoet te komen, onderzochten wij in hoofdstuk 2 mediatie-analyse in crossover studies binnen een 'tegenfeitelijk' denkkader. Zo konden we aan de hand van zogenaamde 'potentiële uitkomsten' niet-parametrische formules opstellen voor het directe en indirecte effect in 1-1-1-designs, wat meteen ook meer duidelijkheid schiep over de assumpties die nodig zijn om deze effecten te identificeren. Wanneer we vervolgens focusten op een data genererend mechanisme dat voldeed aan deze assumpties, konden we aan de hand van de mediatie-formule (Pearl, 2001) parametrische uitdrukkingen opstellen voor het directe en indirecte effect. Zo lieten verschillende data genererende mechanismes (gaande van eenvoudige tot complexe settings met verschillende soorten interacties) ons toe de performantie van statistische modellen te vergelijken in een simulatiestudie. In deze vergelijking beschouwden we onder andere een 'naïef' Lineair Mixed Model (LMM) zonder centrering, een LMM mét centrering binnen-subjecten en een LMM aanpak die de mediator en de uitkomst simultaan of 'joint' modelleert. Naast deze bestaande methoden, implementeerden wij ook een techniek die gebaseerd is op het regresseren van de verschilscores in Y op de verschilscores in M (i.e., een adaptatie van de methode van Judd et al. (2001)), waarbij we eveneens periode-effecten en verschillende types moderatie toelieten.

De simulaties toonden aan dat de 'naïeve' methode niet kan omgaan met ongemeten level-2 confounders van de mediator-uitkomst relatie. Aangezien men de afwezigheid van dergelijke confounders nooit kan garanderen, raden we het gebruik van deze methode dan ten stelligste ook af. Voor de andere statistische modellen konden we zowel het directe als het indirecte effect identificeren, zélfs wanneer er ongemeten tijds-onafhankelijke confounders van de mediator-uitkomst relatie aanwezig zijn. Verder konden we ook vaststellen dat de 'joint' methode iets strengere assumpties veronderstelt,

vergeleken met de centrerings-methode en de verschil-methode (die beiden meestal identieke schatters opleveren). Ten slotte illustreerden we deze conclusies aan de hand van een neurostimulatie-experiment, waarna we ook een sensitiviteits-analyse voorstelden die de impact poogde in te schatten van mogelijke level-1 confounders van M-Y op het directe en indirecte effect.

3 Hoofdstuk 3

In hoofdstuk 3 gaan we verder in op de verschillende centreringstechnieken binnen LMMs. In principe zijn er drie voorname methodes waarmee multilevel data gecentreerd kunnen worden: ofwel wordt er niet gecentreerd, ofwel centreren we over het algemeen gemiddelde, ofwel centreren we binnen clusters. In de huidige literatuur rond multilevel data is er intussen een algemene consensus ontstaan, dat centreren binnen clusters het meest is aangewezen wanneer onderzoekers geïnteresseerd zijn in het effect van level-1 predictoren. Deze conclusie is in overeenstemming met het vorige hoofdstuk, aangezien we hier vaststelden dat de naïeve (of ongecentreerde) methode vertekeningen vertoont wanneer er ongemeten level-2 confounders zijn van mediator en uitkomst, terwijl de methode die centreerde binnen subjecten hiervan bespaard bleef.

Jammer genoeg bleven de meeste discussies binnen de multilevel literatuur tot nu toe beperkt tot het centreren van hoofdeffecten, terwijl het centreren van interacties eerder op de achtergrond bleef. Dit vormt vooral een beperking wanneer zowel de predictor, de moderator (die een interactie vormt met de predictor), als de uitkomst gemeten zijn op het onderste-level (waardoor de interactie tussen de predictor en de moderator zich ook op level-1 bevindt). We kunnen ons dan de vraag stellen hoe we deze onderste-level interactieterm best gaan centreren en wat de eventuele gevolgen zijn van deze keuze: centreren we eerst de predictor en moderator binnen subjecten, waarna we beide gecentreerde variabelen vermenigvuldigen (**c**entreer eerst, neem het nadien het **p**roduct, C1P2), of is het beter om eerst beide -ongecentreerde- variabelen te vermenigvuldigen en deze productterm pas achteraf te centreren (neem eerst het **p**roduct, pas nadien **c**entreren, P1C2)? Want in tegenstelling tot een interactie tussen een onderste- en een bovenste-level variabele, of tussen twee level-2 variabelen, zullen C1P2 en P1C2 verschillende resultaten opleveren wanneer we een level-1 interactie gaan centreren binnen subjecten.

Hoofdstuk 3 trachtte dan ook beide centreringsmethoden (P1C2 en

C1P2) tegen elkaar uit te zetten, om na te gaan of ze het (gemodereerde) effect van predictor op uitkomst al dan niet onvertekend kunnen schatten. Om inzicht te verkrijgen in de prestaties van beide technieken, voerden we een uitgebreide simulatiestudie uit. Hierin beschouwden we verschillende settings waarin we de relatieve vertekening van de schatters en standaardfouten, evenals hun coverage en power trachtten te onderzoeken. We konden vaststellen dat we geen vertekening in de schatters zien, zolang de predictor en moderator onafhankelijk zijn van elkaar. Van zodra de moderator echter beïnvloed werd door de predictor (zodat deze óók een mediator wordt), stelden we vertekening vast in de schatters van het interactie-effect voor C1P2. Een analytische berekening van deze bias toonde aan dat deze vertekening afhangt van de verdeling van de predictor én de grootte van het effect van de predictor op de moderator: wanneer de predictor binair is, neemt deze vertekening toe naarmate het effect van X op M sterker wordt. Bovendien konden we ook vaststellen dat de gemiddelde standaardfout veel kleiner is voor P1C2, vergeleken met de C1P2-centrering. Indien we deze twee conclusies samen in beschouwing nemen, adviseren wij om altijd eerst level-1 predictoren te vermenigvuldigen en pas dan deze product-term te centreren binnen clusters, aangezien: (1) P1C2 resulteert in preciezere schatters van het interactie-effect en (2) P1C2 wordt, in tegenstelling tot C1P2, niet beïnvloed door misspecificatie of omissie van bovenste-level effecten.

Om het hoofdstuk te verduidelijken, illustreerden we de gevonden resultaten aan de hand van een longitudinale dagboek-studie over seksueel gedrag in Vlaanderen. Specifiek focusten we onze op de mannelijke deelnemers en keken naar het effect van intiem gedrag op positieve relatiegevoelens de dag nadien, en in welke mate dit effect veranderde door masturbatie.

4 Hoofdstuk 4

Aangezien in hoofdstukken 2 en 3 enkel werd gekeken naar binnen-subject mediatie in continue settings, kunnen we ons vervolgens de vraag stellen of 1-1-1-mediatie gemakkelijk uit te breiden valt naar binaire settings. Vooraleer we deze vraag kunnen beantwoorden, moeten we echter eerst proberen na te gaan welke schattingsmethoden in staat zijn een simpel (niet-gemediëerd) effect op een binaire uitkomst efficiënt en zonder vertekening te schatten. Dit brengt ons vlekkeloos naar hoofdstuk 4, waar het vooral draait om Gegeneraliseerde Lineaire Mixed Modellen (GLMMs), in tegenstelling

tot de Lineaire Mixed Modellen (LMMs) uit hoofdstukken 2 en 3.

Alhoewel GLMMs heel vaak worden toegepast op geclusterde categorische uitkomsten, wordt hun statistische inferentie gehinderd door problemen tijdens het integreren van de random effecten uit de likelihoodfunctie. Om deze tekortkoming het hoofd te bieden, hebben onderzoekers in de loop der jaren enkele technieken voorgesteld die we grofweg kunnen indelen in twee klasses: likelihood-gebaseerde versus Bayesiaanse methoden. De eerste mogelijkheid kan het integratie-probleem oplossen door ofwel een approximatie van de integrand te berekenen, zoals in de Laplace approximatie (Tierney and Kadane, 1986) of in Penalised Quasi-Likelihood (PQL) (Breslow and Clayton, 1993; Schall, 1991; Stiratelli et al., 1984)), ofwel door de integraal zelf te benaderen door een eindige som, zoals in Adaptive Gaussian Quadrature, AGQ (Pinheiro and Bates, 1995). Bayesiaanse methoden daarentegen, maken gebruik van Markov Chain Monte Carlo (MCMC) implementaties om de posterieure distributies van de gewenste parameters te bekomen, waarbij de likelihood zelf gesimuleerd wordt in plaats van deze te analytisch berekenen. Aangezien MCMC-methoden vaak computationeel heel intensief zijn, werden er ook hybride modellen voorgesteld zoals een Integrated Nested Laplace Approximatie, INLA, (Rue et al., 2009)) die benaderingen gebruikt voor verschillende posterieure distributies.

Aangezien deze voorstellen slechts zelden bevredigende resultaten opleveren voor binaire uitkomsten binnen kleine clusters (Breslow and Clayton, 1993; McMahon et al., 2003), stelden wij Structural Equation Modelling (SEM) voor als alternatief. Hoewel SEM en GLMM op het eerste zicht twee verschillende denkkaders lijken, hebben wetenschappers intussen kunnen vaststellen dat beide vaak equivalent zijn in de afwezigheid van latente variabelen, en dit onder een brede set van condities (Rovine and Molenaar, 2000; Curran, 2003; Bauer, 2003). Binnen het SEM-denkkader zijn er twee belangrijke schattingsmethoden: maximum likelihood (ML) en (diagonally) weighted least squares (DWLS) (Skrondal and Rabe-Hesketh, 2004). Omdat ML-schatting niet veel voorkomt binnen de traditionele SEM-literatuur, en bovendien ook min of meer equivalent blijkt aan ML-schatting via GLMMs, legden wij in dit hoofdstuk de nadruk op DWLS.

Er zijn dus vele mogelijke opties om een binaire geclusterde uitkomst te schatten: de Laplace approximatie, PQL, AGQ, MCMC, en INLA binnen het GLMM-kader, evenals robuuste DWLS binnen SEM. Maar welke methode levert nu de beste en meest efficiënte schatters op? Om deze vraag te beantwoorden, voerden wij in hoofdstuk 4 een uitgebreide simulatiestudie

uit waarin we de performantie nagingen van zes verschillende R-pakketten (R version 3.2.3, R Core Team (2013)). In deze simulaties pasten we bovenstaande methoden toe op random-intercept probit-regressie, waarbij we een groot aantal factoren lieten variëren om de verschillende methoden zo goed en volledig mogelijk te kunnen vergelijken: een clustergrootte van 2, 3 of 5, een steekproefgrootte van 25, 50, 100 of 300, een uitkomst prevalentie van 0.1 of 0.5, een latente intracluster correlatie van 0.1, 0.3 of 0.5, en verschillende types predictoren (continu versus binair, en variërend binnen versus tussen clusters). Dit leverde ons een simulatie-studie op met $3 \times 4 \times 3 \times 2 \times 4$ mogelijke settings, waarbinnen we de convergentie, relatieve vertekening, mean squared error en coverage van de zes verschillende schattingsmethoden met elkaar konden vergelijken. Om na te gaan of de conclusies onafhankelijk zijn van de gebruikte R-pakketten, zijn de belangrijkste simulaties herhaald aan de hand van implementaties in andere software, zoals SAS® (version 9.4 (SAS Institute Inc, 2015)), MPLUS® (version 7 (Muthén and Muthén, 2010)) en JAGS (version 4.1.0. (Plummer, 2003)).

Uit deze simulatiestudie konden we onder meer afleiden dat wanneer we clusters van grootte twee beschouwden, SEM het beste presteerde in termen van vertekening, terwijl AGQ de bovenhand nam in termen van precisie (dit voornamelijk door de robuuste standaardfouten in SEM). Indien de clustergrootte echter toenam, werd AGQ de beste optie voor zowel de vertekening als de precisie. Deze conclusies bleken ook onafhankelijk van het gebruikte software-medium, met als enige uitzondering dat de gebruikte MCMC-implementatie in R sub-optimaal was, vergeleken met de implementatie in JAGS-software.

Tenslotte werden deze conclusies ook nog eens geïllustreerd aan de hand van een dataset over de eetgewoonten van kinderen in Vlaamse kleuterscholen. Hierbij werd er nagegaan of aanmoediging tot het eten van witloof (de interventie X) enige invloed had op het al dan niet lusten ervan (de binaire uitkomst Y).

5 Hoofdstuk 5

Alhoewel de recente literatuur reeds veel aandacht besteedde aan het uitbreiden van mediatie naar multilevel settings, werden dergelijke extensies vaak gelimiteerd tot continue uitkomstmaten. Vandaar dat we in hoofdstuk 5 hierop trachten in te spelen door binnen-subject mediatie uit te breiden naar settings met een binaire uitkomst. Hiervoor bouwden we verder op

het voorbereidende werk van hoofdstuk 4, waarin we aantoonden aan dat GLMMs met AGQ de beste schattingen opleveren, wanneer we het effect van een predictor op een binaire uitkomst nagaan in kleine clusters.

Bovendien willen we ook weer focussen op de gevolgen van ongemeten level-2 confounders van de M-Y relatie, aangezien dit kan zorgen voor vertekening in de schattingen van de regressieparameters, evenals in de schattingen van het indirecte en directe effect. Zoals we reeds aantoonden in hoofdstukken 2 en 3, kan vertekening ten gevolge van zulke confounders in lineaire settings aangepakt worden door de level-1 predictoren te centreren. Helaas vervalt deze oplossing wanneer de uitkomst binair is (Goetgeluk and Vansteelandt, 2008; Brumback et al., 2010), waardoor we een geschiktere manier moeten zoeken die kan omgaan met bovenste-level endogeniteit van de mediator en uitkomst.

Om dit probleem het hoofd te bieden, beschouwden we in hoofdstuk 5 1-1-1 mediatie met een binaire uitkomst vanuit een 'tegenfeitelijk' standpunt, opnieuw met een focus op kleine cluster-groottes. Hoofdstuk 5 stelde voor om dit te doen aan de hand van vier sequentieel te doorlopen stappen. Een eerste stap biedt niet-parametrische definities van de causale mediatie-effecten aan, evenals een opsomming van de assumpties die hun identificatie mogelijk maken. In deze stap focusten wij op expressies die gedefiniëerd zijn op de lineaire schaal, zodat we een contrafactuele definitie op basis van verschillen konden opstellen. Een tweede stap identificeert het directe en indirecte effect op basis van parametrische modellen voor mediator en uitkomst. We deden dit aan de hand van twee link-functies die de binaire uitkomst aan de lineaire predictorterm koppelen: de probit- en de logit-link. Een derde stap schat de regressieparameters van de modellen voor de mediator en de uitkomst. Wij deden dit op basis van drie verschillende modellen: (1) een ongecentreerde methode die de mediator en de uitkomst apart schat, (2) een techniek die de mediator en de uitkomst apart modelleert, waarbij de predictoren gecentreerd zijn binnen subjecten, en (3) een methode die de mediator en uitkomst simultaan of 'joint' modelleert. Ten slotte schatten we in een laatste stap de causale mediatie-effecten zelf, door potentiële uitkomsten voor de mediator en de uitkomst te voorspellen. Dit gebeurde aan de hand van een parametrisch algoritme, waarin de posterieure verdelingen van beide variabelen benaderd werden door hun steekproevenverdeling. Hierbij werden de random effecten op twee verschillende manieren gegenereerd: marginaal of conditioneel.

Op deze manier gingen we in hoofdstuk 5 na welke multilevel schattingsmethoden in staat zijn om op een correcte manier ongemeten level-2

confounding van M en Y te elimineren. Hierbij focusten we op een binaire gerandomiseerde interventie en een binaire uitkomst binnen kleine clusters. Om deze onderzoeksvraag na te gaan voerden we een simulatiestudie uit, waarin we drie schattingsmethoden vergeleken (een ongecentreerde-, een gecentreerde- en een joint- modelleringsmethode), twee link-functies voor de uitkomst (de logit- en de probit-link) en twee manieren om de random effecten te genereren (marginaal versus conditioneel). Om deze verschillende methoden goed met elkaar te kunnen vergelijken, lieten we meerdere factoren binnen de gesimuleerde datasets variëren: de grootte van de clusters, de steekproefgrootte, de intracluster correlatie en de aan- of afwezigheid van ongemeten level-2 confounding van mediator en uitkomst.

Uit deze simulaties konden we afleiden dat het simultaan modelleren van M en Y de beste prestatie vertoonde (gecombineerd met een marginale generatie van de random effecten), vooral in de aanwezigheid van ongemeten level-2 confounding van mediator en uitkomst. Een aparte modelleringsmethode waarbij we de predictoren centreerden binnen subjecten en de random effecten conditioneel gegenereerd werden, leverde ook relatief goede resultaten op. Zoals te verwachten, presteerde de ongecentreerde methode (ongeacht de manier van random effect-generatie) ondermaats in de aanwezigheid van ongemeten level-2 confounding van mediator en uitkomst.

Om deze resultaten te illustreren, pasten we deze methoden toe op een crossover studie die de impact van een geïnduceerde doelconflict-situatie op het hulpgedrag van partners van individuën met chronische pijn. Hierbij trachtten we na te gaan of dit causale effect gemediëerd werd door de hoeveelheid autonoom hulpgedrag bij de partner, zoals waargenomen door de patient.

6 Discussie

Met dit doctoraat hoopten we toegepaste onderzoekers van een concrete set aan richtlijnen te voorzien, over hoe binnen-subject mediatie best te evalueren in de aanwezigheid van: 1) ongemeten level-2 confounding van de relatie tussen mediator en uitkomst , 2) onderste-level interactietermen, 3) binaire uitkomstmaten, en 4) moeilijke of veeleisende settings (e.g., kleine cluster-groottes). Aangezien we de afwezigheid van ongemeten bovenste-level confounding van mediator en uitkomst nooit kunnen garanderen, adviseren wij om er (preventief) van uit te gaan dat dergelijke confounders altijd aanwezig zijn. Wij willen toegepaste wetenschappers ook graag

bewust maken van de beperkingen van verscheidene schattingsmethoden en hun implementaties, wanneer het aantal metingen binnen clusters beperkt zijn. Dit aangezien kleine cluster-groottes vaak uitdagend blijken voor de beschikbare methodologiën.

Ter samenvatting zouden we dit doctoraat graag afsluiten met de volgende suggesties omtrent binnen-subject mediatie. Ten eerste raden we het gebruik van een methode die mediator en uitkomst apart modelleert waarbij de onderste-level predictoren niet gecentreerd worden binnen clusters, ten stelligste af. Deze methode is -in lineaire én niet-lineaire settings- niet uitgerust om met bovenste-level endogeniteit van mediator en uitkomst om te gaan. Ten tweede, in lineaire settings raden we onderzoekers wél aan om alle onderste-level predictoren te centreren binnen-subjecten, opdat mediatie correct kan worden nagegaan in de aanwezigheid van ongemeten confounding van M en Y. Ten derde, wanneer wetenschappers de intentie hebben om onderste-level interacties toe te voegen aan hun lineair mediatiemodel, adviseren wij om de betrokken variabelen eerst te vermenigvuldigen en pas nadien deze productterm te centreren binnen clusters. Ten vierde willen we in binaire settings pleiten voor het gebruik van GLMMs, waarin de integraal van de likelihood-functie benaderd wordt door 'Adaptive Gaussian Quadrature'. Deze suggestie wordt des te belangrijker wanneer de onderste-level steekproefgrootte afneemt. En ten slotte, wanneer binnen-subject mediatie wordt nagegaan in binaire settings, raden we aan om de mediator en de uitkomst gezamelijk te modelleren, zodat de causale mediatie-effecten correct kunnen worden geëvalueerd in de aanwezigheid van bovenste-level endogeniteit van M en Y. We raden het gebruik van binnen-cluster centrering af in settings met een binaire uitkomstmaat, aangezien deze methode dan niet langer onvertekende schatters oplevert.

Betreffende de huidige tekortkomingen van deze thesis en mogelijke richtingen voor toekomstig onderzoek, refereren we naar de assumpties die we gedurende de verschillende hoofdstukken veronderstelden. Onderzoek naar de impact van het schenden van elk van deze veronderstellingen, zou een uiterst waardevolle bijdrage tot de huidige literatuur kunnen leveren. Bovendien promoten we de constructie van een software implementatie, die het toegepaste onderzoekers mogelijk zou maken om (gemodereerde) binnen-subject mediatie te schatten in de aanwezigheid van ongemeten level-2 confounding van M en Y, in zowel lineaire als binaire settings.

Bibliografie

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2):135–167.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brumback, B. A., Dailey, A. B., Brumback, L. C., Livingston, M. D., and He, Z. (2010). Adjusting for confounding by cluster using generalized linear mixed models. *Statistics and Probability Letters*, 80(21-22):1650–1654.

Collins, L. M., Graham, J. J., and Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research*, 33(2):295–312.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38(4):529–569.

Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4):408–420.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, 6(2):115–134.

MacKinnon, D. P., Krull, J. L., and Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1(4):173–181.

McMahon, J. M., Tortu, S., Torres, L., Pouget, E. R., and Hamid, R. (2003). Recruitment of heterosexual couples in public health research: a study protocol. *BMC medical research methodology*, 3:24.

Muthén, L. K. and Muthén, B. O. (2010). *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA, sixth edition.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainy in Artificial Intelligence*, pages 411–420.

Pearl, J. (2012). The causal mediation formula-a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–36.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Plummer, M. (2003). JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling. In Hornik, K., Leisch, F., and Zeileis, A., editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20-22*, pages 1–10, Vienna, Austria.

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1):185–227.

R Core Team (2013). R: A language and environment for statistical computing.

Rovine, M. J. and Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, 35(1):55–88.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392.

SAS Institute Inc (2015). Base SAS® 9.4 Procedures Guide, Fifth Edition.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC, New York.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971.

Tierney, L. and Kadane, J. B. (1986). Acccurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tucker-Drob, E. M. (2011). Individual differences methods for randomized experiments. *Psychological Methods*, 16(3):298–318.

Zhao, X., Lynch Jr., J. G., and Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2):197–206.

9 Data Storage Fact Sheets

1 Data Storage Fact Sheet Chapter 2

```
1. Contact details
==========================================================
1a. Main researcher
----------------------------------------------------------
- name: Haeike Josephy
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Haeike.Josephy@gmail.com

1b. Responsible Staff Member (ZAP)
----------------------------------------------------------
- name: Tom Loeys
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Tom.Loeys@Ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data Management,
Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2,
9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========================================================
* Reference of the publication in which the datasets are reported:
Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., & Loeys, T. (2015).
Within-subject mediation analysis in AB/BA crossover designs.
INTERNATIONAL JOURNAL OF BIOSTATISTICS, 11(1), 1-22.

* Which datasets in that publication does this sheet apply to?
This data storage fact sheet refers to the raw data and SAS-code relating
to the example analysis and simulation study in Josephy et. al. (2015).

3. Information about the files that have been stored
==========================================================
3a. Raw data
----------------------------------------------------------
* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
        If NO, please justify:
```

```
* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server: shared drive 'mediation' in the
         Department of Data Analysis  (file: 'Example_Chapter2.csv')
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of
    another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
------------------------------------------------------------
* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported
         results. Specify: 'Example_Chapter2.sas'
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing the data generating mechanism and analyses.
              Specify: 'Simulations_Chapter2.sas'
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this
         content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server: shared drive 'mediation' in the
         Department of Data Analysis
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention
    of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
```

```
   - address:
   - affiliation:
   - e-mail:
```

2 Data Storage Fact Sheet Chapter 3

```
1. Contact details
============================================================
1a. Main researcher
------------------------------------------------------------
- name: Haeike Josephy
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Haeike.Josephy@gmail.com

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Tom Loeys
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Tom.Loeys@Ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data Management,
Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2,
9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Loeys, T., Josephy, H., Dewitte, M. (2018). More precise estimation of
lower-level interaction effects in multilevel models. MULTIVARIATE
BEHAVIORAL RESEARCH, 53(3), 335-347.

* Which datasets in that publication does this sheet apply to?
This data storage fact sheet refers to the raw data and R-code relating
to the example analysis and simulation study in Loeys et. al. (2018).

3. Information about the files that have been stored
============================================================
3a. Raw data
------------------------------------------------------------
* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
        If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server: shared drive 'mediation' in the
          Department of Data Analysis
        (files: 'Example_Chapter3.txt' and 'Example_Chapter3_c.txt')
```

```
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of
    another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
------------------------------------------------------------
* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported
        results. Specify: 'Example_Chapter3.R'
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing the data generating mechanism and analyses.
            Specify: 'Simulations_Chapter3.R'
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this
        content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server: shared drive 'mediation' in the
        Department of Data Analysis
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention
    of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:
```

3 Data Storage Fact Sheet Chapter 4

```
1. Contact details
============================================================
1a. Main researcher
------------------------------------------------------------
- name: Haeike Josephy
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Haeike.Josephy@gmail.com

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Tom Loeys
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Tom.Loeys@Ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data Management,
Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2,
9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Josephy, H., Loeys, T., & Rosseel, Y. (2016). A review of R-packages for
random-intercept probit regression in small clusters.
FRONTIERS IN APPLIED MATHEMATICS AND STATISTICS, 2(18), 1-13.

* Which datasets in that publication does this sheet apply to?
This data storage fact sheet refers to the raw data, R- and SAS-code
relating to the example analysis and simulation study in
Josephy et. al. (2016).

3. Information about the files that have been stored
============================================================
3a. Raw data
------------------------------------------------------------
* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
        If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server: shared drive 'mediation' in the
          Department of Data Analysis (file: 'Example_Chapter4.csv')
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of
   another person)?
  - [X] main researcher
```

```
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
------------------------------------------------------------
* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported
          results. Specify: 'Example_Chapter4.R'
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing the data generating mechanisms and analyses.
Specify:
1. Simulations for cluster size 2:
        A. For a between-subject predictor:
                'Simulations_Chaprer4_Between.R'
        B. For a within-subject predictor:
                'Simulations_Chapter4_Within.R'
        C. For comparing lavaan in R to MPLUS:
                'Simulations_Chapter4_MPLUS.R'
        D. For comparing MCMCglmm in R tot JAGS:
                'Simulations_Chapter4_JAGS.R'
        E. For exporting the data set to SAS:
                'Simulations_Chapter4_SAS_data.R'
        F. For comparing the Laplace approximation and AGQ in R vs. SAS:
                'Simulations_Chapter4_SAS.sas'
2. Simulations for cluster size 3:
        A. For a between-subject predictor:
                'Simulations_Chapter4_Clustersize3_Between.R'
        B. For a within-subject predictor:
                'Simulations_Chapter4_Clustersize3_Within.R'
3. Simulations for cluster size 5:
        A. For a between-subject predictor:
                'Simulations_Chapter4_Clustersize5_Between.R'
        B. For a within-subject predictor:
                'Simulations_Chapter4_Clustersize5_Within.R'

  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how
          this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server: shared drive 'mediation' in the
          Department of Data Analysis
  - [ ] other: ...
```

```
* Who has direct access to these other files (i.e., without intervention
   of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:
```

4 Data Storage Fact Sheet Chapter 5

```
1. Contact details
============================================================
1a. Main researcher
------------------------------------------------------------
- name: Haeike Josephy
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Haeike.Josephy@gmail.com

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------
- name: Tom Loeys
- address: Henri Dunantlaan 2, 9000 Gent
- e-mail: Tom.Loeys@Ugent.be

If a response is not received when using the above contact details,
please send an email to data.pp@ugent.be or contact Data Management,
Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2,
9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
============================================================
* Reference of the publication in which the datasets are reported:
Josephy, H., Kindt, S., Loeys, T. (in preparation).
Lower-level mediation with a binary outcome.

* Which datasets in that publication does this sheet apply to?
This data storage fact sheet refers to the raw data and R-code
relating to the example analysis and simulation study in
```

```
Josephy et. al. (in preparation).

3. Information about the files that have been stored
============================================================
3a. Raw data
------------------------------------------------------------
* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
        If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server: shared drive 'mediation' in the
          Department of Data Analysis (file: 'Example_Chapter5.txt')
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of
   another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
------------------------------------------------------------
* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported
          results. Specify:
        'Example_Chapter5.R' and 'Example_Chapter5.inp'
  - [ ] file(s) containing processed data. Specify: ...
  - [X] file(s) containing the data generating mechanisms and analyses.
Specify:
1. Simulations for cluster size 2:
        A. For probit-regression:
                'Simulations_Chapter5_cs2_probit.R'
        B. For logit-regression:
                'Simulations_Chapter5_cs2_logit.R'
2. Simulations for cluster size 5:
        A. For probit-regression:
                'Simulations_Chapter5_cs5_probit.R'
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how
          this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server: shared drive mediation' in the
```

```
          Department of Data Analysis
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention
   of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
============================================================
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:
```