methods, data, analyses | 2023, pp. 1-25 DOI: 10.12758/mda.2023.02

Many Roads to Mediation: A Methodological and Empirical Comparison of Different Approaches to Statistical Mediation

Dominik Becker

Federal Institute for Vocational Education and Training (BIBB)

Abstract

This paper provides both a theoretical foundation and a simulation analysis of different statistical approaches to mediation. Regarding theory, a brief sketch of the fundamentals of mechanism-based explanations sets the argument of adhering to a consecutive order of predictor, mediator and outcome in mediation analysis. Having summarized the statistical fundamentals of different approaches to mediation analysis including simple mediation within OLS regressions, fixed-effects (FE) regressions, generalized-method-of-moments (GMM) regressions, causal mediation analysis without (CM) and with fixed effects (CMFE), and fixed-effects cross-lagged panel models (FE-CLPMs), I provide a simulation analysis with known but variable values for the intercorrelations between predictor, mediator and outcome in presence of unobserved heterogeneity and reverse causality. The aim of the simulation study is to examine differences in the relative performance of the aforementioned statistical approaches to mediation under different scenarios of causal order.

Results reveal that OLS estimates are generally upwardly biased, FE and CMFE estimates by trend downwardly biased, and the ones of CM models (without FEs) can be biased in both directions. In contrast, coefficients and confidence intervals estimated by both GMM regressions and FE-CLPMs are most accurate – particularly if the structure of lags in the empirical models met the consecutive order set up in the data-generating process. Furthermore, FE-CLPMs are least sensitive to whether the first lag of the outcome variable is included as an additional predictor. All in all, analyses imply the importance that researchers most carefully translate their theoretical assumptions into an empirical model with the appropriate causal order.

Keywords: Panel data, Mediation, Unobserved heterogeneity, Reverse causality, Simulation analysis

Whether an observed association between two social constructs is based on a causal effect is one of the most fundamental methodological questions in the social sciences. Apart from simply asking *if X* causes *Y*, social scientists are concerned with *how* a causal effect is brought about. From a theoretical perspective, this relates to the idea of a *social mechanism M* (Hedström & Swedberg, 1996) along which an effect of *X* on *Y* is transmitted (*X* => *M* => *Y*). Statistically, this perspective translates into the broad field of *mediation analysis* which investigates whether a significant parameter estimate from some type of regression of *Y* on *X* persists once *M* is controlled for. Also, it is possible to specify the share of the *X* => *Y* effect that is transmitted via *M* ("indirect" effect via the mediator), and the residual part ("direct effect"; Baron & Kenny, 1986).

When it comes to the identification of mediation effects in panel data, (at least) two important challenges need to be considered: First, if *unobserved heterogeneity* of either time-constant or time-varying covariates which are exogenous either to *X* or to *M* is present, the seeming mediation effect may be spurious (Imai et al., 2010). Second, a proper measurement of the causal order underlying the *X* => *M* => *Y* chain must ensure that no *reverse causality* (in terms of current values of *X* and/or *M* being endogenous to prior values of *Y*) is present.

The aim of this paper is to explore how well different statistical approaches to mediation analysis are capable of addressing problems of causal order in the presence of unobserved heterogeneity with simulated data. In a brief theoretical section, I will first outline how the idea of mediation analysis relates to the social mechanisms approach to causality in the social sciences. I will then summarize different statistical approaches to mediation analysis and how they address problems of unobserved heterogeneity and reverse causality. Concretely, I will start with the simple "covariate inclusion" approach to mediation analysis in Ordinary Least Squares (OLS) regression. I will then move on to discuss how the introduction of (person) fixed-effects (FE) may solve problems of time-constant unobserved heterogeneity in panel data. A further extension, the Generalized Method of Moments (GMM), the most prominent of which is the Arrelano-Bond (AB) estimator (Arellano & Bond, 1991), additionally addresses the challenge of reverse causality by instrumenting both predictors and outcome by their respective lagged values of first, second, or higher order. A different approach to mediation is given by the causal mediation (CM) approach (Imai, Keele, Tingley, & Yamamoto, 2011) which advances Rubin's (1986) potential outcomes (PO) model by the introduction of potential outcomes for the mediator variable giving treatment status on the one hand, and for the outcome given treatment and mediator status on the other

Direct correspondence to
Dominik Becker, Federal Institute for Vocational Education and Training,
Division 1.3 „Economics of VET", Robert-Schuman-Platz 3, 53175 Bonn, Germany.
E-mail: dominik.becker@bibb.de

hand. As this model has primarily been developed for cross-sectional data, it will prove useful to investigate its applicability to the analysis of panel data. Finally, I will discuss a more recent version of Fixed-Effects Cross-Lagged Panel Models (FE-CLPMs) which addresses both unobserved heterogeneity and reverse causality in the Structural Equation Modeling (SEM) framework (Allison, Williams, & Moral-Benito, 2017).

As the crucial touchstone of this study, I put all of the aforementioned approaches to mediation analysis to the test of an in-depth simulation analysis. Concretely, I will build on Leszczensky and Wolbring's (2019) simulation study to generate random data with known but variable parameters for intercorrelations between *X*, *M*, and *Y* in the presence of both unobserved heterogeneity and reverse causality. I will then explore how well different statistical approaches to mediation analysis can approximate the 'true' parameters. Finally, in the conclusion section, I will summarize the relative advantages of one analysis method over the other and provide practical recommendations in light of the theoretical idea of mediation.

Theoretical Background

Causality and Social Mechanisms

As statistical techniques matured over the course of the 20th century, it has been criticized that the quantitative approach might have gotten lost in "variable sociology", i.e., a mainly data- and model-driven enterprise that lost sight of trying to 'understand' (e.g., Esser, 1996). Luckily, since the 1990s, mainly quantitative sociologists began to place renewed emphasis on the "understanding" dimension of explanation. One prominent proposition is grounded in the philosophy of social (but also life) science and posits a mechanism-based approach to explanation in the social sciences (Hedström, 2005; Hedström & Swedberg, 1996).

There exist numerous definitions of social mechanisms (Hedström & Ylikoski, 2010), the common denominator of which can be described as follows: "Social mechanisms are abstract and general models of spatially, temporally, and functionally organized entities and activities that explain why and how social phenomena are generated by *preceding* causal factors" (Tranow, Beckers, & Becker, 2016, 5f.; my emphasis).

Methodologically, the conceptual idea of a social mechanism as an explanation of why and how social phenomena are generated by preceding causal factors is closely related to the idea of statistical mediation. Consider the mechanism of "wishful thinking" (Elster, 1989): the *desire* for something to be true influences my *belief* about whether it is actually true and, in consequence, my correspond-

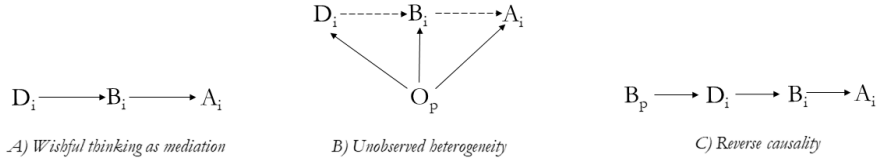

Figure 1 A social mechanism approach to mediation, unobserved heterogeneity and reverse causality.

ing social *action*. For instance, sports betters might overestimate the winning chances of their preferred team (Babad & Katz, 1991).[1]

More generally, the impact of desires D_i on action A_i is brought about via (or, statistically speaking, *mediated* by) beliefs B_i (Figure 1, Panel A). Continuing the above example, the effect of a better's team preference on betting investments would be mediated by the subjective winning chances that the better attributes to their preferred team. But the mechanism approach is also suited to mapping the ideas of unobserved heterogeneity and reverse causality: With respect to unobserved heterogeneity, let O_p refer to an unobserved component of the opportunity structure (O) (e.g., changes in shadow prices) which is *prior* (subscript p) to both individuals' desires D_i, beliefs B_i, and their corresponding action A_i. Let us further assume that O_p brings about D_i, B_i, and A_i. In that case, we would not call desires D_i a social mechanism with causal force (Figure 1, Panel B). Similarly, let us assume that B_p refers to an (even observable) prior instance of belief B_i which brings about desires D_i. In this case of reverse causality and in contrast to the general idea of wishful thinking (cf. panel A), D_i would rather be a mechanism (or statistically: mediator) of B_p effects on A_i (Figure 1, Panel C).[2]

Statistical Approaches to Mediation Analysis

Simple mediation

A seminal definition of mediation analysis was formulated by Baron and Kenny (Baron & Kenny, 1986, p. 1177; also see Figure 2):

1 For the DBO scheme linking individuals' desires and beliefs to situational opportunities see Hedström (2005).

2 There exist of course other forms of heterogeneity that might complicate the identification of mediation effects. Below, I will only briefly touch upon these issues as they surpass what will be covered in the simulation analyses presented below, but I will advise directions for future research in the conclusion section.

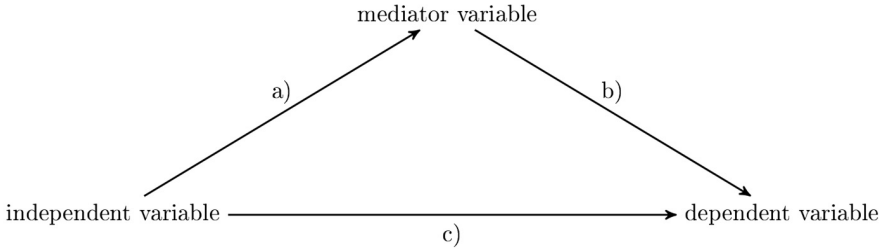

Figure 2 A simple mediation model.

> "A variable functions as a mediator when it meets the following conditions: (a) variations in levels of the independent variable significantly account for variations in the presumed mediator (i.e., Path a), (b) variations in the mediator significantly account for variations in the dependent variable (i.e., Path b), and (c) when Paths a and b are controlled, a previously significant relation between the independent and dependent variables is no longer significant, with the strongest demonstration of mediation occurring when Path c is zero."

It is further common to distinguish between a direct, an indirect, and the total effect of a predictor (or treatment) variable on its outcome. In Figure 2, the direct effect is given by path *c*, the indirect effect is the product of paths *a* and *b*, and the total effect is the sum of both the direct and the indirect effect, i.e. $c + a*b$ (Hayes, Preacher, & Myers, 2011, p. 438).

Consequently, a rigorous application of the simple mediation model in regression analysis would first estimate the effect of an independent variable *X* on the potential mediator variable *M* to ensure that Baron and Kenny's (1986) condition a) is met:

$$M_{(i)} = \beta_{0M} + \beta_1 X_i + \epsilon_{M(i)}. \quad (1)$$

In a second step, the dependent variable of interest *Y* is predicted by *X* (2), and in a third step, by both *X* and *M* (3) to explore whether the effect of *X* on *Y* persists once (2) is controlled for *M*. In practice, both (2) and (3) will often add a vector of covariates *C* to ensure that neither the relation of X nor the one of M to Y is spurious:

$$Y_i = \beta_{0Y} + \beta_2 X_i + \beta_3 C_i + \epsilon_{Y(i)}, \quad (2)$$

$$Y_i = \beta_{0Y} + \beta_2 X_i + \beta_3 C_i + \beta_4 M_i + \epsilon_{Y(i)}. \quad (3)$$

Both unobserved heterogeneity and reverse causality can be addressed in the simple mediation model once we assume to have panel data at our disposal. In that case, unobserved heterogeneity can be addressed using (person-level) fixed

effects (FEs) which 'de-mean' both *X* and *Y* to remove any variation between individuals which is constant over time (e.g., gender, migration background, or the fixed part of personality differences).[3] Adding subscript *t* to refer to observation time, equation (3) amounts to

$$Y_{i(t)} - \bar{Y}_i = \beta_{0Y} + \beta_2\left(X_{i(t)} - \bar{X}_i\right) + \beta_3\left(C_{i(t)} - \bar{C}_i\right) + \beta_4\left(M_{i(t)} - \bar{M}_i\right) + (\alpha_i - \bar{\alpha}_i) + \epsilon_{Yi(t)} - \bar{\epsilon}_{Y(i)}. \quad (4)$$

Since α_i is time-constant by definition, it is identical to its person-specific mean. Consequently, $(\alpha_i - \bar{\alpha}_i)$ amounts to zero, and unobserved heterogeneity is wiped out after demeaning.

FE regressions build on the assumption of strict exogeneity, meaning that current values of $\epsilon_{Yi(t)}$ should not depend on past, present and future values of X_{it} (Brüderl & Ludwig, 2015). This assumption is violated in the case of reverse causality, i.e., when $Y_{i(t)}$ affects $X_{i(t+1)}$ (Leszczensky & Wolbring, 2019). As a consequence, estimates of (4) will be biased if reverse causality is present. To address this issue, researchers often apply 'lags' to *X* or *M*, i.e., they use observations one or even more waves prior to the one in which *Y* is observed. In accordance to the idea of a causal order in terms of changes in *X* affecting changes in *Y* via changes in *M*, one approach could be to predict Y_{it} via $X_{i(t\text{-}2)}$ and $M_{i(t\text{-}1)}$, i.e., applying the first lag to the mediator of interest, and the second lag to the main predictor at hand:

$$Y_{i(t)} - \bar{Y}_i = \beta_{0Y} + \beta_2\left(X_{i(t-2)} - \bar{X}_i\right) + \beta_3\left(C_{i(t)} - \bar{C}_i\right) + \beta_4\left(M_{i(t-1)} - \bar{M}_i\right) + \epsilon_{Yi(t)} - \overline{\epsilon_{Y(i)}}. \quad (5)$$

However, it has been shown both analytically and based on simulations that lags of either variable do not circumvent biased estimates and statistical inference in the case of reverse causality (Reed, 2015). A more generalized approach that also relies on lagged variables, but tries to resolve identification issues of previous approaches, is the Generalized Method of Moments (GMM), a particular version of which is known as the Arellano-Bond (AB) estimator (Arellano & Bond,

3 There are several methods to address the problem of unobserved heterogeneity in panel data: *first-differences,* where each current value of a variable is subtracted by the one of the previous wave, *person dummies*, which include dummy variables for all *n-1* individuals in the sample, and *demeaning*, where each value of a variable is subtracted by its unit-specific mean over time. The latter approach is explained more extensively below and is also the one that will be used in the simulation study to follow.

1991). In its most simplistic form, the AB approach starts from the following model:[4]

$$Y_{i(t)} = \beta_1 Y_{i(t-1)} + \beta_2 X_{i(t)} + \alpha_i + \epsilon_{i(t)}. \quad (6)$$

As a first step, first-differences for all terms in (6) are computed to get rid of time-constant unobserved heterogeneity α_i:

$$\Delta Y_{i(t)} = \beta_1 \Delta Y_{i(t-1)} + \beta_2 \Delta X_{i(t)} + \Delta \epsilon_{i(t)}. \quad (7)$$

As a second step, $Y_{i(t\text{-}2)}$ is used as an instrument for $\Delta Y_{i(t\text{-}1)}$. In practice, and as recommended by the authors, additional higher-order lags of Y ($\Delta Y_{i(t\text{-}3)}$, $\Delta Y_{i(t\text{-}4)}$, …) are often used to instrument $\Delta Y_{i(t\text{-}1)}$ (Arellano & Bond, 1991). Alternatively, or in addition, $\Delta Y_{i(t\text{-}1)}$ may be instrumented by second, third, or even higher-order differences of Y ($\Delta Y_{i(t\text{-}2)}$, $\Delta Y_{i(t\text{-}3)}$, …). By this design, it is possible to separate strictly exogenous from sequentially exogenous or predetermined variables from one another. Consequently, "AB-type panel estimators thus weaken the exogeneity assumption for a subset of regressors, thereby providing consistent estimates even if reverse causality is present" (Leszczensky & Wolbring, 2019, p. 9).

Yet, despite this pleasant statistical property, real-world applications of the AB estimator are not without pitfalls: As Allison et al. (2017) outline, while the AB-estimator provides consistent estimators, "there is evidence that the estimators are not fully efficient, have considerable small-sample bias, and often perform poorly when the autoregressive parameter (the effect of a variable on itself at a later point in time) is near 1.0" (p. 1f.). In my discussion of the FE-CLPM, I will come back to how these drawbacks may be circumvented by a maximum-likelihood approach.

Causal mediation analysis

Imai, Keele, et al. (2011) advance the idea of mediation analysis as a methodological approximation to causal mechanisms within the potential outcomes (PO) framework (Rubin, 1986). In contrast to previous common practice when social scientists tended to interpret each estimate of multivariate analysis as causal, the PO approach focuses on the causal identification of solely *one* effect, called treatment T, on the outcome of interest, Y. Although the question of how a particular individual i in the treatment group would have behaved had they not received the treatment cannot be answered empirically, it can be approximated by comparing outcome Y of the treatment group ($Y_i|T{=}1$) with the non-treatment group ($Y_i|T{=}0$):

4 As a distinct AB-type equation for the mediator is not shown, subscript $_Y$ is omitted for now.

$$T_i \equiv Y_i(1) - Y_i(0). \tag{8}$$

The next step is to introduce the mediator variable into the PO main equation. For dichotomous mediators, outcome *Y* in the treatment group under the condition of *M*=1 ($Y_i|T$=1, *M*=1) is compared to *Y* in the non-treatment group under the condition of M=0 ($Y_i|T$=0, *M*=0):

$$T_i \equiv Y_i\big(1, M_i(1)\big) - Y_i\big(O, M_i(0)\big). \tag{9}$$

Having defined mediation in the PO framework, it is possible to define the indirect or *causal* mediation effect

$$\delta_i(t) \equiv Y_i\big(t, M_i(1)\big) - Y_i\big(t, M_i(0)\big). \tag{10}$$

which refers to paths a) and b) in Figure 2, as well as the direct/residual effect

$$\zeta_i(t) \equiv Y_i\big(1, M_i(t)\big) - Y_i\big(0, M_i(t)\big) \tag{11}$$

which amounts to path c) in Figure 2 .

Another important assumption for causal mediation in the potential outcomes framework is the one of *sequential ignorability* (SIA), which can be decomposed into *ignorability of treatment assignment* (ITA) given *X*,

$$\{Y_i(t', m), M_i(t)\} \perp T_i \vee X_i = x, \tag{12}$$

and *ignorability of mediator status* (IMS) given $T + X$:

$$Y_i(T, m) \perp M_i(t) \vee T_I = t, X_i = x. \tag{13}$$

Concretely, ITA given *X* in (12) means that having controlled for a vector of covariates (which is here denoted *X*), it should be random whether a particular individual *i* belongs to the treatment or to the control group. Furthermore, IMS given *T* and *X* in (13) means once I know whether individual *i* belongs to the treatment or to the control group *and* I have controlled for my set of covariates *X*, there should (by assumption) be no other systematic variation in the mediator variable.

How are unobserved heterogeneity and reverse causality addressed in the causal mediation model? Regarding unobserved heterogeneity, the SIA is crucial: If the set of covariates *C* is exhaustive and both treatment and mediator status are independent of unmeasured confounders, unobserved heterogeneity is no issue by definition. For particular scenarios in which the causal effect of *T* on *Y* is passed on across a second, unobserved mediator *N* that either runs parallel to the observed mediator *M* or is endogenous to the latter (Figure 3, Panel A; taken by Imai, Keele, et al., 2011, p. 786), the SIA is violated but can yet be addressed via sensitivity analyses in which the correlation between the residual terms of

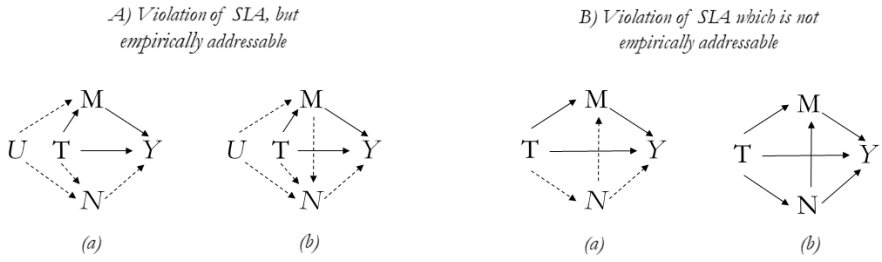

Figure 3 Methodological challenges of the causal mediation model. Summary of Imai, Keele, et al. (2011, 786f.)

both the mediator and the outcome equation is examined (Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010). For that purpose, it is useful to specify mediation in the linear structural equation framework again (Imai, Keele, & Yamamoto, 2010, p. 57; Imai, Keele, et al., 2011, p. 774): In our notation (cf. equations (1) and (2)), the correlation of interest is defined as $\rho = corr(\epsilon_{Y(i)}, \epsilon_{M(i)})$. The magnitude of ρ can be used to measure to what extent the SIA is violated: in the case of no violation, ρ should amount to zero; the more severely the model deviates from this ideal state, the larger ρ. The key element of the sensitivity analysis is now to approximate the unobserved mediator by a random variable whose correlations with *T*, *M* and *Y* are varied over the course of the estimation process. As an alternative measure of potential bias due to an unobserved mediator, relative changes in R^2 can be used. In contrast, the case of *M* being endogenous to an unobserved mediator *N* constitutes a severe threat to the SIA and cannot be addressed by sensitivity analyses (Figure 3, Panel B).

Concerning reverse causality between *T*, *M* and *Y*, the causal mediation proponents simply state that "[l]ongitudinal data with covariates (realized and measured before treatment assignment) and treatment assignment (realized and measured before outcomes) eliminates the possibility of reverse causality and thus provides a clear way to adhere to this prescription of design followed by analysis" (Imai, Jo, & Stuart, 2011, p. 868). Since it is well known, however, that a discrete longitudinal measurement of relevant indicators (i.e., in terms of annual panel waves) is no insurance against *unobserved* forms of reverse causality (Leszczensky & Wolbring, 2019), it remains an open question as to how the causal mediation approach can handle this challenge. I will address this issue in my simulation analysis section.[5]

5 Lutz, Sordillo, Hokanson, Chen Wu, and Lange (2020) provide a first insight into how sensitively the causal mediation approach reacts to reverse causality. However, they do not consider the case in which both unobserved heterogeneity and reverse causality is present simultaneously.

SEM approach to mediation

The SEM approach to mediation advances the simple mediation model both structurally and in terms of measurement: First, as longitudinal data is structurally arranged in 'wide' format, more complex mediational structures (e.g., two mediators at once) can be easily implemented. Second, the SEM approach holds a more elaborate perspective on the measurement component of the constructs at hand, which amounts to the option of using latent variable models for both predictor variable(s), mediator(s), outcome(s), and covariates. As for the ease of comparison between mediation approaches I will refrain from using latent variable models in the simulation models; the formal details to follow will focus on observed variable models which are just a special case of latent variable models.

For a conventional "*x* 'causes' *y*" model without any mediator, the structural part is defined as in conventional OLS regression analysis (cf. Bollen, 1989, 41ff.):

$$Y = \gamma_1 X + \zeta_1, \tag{14}$$

where Y denotes the dependent variable, X the independent variable with regression weight γ_1 on Y, and ζ_1 the error, residual or disturbance term.

As before, a mediator variable M can be introduced by setting it exogenous to Y and endogenous to X:

$$M = \gamma_2 X + \zeta_2, \tag{15}$$

$$Y = \gamma_1 X + \gamma_3 M + \zeta_1. \tag{16}$$

As usual, the indirect effect for observed variable models is defined as the difference between the total effect and the direct effects. For latent variable models, the decomposition of direct, indirect and total effects is more complex (see Bollen, 1989, 376ff.). Luckily, modern statistical software which is capable of estimating SEMs – such as *Stata*, *R* (with *lavaan* in particular) or *Mplus* – provides handsome sub-routines to decompose total, direct and indirect effects in both observed and latent variable models (see, e.g., Mehmetoglu, 2018; Muthén, 2017; Rosseel, 2012).

While the added value of mediation of observed variables within the SEM approach may not be evident at first sight, its advantage becomes more obvious when it comes to addressing the challenge of *reverse causality* in panel data. There is a long tradition within the SEM approach to do so by means of *cross-lagged panel models* (CLPMs; also see Finkel, 1995). Taking advantage of the wide data structure underlying the SEM approach, in case of a predictor X and an outcome Y measured at times t_1 and t_2, a cross-lagged panel model applies the following steps:

$$X_2 = \beta_1 X_1 + \beta_2 Y_1 + \zeta_X, \tag{17}$$

$$Y_2 = \beta_3 Y_1 + \beta_4 X_1 + \zeta_Y. \tag{18}$$

That is, Y_2 is regressed on both X_1 and Y_1, while at the same time, X_2 is regressed on both X_1 and Y_1. Apart from simply controlling for potential reverse causality effects, one appeal of the CLPM is that reciprocal effects which are often assumed by theory can be directly estimated (Selig & Little, 2012, p. 268). A crucial objection that has been raised against the CLPM is that it may lead to biased results in case of unobserved stable individual-level characteristics (Hamaker, Kuiper, & Grasman, 2015). There have already been several approaches to incorporate the FE estimator into the SEM framework both with and without a cross-lagged structure (Allison, 2009; Curran & Bollen, 2001). A more recent approach to Fixed-Effects Cross-Lagged Panel Models (FE-CLPMs) by Allison et al. (2017) draws on previous work of Moral-Benito (2013) who has outlined a maximum-likelihood-based estimation method that circumvents several computational drawbacks of GMM estimators in general and of the AB method in particular. The contribution of Allison et al. (2017) is to integrate Moral-Benito's (2013) approach into the general SEM framework, as a consequence of which it can be estimated using conventional SEM software subroutines.

The FE-CLPM is defined as follows:

$$Y_{i(t)} = \mu_{(t)} + \beta_1 X_{i(t-1)} + \beta_2 Y_{i(t-1)} + \delta_1 W_{i(t)} + \gamma_1 Z_i + \alpha_i + \epsilon_{i(t)}, \tag{19}$$

$$X_{i(t)} = \tau_{(t)} + \beta_3 X_{i(t-1)} + \beta_4 Y_{i(t-1)} + \delta_2 W_{i(t)} + \gamma_2 Z_i + \eta_i + \nu_{i(t)}. \tag{20}$$

where in (19) μ_t describes the intercept of Y that varies across time t, β_1 and β_2 are scalar coefficients assessing how Y is predicted by former values of both X and Y, δ_1 and γ_1 are row vectors of coefficients for both time-variant controls variables W and time-constant control variables Z, α_1 refers to the joint effects of time-constant unsobservables (assuming them to exert constant effects on $Y_{i(t)}$, and $\epsilon_{i(t)}$ is a random error term.

Accordingly, in (20), $\tau_{(t)}$ describes the intercept of X that varies across time t, β_3 and β_4 are scalar coefficients assessing how X is predicted by former values of both X and Y, δ_2 and γ_2 are row vectors of coefficients for both time-variant controls variables W and time-constant control variables Z, η_1 refers to the joint effects of time-constant unsobservables (assuming them to exert constant effects on $X_{i(t)}$), and $\nu_{i(t)}$ is a random error term.

The most notable difference compared to the 'traditional' CLPM presented in (17)-(18) is the inclusion of terms α_1 and η_1 to address time-constant unobserved effects on $Y_{i(t)}$ and $X_{i(t)}$, respectively. In econometric approaches, α_1 and η_1 are often assumed to be "fixed", i.e., exert the same effect for each individual, whereas in other social science disciplines, this assumption might be relaxed (e.g., Hamaker et al., 2015).

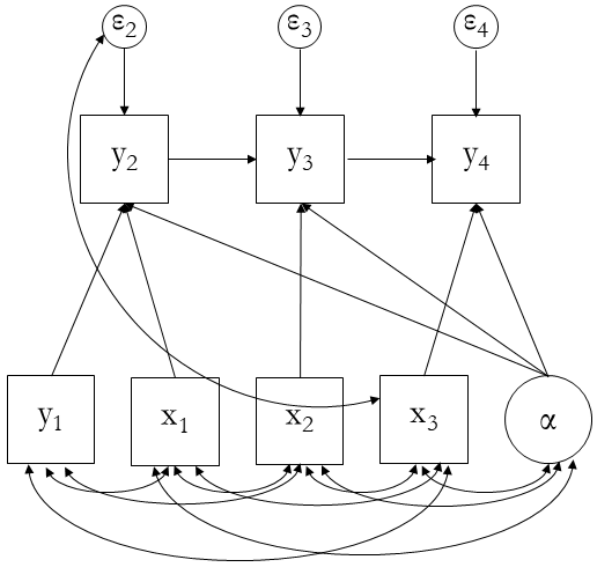

Figure 4 The FE-CLPM. Source: Allison et al. (2017, 6).

To recall, a combination of fixed effects and lagged outcome variables will lead to biased estimates of the β coefficients. Within the AB approach, this has been addressed by, first, removing fixed effects by computing first differences for *X* and *Y*, and then, second, instrumenting these differences by lagged difference scores (cf. eq. (7)), which are finally, third, estimated by GMM. It is well-known, however, that GMM approaches are particularly sensitive to the number of lags and corresponding instruments (Leszczensky & Wolbring, 2019; Roodman, 2009). In contrast, the ML approach to reverse causality produces estimators that are asymptotically equivalent to GMM, but have more preferable finite sample properties in case of weak and/or numerous instruments (Moral-Benito, 2013).

In what follows, Allison et al. (2017) argue that the ML approach to the cross-lagged model with fixed effects is a special case of the general SEM framework outlined in (12) which is illustrated in Figure 4 Leaving aside both *W* and *Z* variables and focusing on the case of manifest *X* and *Y* the latter of which is measured on four occasions, it is evident that while Y_t is predicted by $Y_{t\text{-}1}$, this is not the case for instances of *X* which are simply allowed to correlate with one another. In addition, each Y_t is predicted by $X_{t\text{-}1}$ as well as α_1, which is the FE estimate intended to address time-constant unobserved heterogeneity. Coefficient α_1, in turn, correlates with all instances of *X* (but is not allowed to correlate with

any time-invariant observable Z if the latter is present in the model).[6] Finally, and of crucial importance, x_3 is allowed to correlate with ϵ_2, the error term of Y_2. According to Allison et al. (2017, 6), it is this correlation that makes X predetermined (by Y). In other words, this correlation is the crucial leverage to account for reverse causality between X and Y.

Observed heterogeneity and interim conclusion

Apart from the challenges of reverse causality and *unobserved* heterogeneity, the statistical approaches just discussed can also address several issues of *observed* heterogeneity. There are different terms by which this kind of heterogeneity is referred to, the most prominent of which are interaction effects, moderator effects, multiplicative effects, and treatment effect heterogeneity (Baron & Kenny, 1986; Brambor, Clark, & Golder, 2006; Xie, Brand, & Jann, 2012). As a common denominator, a predictor (or treatment) variable is multiplied (i.e., "interacted"), with an observed variable Z. In our case, we can generally distinguish three possible interaction terms: *i)* between the main predictor (or treatment) variable (usually denoted X or T) and another moderating variable Z; *ii*) between the mediator M and Z, and between X (or T) and M. It can be formally outlined that the above approaches are generally capable to address either form of observed heterogeneity (available upon request). In contrast, and as outlined above, they differ in their capacity to address *unobserved* heterogeneity and reverse causality. The essence of this methodological comparison is tabulated in Table 1.

6 As a consequence of this identificatory step, it is advised to exclude all time-constant variables from the estimation model (Allison et al. 2017: 6).

Table 1 Comparison of different statistical approaches to mediation analysis in their capacity to address several methodological challenges

	Observed heterogeneity	Unobserved heterogeneity	Reverse causality
OLS	Can incorporate interactions of type *XZ*, *MZ*, and *XM*	Not in baseline model, but can be advanced to FE estimator by manual demeaning	May incorporate lags of *X* and *Y*, but results will be biased
FE	Can incorporate interactions of type *XZ*, *MZ*, and *XM*	Rules out time-constant unobserved heterogeneity by demeaning all variables	May incorporate lags of *X* and *Y*, but results will be biased
AB/GMM	Can incorporate interactions of type *XZ*, *MZ*, and *XM*	See FE	First-differences for *X* and *Y* instrumented by higher-order lags
CM	Can incorporate interactions of type *XZ* and *XM* (unclear if *MZ* identified)	See OLS. Yet, empirical performance of manual approach untested hitherto.	May incorporate lags of *X* and *Y*, but empirical performance of this approach untested hitherto.
SEM	Can incorporate interactions of type *XZ*, *MZ*, and *XM*	Not in baseline model	Addressed by cross-lagged panel-model
FE-CLPM	Can incorporate interactions of type *XZ*, *MZ*, and *XM*	Introduces variables α and η to capture unobserved heterogeneity effects on *X* and *Y*, respectively	See SEM

Simulation Analysis

The Present Study

Previous simulation studies have revealed that both OLS and FE analysis are biased when both unobserved heterogeneity and reverse causality are present (Leszczensky & Wolbring, 2019). Other research based on simulation analysis suggests that GMM strategies such as the AB estimator can run into problems, for instance, when the number of waves is small and lags are long (Newey & Windmeijer, 2009; Windmeijer, 2005). Further simulation studies suggest that the FE-CLPM can keep up with the GMM approach in the presence of both unobserved heterogeneity and reverse causality (Allison et al., 2017; Moral-Benito, Allison,

& Williams, 2019; also see Leszczensky & Wolbring, 2019). Yet, two gaps in research can be identified which the present contribution aims to address.

First, it has not yet been explored if these results generalize to the inclusion of a mediator variable which, in an ideal-world data-generating process (DGP), will be preceded by the main predictor but succeeded by the outcome (see below). Second, it has not been tested how the gold standard in mediation analysis, the causal mediation model in the potential-outcomes framework, performs if the challenges of unobserved heterogeneity and reverse causality are addressed by "on-board resources" in terms of demeaning and lagging all relevant variables.

Consequently, I will now present a simulation analysis to evaluate which of the statistical approaches to mediation analysis identifies the parameter values of predictor *X*, a mediator *M*, and their corresponding lags – which have been specified in the DGP prior to the simulation analysis – with minimal bias.

Parameters and scenarios of the simulation model

My simulation analysis builds on the one by Leszczensky and Wolbring (2019) but advances it by including an additional variable *M* which shall mediate the effect of *X* on *Y* in the simulated data set. I first generated data with intercorrelations of $\rho_{\{X,M,Y\}} = .5$ and standard normally distributed independent error terms at t_0, respectively. This data was expanded to waves 1-5 in a second step by the following data-generating process (DGP):

$$Y_{it} = \beta_1 Y_{it-1} + \beta_2 X_{it-2} + \beta_3 M_{it-1} + \beta_4 Z_i + \epsilon_{it} \quad \text{with} \quad \epsilon_{it} \sim N(0;1),$$

$$X_{it} = \beta_5 Y_{it-1} + \beta_6 Z_i + \mu_{it} \quad \text{with} \quad \mu_{it} \sim N(0;1),$$

$$M_{it} = \beta_7 Y_{it-1} + \beta_8 X_{it-1} + \beta_9 Z_i + \nu_{it} \quad \text{with} \quad \nu_{it} \sim N(0;1).$$

Above, β_1 refers to the extent of autocorrelation for outcome *Y*. As the variation of β_1 had no substantial impact on the simulation results by Leszczensky and Wolbring (2019), I set the parameter to be constant ($\beta_1 = .5$). Most important, Y_{it} is modeled as an outcome of both $X_{it\text{-}2}$ (with effect β_2) and $M_{t\text{-}1}$ (with effect β_3). That is, in accordance to the idea of a social mechanism which is by definition situated *between* a cause and its outcome, the DGP understands the mediation model as the statistical pendent of a mechanism-based explanation. Consequently, the consecutive order of *X, M* and *Y* is of vital importance here. While Leszczensky and Wolbring (2019) switch between contemporaneous and lagged effects of *X* on *Y*, my model is more simplistic in assuming constant effects of $X_{it\text{-}2}$ on $Y_{(t)}$.

In addition, *Z* denotes an unmeasured, time-constant normally-distributed variable that addresses the challenge of unobserved heterogeneity. *Z* is associated with *Y*, *X*, and *M* by parameters β_4, β_6 and β_9, respectively. To simplify the

Table 2 Parameter values of the simulation analysis

Parameter	Concept	Values
β_1	Autocorrelation of Y	0.5
β_2	Effect of X_{t-2} on Y_t	0, 0.5
β_3	Effect of M_{t-1} on Y_t	0, 0.5
β_8	Effect of X_{t-1} on M_t	0, 0.5
β_4 / β_6 / β_9	Unobserved heterogeneity on Y, X, M, respectively	0.5
β_5 / β_7	Reverse causality on X and M, respectively	0.5

simulation model, these were set to 0.5 (unobserved heterogeneity moderately present), respectively. For all possible combinations of parameters (which are summarized in Table 2), 500 datasets with 500 observations each were generated.

Models

To compare point estimates and corresponding confidence intervals of the aforementioned mediation approaches, for either of them, the same set of sub-models will be estimated. Concretely, for both 1) FE regressions, 2) the GMM approach, 3) the causal mediation (CM) approach, and 4) the FE-CLPM, the following scenarios will be compared (see Table 3): *Scenario A)* employs a simultaneous analysis of *Y* predicted by the variables at the same point in time. *Scenario B)* takes the first lag of all variables to predict later instances of Y. *Scenario C)* follows the idea of a consecutive order between *X*, *M*, and *Y* (which is inspired by the rationale of mechanism-based explanations) by modeling *Y* by the second lag of *X* and the first lag of *M*. Finally, *scenario D)* amends *scenario C)* by adding the first lag of *Y* to account for potential reverse causality between *X* and *Y*.

Moreover, for each scenario, the following two submodels are estimated: *Submodel i)* predicts *Y* only by *X* (or its first or second lag) or, as in *scenario D)*, the first lag of *Y*, and *submodel ii)* adds the mediator variable *M* (or its first lag).

Table 3 Scenarios for the simulation study

Scenario	*Submodel i)*	*Submodel ii)*
A) Simultaneous scenario (no lags)	$Y_t = X_t$	$Y_t = X_t + M_t$
B) Lagged scenario	$Y_t = X_{t\text{-}1}$	$Y_t = X_{t\text{-}1} + M_{t\text{-}1}$
C) Consecutive scenario	$Y_t = X_{t\text{-}2}$	$Y_t = X_{t\text{-}2} + M_{t\text{-}1}$
D) Consecutive scenario + L1(Y)	$Y_t = Y_{t\text{-}1} + X_{t\text{-}2}$	$Y_t = Y_{t\text{-}1} + X_{t\text{-}2} + M_{t\text{-}1}$

Results

Tables 4-6 show the results of the simulation study. Table 4 lists the predicted β coefficients and their corresponding standard errors for both OLS and FE analyses of the simulated data. Between columns, it is differentiated between the four data simulation scenarios (see Table 3). Between rows, the values for the regression parameters are varied (see Table 2), and it is differentiated between two submodels one of which predicts *Y* only by *X*, and the other one by both *X* and *M*. If the predicted β coefficients of *X* and/or *M* are subject to a bias of $|\varepsilon_\beta| > 0.1$, the background color of the corresponding table cell is highlighted in different shades of green for upward bias, and in different shades of red for downward bias (see the explanatory notes below Tables 4-6). In addition, Figures A1-A6 in Appendix A show coefficient plots of all parameter estimates and corresponding confidence intervals. These plots may provide visual aid to answer the question of if the statistical approaches applied to the simulation models correctly identify mediation effects which may or may not have been set in the underlying DGP.

For the *OLS approach*, when all β coefficients have been set to zero, the predicted effects of *X* and *M* on *Y* are *overestimated* given they have been set to be absent in the DGP (see left panel of Table 4). The upward bias within this particular setting is largest in the lagged scenario, and smallest in the consecutive scenario controlled for the first lag of *Y*. Once β_2 and/or β_3 are set to .5, this pattern persists for most of the predicted effects of *X*, and their bias is generally larger as long as the analyses have not controlled for *M*. If they do, the OLS approach incorrectly identifies mediation effects of *M* although β_8 is still set to zero (see Appendix A, Figure A1a). Furthermore, if β_8 is set to .5, the amount of mediation predicted by the OLS approach is way too high particularly in case of β_3 =.5 (Appendix A, Figure A1b). The general upward bias of the OLS approach is most pronounced if both β_2 and β_3 are set to .5. In contrast, when both β_2 and β_8 are set to .5, predicted effects of *X* may be slightly downwardly biased in the contemporary and lagged scenarios given they have not been controlled for *M*.

Table 4 Results of Ordinary Least Squares (OLS) and Fixed-Effects (FE) regressions of simulated data

			OLS Regressions								**FE Regressions**							
			contemporary		*lagged*		*consecutive*		*consecutive + Ly*		*contemporary*		*lagged*		*consecutive*		*consecutive + Ly*	
			$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)
$\beta_2 = \beta_3 = \beta_8 = 0$	Y=|X|	X (β_2)	**0.375**	0.018	**0.385**	0.019	**0.375**	0.021	**0.236**	0.019	**-0.063**	0.019	-0.043	0.023	0.001	0.027	-0.017	0.026
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.252**	0.019	**0.263**	0.021	**0.256**	0.021	**0.178**	0.019	**-0.046**	0.019	-0.027	0.024	-0.006	0.027	-0.031	0.026
		M (β_3)	**0.221**	0.018	**0.219**	0.020	**0.244**	0.020	**0.169**	0.019	**-0.088**	0.020	**-0.087**	0.023	**-0.081**	0.028	**-0.137**	0.026
$\beta_2 = 0.5, \beta_3 = 0, \beta_8 = 0$	Y=|X|	X (β_2)	**0.637**	0.018	**0.661**	0.020	**0.831**	0.019	**0.627**	0.020	**-0.129**	0.020	**-0.182**	0.023	**0.493**	0.026	**0.433**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.452**	0.022	**0.463**	0.024	**0.695**	0.020	**0.585**	0.020	**-0.114**	0.021	**-0.159**	0.024	**0.481**	0.026	**0.406**	0.027
		M (β_3)	**0.294**	0.020	**0.317**	0.023	**0.236**	0.019	**0.140**	0.019	**-0.073**	0.022	**-0.112**	0.024	**-0.087**	0.027	**-0.147**	0.026
$\beta_2 = 0, \beta_3 = 0.5, \beta_8 = 0$	Y=|X|	X (β_2)	**0.645**	0.021	**0.708**	0.021	**0.681**	0.025	**0.301**	0.024	**-0.117**	0.022	**0.051**	0.024	**-0.062**	0.029	-0.041	0.026
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.349**	0.021	**0.254**	0.020	**0.230**	0.020	**0.128**	0.020	**-0.069**	0.021	-0.025	0.023	-0.014	0.026	-0.013	0.025
		M (β_3)	**0.430**	0.022	**0.669**	0.019	**0.705**	0.018	**0.600**	0.020	**-0.231**	0.024	**0.410**	0.023	**0.416**	0.028	**0.284**	0.029
$\beta_2 = 0.5, \beta_3 = 0.5, \beta_8 = 0$	Y=|X|	X (β_2)	**0.923**	0.020	**1.012**	0.022	**1.172**	0.023	**0.678**	0.026	**-0.094**	0.024	-0.041	0.024	**0.427**	0.028	**0.392**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.552**	0.023	**0.453**	0.023	**0.674**	0.019	**0.545**	0.021	**-0.064**	0.025	**-0.133**	0.024	**0.479**	0.026	**0.440**	0.025
		M (β_3)	**0.502**	0.023	**0.773**	0.022	**0.700**	0.018	**0.576**	0.019	**-0.121**	0.028	**0.402**	0.025	**0.412**	0.027	**0.314**	0.027
$\beta_2 = 0, \beta_3 = 0, \beta_8 = 0.5$	Y=|X|	X (β_2)	**0.375**	0.018	**0.385**	0.019	**0.375**	0.021	**0.236**	0.019	**-0.063**	0.019	-0.043	0.023	0.001	0.027	-0.017	0.026
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.170**	0.019	**0.187**	0.021	**0.134**	0.027	**0.093**	0.024	**-0.054**	0.019	-0.037	0.023	0.034	0.030	0.037	0.029
		M (β_3)	**0.237**	0.014	**0.230**	0.016	**0.244**	0.020	**0.169**	0.019	**-0.090**	0.017	**-0.083**	0.020	**-0.081**	0.028	**-0.137**	0.026
$\beta_2 = 0.5, \beta_3 = 0, \beta_8 = 0.5$	Y=|X|	X (β_2)	**0.637**	0.018	**0.661**	0.020	**0.831**	0.019	**0.627**	0.020	**-0.129**	0.020	**-0.182**	0.023	**0.493**	0.026	**0.433**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.321**	0.021	**0.243**	0.024	**0.577**	0.026	**0.515**	0.025	**-0.117**	0.020	**-0.188**	0.024	**0.524**	0.028	**0.480**	0.029
		M (β_3)	**0.328**	0.015	**0.442**	0.017	**0.236**	0.019	**0.140**	0.019	**-0.124**	0.020	**0.087**	0.022	**-0.087**	0.027	**-0.147**	0.026
$\beta_2 = 0, \beta_3 = 0.5, \beta_8 = 0.5$	Y=|X|	X (β_2)	**0.788**	0.021	**0.863**	0.022	**0.941**	0.025	**0.487**	0.026	**-0.115**	**0.022**	0.002	0.023	**0.182**	0.028	**0.172**	0.026
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.281**	0.021	**0.172**	0.020	**0.109**	0.026	**0.048**	0.024	**-0.098**	0.022	-0.038	0.022	0.026	0.029	**0.064**	0.027
		M (β_3)	**0.482**	0.016	**0.675**	0.015	**0.702**	0.018	**0.581**	0.020	**-0.132**	0.022	**0.411**	0.020	**0.412**	0.028	**0.292**	0.028
$\beta_2 = 0.5, \beta_3 = 0.5, \beta_8 = 0.5$	Y=|X|	X (β_2)	**1.054**	0.020	**1.170**	0.023	**1.429**	0.022	**0.919**	0.027	-0.048	0.029	-0.049	0.026	**0.674**	0.027	**0.632**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.421**	0.024	**0.215**	0.024	**0.563**	0.025	**0.494**	0.025	-0.049	0.027	**-0.168**	0.023	**0.519**	0.027	**0.516**	0.027
		M (β_3)	**0.568**	0.018	**0.887**	0.017	**0.692**	0.017	**0.575**	0.018	0.005	0.029	**0.627**	0.021	**0.416**	0.026	**0.344**	0.026

Note: Bold coefficients are at least 1.96 their standard errors. Average bias ε_β of predicted β parameters: $0.1 \leq \varepsilon_\beta < 0.3$; $-0.1 \geq \varepsilon_\beta > -0.3$; $0.3 \leq \varepsilon_\beta < 0.5$; $-0.3 \geq \varepsilon_\beta > -0.5$; $\varepsilon_\beta > 0.5$; $\varepsilon_\beta < -0.5$.

Table 5 Generalized Method of Moments (GMM) and Fixed-Effects Cross-Lagged Panel-Model Regressions (FE-CLPM) of simulated data

			GMMs								**FE-CLPMs**							
			contemporary		*lagged*		*consecutive*		*consecutive + Ly*		*contemporary*		*lagged*		*consecutive*		*consecutive + Ly*	
			$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)
$\beta_2 = \beta_3 = \beta_8 = 0$	Y=|X|	X (β_2)	**0.375**	0.013	0.022	0.036	0.015	0.048	0.005	0.039	**-0.061**	0.019	-0.043	0.023	0.000	0.027	0.002	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.252**	0.016	0.022	0.034	-0.004	0.049	-0.005	0.041	**-0.044**	0.019	-0.028	0.024	-0.001	0.027	0.001	0.028
		M (β_3)	**0.221**	0.016	-0.022	0.039	-0.014	0.053	-0.015	0.050	**-0.088**	0.020	**-0.087**	0.023	-0.007	0.029	-0.001	0.033
$\beta_2 = 0.5, \beta_3 = 0, \beta_8 = 0$	Y=|X|	X (β_2)	**0.637**	0.014	**-0.592**	0.043	**0.501**	0.052	**0.517**	0.044	**-0.118**	0.020	**-0.183**	0.024	**0.492**	0.026	**0.496**	0.028
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.452**	0.019	**-0.490**	0.038	**0.482**	0.049	**0.501**	0.053	**-0.107**	0.021	**-0.158**	0.024	**0.490**	0.026	**0.494**	0.029
		M (β_3)	**0.294**	0.020	**-0.195**	0.043	-0.023	0.053	-0.012	0.054	**-0.070**	0.022	**-0.112**	0.024	-0.015	0.029	-0.010	0.032
$\beta_2 = 0, \beta_3 = 0.5, \beta_8 = 0$	Y=|X|	X (β_2)	**0.645**	0.015	**0.090**	0.044	**-0.134**	0.063	**-0.139**	0.036	**-0.115**	0.021	**0.051**	0.024	-0.055	0.029	**-0.055**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.349**	0.021	0.023	0.035	-0.024	0.048	-0.031	0.041	**-0.075**	0.020	-0.026	0.023	-0.007	0.027	-0.006	0.027
		M (β_3)	**0.430**	0.021	**0.477**	0.039	**0.473**	0.051	**0.438**	0.091	**-0.245**	0.024	**0.410**	0.023	**0.493**	0.030	**0.505**	0.040
$\beta_2 = 0.5, \beta_3 = 0.5, \beta_8 = 0$	Y=|X|	X (β_2)	**0.923**	0.015	**-0.612**	0.056	**0.340**	0.069	**0.293**	0.036	**-0.113**	0.023	-0.044	0.024	**0.433**	0.028	**0.408**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.552**	0.024	**-0.450**	0.038	**0.471**	0.046	**0.470**	0.042	**-0.088**	0.023	**-0.135**	0.025	**0.487**	0.026	**0.488**	0.027
		M (β_3)	**0.502**	0.024	**0.338**	0.043	**0.470**	0.047	**0.461**	0.063	**-0.162**	0.029	**0.402**	0.025	**0.485**	0.028	**0.490**	0.035
$\beta_2 = 0, \beta_3 = 0, \beta_8 = 0.5$	Y=|X|	X (β_2)	**0.375**	0.013	0.022	0.036	0.015	0.048	0.005	0.039	**-0.061**	0.019	-0.043	0.023	0.000	0.027	0.002	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.170**	0.017	0.018	0.039	0.008	0.043	0.005	0.039	**-0.052**	0.019	-0.038	0.023	0.003	0.030	0.002	0.030
		M (β_3)	**0.237**	0.013	-0.002	0.041	-0.009	0.051	-0.011	0.050	**-0.089**	0.017	**-0.084**	0.020	-0.007	0.029	-0.001	0.033
$\beta_2 = 0.5, \beta_3 = 0, \beta_8 = 0.5$	Y=|X|	X (β_2)	**0.637**	0.014	**-0.592**	0.043	**0.501**	0.052	**0.517**	0.044	**-0.118**	0.020	**-0.183**	0.024	**0.492**	0.026	**0.496**	0.028
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.321**	0.020	**-0.441**	0.042	**0.498**	0.041	**0.507**	0.042	**-0.113**	0.020	**-0.187**	0.024	**0.497**	0.029	**0.499**	0.029
		M (β_3)	**0.328**	0.016	0.075	0.045	-0.016	0.049	-0.014	0.056	**-0.121**	0.021	**0.086**	0.022	-0.015	0.029	-0.010	0.032
$\beta_2 = 0, \beta_3 = 0.5, \beta_8 = 0.5$	Y=|X|	X (β_2)	**0.788**	0.015	**-0.293**	0.051	0.087	0.069	0.065	0.036	**-0.121**	0.021	0.002	0.023	**0.189**	0.028	**0.174**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.281**	0.023	-0.002	0.040	-0.010	0.041	-0.006	0.037	**-0.115**	0.022	-0.038	0.022	-0.004	0.029	-0.007	0.030
		M (β_3)	**0.482**	0.017	**0.473**	0.038	**0.469**	0.047	**0.439**	0.080	**-0.164**	0.024	**0.410**	0.020	**0.488**	0.029	**0.497**	0.038
$\beta_2 = 0.5, \beta_3 = 0.5, \beta_8 = 0.5$	Y=|X|	X (β_2)	**1.054**	0.016	**-0.820**	0.060	**0.605**	0.076	**0.549**	0.036	**-0.087**	0.027	**-0.064**	0.026	**0.681**	0.027	**0.654**	0.027
		M (β_3)																
	Y=|X|+|M|	X (β_2)	**0.421**	0.028	**-0.333**	0.036	**0.488**	0.038	**0.490**	0.034	**0.435**	0.074	**-0.169**	0.024	**0.493**	0.028	**0.492**	0.028
		M (β_3)	**0.568**	0.021	**0.717**	0.032	**0.476**	0.040	**0.464**	0.046	**0.563**	0.060	**0.627**	0.021	**0.483**	0.027	**0.486**	0.031

Note: Bold coefficients are at least 1.96 their standard errors. Average bias ε_β of predicted β parameters: $0.1 \leq \varepsilon_\beta < 0.3$; $-0.1 \geq \varepsilon_\beta > -0.3$; $0.3 \leq \varepsilon_\beta < 0.5$; $-0.3 \geq \varepsilon_\beta > -0.5$; $\varepsilon_\beta > 0.5$; $\varepsilon_\beta < -0.5$.

Table 6 Causal mediation (CM) and causal mediation regression analysis with fixed effects (CM-FE) of simulated data

		Causal mediation (without fixed effects)								Causal mediation (with fixed effects)							
		contemporary		*lagged*		*consecutive*		*consecutive* + L_y		*contemporary*		*lagged*		*consecutive*		*consecutive* + L_y	
		$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)	$\hat{\beta}$	(se)
$\beta_2 = \beta_3 = \beta_8 = 0$	Direct (β_2)	**0.252**	0.019	**0.213**	0.019	**0.151**	0.017	**0.097**	0.015	**-0.046**	0.019	0.012	0.022	0.015	0.025	0.019	0.023
	M (β_3)	**0.221**	0.018	**0.175**	0.019	**0.237**	0.017	**0.160**	0.016	**-0.088**	0.020	0.023	0.023	0.026	0.026	0.001	0.025
	Total	**0.375**	0.016	**0.310**	0.016	**0.241**	0.017	**0.158**	0.016	**-0.063**	0.020	0.017	0.023	0.014	0.025	0.019	0.023
	ACME	**0.123**	0.011	**0.098**	0.011	**0.091**	0.008	**0.061**	0.008	**-0.017**	0.004	0.005	0.005	-0.001	0.002	0.000	0.002
$\beta_2 = 0.5, \beta_3 = 0, \beta_8 = 0$	Direct (β_2)	**0.452**	0.021	**0.369**	0.024	**0.341**	0.022	**0.248**	0.020	**-0.114**	0.021	0.031	0.027	0.035	0.030	0.043	0.027
	M (β_3)	**0.294**	0.020	**0.245**	0.023	**0.332**	0.021	**0.204**	0.021	**-0.073**	0.022	0.018	0.026	0.028	0.029	0.003	0.028
	Total	**0.637**	0.017	**0.524**	0.020	**0.490**	0.021	**0.339**	0.022	**-0.129**	0.021	0.036	0.028	0.033	0.029	0.043	0.027
	ACME	**0.185**	0.014	**0.154**	0.015	**0.149**	0.011	**0.091**	0.011	**-0.015**	0.005	0.004	0.005	-0.002	0.002	0.000	0.002
$\beta_2 = 0, \beta_3 = 0.5, \beta_8 = 0$	Direct (β_2)	**0.348**	0.021	**0.203**	0.023	**0.108**	0.019	**0.037**	0.017	**-0.069**	0.021	0.018	0.024	0.031	0.027	0.037	0.025
	M (β_3)	**0.430**	0.022	**0.520**	0.022	**0.602**	0.019	**0.483**	0.020	**-0.231**	0.024	**0.060**	0.026	**0.067**	0.028	0.004	0.028
	Total	**0.645**	0.018	**0.561**	0.021	**0.407**	0.023	**0.277**	0.022	**-0.116**	0.022	0.031	0.025	0.028	0.027	0.036	0.025
	ACME	**0.296**	0.018	**0.359**	0.018	**0.299**	0.015	**0.240**	0.015	**-0.047**	0.006	**0.013**	0.006	-0.003	0.002	0.000	0.002
$\beta_2 = 0.5, \beta_3 = 0.5, \beta_8 = 0$	Direct (β_2)	**0.551**	0.022	**0.346**	0.029	**0.259**	0.025	**0.160**	0.022	**-0.064**	0.025	0.019	0.029	0.027	0.032	0.033	0.029
	M (β_3)	**0.502**	0.023	**0.580**	0.029	**0.698**	0.023	**0.521**	0.026	**-0.121**	0.028	0.035	0.029	0.044	0.032	0.008	0.031
	Total	**0.923**	0.017	**0.776**	0.026	**0.639**	0.028	**0.443**	0.028	**-0.093**	0.026	0.028	0.030	0.025	0.032	0.033	0.029
	ACME	**0.372**	0.019	**0.430**	0.023	**0.380**	0.018	**0.283**	0.018	**-0.029**	0.007	0.009	0.007	-0.003	0.002	0.000	0.002
$\beta_2 = 0, \beta_3 = 0, \beta_8 = 0.5$	Direct (β_2)	**0.170**	0.018	**0.148**	0.020	**0.055**	0.018	**0.034**	0.017	**-0.053**	0.019	0.012	0.022	0.016	0.025	0.020	0.023
	M (β_3)	**0.237**	0.014	**0.187**	0.015	**0.237**	0.015	**0.168**	0.014	**-0.090**	0.017	0.023	0.021	0.025	0.023	0.000	0.023
	Total	**0.375**	0.016	**0.310**	0.017	**0.242**	0.017	**0.166**	0.016	**-0.063**	0.020	0.015	0.023	0.015	0.025	0.020	0.023
	ACME	**0.205**	0.014	**0.162**	0.014	**0.186**	0.012	**0.132**	0.012	**-0.009**	0.003	0.003	0.003	-0.001	0.001	0.000	0.001
$\beta_2 = 0.5, \beta_3 = 0, \beta_8 = 0.5$	Direct (β_2)	**0.321**	0.021	**0.193**	0.024	**0.185**	0.022	**0.151**	0.020	**-0.117**	0.020	0.031	0.026	0.034	0.030	0.043	0.027
	M (β_3)	**0.328**	0.015	**0.343**	0.019	**0.361**	0.017	**0.267**	0.019	**-0.124**	0.020	0.032	0.023	0.038	0.026	0.003	0.025
	Total	**0.637**	0.016	**0.524**	0.021	**0.490**	0.021	**0.377**	0.021	**-0.129**	0.021	0.034	0.027	0.033	0.030	0.043	0.027
	ACME	**0.316**	0.017	**0.331**	0.019	**0.305**	0.016	**0.226**	0.018	**-0.012**	0.003	0.003	0.003	-0.001	0.002	0.000	0.002
$\beta_2 = 0, \beta_3 = 0.5, \beta_8 = 0.5$	Direct (β_2)	**0.281**	0.021	**0.131**	0.026	-0.007	0.020	-0.034	0.019	**-0.098**	0.022	0.026	0.025	0.030	0.028	0.038	0.026
	M (β_3)	**0.482**	0.016	**0.513**	0.020	**0.591**	0.017	**0.490**	0.020	**-0.132**	0.022	0.038	0.024	0.044	0.026	0.006	0.026
	Total	**0.788**	0.017	**0.670**	0.023	**0.531**	0.024	**0.412**	0.026	**-0.115**	0.023	0.031	0.026	0.029	0.028	0.038	0.026
	ACME	**0.507**	0.020	**0.539**	0.024	**0.538**	0.020	**0.446**	0.022	**-0.017**	0.004	0.005	0.004	-0.001	0.002	0.000	0.002
$\beta_2 = 0.5, \beta_3 = 0.5, \beta_8 = 0.5$	Direct (β_2)	**0.421**	0.024	**0.156**	0.034	**0.081**	0.024	**0.057**	0.022	-0.049	0.027	0.014	0.034	0.017	0.037	0.022	0.034
	M (β_3)	**0.568**	0.018	**0.643**	0.027	**0.699**	0.020	**0.609**	0.026	0.005	0.029	0.010	0.030	0.017	0.033	0.011	0.032
	Total	**1.054**	0.017	**0.874**	0.029	**0.733**	0.029	**0.626**	0.033	-0.047	0.030	0.017	0.035	0.016	0.037	0.022	0.034
	ACME	**0.633**	0.023	**0.718**	0.031	**0.653**	0.024	**0.569**	0.029	0.002	0.007	0.003	0.007	-0.001	0.002	**0.018**	0.002

Note: Bold coefficients are at least 1.96 their standard errors. Average bias ε_β of predicted β parameters: $0.1 \leq \varepsilon_\beta < 0.3$; $0.3 \leq \varepsilon_\beta < 0.5$; $\varepsilon_\beta > 0.5$; $-0.1 \geq \varepsilon_\beta > -0.3$; $-0.3 \geq \varepsilon_\beta > -0.5$; $\varepsilon_\beta < -0.5$.

In contrast to the OLS approach, the estimates of the *FE approach* (see right panel of Table 4) tend to be downwardly biased (though its bias is generally smaller compared to OLS). While some of the estimates are significantly negative although the respective β coefficients have been set to zero, several predicted values of both β_2 and β_3 get remarkably close to the generated ones in the *consecutive* scenario *(C)* without modeling an effect of *L.Y* (which had been set in the DGP, though) in particular. Moreover, on the one hand, the FE approach does not stand at risk to erroneously predict a mediation effect that has not been introduced in the DGP (Appendix A, Figure A2a). On the other hand, however, once a mediation effect is considered in the DGP, it is correctly identified by *scenario C)* only (Appendix A, Figure A2b).

The average bias of the *GMM approach* is even smaller compared to the FE approach (see left panel of Table 5). Note that the *contemporary* scenario *(A)* of the GMM approach is a replication of the corresponding OLS approach modeled as a special case of GMM – which is why the respective point estimates are almost identical to the *contemporary* scenario from the OLS approach (left panel in Table 4); with smaller standard errors, though. While there is some amount of downward bias in the effect of *X* in the *lagged* scenario *(B)*, the *consecutive* scenario *(C)* in particular performs very well to detect the coefficients modeled in the DGP (although their corresponding confidence intervals still overlap in case of $\beta_2 = \beta_3 = \beta_8 = .5$; see Appendix A, Figure 3b). Interestingly, the *consecutive* scenario which controls for the lag of *Y (D)* is also slightly biased downwardly once β_3 has been set to .5.

The FE-CLPM approach (right panel of Table 5) yields results that are, on average, similarly accurate as the ones produced by the GMM approach – but with a few differences that deserve to be carved out: First, while most parameter estimates from the *contemporary* scenario *(A)* of the GMM approach (which is equivalent to the one by the OLS approach) are upwardly biased, most parameter estimates from the *contemporary* scenario of the FE-CLPM approach are downwardly biased. Second, the downward bias in the *lagged* scenario *(B)* of the FE-CLPM approach is comparable to the one of the GMM approach. Third, in the *consecutive* scenarios *(C)* and *(D)*, the FE-CLMPs correctly identify both the effects of *X* and *M* on *Y* as well as the mediation effects once they have been modeled in the DGP (also see Appendix A, Figure A4). Fourth, as an advantage to the GMM approach, the predicted coefficients from the FE-CLMP approach are less sensitive towards the specification of the first lag of *Y* in the estimation process.

In the CM models without fixed effects (left panel of Table 6), the parameter estimates can be biased in both directions: On the one hand, in case of $\beta_2 = \beta_3 = \beta_5 = 0$, significant positive effects are predicted for all parameters (including the ACME) although they have been absent in the DGP. On the other hand, in case of $\beta_2 = .5$ and $\beta_8 = .5$, the direct effect of *X* on *Y* is notably underestimated within

all scenarios, while the effect of M is still overestimated.[7] The coefficient plots of the CM models without fixed effects are displayed in Figure A5 of Appendix A.

Finally, in the CMFE models (right panel of Table 6), most predicted parameters suffer from a considerable downward bias. For instance, in case of $\beta_2 = \beta_3 = \beta_8 = 0$, all parameter estimates of the *contemporary* scenario *(A)* are negative. While the other scenarios correctly identify the above effects to be absent, for other values of β_2, β_3 and β_8, they likewise fail to identify effects that should be present according to the DGP (i.e., the corresponding coefficients are not significant). This bias of the CMFE approach is most pronounced in case of $\beta_2 = \beta_3 = 8 = .5$. The coefficient plots of the CMFE models are displayed in Figure A6 of Appendix A. A concise summary and corresponding interpretation of all findings will be given in the conclusion section.

Conclusion

The aim of this paper was to provide both a theoretical foundation and an empirical examination of different statistical approaches to mediation analysis. Regarding theory, a brief sketch of the fundamentals of mechanism-based explanations set the argument of adhering to a consecutive order of predictor, mediator and outcome in mediation analysis. Having summarized the statistical fundamentals of different approaches to mediation analysis, I provided a simulation analysis of the data-generating process (DGP) which could be actively manipulated to examine differences in relative performance under different scenarios: A) all-simultaneous, B) first lag of all coefficients; C) consecutive order; D) consecutive order plus first lag of Y as a predictor. Each scenario was analyzed by the following methods: OLS regressions, fixed effects (FE) regressions, generalized method of moments (GMM) regressions, causal mediation analysis without (CM) and with fixed effects (CMFE), and fixed-effects cross-lagged panel models (FE-CLPMs).

The results of the simulation study suggest that the estimates of the OLS approach are generally upwardly biased, the ones of the FE and CMFE regressions are by trend downwardly biased, and the ones of the CM models (without FEs) can be biased in both directions. In contrast, the coefficients and confidence intervals estimated by both GMM regressions and FE-CLPMs are most accurate, in particular if the structure of lags in the empirical models met the consecutive order which had been set up in the underlying DGP. Most interestingly, while the GMM approach tended to be sensitive against whether or not the first lag of Y ($L.Y$) was modeled as an additional predictor (the autocorrelation of Y was set

7 For ease of interpretation, recall that the total effect of X on Y (TE_{XY}) is computed as follows: $TE_{XY} = \beta_2 + (\beta_3 \cdot \beta_8)$.

to .5 in all models), the FE-CLPMs appeared to be insensitive in this respect. As a first practical implication, FE-CLPMs could be more applicable in cases of mediation analysis where the researcher is not sure whether or not *L.Y* should be included as a predictor. A second practical implication is that *even* GMM regressions and FE-CLPMs can only detect the true parameter values when the order of the DGP is met. Consequently, it is of utmost importance that researchers most carefully translate their theoretical assumptions into an empirical model with the appropriate causal order: if a researcher is theoretically convinced that the causal order of the hypothesized effect is $X_{(t\text{-}2)} \rightarrow M_{(t\text{-}1)} \rightarrow Y_t$, then naïvely predicting Y_t by X_t and M_t or even by $X_{(t\text{-}1)}$ and $M_{(t\text{-}1)}$ in any applied data might yield biased results irrespective of the statistical method used.

Concerning directions for future research, one direct advancement would be to shed more light on how different values for the autocorrelation of *Y* affect the extent to which the results of the GMM approach depend on the inclusion of *L.Y* as an additional predictor of *Y*. A second, more challenging direction could be to consider more complex data structures (such as time nested in individuals nested in additional contexts) or modeling purposes (such as moderated mediation). As a third, related, direction, future simulation studies could manipulate different forms of *observed* heterogeneity (between *X* and *Z*, *M* and *Z*, and/or *X* and *M*) to explore the performance of each approach to *mediation* under different scenarios of *moderation*.

All in all, analyzing various DGP scenarios by different statistical approaches to mediation analysis will yield important implications for applied researchers who aim to translate particular mechanism-based explanations in statistical mediation models.

References

Allison, P. D. (2009). *Fixed effects regression models. Quantitative applications in the social sciences: Vol. 160*. Los Angeles: Sage Publications.

Allison, P. D., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, *3*, 2378023117710578.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277–297.

Babad, E., & Katz, Y. (1991). Wishful Thinking—Against All Odds. *Journal of Applied Social Psychology*, *21*(23), 1921–1938. https://doi.org/10.1111/j.1559-1816.1991.tb00514.x

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

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. New York: Wiley.

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14*(1), 63–82.

Brüderl, J., & Ludwig, V. (2015). Fixed-effects panel regression. *The Sage Handbook of Regression Analysis and Causal Inference*, *327*, 357.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 107–135). Washington: American Psychological Association. https://doi.org/10.1037/10409-004

Elster, J. (1989). *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Esser, H. (1996). What is wrong with 'variable sociology'? *European Sociological Review*, *12*(2), 159–166.

Finkel, S. (1995). *Causal Analysis with Panel Data*. Thousand Oaks, California. Retrieved from https://methods.sagepub.com/book/causal-analysis-with-panel-data https://doi.org/10.4135/9781412983594

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102.

Hayes, A. F., Preacher, K. J., & Myers, T. A. (2011). Mediation and the estimation of indirect effects in political communication research. In E. P. Bucy & R. L. Holbert (Eds.), *Sourcebook for political communication research: Methods, measures, and analytical techniques* (pp. 434–465). New York: Routledge.

Hedström, P. (2005). *Dissecting The Social. On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.

Hedström, P., & Swedberg, R. (1996). Social mechanisms. *Acta Sociologica*, *39*(3), 281–308.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*, 49–67.

Imai, K., Jo, B., & Stuart, E. A. (2011). Commentary: Using Potential Outcomes to Understand Causal Mediation Analysis. *Multivariate Behavioral Research*, *46*(5), 861–873. https://doi.org/10.1080/00273171.2011.606743

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

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(04), 765–789.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71.

Leszczensky, L., & Wolbring, T. (2019). How to deal with reverse causality using panel data? Recommendations for researchers based on a simulation study. *Sociological Methods & Research*, 0049124119882473.

Lutz, S. M., Sordillo, J. E., Hokanson, J. E., Chen Wu, A., & Lange, C. (2020). The effects of misspecification of the mediator and outcome in mediation analysis. *Genetic Epidemiology*, *44*(4), 400–403. https://doi.org/10.1002/gepi.22289

Mehmetoglu, M. (2018). medsem: a Stata package for statistical mediation analysis. *International Journal of Computational Economics and Econometrics*, *8*(1), 63–78. Retrieved from https://EconPapers.repec.org/RePEc:ids:ijcome:v:8:y:2018:i:1:p:63-78

Moral-Benito, E. (2013). Likelihood-based estimation of dynamic panels with predetermined regressors. *Journal of Business & Economic Statistics*, *31*(4), 451–472.

Moral-Benito, E., Allison, P., & Williams, R. (2019). Dynamic panel data modelling using maximum likelihood: an alternative to Arellano-Bond. *Applied Economics*, *51*(20), 2221–2232. https://doi.org/10.1080/00036846.2018.1540854

Muthén, B. O. (2017). *Regression and mediation analysis using Mplus.*

Newey, W. K., & Windmeijer, F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, *77*(3), 687–719.

Reed, W. R. (2015). On the practice of lagging variables to avoid simultaneity. *Oxford Bulletin of Economics and Statistics*, *77*(6), 897–905.

Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, *71*(1), 135–158.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961–962.

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. *16062360*.

Tranow, U., Beckers, T., & Becker, D. (2016). Explaining and Understanding by Answering 'Why' and 'How' Questions: A Programmatic Introduction to the Special Issue Social Mechanisms. *Analyse & Kritik*, *38*(1), 1–30.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics*, *126*(1), 25–51. https://doi.org/10.1016/j.jeconom.2004.02.005

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, *42*(1), 314–347.

Date of online first publication: 2023-05-22