Haeike Josephy*, Stijn Vansteelandt, Marie-Anne Vanderhasselt and Tom Loeys

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 neurobehavioural 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.

Keywords: mediation; crossover designs; causal inference; direct and indirect effects; counterfactual framework

DOI 10.1515/ijb-2014-0057

1 Introduction

The concept of mediation has received a great deal of attention during the last couple of decades, with Baron and Kenny [1] 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 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:

***Corresponding author: Haeike Josephy**, Department of Data Analysis, Ghent University, Henri Dunantlaan 1, Ghent, East Flanders 9000, Belgium, E-mail: Haeike.Josephy@UGent.be

Stijn Vansteelandt, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, East Flanders 9000, Belgium, E-mail: stijn.vansteelandt@ugent.be

Marie-Anne Vanderhasselt, Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, East Flanders 9000, Belgium, E-mail: MarieAnne.Vanderhasselt@UGent.be

Tom Loeys, Department of Data Analysis, Ghent University, Henri Dunantlaan 1, Ghent, East Flanders 9000, Belgium, E-mail: Tom.Loeys@UGent.be

$$\begin{aligned}
 Y &= i_{Y1} + cX + e_{Y1} \\
 M &= i_M + aX + e_M \\
 Y &= i_{Y2} + c'X + bM + e_{Y2}
 \end{aligned} \tag{1}$$

Figure 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 1 are linear and satisfy a specific set of “no unmeasured confounders” assumptions [2]. 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 [3].

In AB/BA crossover studies, where each participant is observed exactly twice (once under exposure A and once under B , see Figure 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 [4]. 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 [5–10], 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: 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 randomization.

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 [11–15]. In Section 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 3, we discuss Judd et al. [6]’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 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 5). 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 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 Specification of the natural direct and indirect effects in within-subject mediation models

2.1 The counterfactual framework

In order to formalize 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 [16]. 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 (TCE) $E[Y_{it}(1) - Y_{it}(0)]$ 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 [12, 17]. 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 TCE into a direct and indirect effect. Such effect decomposition can occur in two ways: one possibility is to decompose the 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) [18, 19]:

$$\begin{aligned}
 \text{TCE} &= E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0))] \\
 &= E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0)) + Y_{it}(1, M_{it}(0)) - Y_{it}(0, M_{it}(0))] \\
 &= \text{TNIE} + \text{PNDE} \\
 &= E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(1)) + Y_{it}(0, M_{it}(1)) - Y_{it}(0, M_{it}(0))] \\
 &= \text{TNDE} + \text{PNIE}
 \end{aligned} \tag{2}$$

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

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 generalized as follows [15]:

- (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 randomized, 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 summarized by the (lack of) arrows in the directed acyclic graph of Figure 3, which we will interpret as a nonparametric structural equation model with independent errors [20]. In this causal diagram, X_{i0} , M_{i0} and Y_{i0} represent the exposure, mediator and outcome values for subject i , measured during the first period ($t = 0$), while X_{i1} , M_{i1} and Y_{i1} denote these values during the second period ($t = 1$).

Figure 3: Causal diagram, graphically representing the causal 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 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 [21] (e.g. seasonal effects, changes in measurement conditions, 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 [4, 22]. 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 [4].

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):

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + g(U_i) + \varepsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \kappa_Y t + U_i + \varepsilon_{Yit} \end{aligned} \quad (3)$$

Under this mechanism, in correspondence to Figure 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 section will first describe the identification of the natural direct and indirect effect. Next, we will summarize four existing approaches that can assess within-subject mediation in AB/BA crossover designs (Sections 3.2–3.5).

3.1 Identification of the direct and indirect effect

Under the above-described data-generating mechanism, the assumptions introduced in Section 2.2 are met. This enables us to operate Pearl's mediation formula [12, 23] in deriving the total, PNDE and TNIE for each subject i (conditional on i) during period t . Based on eq. (3), the subject- and period-specific TCE equals $\alpha\beta + \zeta'$, the individual- and period-specific TNIE equals $\alpha\beta$ and the PNDE in turn equals ζ' (detailed calculations can be found in the supporting Appendix A.1).

3.2 The difference approach for the AB/BA design

Judd et al. [6] proposed a straightforward method to evaluate mediation in AB/BA crossover designs specifically. 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 :

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

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.

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

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}):

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

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. [6] 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 discuss from Section 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 [24–28], 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 [25]. A second drawback is that it does not account for period effects [4, 21].

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 [29]. Allowing for a period effect, the following lower-level equations would typically be considered:

$$\begin{aligned} M_{it} &= d_{Mi} + aX_{it} + k_M t + e_{Mit} \\ Y_{it} &= d_{Yi} + c'X_{it} + bM_{it} + k_Y t + e_{Yit} \end{aligned} \quad (4)$$

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

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

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 $\perp\!\!\!\perp$) of the predictors in their respective equations. Both upper- and lower-level error terms are assumed to be independently and normally distributed with mean zero.

Throughout this paper, maximum likelihood estimators for the parameters from the working models are denoted with “ $\wedge$ ”. The TNIE is estimated from eq. (4) as $\hat{a}\hat{b}$, while the PNDE is estimated from eq. (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 (3). This may result in an “omitted variable bias” [30] for β and ζ' when such confounding is indeed present (as is the case in data-generating mechanism (3)), resulting in biased estimators for direct and indirect effect ζ' and $a\beta$. To this end, we will henceforth refer to this approach as the *naive modelling approach*.

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 [8, 9, 31–35]. 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 [35, 36]. This can result from unmeasured upper-level confounding, which is absorbed in the between-subject effect [37]. 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 focus on the within-subject effects, as these are of primary interest in crossover studies.

Following MacKinnon [29], the within-subject effect can be estimated by regressing Y_{it} on the subject-mean-centred mediator ($M_{it} - \bar{M}_{it}$). Here $\bar{M}_{it}$ 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:

$$\begin{aligned} M_{it} &= d_{Mi} + aX_{it} + k_M t + e_{Mit} & \text{with } d_{Mi} &= d_M + u_{Mi} \\ Y_{it} &= d_{Yi} + c'X_{it} + b(M_{it} - \bar{M}_{it}) + k_Y t + e_{Yit} & \text{with } d_{Yi} &= d_Y + u_{Yi} \end{aligned} \quad (5)$$

The lower-level residuals e_{Mit} and e_{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 assumed data-generating mechanism (eq. (3)), the unmeasured confounder U_i is uncorrelated with $M_{it} - \bar{M}_{it}$. 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 (3) holds, the TNIE can be estimated unbiasedly from eq. (5) as $\hat{a}\hat{b}$ and the PNDE 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 [29]. 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 $\bar{M}_i$ being uncorrelated with $(M_{it} - \bar{M}_i)$ and is, as such, not further considered.

3.5 A joint modelling approach

Another multilevel approach, described by Bauer, Preacher, and Gil [5], 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:

$$Z_{it} = S_M(d_{Mi} + aX_{it} + k_M t) + S_Y(d_{Yi} + c'X_{it} + bM_{it} + k_Y t) + e_{Zit} \quad (6)$$

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 (3) is estimated from eq. (6) as $\hat{a}\hat{b}$ and the PNDE 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 the supporting 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 M_{it} in the outcome model.

4 Estimating direct and indirect effect in more complex settings involving interactions

The data-generating mechanism that we considered so far (eq. (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:

$$\begin{aligned}
M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + \omega_M D_i + v_M X_{it} D_i + g(U_i) + \varepsilon_{Mit} \\
Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \phi X_{it} M_{it} + \kappa_Y t + \omega_Y D_i + v_Y X_{it} D_i \\
&\quad + \eta M_{it} D_i + U_i + \varepsilon_{Yit}
\end{aligned} \tag{7}$$

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 v_M and v_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 $g(U_i)$.

4.1 Identification of the direct and indirect effect in complex settings

Under data-generating mechanism (7) the assumptions introduced in Section 2.2 continue to apply, which enables use of the mediation formula to derive the TCE, PNDE and TNIE for subject i during period t . Here, the subject- and period-specific TNIE equals (detailed calculations can be found in the supporting Appendix A.3):

$$\begin{aligned}
&E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0)) | D_i = d, U_i] \\
&= \sum_m \{E[Y_{it} | X_{it} = 1, M_{it} = m, D_i = d, U_i] P(M_{it} = m | X_{it} = 1, D_i = d, U_i) \\
&\quad - E[Y_{it} | X_{it} = 1, M_{it} = m, D_i = d, U_i] P(M_{it} = m | X_{it} = 0, D_i = d, U_i)\} \\
&= (\alpha + v_M d)(\beta + \phi + \eta d)
\end{aligned} \tag{8}$$

Since this expression does not depend on U_i (see Appendix A.3), the subject- and period-specific TNIE can be marginalized 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 PNDE and TCE; 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 *PNDE* at $g(U_i) = 0$:

$$\begin{aligned}
&E[Y_{it}(1, M_{it}(0)) - Y_{it}(0, M_{it}(0)) | D_i = d, U_i] \\
&= \sum_m \{E[Y_{it} | X_{it} = 1, M_{it} = m, D_i = d, U_i] P(M_{it} = m | X_{it} = 0, D_i = d, U_i) \\
&\quad - E[Y_{it} | X_{it} = 0, M_{it} = m, D_i = d, U_i] P(M_{it} = m | X_{it} = 0, D_i = d, U_i)\} \\
&= \zeta' + v_Y d + \phi(\delta_M + \kappa_M t + \omega_M d)
\end{aligned} \tag{9}$$

and *TCE* at $g(U_i) = 0$:

$$\begin{aligned}
&E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0)) | D_i = d, U_i] \\
&= \sum_m \{E[Y_{it} | X_{it} = 1, M_{it} = m, D_i = d, U_i] P(M_{it} = m | X_{it} = 1, D_i = d, U_i) \\
&\quad - E[Y_{it} | X_{it} = 0, M_{it} = m, D_i = d, U_i] P(M_{it} = m | X_{it} = 0, D_i = d, U_i)\} \\
&= (\alpha + v_M d)(\beta + \phi + \eta d) + \zeta' + v_Y d + \phi(\delta_M + \kappa_M t + \omega_M d)
\end{aligned} \tag{10}$$

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 TCE marginally over U_i , for example by regressing the outcome Y_{it} on X_{it} and D_i . The PNDE marginalized over U_i can subsequently be estimated by subtracting the TNIE 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.

4.2 A more flexible difference approach

The previously discussed difference approach by Judd et al. [6] 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: $Y_i^{\text{Dif}} = c' + b_{\text{Dif}}M_i^{\text{Dif}} + b_{\text{Sum}}M_i^{\text{Sum}} + e_{Yi}^*$. 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 [27, 38, 39]. To allow for such moderation, as well as the above-mentioned period effects, we will extend the approach proposed by Judd et al. [6] as follows:

$$\begin{aligned} M_i^{\text{Dif}} &= a + k_M t_i^{\text{Dif}} + v_M D_i + e_{Mi} \\ Y_i^{\text{Dif}} &= c' + b M_i^{\text{Dif}} + f X M_i^{\text{Dif}} + k_Y t_i^{\text{Dif}} + v_Y D_i + n D_i M_i^{\text{Dif}} + e_{Yi} \end{aligned} \quad (11)$$

In eq. (11), $X M_i^{\text{Dif}}$ equals $M_i^{x=1}$ and $t_i^{\text{Dif}} = t_i^{x=1} - t_i^{x=0}$, 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 (7):

$$\begin{aligned} M_i^{\text{Dif}} &= \alpha + \kappa_M t_i^{\text{Dif}} + v_M D_i + \varepsilon_{Mi} \\ Y_i^{\text{Dif}} &= \zeta' + \beta M_i^{\text{Dif}} + \phi X M_i^{\text{Dif}} + \kappa_Y t_i^{\text{Dif}} + v_Y D_i + \eta D_i M_i^{\text{Dif}} + \varepsilon_{Yi} \end{aligned} \quad (12)$$

the difference approach will allow unbiased estimation of the indirect effect in this setting. The indirect effect can be estimated from eq. (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 marginalized over U_i).

A final remark can be made regarding eq. (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 eq. (12) when $v_Y \neq 0$ or $\eta \neq 0$, as the interaction term between the 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 $v_Y = 0$ and $\eta = 0$, but $v_M \neq 0$, 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 $v_Y = \eta = 0$, 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 .

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 (7), by adding the respective interaction terms to the models. These become

$$\begin{aligned} M_{it} &= d_{Mi} + a X_{it} + k_M t + w_M D_i + v_M X_{it} D_i + e_{Mit} \\ Y_{it} &= d_{Yi} + c' X_{it} + b M_{it} + f X_{it} M_{it} + k_Y t + w_Y D_i + v_Y X_{it} D_i \\ &\quad + n M_{it} D_i + e_{Yit} \end{aligned} \quad (13)$$

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

$$\begin{aligned} M_{it} &= d_{Mi} + \alpha X_{it} + \kappa_M t + w_M D_i + v_M X_{it} D_i + e_{Mit} \\ Y_{it} &= d_{Yi} + \zeta' X_{it} + \beta(M_{it} - \bar{M}_{it}) + f(X_{it} M_{it} - \bar{X}_{it} \bar{M}_{it}) + \kappa_Y t \\ &\quad + w_Y D_i + v_Y X_{it} D_i + n(M_{it} - \bar{M}_{it}) D_i + e_{Yit} \end{aligned} \quad (14)$$

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 $X_{it}(M_{it} - \bar{M}_{it})$) 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 effects under data-generating mechanism (7), but in contrast to the difference approach, the separate W -only approach and Joint modelling approach also require modelling (and hence potentially correct specification of) the main effect of D_i .

5 Simulation study

To gain insight into the finite sample performance of all four modelling approaches represented by eqs (11), (13) and (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 :

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + v_M V_i X_{it} + V_i + \varepsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \phi X_{it} M_{it} + \kappa_Y t + U_i + \varepsilon_{Yit} \end{aligned} \quad (15)$$

Here, V_i and U_i represent zero-mean bivariate normally distributed unmeasured 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 [40, 41], 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 (v_M). Note that when $v_M \neq 0$, assumption (vi) is violated.

For the first simulation setting, the M and Y values are generated according to eq. (15), but with v_M and σ_{V_i, U_i} both set to zero (thus satisfying the assumptions in Section 2.2). 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 v_M to 1 (thus violating assumption (vi) in Section 2.2).

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 six conditions (three simulation settings and two 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 (eqs (11) (13), and (14)). Additionally, the square root of the mean

Table 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).

N	Method	$\beta = -1.00$			$\zeta' = 2.00$			$\phi = 2.00$		
		Mean (se)	Cov	$\sqrt{\text{MSE}}$	Mean (se)	Cov	$\sqrt{\text{MSE}}$	Mean (se)	Cov	$\sqrt{\text{MSE}}$
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

Note: Simulations (1)–(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 ($\sqrt{\text{MSE}}$). 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.

squared error (MSE) is provided for these estimators. The results for $N = 50$ and $N = 200$ are shown in Table 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.

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 from the difference and separate W-only modelling

approach is expected when there are no random slopes [37]. 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 3.5, as the assumption of no moderation of M_{it} by u_{Yi} (condition (c)) is violated under the third setting ($v_M \neq 0$).

5.2 Coverage and MSE

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 approaches 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.

Second, as long as there is no unmeasured confounding of the M – Y relation, we observe the lowest 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 MSEs of the other three approaches. Overall, of these three methods the joint modelling approach seems to provide the lowest MSEs, 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.

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 [42]. 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 min (through the use of electrodes), which induces polarization shifts in the resting membrane potential [43]. 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 washout period lasted for a minimum of 48 h, and since current research suggests an intersession interval of 48 h after a long stimulation is more than sufficient [44], the absence of carryover 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, 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 their 95% percentile-based confidence intervals were obtained through bias-corrected bootstrapping, based on 1,000 bootstrap samples), can be found at the left panel of Figure 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.

Figure 4: (a) Results of fitting each of the four within-subject modelling approaches for the neurostimulation data, according to the settings summarized by eq. (3). “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 estimated direct effect and indirect effect are given, alongside the 95% confidence intervals (these estimates and confidence intervals are obtained by percentile-based bias-corrected bootstrapping, based on 1,000 samples). (b) Results of fitting each of the four within-subject modelling approaches for the neurostimulation data, according to the settings summarized by eq. (7) (with $\phi = 0$).

In a second step, we check for moderation effects of a centred subject-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 effect of D , nor on the period effects 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 on the right side of Figure 4 (b). We observe no significant indirect and direct effects 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 approaches 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 M – 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 of -0.21 ($p = 0.58$) between both random intercepts, 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 5(a)). As hypothesized, significant indirect effects are only observed for high levels of trait rumination [42]. A period-averaged direct effect, on the other hand, remains absent over the entire range of D -values.

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

Figure 5: (a) The average direct (on the left) and indirect effect (on the right) and their 95% confidence intervals (95% CI) obtained from the difference approach (by percentile-based bias-corrected bootstrapping, based on 1,000 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 effect (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 through percentile-based bias-corrected bootstrapping, based on 1,000 samples).

such lower-level M – Y confounding. More precisely, we assume the following extension of data-generating mechanism (7):

$$\begin{aligned} M_{it} &= \delta_M + \alpha X_{it} + \kappa_M t + \omega_M D_i + v_M D_i X_{it} + g(U_i) + \varepsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta' X_{it} + \beta M_{it} + \phi X_{it} M_{it} + \kappa_Y t + \omega_Y D_i + v_Y D_i X_{it} + \eta D_i M_{it} \\ &\quad + U_i + \theta \varepsilon_{Mit} + \varepsilon_{Yit} \end{aligned} \quad (16)$$

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 neurobehavioural example, ϕ is assumed to be zero, but is included here to allow for generalization. In terms of differences it follows from eq. (16) that

$$\begin{aligned} M_i^{\text{Dif}} &= \alpha + \kappa_M t_i^{\text{Dif}} + v_M D_i + \varepsilon_{Mi} \\ Y_i^{\text{Dif}} &= \zeta' + \beta M_i^{\text{Dif}} + \phi X M_i^{\text{Dif}} + \kappa_Y t_i^{\text{Dif}} + v_Y D_i + \eta D_i M_i^{\text{Dif}} \\ &\quad + \theta \varepsilon_{Mi} + \varepsilon_{Yi} \end{aligned} \quad (17)$$

with ε_{Mi} and ε_{Yi} encoding the difference in M - and Y -residuals from eq. (16). When ε_{Mi} is substituted by $M_i^{\text{Dif}} - \alpha - \kappa_M t_i^{\text{Dif}} - v_M D_i$ in the outcome equation, we obtain

$$\begin{aligned} Y_i^{\text{Dif}} &= (\zeta' - \theta \alpha) + (\beta + \theta) M_i^{\text{Dif}} + \phi X M_i^{\text{Dif}} + (\kappa_Y - \theta \kappa_M) t_i^{\text{Dif}} \\ &\quad + (v_Y + \theta v_M) D_i + \eta D_i M_i^{\text{Dif}} + \varepsilon_{Yi} \end{aligned} \quad (18)$$

which can be rewritten as

$$Y_i^{\text{Dif}} = \zeta'^* + \beta^* M_i^{\text{Dif}} + \phi(XM)_i^{\text{Dif}} + \kappa_Y^* t_i^{\text{Dif}} + \nu_Y^* D_i + \eta D_i M_i^{\text{Dif}} + \varepsilon_{Yi}$$

$$\text{where : } \begin{cases} \zeta' = \zeta'^* + \theta\alpha \\ \beta = \beta^* - \theta \\ \kappa_Y = \kappa_Y^* + \theta\kappa_M \\ \nu_Y = \nu_Y^* + \theta\nu_M \end{cases} \quad (19)$$

As such, under data-generating mechanism (16), the difference approach based on model (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 eq. (17) (ε_{Mi} and $\theta\varepsilon_{Mi} + \varepsilon_{Yi}$), rather than θ [11]. It can be shown that

$$\theta = \frac{\rho}{\sqrt{1 - \rho^2}} \frac{\sigma_{\varepsilon Y}}{\sigma_{\varepsilon M}} \quad (20)$$

Under the above setting, the sensitivity analysis then proceeds as follows. First, all parameters in the difference eq. (11) are estimated and the estimates for the residual error variances, $\sigma_{\varepsilon M}$ and $\sigma_{\varepsilon Y}$, are determined, assuming that $\theta = 0$. Next, a plausible range of values of ρ (varying between -1 and 1) is considered while keeping $\sigma_{\varepsilon M}$ and $\sigma_{\varepsilon Y}$ fixed, so that θ can be calculated by applying expression (20). Then, relying on the estimate for θ , the estimated parameters from eq. (11) and the equalities on the right side of eq. (19), estimates for the true parameters can be obtained. Additionally, precision of the resulting direct and indirect effect at different values of ρ can be assessed by bootstrapping procedures.

We will now illustrate the above proposed sensitivity analysis on our neurobehavioural 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 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 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 .

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 exposure–mediator interactions. 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 [45], 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. [11] yields unbiased estimators for the direct and indirect effects 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 [37]. We conjecture that the *mediation* package approach in the multilevel setting may require assumptions that are too stringent, even if centred predictors were used. Other estimation approaches may thus be indicated: e.g. conditional generalized estimating equations (CGEE) provide a more general framework for sheltering the estimation of within-subject effects from unmeasured between-subject confounding factors [37].

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 [46] or G-estimation [47], that can deal with intermediate confounding concerning the estimation of the controlled direct effect in single-level settings, could be applied to multilevel settings.

Appendix

A.1 Identification of the causal effects in simple settings

We start from a more general data-generating mechanism (compared to model (3)), which relaxes assumption (vii):

$$\begin{aligned} M_{it} &= \delta_M + \alpha_i X_{it} + \kappa_{Mi} t + g(U_i) + \varepsilon_{Mit} \\ Y_{it} &= \delta_Y + \zeta'_i X_{it} + \beta_i M_{it} + \kappa_{Yi} t + U_i + \varepsilon_{Yit} \end{aligned} \quad (21)$$

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 eq. (21), however, encompasses the restrictions dictated by this design (and therefore also eq. (3)). Based on this generalized data-generating mechanism summarized in expression (21), the “*it*-th”-specific *TNIE* can be identified, when the assumptions (i)–(vi) from Section 2.2 are satisfied:

$$\begin{aligned} & E[Y_{it}(x, M_{it}(x)) - Y_{it}(x, M_{it}(x^*)) | \alpha_i, \kappa_{Mi}, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] \\ &= \sum_m \{ E[Y_{it} | X_{it} = x, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\ &\quad - E[Y_{it} | X_{it} = x, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \} \\ &= \sum_m (d_Y + \zeta'_i x + \beta_i m + \kappa_{Yi} t + U_i) \{ P(M_{it} = m | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\ &\quad - P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \} \\ &= \beta_i \left\{ \sum_m m P(M_{it} = m | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) - \sum_m m P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \right\} \\ &= \beta_i \{ E[M_{it} | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i] - E[M_{it}^* | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i] \} \\ &= \beta_i \{ d_M + \alpha_i x + \kappa_{Mi} t + g(U_i) - d_M - \alpha_i x^* - \kappa_{Mi} t - g(U_i) \} \\ &= \alpha_i \beta_i (x - x^*) \end{aligned} \quad (22)$$

Similarly, the “*it*-th” specific *PNDE* can be identified (based on eq. (21)):

$$\begin{aligned}
& E[Y_{it}(x, M_{it}(x^*)) - Y_{it}(x^*, M_{it}(x^*)) | \alpha_i, \kappa_{Mi}, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] \\
&= \sum_m \{ E[Y_{it} | X_{it} = x, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it} | X_{it} = x^*, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \} \\
&= \sum_m P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad (\delta_Y + \zeta'_i x + \beta_i m + \kappa_{Yi} t + U_i - \delta_Y - \zeta'_i x^* - \beta_i m - \kappa_{Yi} t - U_i) \\
&= \sum_m P(M_{it} = m | X_{it} = x^*, i) (x - x^*) (\zeta'_i) \\
&= \zeta'_i (x - x^*)
\end{aligned} \tag{23}$$

Finally, the “it-th”-specific *TCE* can be identified as (based on eq. (21)):

$$\begin{aligned}
& E[Y_{it}(x, M_{it}(x)) - Y_{it}(x^*, M_{it}(x^*)) | \alpha_i, \kappa_{Mi}, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] \\
&= \sum_m \{ E[Y_{it} | X_{it} = x, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it} | X_{it} = x^*, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \} \\
&= \sum_m (\delta_Y + \beta_i m + \zeta'_i x + \kappa_{Yi} t + U_i) P(M_{it} = m | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - \sum_m (\delta_Y + \beta_i m + \zeta'_i x^* + \kappa_{Yi} t + U_i) P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&= \zeta'_i (x - x^*) + \beta_i \sum_m m P(M_{it} = m | X_{it} = x, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - \beta_i \sum_m m P(M_{it} = m | X_{it} = x^*, \alpha_i, \kappa_{Mi}, U_i) \\
&= \zeta'_i (x - x^*) + \beta_i (d_M + \alpha_i x + \kappa_{Mi} t + g(U_i)) - \beta_i (d_M + \alpha_i x^* + \kappa_{Mi} t + g(U_i)) \\
&= (\alpha_i \beta_i + \zeta'_i) (x - x^*)
\end{aligned} \tag{24}$$

Consequently, the subject- and period-specific *TCE* equals:

$$\begin{aligned}
& E[Y_{it}(1, M_{it}(1)) - Y_{it}(0, M_{it}(0)) | \alpha_i, \kappa_{Mi}, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] \\
&= \sum_m \{ E[Y_{it} | X_{it} = 1, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = 1, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it} | X_{it} = 0, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \} \\
&= \alpha_i \beta_i + \zeta'_i
\end{aligned} \tag{25}$$

The individual- and period-specific *TNIE* and *PNDE* in turn equal

$$\begin{aligned}
& E[Y_{it}(1, M_{it}(1)) - Y_{it}(1, M_{it}(0)) | \alpha_i, \kappa_{Mi}, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] \\
&= \sum_m \{ E[Y_{it} | X_{it} = 1, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = 1, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it} | X_{it} = 1, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \} \\
&= \alpha_i \beta_i
\end{aligned} \tag{26}$$

and

$$\begin{aligned}
& E[Y_{it}(1, M_{it}(0)) - Y_{it}(0, M_{it}(0)) | \alpha_i, \kappa_{Mi}, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] \\
&= \sum_m \{ E[Y_{it} | X_{it} = 1, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \\
&\quad - E[Y_{it} | X_{it} = 0, M_{it} = m, \beta_i, \zeta'_i, \kappa_{Yi}, U_i] P(M_{it} = m | X_{it} = 0, \alpha_i, \kappa_{Mi}, U_i) \} \\
&= \zeta'_i
\end{aligned} \tag{27}$$

These effects are in line with results from traditional lower-level mediation analysis in linear settings [7]. When we marginalize these effects over individuals, we obtain a TNIE of $E[\alpha_i\beta_i] = E[\alpha_i]E[\beta_i] + \text{Cov}(\alpha_i, \beta_i) = E[\alpha_i]E[\beta_i] + \sigma_{\alpha_i\beta_i}$, a PNDE of $E[\zeta'_i]$ and a TCE 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 (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 (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

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

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 $E(u_{Yi}|M_i^{\text{Sum}}, X_{it})$ 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 $E(u_{Yi}|M_i^{\text{Sum}}, X_{it})$ 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 eq. (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, summarized by eq. (7), the “it-th”-specific TNIE can also be identified, when the assumptions (i)–(vii) from Section 2.2 are satisfied:

$$\begin{aligned} &E[Y_{it}(x, M_{it}(x)) - Y_{it}(x, M_{it}(x^*))|D_i = d, U_i] \\ &= \sum_m \{E[Y_{it}|X_{it} = x, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = x, D_i = d, U_i) \\ &\quad - E[Y_{it}|X_{it} = x, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = x^*, D_i = d, U_i)\} \\ &= \sum_m (\delta_Y + \zeta'x + \beta m + \phi x m + \kappa_Y t + \omega_Y d + v_Y dx + \eta dm + U_i) \\ &\quad \{P(M_{it} = m|X_{it} = x, D_i = d, U_i) - P(M_{it} = m|X_{it} = x^*, D_i = d, U_i)\} \\ &= (\beta + \phi x + \eta d) \left\{ \sum_m m P(M_{it} = m|X_{it} = x, D_i = d, U_i) \right. \end{aligned} \tag{28}$$

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

Similarly, the “it-th”-specific *PNDE* can be identified (based on eq. (7)):

$$\begin{aligned}
& E[Y_{it}(x, M_{it}(x^*)) - Y_{it}(x^*, M_{it}(x^*))|D_i = d, U_i] \\
& = \sum_m \{E[Y_{it}|X_{it} = x, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = x^*, D_i = d, U_i) \\
& \quad - E[Y_{it}|X_{it} = x^*, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = x^*, D_i = d, U_i)\} \\
& = \sum_m P(M_{it} = m|X_{it} = x^*, D_i = d, U_i)(\delta_Y + \zeta'x + \beta m + \phi x m + \kappa_Y t + \omega_Y d \\
& \quad + v_Y dx + \eta dm + U_i - \delta_Y - \zeta'x^* - \beta m - \phi x^* m - \kappa_Y t - \omega_Y d - v_Y dx^* - \eta dm - U_i) \\
& = \{(\zeta' + v_Y d) \sum_m P(M_{it} = m|X_{it} = x^*, D_i = d, U_i) \\
& \quad + \phi \sum_m mP(M_{it} = m|X_{it} = x^*, D_i = d, U_i)\}(x - x^*) \\
& = (\zeta' + v_Y d + \phi E[M_{it} = m|X_{it} = x^*, D_i = d, U_i])(x - x^*) \\
& = (\zeta' + v_Y d + \phi(\delta_M + \alpha x^* + \kappa_M t + \lambda_M c + \omega_M d + v_M dx^* + g(U_i)))(x - x^*)
\end{aligned} \tag{29}$$

Finally, the “it-th”-specific *TCE* can be identified as (based on eq. (7)):

$$\begin{aligned}
& E[Y_{it}(x, M_{it}(x)) - Y_{it}(x^*, M_{it}(x^*))|D_i = d, U_i] \\
& = \sum_m \{E[Y_{it}|X_{it} = x, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = x, D_i = d, U_i) \\
& \quad - E[Y_{it}|X_{it} = x^*, M_{it} = m, D_i = d, U_i]P(M_{it} = m|X_{it} = x^*, D_i = d, U_i)\} \\
& = \sum_m (\delta_Y + \zeta'x + \beta m + \phi x m + \kappa_Y + \omega_Y d + v_Y dx + \eta dm + U_i) \\
& \quad P(M_{it} = m|X_{it} = x, D_i = d, U_i) \\
& \quad - \sum_m (\delta_Y + \zeta'x^* + \beta m + \phi x^* m + \kappa_Y t + \omega_Y d + v_Y dx^* + \eta dm + U_i) \\
& \quad P(M_{it} = m|X_{it} = x^*, D_i = d, U_i) \\
& = (\zeta' + v_Y d)(x - x^*) + (\beta + \phi x + \eta d) \sum_m mP(M_{it} = m|X_{it} = x, D_i = d, U_i) \\
& \quad - (\beta + \phi x^* + \eta d) \sum_m mP(M_{it} = m|X_{it} = x^*, D_i = d, U_i) \\
& = (\zeta' + v_Y d)(x - x^*) + (\beta + \phi x + \eta d)(\delta_M + \alpha x + \kappa_M t + \omega_M d + v_M dx + g(U_i)) \\
& \quad - (\beta + \phi x^* + \eta d)(\delta_M + \alpha x^* + \kappa_M t + \omega_M d + v_M dx^* + g(U_i)) \\
& = ((\alpha + v_M d)(\beta + \eta d) + \zeta' + v_Y d + \phi(d_M + \kappa_M t + \omega_M d + g(U_i)))(x - x^*) \\
& \quad + \phi(\alpha + v_M d)(x^2 - x^{*2})
\end{aligned} \tag{30}$$

References

1. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986;51:1173–82.
2. Loeys T, Moerkerke B, De Smet O, Buysse A, Steen J, Vansteelandt S. Flexible mediation analysis in the presence of nonlinear relations: beyond the mediation formula. *Multivar Behav Res* 2013;48:871–94.
3. Mackinnon DP, Dwyer JH. Estimating mediated effects in prevention studies. *Eval Rev* 1993;17:144–58.

4. Senn S. Cross-over trials in clinical research. New York: John Wiley & Sons, 2002.
5. Bauer DJ, Preacher KJ, Gil KM. Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychol Methods* 2006;11:142–63.
6. Judd CM, Kenny DA, McClelland GH. Estimating and testing mediation and moderation in within-subject designs. *Psychol Methods* 2001;6:115–34.
7. Kenny DA, Korchmaros JD, Bolger N. Lower level mediation in multilevel models. *Psychol Methods* 2003;8:115–28.
8. Pituch KA, Stapleton LM. Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Soc Methods Res* 2012;41:630–70.
9. Preacher KJ, Zyphur MJ, Zhang Z. A general multilevel SEM framework for assessing multilevel mediation. *Psychol Methods* 2010;15:209–33.
10. Raykov T, Mels G. Lower level mediation effect analysis in two-level studies: a note on a multilevel structural equation modeling approach. *Struct Equ Model* 2007;14:636–48.
11. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods* 2010;15:309–34.
12. Pearl J. Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann 2001, 411–20.
13. Pearl J. The causal mediation formula—a guide to the assessment of pathways and mechanisms. *Prev Sci* 2012;13:426–36.
14. VanderWeele T, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface* 2009;2:457–68.
15. VanderWeele TJ. Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Soc Methods Res* 2010;38:515–44.
16. Rubin D. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 1978;6:34–58.
17. Robins J, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;3:153–5.
18. Hafeman DM, Schwartz S. Opening the black box: a motivation for the assessment of mediation. *Int J Epidemiol* 2009;38:838–45.
19. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology* 2013;24:224–32.
20. Robins J, Richardson T. Alternative graphical causal models and the identification of direct effects. In: Shroff K, Keyes P, Ornstein K, editors. *Causality and psychopathology: finding the determinants of disorders and their cures*, 100. Oxford: Oxford University Press, 2011, chapter 6: 1–52.
21. Tucker-Drob EM. Individual differences methods for randomized experiments. *Psychol Methods* 2011;16:298–318.
22. Palta M, Yao T, Velu R. Testing for omitted variables and non-linearity in regression models for longitudinal data. *Stat Med* 1994;13:2219–31.
23. Pearl J. An introduction to causal inference. *Int J Biostat* 2010;6:Article 7.
24. Collins L, Graham J, Flaherty B. An alternative framework for defining mediation. *Multivar Behav Res* 1998;33:295–312.
25. Hayes AF. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Commun Monogr* 2009;76: 408–20.
26. MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. *Prev Sci* 2000;1:173–81.
27. Preacher KJ, Rucker DD, Hayes AF. Addressing moderated mediation hypotheses: theory, methods, and prescriptions. *Multivar Behav Res* 2007;42:185–227.
28. Zhao X, Lynch Jr JG, Chen Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. *J Consumer Res* 2010;37:197–206.
29. MacKinnon DP. Introduction to statistical mediation analysis. New York: Taylor & Francis Group, LLC, 2008.
30. Tofighi D, West SG, MacKinnon DP. Multilevel mediation analysis: the effects of omitted variables in the 1-1-1 model. *Br J Math Stat Psychol* 2013;66:290–307.
31. Begg MD, Parides MK. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Stat Med* 2003;22:2591–602.
32. Kenward MG, Roger JH. The use of baseline covariates in crossover studies. *Biostatistics (Oxford, England)* 2010;11:1–17.
33. Louis T. General methods for analyzing repeated measures. *Stat Med* 1988;7:29–45.
34. Neuhaus J, Kalbfleisch J. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998;54:638–45.
35. Zhang Z, Zyphur MJ, Preacher KJ. Testing multilevel mediation using hierarchical linear models. *Organ Res Methods* 2009;12:695–719.
36. Davis J, Spaeth J, Huson C. A technique for analyzing the effects of group composition. *Am Soc Rev* 1961;26:215–25.
37. Goetgheul S, Vansteelandt S. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* 2008;64:772–80.
38. Edwards JR, Lambert LS. Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis. *Psychol Methods* 2007;12:1–22.
39. Fairchild AJ, MacKinnon DP. A general model for testing mediation and moderation effects. *Prev Sci* 2009;10:87–99.

40. McCulloch CE, Neuhaus JM. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Stat Sci* 2011;26:388–402.
41. Verbeke G, Lesaffre E. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Comput Stat Data Anal* 1997;23:541–56.
42. Vanderhasselt M-A, Brunoni AR, Loeys T, Boggio PS, De Raedt R. Nosce te ipsum – Socrates revisited? Controlling momentary ruminative self-referent thoughts by neuromodulation of emotional working memory. *Neuropsychologia* 2013;51:2581–9.
43. Brunoni AR, Nitsche MA, Bolognini N, Bikson M, Wagner T, Merabet L, et al. Clinical research with transcranial direct current stimulation (tDCS): challenges and future directions. *Brain Stimulation* 2012;5:175–95.
44. Nitsche MA, Cohen LG, Wassermann EM, Priori A, Lang N, Antal A, et al. Transcranial direct current stimulation: state of the art 2008. *Brain Stimulation* 2008;1:206–23.
45. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software* 2014;59:1–38.
46. Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Glymour C, Cooper G, editors. *Computation, causation, and discovery*. Menlo Park, CA and Cambridge, MA: AAAI Press/The MIT Press, 1999: 349–405.
47. Goetgeluk S, Vansteelandt S, Goetghebeur E. Estimation of controlled direct effects. *J R Stat Soc* 2009;70:1049–66.