mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology

Volume 18, 2024 | 1

Recent Developments and Current Approaches to the Analysis of Panel Data

Henrik Kenneth Andersen, Jochen Mayerl & Elmar Schlüter (Editors)

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editor: Sabine Häder
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Tel.: + 49.621.1246526
E-mail: mda@gesis.org
Internet: www.mda.gesis.org

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession‘s knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.
All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (January and July).

Layout: Bettina Zacharias (GESIS)
Print: Bonifatius Druck GmbH Paderborn, Germany

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

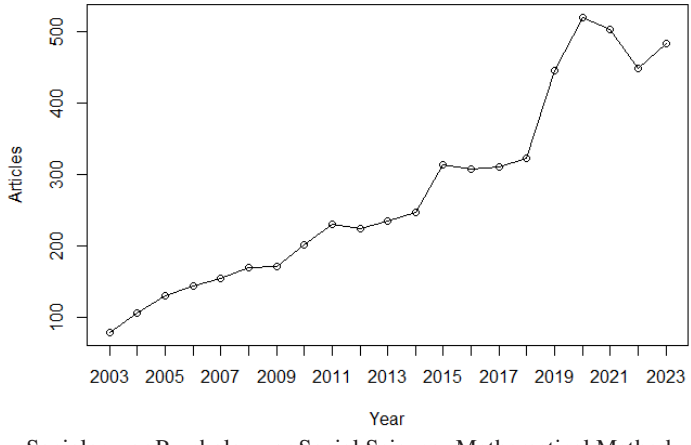
methods, data, analyses | Vol. 18(1), 2024, pp. 3-6

Editorial
Recent Developments and Current Approaches to the Analysis of Panel Data

Henrik Kenneth Andersen[1], *Jochen Mayerl*[1] & *Elmar Schlüter*[2]

[1] *Chemnitz University of Technology*
[2] *University of Giessen*

Panel data refer to repeated observations of the same units over time. Due to the growing interest in causal inference in the social sciences, and the increasing feasibility of collecting (intensive) longitudinal data, interest in panel data has grown steadily in the social sciences (Rohrer & Murayama, 2023). Figure 1 shows the number of articles containing the term "panel data" published just in the fields of Sociology, Psychology, and Social Sciences Mathematical Methods over the last 20 years (according to Web of Science, as of January 2024).



Categories: Sociology or Psychology or Social Sciences Mathematical Methods

Figure 1 Articles featuring keywords "panel data" (all fields), Web of Science years 2003-2023

Panel data offer a wide variety of advantages over cross-sectional data or even other types of longitudinal data. For one, they are valuable for the purposes of causal inference, that is, drawing causal conclusions from observational (rather than experimental) data. Indeed, as Hamaker (2012) notes, most social science theories are implicitly formulated at the *within-person* level. And the potential outcomes framework always begins with formulating a unit-specific causal effect: a contrast between realized and counterfactual states at the individual level (Rohrer & Murayama, 2023). For example, when we think of the relationship between typing speed and typing errors, most of us would probably expect the effect to be positive: the faster one types, the more mistakes she or he makes (Hamaker, 2012). This is exactly because we are thinking at the *within-person* level rather than the between-person level: if an individual increases her or his typing speed (holding all else constant), she or he is likely to make more errors. Panel data allows us to get closer to this ideal. By comparing the same individuals over time, we can be sure that we're holding constant all the things that don't change for a given individual, such as place and time of birth, upbringing, and potentially even psychological traits.

With panel data, researchers can respect the fact that processes and effects "unfold over time" (Hamaker & Wichers, 2017). Thus, social change over time can be analyzed at the individual rather than aggregate level, avoiding ecological fallacies. As technology evolves to make (intensive) longitudinal data collection more feasible, and as causal inference becomes the focus of many social science studies (e.g. fixed effects panel regressions), panel data are becoming increasingly important (Rohrer & Murayama, 2023).

The field of panel data research is still growing, addressing the need for research on innovative panel data collection methods as well as panel data analysis techniques. On the methodological side, the quality of panel data collection is challenged by issues such as panel conditioning (e.g., learning effects), the question of optimal lags for identifying causal effects, and high attrition rates that require missing value treatment techniques or weighting procedures. To further improve panel data analysis, research is needed on issues such as dealing with violations of the parallel assumption and heterogeneous growth, comparing different statistical approaches to panel data analysis, mediation analysis based on panel data, estimation of treatment effect dynamics and dealing with negative weighting bias, the challenges of dynamic panel models and the inclusion of bidirectional effects and lagged dependent variables, and continuous versus discrete time modeling, to name just a few current research issues.

This special issue contains applications to methodological issues and statistical problems in panel data analysis in a variety of content-related areas:

The contribution from *Dominik Becker*, entitled "Many Roads to Mediation: A Methodological and Empirical Comparison of Different Approaches to Statistical Mediation", examines the use of panel data to investigate social mechanisms in the form of mediation analyses. While mediation analysis is often done using cross-sectional data, the use of panel data has several interesting advantages. For one, mediation analysis with panel data allows for drawing causal inference under less strict assumptions. If confounders of the effects of interest are stable within individuals over time, then the broad category of panel fixed effects panel models can eliminate unobserved time-invariant heterogeneity. Second, panel data allow researchers to empirically establish the theoretical causal order of cause, mediator, and outcome. In particular, the specification of lagged effects between variables helps to rule out reverse causality. The article constructs a simulation study and compares a variety of modeling techniques with respect to their ability to recover the true parameter values, and provides researchers with valuable recommendations for approaching questions of causal mechanisms with panel data.

Judith Lehmann contributes an article entitled "Analyzing the Causal Effect of Obesity on Socioeconomic Status – the Case for Using Difference-in-Differences Estimates in Addition to Fixed Effects Models" in which she compares Difference-in-Differences (DiD) with Fixed Effects (FE) models to investigate the empirically well-established obesity penalty with respect to labor market outcomes. Like other articles in this issue, this one also combines strong substantive and methodological components. Substantively, the author finds no effects of obesity on socioeconomic status in either the FE or the DiD model. However, the DiD estimator explicitly models the development of the control group, providing a deeper understanding of the relationships. Namely, the non-obese individuals in the analysis showed stronger socioeconomic development over time compared to the group of obese individuals.

Manuel Holz and Jochen Mayerl compare health outcomes of migrants and native Germans over time in a contribution entitled "Migrant health inequalities or unequal measurements? Testing for cross-cultural and longitudinal measurement invariance of subjective physical and mental health". The so-called healthy migrant effect describes both the self-selection of comparatively healthy individuals to migrate from their home countries and the greater decline in health among migrants compared to the native population. The paper draws attention to an aspect of cross-cultural comparisons of health outcomes that has been overlooked in the previous research: to make valid comparisons of (especially) subjective measures of health, one must establish that components of the measurement instrument have the same meaning and importance across cultures and time. Thus, this article compares the trajectories of subjective health (SF-12 for physical and mental health)

of migrants and native-born Germans, testing for measurement invariance across groups and over time.

Christina Beckord tackles an interesting methodological topic in her contribution entitled "Challenges in Assigning Panel Data with Cryptographic Self-generated Codes – Between Anonymity, Data Protection and Loss of Empirical Information". The article examines the difficulties of linking data across 13 survey waves of the "Crime in the Moden City" (CrimoC) study and details a unique strategy for dealing with ambiguous user-generated codes. The author describes a meticulous, error-tolerant matching process, involving manual handwriting comparison, to merge individual data over time. The matching process resulted in 3,589 filled missing units.

The final contribution by *Jost Reinecke, Anke Erdmann, & Manuel Voelkle* entitled "Continuous Time Modeling with Criminological Panel Data: An Application to the Longitudinal Association between Victimization and Offending" re-examines the well-known victim-offender overlap – that offenders tend to have been victimized themselves – with novel panel data from the Crime in the Modern City (CrimoC) study. Methodologically, this paper adds to the new but growing literature on so-called continuous time panel models. Unlike the more commonly applied discrete time models (e.g., cross-lagged panel models, latent growth curves), continuous time model recognize that panel data provide multiple discrete snapshots of constructs over time. Yet effects between constructs over time are highly sensitive to the time interval between these snapshots, which is often chosen arbitrarily (e.g., one panel wave per year) or set based on time and budget constraints. The article discusses the results of the continuous time models, explains how researchers can transform continuous parameters into discrete parameters and visualizes the dynamic effects of constructs on each other (and themselves) as time unfolds.

References

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), Handbook of research methods for studying daily life (pp. 43–61). The Guilford Press.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. Current Directions in Psychological Science, 26(1), 10-15.

Rohrer, J. M., & Murayama, K. (2023). These are not the effects you are looking for: causality and the within-/between-persons distinction in longitudinal data analysis. Advances in methods and practices in psychological science, 6(1), 25152459221140842.

methods, data, analyses | Vol. 18(1), 2024, pp. 7-32 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-

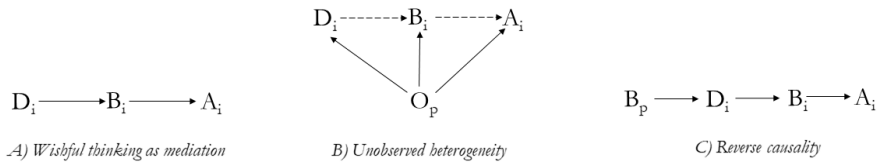


Figure 1 A social mechanism approach to mediation, unobserved heterogeneity and reverse causality.

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.

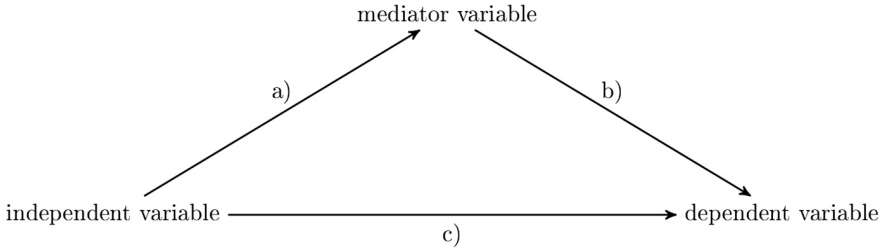


Figure 2 A simple mediation model.

> "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:

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

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:

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

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

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

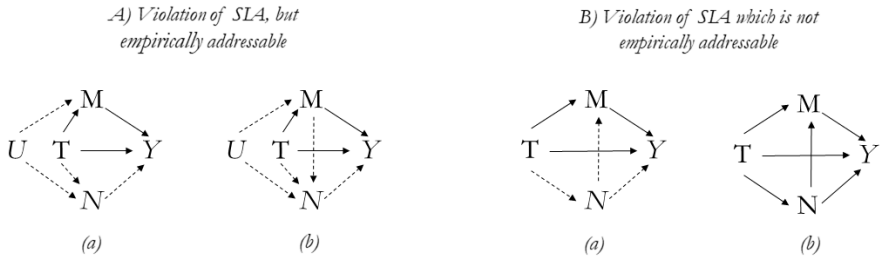


Figure 3 Methodological challenges of the causal mediation model. Summary of Imai, Keele, et al. (2011, 786f.)

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).

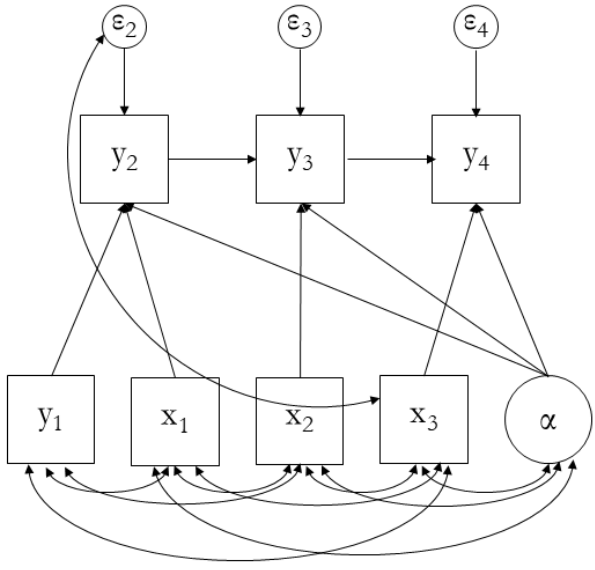


Figure 4 The FE-CLPM. Source: Allison et al. (2017, 6).

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: $0.1 \le \varepsilon_\beta < 0.3$; $0.3 \le \varepsilon_\beta < 0.5$; $\varepsilon_\beta > 0.5$; $-0.1 \ge \varepsilon_\beta > -0.3$; $-0.3 \ge \varepsilon_\beta > -0.5$; $\varepsilon_\beta < -0.5$.

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.

methods, data, analyses | Vol. 18(1), 2024, pp. 33-58 DOI: 10.12758/mda.2022.14

Analyzing the Causal Effect of Obesity on Socioeconomic Status – the Case for Using Difference-in-Differences Estimates in Addition to Fixed Effects Models

Judith Lehmann
Otto-Friedrich-Universität Bamberg

Abstract

Recent studies use Fixed Effects (FE) models to estimate the causal effect of obesity on socioeconomic status, the so-called obesity penalty. In this paper, I will illustrate the advantages of using a Difference in Differences (DID) approach as an alternative method of causal analysis. Combining the German National Health Interview and Examination Survey 1998 (GNHIES98) and the German Health Interview and Examination Survey for Adults 2008 (DEGS1) allowed for a panel analysis of 3934 respondents. The dependent variable is a socioeconomic status score that integrates level of education, occupation and household income. The binary treatment variable is abdominal obesity. To estimate the causal effect of the treatment, FE and DID approaches were used.

Both the FE model and the DID estimate show no statistically significant causal effect of abdominal obesity on socioeconomic status for adults in Germany. However, both the respondents who became obese and those who stayed non-obese experience a rise in socioeconomic status over time. Nonetheless, the non-obese group had a more substantial increase in socioeconomic status than the obese group. Therefore, the obesity penalty does not necessarily have to be a decrease in socioeconomic status but could instead be a slowed growth or stagnation in status. The advantage of the DID approach is that the development in the control group is explicit. If obese individuals are more likely to have less favorable positive trends in socioeconomic status over time than other individuals, using DID estimates demonstrates the obesity penalty more effectively than using only FE models.

Keywords: Difference-in-Differences, Propensity Score Matching, Fixed-Effects, Obesity penalty, Germany

Fixed Effects (FE) models have become a popular and widely used method of panel analysis. Researchers apply FE models to identify causal effects through the within-comparison of cases (Brüderl, 2010). However, there are alternative methods for identifying causal effects with observational data that can add important insights into the topic under study (Gangl, 2010). Hence, in this paper I will illustrate the advantages of supplementing fixed effects analyses with Difference-in-Differences (DID) approaches.

To highlight the differences and advantages of both FE models and DID estimators, I chose the example of the obesity penalty (Averett & Korenman, 1996). The obesity penalty describes the finding that obese people earn lower wages and report more adverse labor market outcomes than non-obese people (Caliendo & Gehrsitz, 2016). To support these findings theoretically, different mechanism such as lower human capital, lower productivity and higher probability of health issues of the obese as well as discrimination and negative stereotyping are discussed (Bozoyan & Wolbring, 2018).

The obesity penalty is an interesting example to consider. On the one hand, FE models are frequently applied in research on the effect of body weight on socioeconomic status. Many previous studies focus on the question of what happens to an individual's socioeconomic status when they become obese. On the other hand, research on labor market outcomes has shown that it is important to observe an adequate control group (i.e. Angrist & Pischke, 2008). Knowledge concerning the development of socioeconomic status in the control group can change the interpretation of the development of status in the treatment group. Hence, DID estimators may contribute important new information on the obesity penalty. As a result, the main question of this study is: What are the advantages of using DID estimators in addition to using FE models in regards to the obesity penalty?

In this study, I will use FE models and DID estimators to identify the causal effect of abdominal obesity on socioeconomic status in a sample of adults in Germany. Propensity Score Matching will be applied to create an adequate control group for the DID estimator since treatment is not assigned randomly in observational data. I will use these two methods to the full sample and to female and male respondents separately. In the discussion, I will highlight the advantages of including DID approaches in this line of research and the additional information gained by introducing an adequate control group. I will also discuss the different perspectives offered by FE models and DID estimators to further evaluate their benefits.

Direct correspondence to
Judith Lehmann, M.A., Otto-Friedrich-Universität Bamberg, Lehrstuhl für Soziologie, insbesondere Soziale Ungleichheit, Feldkirchenstr. 21, 96052 Bamberg
E-mail: judith.lehmann@uni-bamberg.de

Previous Research

Studies on the causal relationship of obesity and socioeconomic status focus on both the social causation hypothesis and the health selection hypothesis. The social causation hypothesis states that socioeconomic status influences body weight and the probability of becoming obese. For example, Ball and Crawford (2005) show in their review that lower job position increases the probability of becoming obese compared to higher job position. Gebremariam et al. (2017) conclude that socioeconomic position of the parents influences body weight of their children through mediators such as food consumption and TV usage. In a meta-analysis, Kim et al. (2017) show that both social causation and health selection exist in regards to education. However, the evidence for the health selection hypothesis is more consistent.

The health selection hypothesis states that obesity leads to lower socioeconomic status. Studies that focus on the health selection hypothesis overwhelmingly use FE models to identify the causal effect. For example, many studies used the National Longitudinal Survey of Youth, which is a panel study in the United States of America (US). Baum and Ford (2004) find negative effects of obesity on wages for men and women using this data. In this study, women experience stronger and more consistent negative effects than men. Cawley (2004) identifies a negative effect of Body Mass Index (BMI) on wages for white women. The effect for men is non-linear, with overweight men earning more than normal-weight or obese men. Han et al. (2009) report similar findings with overweight and obese white women and obese Black women earning less than their normal-weight peers. They report no causal effect for men. However, the authors can show that respondents in jobs with social interactions are especially affected. Harris (2019) builds on that and finds that high body weight leads to lower wages in jobs that are socially and mentally intensive and to higher wages in physically challenging jobs. He concludes that gender differences in the effect of body weight on wages can be explained through differing occupational positions.

Other recent studies use different data to analyze the obesity penalty. Bozoyan and Wolbring (2011) use fat free mass and body fat instead of BMI to model body weight. Using FE models, they cannot identify a significant effect of body weight on wages. Ahn et al. (2019) conclude that obese women and underweight men are disadvantaged on the labor market even if employment efforts are controlled for. When obese women find a job, Lee et al. (2019) show that their wages and other characteristics of the job (i.e. getting a bonus or having a job in a company with a labor union) are inferior to those of other women.

Another popular method of causal inference are instrument variables (IV) because exogenous instruments allow the identification of the treatment effect even in the presence of unobserved heterogeneity (Gangl, 2010). The IV method is applied in the context of the obesity penalty as well. For example, Cawley et al.

(2005) use this method and identify a negative effect of obesity on wages only for women in the US. Morris (2007) finds a negative effect of obesity on employment for men and women. Sari and Acan Osman (2018) show that obesity negatively influences labor market participation for women. Böckerman et al. (2019) identify negative effects of body weight on multiple dimensions of socioeconomic status, such as earnings, employment and social income transfers.

Some studies have combined FE models and IV methods to strengthen their findings. While IV methods control for unobserved time-varying heterogeneity in theory, it is challenging to find good instruments in practice. Therefore, results gained through IV methods are often viewed with caution and FE models are added as an alternative method of causal inference. Sabia and Rees (2012) find effects of body weight on wages only for white women using FE models. Their IV analyses confirm this finding; therefore, they conclude that this result is not influenced by unobserved time-varying heterogeneity. Katsaiti and Shamsuddin (2016) analyze a number of aspects of the socioeconomic position and find negative effects of higher body weight on wages, employment, promotions and a positive effect on duration of unemployment for women. There are no causal effects for men. Both FE models and IV method produce these findings. Wada and Tekin (2010) analyze the effects of body fat and fat free mass on wages. Their FE models find positive effects of fat free mass and negative effects of body fat on wages for white men and women. Using the IV method, only the effects for men can be confirmed. The authors do not interpret this further because of small sample sizes and restrictions of the IV. These studies mostly report similar findings for both methods, however, none compare the methods directly.

So far, DID approaches have not been applied to this field of research. While FE models focus on changes within the cases of the treatment group and use a control group implicitly when confounders are controlled for, DID estimators explicitly use the changes in the control group in addition to the ones in the treatment group to estimate the causal effect. For identifying causal effects, Angrist and Pischke (2008) have shown the importance of a comparable control group. They use examples from educational and labor market research to illustrate the advantages of DID estimators. Using an adequate control group can help identify the causal effect of abdominal obesity on socioeconomic status and provide important new insights. Therefore, I will address this research gap by using DID estimators in addition to FE models to show which additional information can be gained concerning the obesity penalty.

Research Question and Hypotheses

The aim of this study is to apply FE models and DID estimators to the same research question to illustrate the value of using both methods. To do this, I focus on the research question: Is there a causal effect of obesity on socioeconomic status? Previous research has reported mixed results on this question, especially for Germany. By using two different approaches to estimate the causal effect, I will strengthen the results and show the advantages of each method.

From previous research, I derived two hypotheses concerning the causal effect of obesity on socioeconomic status. First, I expect that obesity decreases socioeconomic status. This hypothesis is usually referred to as the obesity penalty. It is assumed that because of different mechanisms such as discrimination, differences in human capital and productivity or health problems obesity leads to a lower socioeconomic status (Bozoyan & Wolbring, 2018).

Second, I expect that the negative effect of obesity on socioeconomic status is stronger for women than for men. It is often assumed that women are judged more harshly for their appearance and body weight than men (Caliendo & Gehrsitz, 2016). Women have to comply with the norm for thinness more than men do (Magallares, 2016). Some research even indicates that overweight men are more privileged than other men (Cawley, 2004).

While the example of this study is the obesity penalty, the focus lies on the exploration of the benefits of combining FE models and DID estimators. Therefore, the main research question is: What are the advantages of using DID estimators in combination with Propensity Score Matching in addition to using FE models?

I argue that the DID approach can offer more information on the causal relationship between obesity and socioeconomic status. As will be shown in this study, the DID approach uses an explicit control group. Hence, it provides researchers with the chance to compare the development of the outcome variable in the treatment and control group. Furthermore, it requires a theoretical discussion of the comparability of treatment and control group.

Data & Methods

The next section will give an overview of the data and methods used. Since I will focus on discussing the use of FE models and DID estimators as methods of causal inference, I will present their advantages and disadvantages as well as their general logic and assumptions in more detail than usual.

Data

Both DID estimators and FE models usually require longitudinal data to estimate the causal effect of obesity on socioeconomic status. However, representative longitudinal data of the German adult population with a focus on health and the socioeconomic position of households or individuals is still sparse. Therefore, the German National Health Interview and Examination Survey 1998 (GNHIES98) and the German Health Interview and Examination Survey for Adults 2008 (DEGS1) conducted by the Robert Koch-Institute were combined to allow for panel analyses.

The German Health Interview and Examination Survey for Adults (DEGS) is the first representative longitudinal survey focusing on health and the socioeconomic position of adult respondents in Germany (Gößwald et al., 2012). The first wave of data was collected between 2008 and 2011. Respondents for DEGS1 were selected based on a previous study conducted by the Robert Koch-Institute: the German National Health Interview and Examination Survey 1998 (GNHIES98). Between 1997 and 1999, 7124 respondents were interviewed and examined for GNHIES98 (Thefeld et al., 1999). Those respondents who were still alive in 2008 were invited to also participate in DEGS1. Therefore, a longitudinal sample of 3959 respondents exists, covering two waves and a period of around ten years between the waves (Gößwald et al., 2012). Restricting the sample to cases with valid values for both the dependent and the central explanatory variable leads to 2835 cases that can be included in the analyses.

Both GNHIES98 and DEGS1 include medical interviews and medication history, health questionnaires and nutrition interviews, and laboratory and physical examinations. The data set contains anthropometric data such as height, body weight and waist circumference measured by health professionals as well as information on the socioeconomic situation of individuals and households (Scheidt-Nave et al., 2012, Gößwald et al., 2012). Hence, I chose this data set for the following analyses.

Variables

The dependent variable for the analyses is a socioeconomic status score that is provided by the Robert Koch-Institute and integrates information on the level of education, occupation and household income of the respondents. The socioeconomic status score is not a variable on the individual level, because it combines individual and household information. It creates a scale of socioeconomic status that integrates three different dimensions of social status (Lampert et al., 2013). For the subscale of education, schooling and vocational training of the respondents are combined and ranked from 1 (lowest education) to 7 (highest education). For the subscale of occupation, the jobs of the respondents and the main earners of their households

are compared and the higher occupational position of the two is ranked from 1 (lowest occupational position) to 7 (highest occupational position) according to the average wages earned in that profession. For the subscale of household income, weighted net household income was ranked from 1 (lowest income) to 7 (highest income). The three subscales were summed up to form a socioeconomic status score ranging from 3 to 21 (Winkler & Stolzenberg, 1999; Lampert et al., 2013). The socioeconomic status score is considered a quasi-metric variable (Lampert et al., 2013). Since the socioeconomic status score is a relative measure of the social position, changes in the score over time can occur even if educational level, occupation and household income of the respondents did not change between waves. The socioeconomic status score is normally distributed. For the analyzed sample of this study, the mean of the socioeconomic status score was 11.5 in GNHIES98 and 11.6 in DEGS1.

The central explanatory variable is abdominal obesity defined by waist circumference. Health professionals measured waist circumference during both waves of data collection. The measurement was standardized as much as possible. For GNHIES98, waist circumference was measured at the midway point between the lowest rib and the pelvic crest while respondents wore a light layer of clothing (Bergmann, 1999). The same method was used for DEGS1; however, respondents were measured wearing only their underwear (Haftenberger et al., 2016). This change in measurement between the two waves might lead to small differences in waist circumference, even if the body weight of the respondents did not vary. The mean waist circumference of the analyzed sample is 89.8cm in GNHIES98 and 93.8cm in DEGS1, so in general, the respondents gained weight between the two waves.

Using waist circumference, I created a binary variable for abdominal obesity as the treatment variable. The binary variable allows for an easy separation of the sample into treatment and control group. The World Health Organization provides the following cut-offs to define abdominal obesity by waist circumference: 88cm for women and 102cm for men (WHO, 2011). These cut-offs were employed to generate the binary variable for abdominal obesity. According to this new variable, 31% of the sample were obese in GNHIES98 and 44% in DEGS1. Since the focus of this paper is on the causal effect of obesity, respondents who were not obese in GNHIES98, but were obese in DEGS1 constitute the treatment group. Approximately 24% of the non-obese respondents in the first wave became obese by the second wave. While respondents who were not obese in both waves constitute the control group, respondents who were already obese in the first wave were excluded from the analyses (888 cases).

The Counterfactual Framework

This study analyzes the causal effect of obesity on socioeconomic status using observational data. For this purpose, the counterfactual framework allows the integration of causal analyses and observational data (Gangl, 2010). This is necessary since most research questions in the social sciences cannot be analyzed using randomized experiments due to ethical and practical restrictions (Leszczensky & Wolbring, 2019). Randomized experiments are usually considered the golden standard of causal inference because respondents are randomly selected into the treatment or the control group and thus selection bias is eliminated (Gangl, 2010).

In contrast, in observational studies treatments are assigned in a socially structured way and therefore treatment assignment and the expected outcome might be correlated (Gangl, 2010). To estimate the causal effect, it is necessary to disrupt this correlation by conditioning on covariates. The aim is to achieve conditional independence, which states, "conditional on covariates, variation in [treatment variable] D is as good as randomly assigned" (Gangl, 2010, p. 27). If the conditional independence assumption (CIA) holds, conditioning on the covariates will lead to unbiased causal effects.

The identification of causal effects is complex because, in theory, the individual causal effect is calculated by subtracting the outcome of a person i who receives the treatment (Y_i^1) from the outcome of the same person i if they do not receive the treatment (Y_i^0) (Rosenbaum & Rubin, 1983). The only difference between the two states of person i is the treatment status so that any changes in the outcome can be attributed to the treatment. In practice, it is not possible to observe the outcome of person i in both treatment states at the same time – so it is not possible to calculate the individual causal effect (Holland, 1986; Dehejia & Wahba, 1999).

Therefore, Holland introduced the following statistical solution: replace "the impossible-to-observe causal effect of "X" on a specific unit with the possible-to-estimate average causal effect of "X" over a population of units" (Holland, 1986, p. 947). This is unproblematic since the research interests in the social sciences usually focus more on average causal effects in groups than individual causal effects. Still, with the methods of causal analyses of observational data we can only explore the causal effect indirectly and under the validity of certain assumptions (Brüderl, 2010; Gangl, 2010).

The counterfactual framework proposes the use of counterfactuals to identify the average causal effect. Counterfactuals are defined as the unobservable outcome of person i if their treatment status had been different (Gangl, 2010; Pearl, 2009). In place of the unobservable outcome, observational data can be used to estimate the outcome the treatment group would have had, if they had not received the treatment (Oakes & Johnson, 2006). The estimation strategy of this counterfactual outcome is a very crucial decision because different methods use different

approaches. If plausible counterfactuals are estimated, they can be used to calculate the average causal effect, which is usually expressed as Average Treatment Effect on the Treated (ATT). To estimate the ATT for the treatment group, the counterfactual outcome Y_i^0 is subtracted from the observed outcome Y_i^1 (Dehejia & Wahba, 1999). The most important assumption is that no factors other than the treatment are responsible for the differences in the outcome of treatment and control group (Brüderl, 2010).

In this paper, I will highlight two different approaches of causal analysis: Fixed Effects (FE) models and Difference-in-Differences (DID) estimators. These methods use different approaches to estimate the counterfactuals and therefore underlie different assumptions. The aim of this paper is to show the advantages and disadvantages of both methods using a practical example from research on health and social inequalities.

Fixed Effects Models

Fixed Effects models have been employed widely in recent studies using panel data in the social sciences and are often used to evaluate the obesity penalty. FE models are appealing because they automatically condition on all time-constant unobserved heterogeneity (Gangl, 2010). Therefore, time-constant covariates cannot bias the causal effect. Thus, using FE models has clear advantages over traditional regressions (Brüderl, 2010).

Returning to the counterfactual framework of causality, the question is how FE models create the counterfactual to estimate the causal effect. In short, FE models estimate the causal effect within person i over time. The outcome of person i at time t1 before the treatment is used to construct the counterfactual. To estimate the causal effect this counterfactual is subtracted from the outcome of person i at time t2 after the treatment (Brüderl, 2010). Hence, the difference in the outcome between t2 and t1 is viewed as the causal effect.

Since FE models compare person i with itself, they automatically control for all unobserved heterogeneity that is time-constant (Brüderl, 2010). This is achieved through within transformation of the data. Within transformation removes the person-specific time-constant error by using only variation within individuals over time for the estimation of the treatment effect (Brüderl & Ludwig, 2015). Due to within transformation, the effect of characteristics of respondents that are stable over time is removed and thus changes in outcome are influenced only by treatment status, time-varying covariates and time-varying idiosyncratic error (Gangl, 2010). Therefore, the CIA is weaker than in traditional regression analysis; however, the assumption that time-varying unobserved characteristics do not bias the causal effect is still a strong one (Gangl, 2010).

Hence, the problem of unobserved heterogeneity that is not time-constant still remains (Hill et al., 2019). As long as information on the influencing factors that change over time is available in the data, conditioning on these variables will lead to unbiased estimates. Beyond that, the assumption of FE models is that there is no unobserved time-varying heterogeneity. Therefore, to strengthen the results of FE models, it is necessary to discuss explicitly which influencing factors might bias the causal effect and whether they can be controlled for in the model. Hence, substantive theoretical models must be the base of causal analysis (Gangl, 2010).

Further, within transformation of the data can also lead to higher risk of bias because only a selective group – the treated – contribute within information for the estimation (Gangl, 2010). This also enhances problems of measurement error due to misreporting or miscoding because small changes can lead to a big bias in the estimates (Angrist & Pischke, 2008). Additionally, FE models might also remove valuable information on the causal relationship of interest because of the within transformation of data.

Difference-in-Differences Estimator and Propensity Score Matching

Difference-in-Differences (DID) estimators use the same logic of comparing cases before and after treatment to estimate the causal effect (Gangl, 2010). However, DID estimators use the aggregate level, not the individual level (Angrist & Pischke, 2008). In contrast to FE models, DID estimators employ a control group to identify the causal effect of the treatment. Thus, the development of the outcome variable over time in the control group is used as the counterfactual for the changes in outcome the treatment group would have had, if they had not received the treatment (Halaby, 2004). DID estimates subtract the average change over time in the outcome variable of the control group from the average change over time in the outcome variable of the treatment group (Halaby, 2004; Stuart et al., 2014). Consequently, DID estimators condition on all group-specific time-constant unobserved heterogeneity (Gangl, 2010).

Additionally, DID estimators can reduce time-varying unobserved heterogeneity by using a control group. However, the central assumption of the DID approach is the parallel trends assumption: it is assumed that treatment and control group would have had the same development over time if the treatment had not happened in the treatment group (Caniglia & Murray, 2020; Cataife & Pagano, 2017). Therefore, the choice of control group is of utmost importance, as is shown by Angrist and Pischke (2008). If the parallel trends assumption does not hold, the DID estimate will be biased because the effect of time-varying unobserved heterogeneity is not statistically controlled for (Cataife & Pagano, 2017).

Similar to CIA in the case of FE models, the parallel trends assumption cannot be proven in a mathematical sense; however, it can be made plausible through theoretical arguments. One way to strengthen the assumption is to use Propensity Score Matching to weight the control group so it matches the treatment group in all relevant aspects (Godard-Sebillotte et al., 2019, Stuart et al., 2014; Heckman et al., 1997). Due to Propensity Score Matching, the treatment and control group are comparable to each other before the treatment. Therefore, it is more plausible that their further development would have been similar if the treatment had not happened (Caniglia & Murray, 2020; Cataife & Pagano, 2017). However, the selection of covariates chosen for Propensity Score Matching must be based on a strong theoretical model.

Since Propensity Score Matching was employed in the following analyses, a brief description of this method will be provided. The Propensity Score is "defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates" (Rosenbaum & Rubin, 1983, p. 41). Conditioning on the Propensity Score, there should be no difference in the probability of receiving the treatment between the treatment and control group. Thus, it is used to reduce the bias that exists in observational data due to self-selection into the treatment (Austin, 2007).

The Propensity Score is usually estimated via logit models, using the treatment as the dependent variable (Gangl, 2010). The relevant covariates that influence the probability of receiving the treatment are used as independent variables in these models (Oaks & Johnson, 2006). The covariates are chosen based on theoretical considerations (Rosenbaum & Rubin, 1983), usually based on the idea of d-separation (Pearl, 2009, p. 106): all paths that could bias the effect of the treatment on the outcome are closed conditioning on the Propensity Score. Thus, the causal effect can be estimated (Pearl, 2009).

Once the Propensity Score has been estimated, treatment and control group can be matched accordingly. The aim is to pair a treated and a control case with very similar Propensity Score values and compare their outcome. In practice, Propensity Score Matching is a way of weighting the data so that treatment and control group are comparable (Dehejia & Wahba, 2002). After Propensity Score Matching, the DID estimate can be used to calculate the causal effect.

In conclusion, DID estimators condition on group-specific time-constant unobserved heterogeneity. Propensity Score Matching will provide an adequate control group for the estimation if all relevant covariates are available in the data and a good matching quality can be achieved. Therefore, the DID estimator after Propensity Score Matching will also reduce unobserved time-varying heterogeneity, as long as the parallel trends assumption holds. However, the assumption that time-varying unobserved characteristics influence treatment and control group in

exactly the same way and therefore do not bias the DID estimator is still a strong one (Cataife & Pagano, 2017).

Analytical Strategy

After this brief overview of Fixed Effects models and Difference-in-Differences estimators, I will discuss the concrete analytical approach in this section.

First, I estimated several FE models. In these models, the dependent variable is the socioeconomic status score. Abdominal obesity constitutes the treatment variable. I excluded respondents that were pregnant during one of the interviews (4 cases) and disabled respondents (600 cases) from the analyses.

FE models control for time-invariant heterogeneity, however, time-varying heterogeneity might bias the effect. Therefore, the following time-variant control variables were chosen: marital status, number of adults and children in the household, years of education and age as well as age squared. Changes in marital status and household composition can directly affect socioeconomic status on the household level. At the same time, changes in marital status and household composition can influence body weight (Huyer-May, 2018). Changes in education directly affect socioeconomic status and can influence body weight indirectly through changes in health behavior (Brunello et al., 2013). Age affects both body weight and socioeconomic status positively but not necessarily linearly, thus it is a confounder of the causal relationship under study (Schienkiwitz et al., 2017; Krause & Schäfer, 2005). Respondents who had missing values on any of the control variables were excluded. I will present the results of the FE models with and without control variables. Separate models were estimated for men and women since the effect of obesity on socioeconomic status could vary by gender.

Second, I used Propensity Score Matching to prepare the data for the DID estimator. The choice of covariates to include in the estimation of the Propensity Score is of utmost importance. To allow for the interpretation of the DID estimator as a causal effect, all relevant variables need to be included as covariates in the estimation of the Propensity Score. Following a method introduced by Shrier and Platt (2008), I developed an explanatory model for the causal effect of obesity on socioeconomic status (Figure 1). Going through the six steps of the method led to the following list of covariates to include in the Propensity Score: gender, age, educational level, marital status and number of adults and children in the household, disability, diet and exercise. I estimated the Propensity Score using logit models including these covariates that were measured before the treatment. Respondents who had missing values on any of these variables were excluded.

After estimating the Propensity Score, I chose a matching algorithm. To identify the causal effect, a high matching-quality must be achieved. On the one hand, the overlap of the treatment and control group must be sufficient (Gangl, 2010).

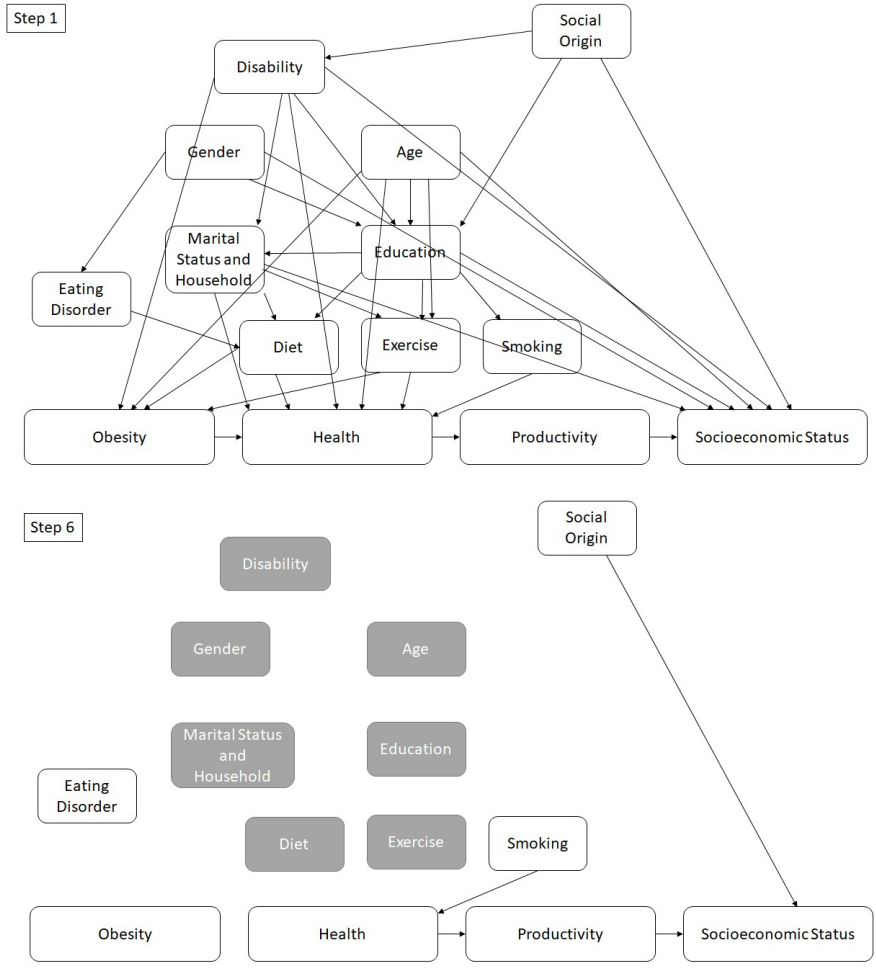


Figure 1 Explanatory model for the causal effect of obesity on socioeconomic status to choose covariates for Propensity Score Matching according to the method of Shrier and Platt (2008)

Therefore, there must be a reasonable number of respondents in each group that have a comparable Propensity Score (Figure 2, bottom). On the other hand, the matching algorithm that achieves the highest similarity in the chosen covariates between treatment and control group must be chosen. The best fit in this case was achieved using Radius Matching (Figure 2, top). Radius Matching is a variation of Caliper Matching where all possible matches with a certain maximum distance in the Propensity Score are used to create the counterfactual of each treated case (Caliendo & Kopeinig, 2008).

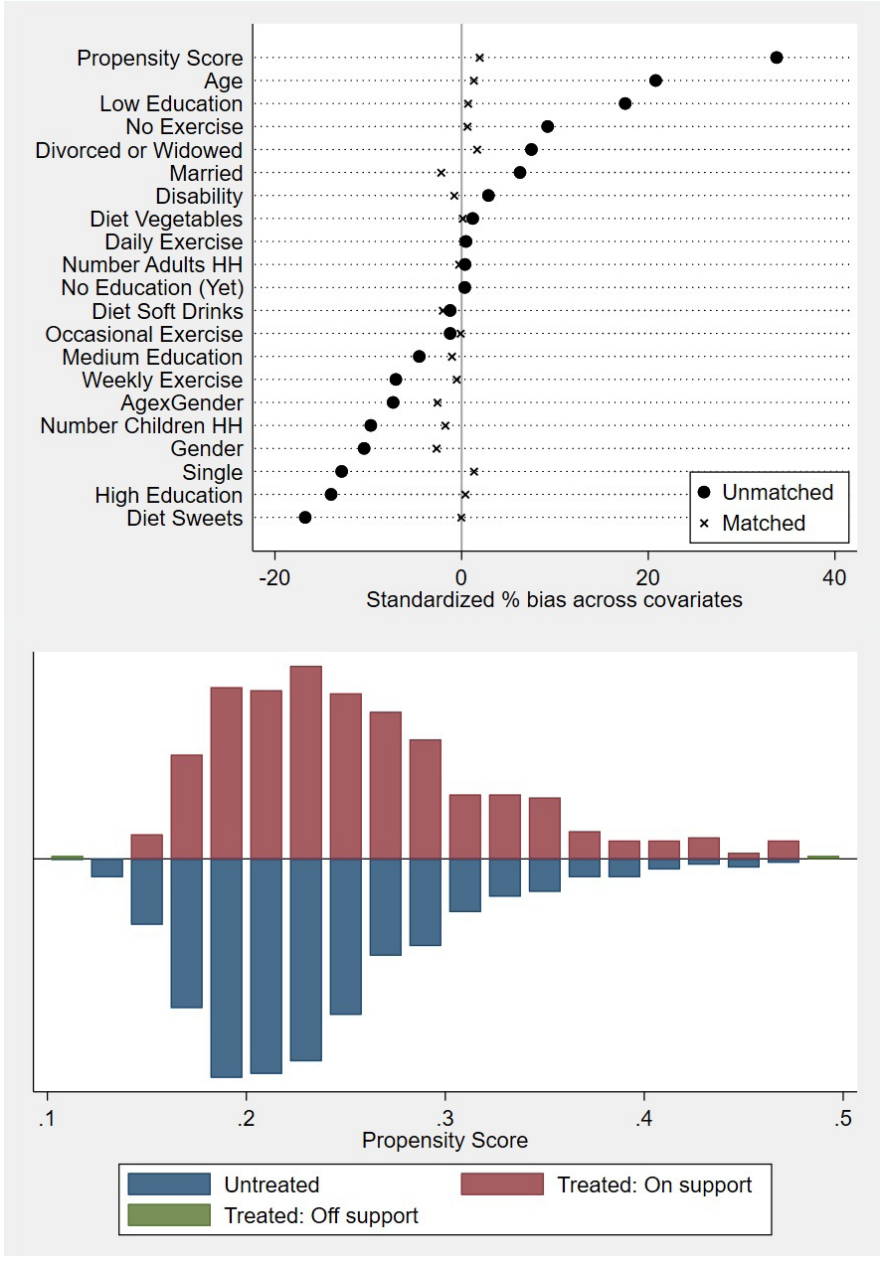


Figure 2 Bias reduction due to Propensity Score Matching (top) and overlap in the Propensity Score in treatment and control group (bottom) achieved through Radius Matching (Data: DEGS1 & GNHIES98)

Third, I calculated the DID estimator after Propensity Score Matching using the PSMATCH2 Stata module by Leuven and Sianesi (2003). Respondents who were already obese in the first wave of data collection and pregnant respondents were excluded. As the dependent variable, I used the socioeconomic status score and calculated the difference in the score between the first and second wave. This difference represents the change in socioeconomic status for each respondent during the observation period. The DID estimator then shows the difference in the changes over time between treatment and control group. I used bootstrapping to calculate standard errors (Gangl, 2010). Men and women were analyzed separately in case the causal effect varies by gender. The Propensity Score Matching process was repeated for each DID estimation.

Results

First, I will present the results of the FE models (Table 1). Model 1 represents the full sample and only includes the treatment variable without other covariates. Abdominal obesity has a non-significant positive effect on socioeconomic status according to Model 1 ($b = .197$, $p = .138$). We can see a non-significant .197 scale-points increase in socioeconomic status (on a scale ranging from 3-21) when respondents become obese. However, this result can be biased due to time-varying heterogeneity. Model 2 shows the results for the whole sample after conditioning on the time-varying control variables. The effect is still not statistically significant; however, it is now negative ($b = -.05$, $p = .719$). Controlling for changes in education, age, marital status, and composition of household, we see a slight decrease of socioeconomic status in respondents who become obese.

We observe the same pattern in the separate models for men and women. Model 3 shows a non-significant positive effect of abdominal obesity for female respondents ($b = .154$, $p = .368$). However, after conditioning on the control variables, in Model 4 the non-significant effect is negative ($b = -.097$, $p = .607$). For male respondents, the effect is not significant and positive in Model 5 ($b = .256$, $p = .222$) and not significant and negative in Model 6 after controlling for confounders ($b = -.067$, $p = .747$).

For both female and male respondents, FE models do not identify a significant causal effect of abdominal obesity on socioeconomic status. While the effect appears positive when confounders are not controlled for, it is negative after conditioning on the control variables. Respondents who become obese see a small decrease in socioeconomic status because of their obesity. However, this finding is not statistically significant and may therefore be due to chance.

Second, I will present the results of the DID estimator after Propensity Score Matching. Table 2 shows the findings for the full sample. The results are more illu-

Table 1 Fixed Effects models, dependent variable: Socioeconomic Status Score

	Model 1 Full Sample	Model 2 Full Sample	Model 3 Female Respondents	Model 4 Female Respondents	Model 5 Male Respondens	Model 6 Male Respondents
Obesity	.197 (.138)	-.050 (.719)	.154 (.368)	-.097 (.607)	.256 (.222)	-.067 (.747)
Conditioning on Controls		X		X		X
σ_u	3.324	2.665	3.121	2.592	3.523	2.754
σ_e	1.804	1.630	1.764	1.627	1.848	1.627
ρ	.773	.728	.758	.717	.784	.741
Within-R^2	.001	.192	.001	.163	.002	.240
Observations	3,761	3,761	1,972	1,972	1,789	1,789
Groups	2,209	2,209	1,157	1,157	1,052	1,052

Note. Data: DEGS1 & GNHIES98; Obesity: abdominal obesity (>88cm Waist Circumference for women, >102cm Waist Circumference for men); Control variables: years of education, marital status, number of adults and children in household, age, age²; p-values in parentheses; σ_u error due to differences between units, σ_e error due to differences within units, ρ proportion of variance due to unit effects

Table 2 Difference-in-Differences estimator of the full sample; dependent variable: Socioeconomic Status Score

Propensity Score Matching	Treatment Group	Control Group	DID Estimator	S.E.
Before	.107	.276	-.169	.135
After	.100	.085	.015	.128[1]
N (on support)	455	1,446		
N (off support)	2	0		

Note. Data: GNHIES98 & DEGS1, Treatment: abdominal obesity (>88cm Waist Circumference for women, >102cm Waist Circumference for men); [1] S.E. boot-strapped (1000 repetitions)

minating than in the FE models. We can see the changes in socioeconomic status in the treatment and control group before and after Propensity Score Matching as well as the DID estimator.

Before Propensity Score Matching, both treatment and control group see an increase in socioeconomic status over time. However, the increase of .276 points for the control group is larger than the increase of .107 in the treatment group. Therefore, the DID estimator before Propensity Score Matching is negative with -.169 points on the socioeconomic status score. This effect cannot be interpreted as causal, though, because differences in the composition of treatment and control group bias the results.

The bias becomes apparent when we consider the findings after Propensity Score Matching. While the increase in socioeconomic status for the treated group is only marginally smaller with .1 points, the increase of the control group is reduced to .085 points. The DID estimator is now positive with .015; however, it is not statistically significant. The finding that respondents who become obese gain less socioeconomic status over time than people who stay non-obese is explained by differences in the composition of both groups.

Table 3 shows the results for the female respondents. In this subgroup, the DID estimator is negative both before and after Propensity Score Matching. We can see that both the treatment and the control group experience an increase in socioeconomic status over time; however, the increase is only .041 in the treated group and .291 in the untreated group before Propensity Score Matching. After Propensity Score Matching, the DID estimator is not statistically significant with -.042 points. Among the female respondents, the differences in the changes in socioeconomic status over time between treatment and control group can be mostly explained by the different composition of the groups.

Table 3 Difference-in-Differences estimator for female respondents; dependent variable: Socioeconomic Status Score

Propensity Score Matching	Treatment Group	Control Group	DID Estimator	S.E.
Before	.041	.291	-.249	.182
After	.033	.076	-.042	.173[1]
N (on support)	251	722		
N (off support)	1	0		

Note. Data: GNHIES98 & DEGS1, Treatment: abdominal obesity (>88cm Waist Circumference for women, >102cm Waist Circumference for men); [1] S.E. boot-strapped (1000 repetitions)

Considering the male respondents, the findings are very similar. Table 4 shows that both treatment and control group see an increase in socioeconomic status over time both before and after Propensity Score Matching. The DID estimator after Propensity Score Matching is not statistically significant with .029 points. The differences in growth of socioeconomic status between treatment and control group over time can be explained by the different composition of the groups.

In conclusion, both FE models and DID estimators after Propensity Score Matching do not identify a causal effect of obesity on socioeconomic status. This is surprising because most previous studies find a negative effect of obesity on different aspects of socioeconomic status for women (Cawley, 2004; Han et al., 2009; Sabia & Rees, 2012, Katsaiti & Shamsuddin, 2016; Ahn et al., 2019; Lee et al., 2019). Others confirmed the obesity penalty for men as well (Baum & Ford, 2004; Wada & Tekin, 2010; Harris, 2019). Studies that cannot find a significant effect of body weight on socioeconomic status are rare. Bozoyan and Wolbring (2011) also do not find a significant effect of body weight on socioeconomic status. They use data from Germany and wages as dependent variable. Similarly, Cawley et al. (2005) use the IV method and find no significant effect of obesity on wages with German data. Thus, the presented results of this study are consistent with some previous research.

Table 4 Difference-in-Differences estimator for male respondents; dependent variable: Socioeconomic Status Score

Propensity Score Matching	Treatment Group	Control Group	DID Estimator	S.E.
Before	.187	.262	-.075	.202
After	.187	.159	.029	.198[1]
N (on support)	205	724		
N (off support)	0	0		

Note. Data: GNHIES98 & DEGS1, Treatment: abdominal obesity (>88cm Waist Circumference for women, >102cm Waist Circumference for men); [1] S.E. boot-strapped (1000 repetitions)

Discussion

The findings do not lend support to the first two hypotheses. Neither method shows a significant negative effect of abdominal obesity on socioeconomic status. Considering men and women separately, there is no significant effect of abdominal obesity on socioeconomic status for either gender. In conclusion, there is no evidence for an obesity penalty for adults in Germany. While respondents who become obese in general have fewer points on the socioeconomic status score than respondents who are not obese, this difference does not change over time because of obesity.

However, the aim of this study was to discuss the potential of using DID estimators combined with Propensity Score Matching in addition to FE models. The results indicate that using DID estimators can lead to more information on the obesity penalty because it explicitly estimates the outcome changes of the control group in addition to the treatment group. I find an increase of socioeconomic status for both the treatment and control group over time and differences in this increase are due to the different composition of these groups. Further, the use of Propensity Score Matching strengthens the focus on the correct choice of covariates based on theoretical considerations to achieve an unbiased causal effect.

FE models and DID estimators mainly differ in the way they construct the counterfactual to estimate the causal effect. While FE models use comparisons within individuals before and after treatment and construct the counterfactual from the before-measurement of the outcome, DID estimates compare the development in the outcome over time between a treatment and a control group. This has important implications for the results.

Since FE models compare the same individual before and after treatment, all time-constant heterogeneity cannot bias the causal effect (Brüderl, 2010). Therefore, these characteristics cannot be included and furthermore they need not be measured or even known (Angrist & Pischke, 2008). However, time-variant heterogeneity must be controlled for or it will bias the results (Hill et al., 2019). In comparison, DID estimators automatically control for time-variant heterogeneity, assuming it is the same in the treatment and control group (Cataife & Pagano, 2017). As long as the parallel trends assumption holds, these characteristics need not be measured or known. Thus, it is of utmost importance for the DID estimator that an adequate control group is found or constructed (Angrist & Pischke, 2008). One way of achieving such a control group is Propensity Score Matching (Godard-Sebillotte et al., 2019; Stuart et al., 2014). A drawback of this approach is the amount of covariates necessary to estimate the Propensity Score.

To produce unbiased causal effects, both methods need covariates based on theoretical considerations (Gangl, 2010). Usually, the theoretical model behind the chosen covariates stays implicit in many studies. None of the previous studies on the obesity penalty presents theoretical considerations as a base for their control

variables. For example, some studies use general health status as a control variable (Wada & Tekin, 2010; Bozoyan & Wolbring, 2011; Katsaiti & Shamsuddin, 2016; Lee et al., 2019; Ahn et al., 2020) even though it can be argued that general health is a causal link through which obesity influences socioeconomic status. This issue is not discussed in the studies. Some studies also include information on perceived discrimination without discussing the theoretical implications (Lee et al., 2019; Ahn et al., 2020). In general, none of the previous studies that use FE models discusses the explanatory model that their chosen covariates are based on.

While any causal analysis should make these decisions explicit, it is much more common in studies that use DID estimators because they have to discuss the parallel trends assumption. In addition, using Propensity Score Matching increases the need to describe the theoretical model and the method of choosing the covariates explicitly (Imbens, 2019; Gangl, 2010). Furthermore, with Propensity Score Matching there exist different methods to confirm matching quality. For example, figures showing the overlap in the Propensity Score of treatment and control group illustrate whether the groups are even similar enough to be compared (Dehejia & Wahba, 2002; Gangl, 2010). Usually there is no similar discussions about FE models and their quality in bias reduction.

Another way of looking at this is through considering the assumptions behind these two methods. The main assumption for FE models is that there would be no change in the outcome variable if there were no treatment (Brüderl, 2010). Meanwhile the main assumption if DID estimators is that the change in the treatment and control group would be the same if there were no treatment. Both are strong assumptions, even though some might argue that the one in FE models is stronger than the DID one because the counterfactual outcome at t2 itself has to be equal to the observed one, not only the counterfactual difference in outcome (Caniglia & Murray, 2020, p. 209).

However, it all comes down to theoretical considerations and well-chosen covariates. Using FE models, we must focus on the changes over time that occur simultaneously as respondents become obese. If important time-variant confounders cannot be controlled for, then the causal effect cannot be estimated. Employing DID estimators, we must concentrate on the differences of people who become obese and those who do not. If there is no adequate control group and none can be constructed using methods like Propensity Score Matching, then DID estimators will be biased. Thus, both methods have slightly different perspectives on causal effects and can therefore be considered complementary.

The main point of this study is to show the potential of adding DID approaches in combination with Propensity Score Matching in future research on the obesity penalty. Apart from the advantages already discussed, the explicit look at the control group within the DID approach is a great benefit.

Considering the results of the DID estimators again, we can derive some important new information. If we look at the DID estimator before Propensity Score Matching, we find a negative effect of abdominal obesity on socioeconomic status. We expect this finding according to the framework of the obesity penalty. However, after Propensity Score Matching this finding does not hold. Since there is no significant DID estimator after Propensity Score Matching, I conclude that the differences in the development of socioeconomic status between treatment and control group can be explained by their different composition. One important aspect is educational level: while higher educational level decreases the risk of becoming obese, it also leads to better job opportunities and higher income. If treatment and control group were comparable in their composition, becoming obese would not lead to differences in the growth of socioeconomic status. Future research into these characteristics could employ decomposition analysis to gain more knowledge about the relative importance of factors that influence the probability of becoming obese and the growth of socioeconomic status over time.

Additionally, we can also see from the results presented with the DID estimator, that both treatment and control group increased their socioeconomic status over time. The negative DID estimator shows us, that the increase in the treated group is smaller than the increase in the untreated group; however, both groups in general gain more socioeconomic status between the two waves. This is also valuable information concerning the obesity penalty. Potentially, the obesity penalty is not a decrease in socioeconomic status of the obese, but rather a slowed growth or stagnation in status. Looking closely at the DID results illustrates that well.

To sum up, the following advantages of the DID approach should be noted: First, the development in the outcome variable for treatment and control group is made explicit. Second, the DID estimator can be calculated before and after Propensity Score Matching to reduce bias due to the different composition of the groups. Third, the theoretical framework behind the choice of covariates for Propensity Score Matching and the parallel trend assumption must be made explicit and discussed.

The aim of this study is to show the advantages of combining FE models and DID estimates, and I have applied these methods in an example concerning obesity and socioeconomic status. I used data collected by the Robert Koch-Institute that have some clear limitations. First, so far there are only two waves of data available. Both FE models and DID estimators would benefit from a dataset with more waves included. Second, the two waves cover a period of about ten years. While this leads to a sufficient number of people who become obese between the two waves, it also leads to a lot of uncertainty about what happened within those ten years. For example, people could have become obese and then lost weight again before the second wave of data collection. We also have no information on when exactly respondents became obese within those ten years. This could influence whether

and how their socioeconomic status changed. Third, socioeconomic status is a variable on the household level. Therefore, other members of the household might level out changes in wages, income or job position that occur because of weight gain, especially for female respondents. Unfortunately, this is the only variable for socioeconomic status that has been measured for both waves of data. Thus, the presented results concerning the causal effect of obesity on socioeconomic status in Germany should be interpreted with caution and further research and better data on this topic are necessary.

In conclusion, the DID approach offers a new perspective and new insights in the obesity penalty. Evidently, the obesity penalty can be understood as a slowed growth or stagnation instead of a decrease in socioeconomic status. If obese individuals are more likely to have less favorable positive trends in socioeconomic status over time than other individuals, using DID estimates demonstrates the obesity penalty more effectively than using only FE models. Therefore, future research should employ the DID approach in addition to FE models to gain more information on the complex relationship of obesity and socioeconomic status.

References

Ahn, R., Kim, T. H., & Han, E. (2019). The Moderation of Obesity Penalty on Job Market Outcomes by Employment Efforts. *International journal of environmental research and public health 16* (16). DOI: 10.3390/ijerph16162974.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly harmless econometrics. An empiricist's companion.* Princeton, NJ: Princeton Univ. Press.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine 27* (12), 2037–2049. DOI: 10.1002/sim.3150.

Averett, S. & Korenman, S. (1993). The Economic Reality of the Beauty Myth. National Bureau of Economic Research. Cambridge, MA (NBER Working Paper Series, Working Paper No. 4521).

Ball, K. & Crawford, D. (2005). Socioeconomic status and weight change in adults: a review. *Social science & medicine (1982) 60* (9), 1987–2010. DOI: 10.1016/j.socscimed.2004.08.056.

Baum, C. L. & Ford, W. F. (2004). The wage effects of obesity: a longitudinal study. *Health economics 13* (9), 885–899. DOI: 10.1002/hec.881.

Bergmann, K. E. & Mensink, G. B. M. (1999). Körpermaße und Übergewicht. *Gesundheitswesen 61* (Sonderheft 2), S115-S120.

Böckerman, P., Cawley, J., Viinikainen, J., Lehtimäki, T., Rovio, S., Seppälä, I., Pehkonen, J., & Raitakari, O. (2019). The effect of weight on labor market outcomes: An application of genetic instrumental variables. *Health economics 28* (1), 65–77. DOI: 10.1002/hec.3828.

Bozoyan, C. & Wolbring, T. (2011). Fat, muscles, and wages. *Economics and human biology 9* (4), 356–363. DOI: 10.1016/j.ehb.2011.07.001.

Bozoyan, C. & Wolbring, T. (2018). The Weight Wage Penalty: A Mechanism Approach to Discrimination. *European Sociological Review 34* (3), 254–267. DOI: 10.1093/esr/jcy009.

Brüderl, J. (2010). Kausalanalyse mit Paneldaten. In C. Wolf & H. Best (Eds.). *Handbuch der sozialwissenschaftlichen Datenanalyse* (963–994). VS Verlag für Sozialwissenschaften.

Brüderl, J. & Ludwig, V. (2014). Fixed-effects panel regression. In H. Best und C. Wolf (Eds.). *The SAGE handbook of regression analysis and causal inference* (327–357). London: SAGE Publications.

Brunello, G., Fabbri, D., & Fort, M. (2013). The Causal Effect of Education on Body Mass: Evidence from Europe. *Journal of Labor Economics 31* (1), S. 195–223.

Caliendo, M. & Gehrsitz, M. (2016). Obesity and the labor market: A fresh look at the weight penalty. *Economics & Human Biology 23*, 209–225.

Caliendo, M. & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *J Economic Surveys 22* (1), 31–72. DOI: 10.1111/j.1467-6419.2007.00527.x.

Caniglia, E. C. & Murray, E. J. (2020). Difference-in-Difference in the Time of Cholera: a Gentle Introduction for Epidemiologists. *Current epidemiology reports 7* (4), 203–211. DOI: 10.1007/s40471-020-00245-2.

Cataife, G. & Pagano, M. B. (2017). Difference in difference: simple tool, accurate results, causal effects. *Transfusion 57* (5), 1113–1114. DOI: 10.1111/trf.14063.

Cawley, J. (2004). The Impact of Obesity on Wages. *J. Human Resources XXXIX* (2), 451–474. DOI: 10.3368/jhr.xxxix.2.451.

Cawley, J. H., Grabka, M. M. & Lillard, D. R. (2005). A Comparison of the Relationship between Obesity and Earnings in the U.S. and Germany. *Schmollers Jahrbuch 125* (1), 119–129.

Dehejia, R. H. & Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics 84* (1), 151–161. DOI: 10.1162/003465302317331982.

Dehejia, R. H. & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association 94* (448), S. 1053–1062.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annu. Rev. Sociol. 36* (1), 21–47. DOI: 10.1146/annurev.soc.012809.102702.

Gebremariam, M. K., Lien, N., Nianogo, R. A. & Arah, O. A. (2017). Mediators of socioeconomic differences in adiposity among youth: a systematic review. *Obesity Reviews 18* (8), 880–898. DOI: 10.1111/obr.12547.

Godard-Sebillotte, C., Karunananthan, S. & Vedel, I. (2019). Difference-in-differences analysis and the propensity score to estimate the impact of non-randomized primary care interventions. *Family practice 36* (2), 247–251. DOI: 10.1093/fampra/cmz003.

Gößwald, A., Lange, M., Kamtsiuris, P., & Kurth, B.-M. (2012). DEGS: Studie zur Gesundheit Erwachsener in Deutschland. Bundesweite Quer- und Längsschnittstudie im Rahmen des Gesundheitsmonitorings des Robert Koch-Instituts. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz 55* (6-7), 775–780. DOI: 10.1007/s00103-012-1498-z.

Haftenberger, M., Mensink, G. B. M., Vogt, S., Thorand, B., Peters, A., & Herzog, B. et al. (2016). Changes in Waist Circumference among German Adults over Time - Compil-

ing Results of Seven Prospective Cohort Studies. *Obesity facts 9* (5), 332–343. DOI: 10.1159/000446964.

Halaby, C. N. (2004). Panel Models in Sociological Research: Theory into Practice. *Annu. Rev. Sociol. 30* (1), 507–544. DOI: 10.1146/annurev.soc.30.012703.110629.

Han, E., Norton, E. C., & Stearns, S. C. (2009). Weight and wages: fat versus lean paychecks. *Health economics 18* (5), 535–548. DOI: 10.1002/hec.1386.

Harris, M. C. (2019). The Impact of Bofy Weight on Occupational Mobility and Career Development. *International Economic Review 60* (2), 631–660. DOI: 10.1111/iere.12364.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *Rev Econ Stud 64* (4), 605–654. DOI: 10.2307/2971733.

Hill, T. D., Davis, A. P., Roos, J. M., & French, M. T. (2020). Limitations of Fixed-Effects Models for Panel Data. *Sociological Perspectives 63* (3), 357–369. DOI: 10.1177/0731121419863785.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association 81* (396), 945–960.

Huyer-May, B. (2018). Do relationship transitions affect body weight? Evidence from German longitudinal data. *Journal of Family Research 30* (3), S. 316–338.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature 58* (4), 1129–1179. DOI: 10.1257/jel.20191597.

Katsaiti, M.-S. & Shamsuddin, M. (2016). Weight discrimination in the German labour market. *Applied Economics 48* (43), 4167–4182. DOI: 10.1080/00036846.2016.1153791.

Kim, T. J., Roesler, N. M., & von dem Knesebeck, O. (2017). Causation or selection - examining the relation between education and overweight/obesity in prospective observational studies: a meta-analysis. *Obesity Reviews 18* (6), 660–672. DOI: 10.1111/obr.12537.

Krause, P. & Schäfer, A. (2005). Verteilung von Vermögen und Einkommen in Deutschland: große Unterschiede nach Geschlecht und Alter. *DIW Wochenbericht 72* (11), S. 199–207.

Lampert, T., Kroll, L., Müters, S., & Stolzenberg, H. (2013). Messung des sozioökonomischen Status in der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1). *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz 56* (5-6), 631–636. DOI: 10.1007/s00103-012-1663-4.

Lee, H., Ahn, R., Kim, T. H., & Han, E. (2019). Impact of Obesity on Employment and Wages among Young Adults: Observational Study with Panel Data. *International journal of environmental research and public health 16* (1). DOI: 10.3390/ijerph16010139.

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*. DOI: 10.1177/0049124119882473.

Leuven, E. & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components S432001, Boston College Department of Economics, revised 01 Feb 2018.

Magallares, A. (2016). Drive for thinness and pursuit of muscularity: the role of gender ideologies. *Universitas Psychologica, 15*(2), 353-360.

Morris, S. (2007). The impact of obesity on employment. *Labour Economics 14* (3), 413–433. DOI: 10.1016/j.labeco.2006.02.008.

Oakes, J. M. & Johnson, P. J. (2006). Propensity Score Matching for Social Epidemiology. In J. M. Oakes und J. S. Kaufman (Eds.). *Methods in social epidemiology* (pp. 370–392). San Franscisco: John Wiley & Sons.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statist. Surv. 3* (none), 96–146 DOI: 10.1214/09-SS057.

Robert Koch-Institut, Abteilung Für Epidemiologie Und Gesundheitsmonitoring (2000). Bundes-Gesundheitssurvey 1998 (BGS98 1997-1999).

Robert Koch-Institut, Abteilung Für Epidemiologie Und Gesundheitsmonitoring (2015). Gesundheit von Erwachsenen in Deutschland (DEGS1 2008-2012).

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

Sabia, J. J. & Rees, D. I. (2012). Body weight and wages: evidence from Add Health. *Economics and human biology 10* (1), 14–19. DOI: 10.1016/j.ehb.2011.09.004.

Sari, N. & Acan Osman, B. (2018). The effect of body weight on employment among Canadian women: evidence from Canadian data. *Canadian journal of public health = Revue canadienne de sante publique 109* (5-6), 873–881. DOI: 10.17269/s41997-018-0097-7.

Scheidt-Nave, C., Kamtsiuris, P., Gößwald, A., Hölling, H., Lange, M., & Busch, M. A. et al. (2012). German health interview and examination survey for adults (DEGS) - design, objectives and implementation of the first data collection wave. *BMC Public Health 12* (730). DOI: 10.1186/1471-2458-12-730.

Schienkiwitz, A., Mensink, G. B. M., Kuhnert, R., & Lange, C. (2017). Übergewicht und Adipositas bei Erwachsenen in Deutschland. *Journal of Health Monitoring 2* (2), S. 21–28.

Shrier, I. & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC medical research methodology 8* (70). DOI: 10.1186/1471-2288-8-70.

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M., & Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health services & outcomes research methodology 14* (4), 166–182. DOI: 10.1007/s10742-014-0123-z.

Thefeld, W., Stolzenberg, H., & Bellach, B. M. (1999). Bundes-Gesundheitssurvey: Response, Zusammensetzung der Teilnehmer und Non-Responder-Analyse. *Gesundheitswesen 61* (Sonderheft 2), S57–S61.

Wada, R. & Tekin, E. (2010). Body composition and wages. *Economics and human biology 8* (2), 242–254. DOI: 10.1016/j.ehb.2010.02.001.

Winkler, J. & Stolzenberg, H. (1999). Der Sozialschichtindex im Bundes-Gesundheitssurvey. *Gesundheitswesen 61* (Sonderheft 2), S178-S183.

World Health Organization (2011). *Waist circumference and waist-hip ratio: report of a WHO expert consultation, Geneva, 8-11 December 2008*. Online: https://apps.who.int/iris/bitstream/handle/10665/44583/9789241501491_eng.pdf

methods, data, analyses | Vol. 18(1), 2024, pp. 59-78 DOI: 10.12758/mda.2024.01

Migrant Health Inequalities or Unequal Measurements? Testing for Cross-cultural and Longitudinal Measurement Invariance of Subjective Physical and Mental Health

Manuel Holz & Jochen Mayerl

Chemnitz University of Technology, Faculty of Behavioral and Social Sciences

Abstract

Background: The aim of the study is to investigate the longitudinal and cross-cultural measurement invariance of the Short-Form 12-Item Health Survey (SF-12) between Native Germans, European migrants and Non-European Migrants. Further, we test for differences in latent means dependent on invariance restrictions.

Methods: We include 7 waves (2006-2018) from a representative panel study in Germany. We apply Multigroup Confirmatory Factor Analysis via a Structural Equation Modelling approach. Finally, we compare gender and age adjusted latent means between different settings of invariance assumptions.

Results: The decrease in model fit measures by increasing equality constraints on the SF-12 factor structure of both physical and mental health between origin groups and across time is within common thresholds for good model fit. Latent means of both health factors differ, dependent on whether scalar invariance is set longitudinally and cross-culturally, or only longitudinally.

Conclusion: We conclude acceptable longitudinal and cross-cultural measurement invariance of the SF-12 for a period of 12 years. Yet, ignoring multigroup scalar invariance constraints produces bias in the latent means of both health factors, where migrant health is shown to be overestimated, especially for Non-European migrants if indicator intercepts are not sufficiently constrained.

Keywords: measurement invariance, cross-cultural comparison, longitudinal study, health inequality, migration, structural equation modelling, SF-12

The study of migrant health inequalities is a crucial and timely issue in post-industrial countries.

The complex nature of health inequalities in migrants is influenced by both subjective and objective factors. In terms of objective measures, migrants often exhibit a higher prevalence of chronic conditions like cardiovascular disease and obesity compared to the native populations (Raza et al., 2017; Rellstab et al., 2016). However, depending on how comparison groups are defined, e.g. with respect to duration of stay, results might differ. It was shown that recent migrants may actually show health advantages in chronic conditions, a phenomenon known as the "Healthy Migrant Effect" (HME) (McDonald et al., 2004). When comparing different countries of origin, variations in prevalence levels and differences compared to native populations have been observed in metrics like obesity (Campostrini et al. 2019), adverse cholesterol levels (Hergenç et al., 1999) and mortality rates (Weitoft et al., 1999).

The examination of subjective measures of health adds further complexity to the picture. On the one hand, there is evidence that newly arrived migrants experience health advantages in terms of subjective physical and mental health scores (Holz, 2022). On the other hand, when all migrants are compared with the native population, only minimal differences in physical and mental health scores persist (Metzing et al., 2019; Wengler, 2011). In particular, migrants from Western countries (Europe, Canada, the United States, etc.) tend to report higher self-rated health outcomes than their counterparts from non-Western countries (Acevedo-Garcia et al., 2010; Holz, 2022).

However, assessing subjective health measures in a cross-cultural context raises certain methodological challenges. Comparative social research has extensively demonstrated the impact of cultural contexts on cognition (Schwarz et al., 2010). Culture variant elements such as value orientations (e.g., individualism vs. collectivism) and other contextual information are strongly linked to cognitive processes during the survey response phase (Schwarz et al., 2010; Sudman et al., 1996; Tourangeau et al., 1988) and can therefore potentially induce bias, leading to variations in the interpretation of results of survey data.

In order to draw valid conclusions about differences in aspects of health between respondents from different cultural contexts, two important aspects need to be considered: firstly, the potentially different ways in which issues of illness, health and disease are expressed need to be taken into account. Secondly, it is necessary to test whether respondents consider the same aspects with the same

Direct correspondence to
Manuel Holz, Chemnitz University of Technology, Faculty of Behavioral and Social Sciences, Institute for Sociology, Thüringer Weg 9, 09126 Chemnitz, Germany
E-mail: manuel.holz@hsw.tu-chemnitz.de

importance and meaning when confronted with a particular object of thought. The existence of differences in meanings, cognition and response behavior can be empirically demonstrated by testing for measurement invariance (also known as measurement equivalence) (Cheung et al., 2000).

This article contributes to the field of comparative social research by addressing a crucial question: whether subjective health measures are genuinely comparable across groups and time periods in Germany. More specifically, our study focuses on assessing the longitudinal and cross-cultural measurement invariance of the Short-Form 12-Item Health Related Quality of Life Questionnaire (SF-12) in its physical and mental health components. The study spans 12 years, from 2006 to 2018, and includes three different groups of origin: European migrants, non-European migrants and native-born Germans without a migration background.

Conceptual Background

Culture, Health and Bias

The formation of health attitudes is significantly influenced by differences in cognition and cultural factors, as they are strongly determined by information from the social, institutional and media environment (Bakanauskas et al., 2020). This influence can lead to differences in attitudes, their conceptualization and survey response behavior. For example, the attribution of causes of disease and illness differs between 'Western' and non-'Western' populations. The Western perspective tends to follow the biomedical model, emphasizing individual responsibility and secular empirical explanations in the field of health and illness. In contrast, non-Western societies often additionally draw on socio-environmental explanations ('holistic' approaches) and may include magico-religious thinking (Bates et al., 1993; Anderson, 1999; Lee et al., 1996).

More precisely, cultural differences play an important role, e.g. in the conceptualization of chronic pain. Hispanic respondents have been shown to be more likely to perceive chronic pain as being beyond the individual's control, whereas non-Hispanic Caucasians, Italians, French Canadians, Irish or Polish respondents tend to believe that the variation of chronic pain can be influenced by the individual (Bates et al., 1993).

Religion, as a cultural factor, introduces additional bias in the pattern of missing values in survey responses on individual health levels. For example, some highly religious respondents in rural Lebanon refused to rate their future health using the SF-36 questionnaire (a related questionnaire to the SF-12) because it was considered blasphemous to make predictions about the future (Sabbah et al., 2003).

Furthermore, migration-specific issues can bring additional challenges. Migration to post-industrial countries is characterized by positive self-selection in terms of health (Holz, 2022), but variations in general health levels exist among different countries of origin (Jürges, 2007). This raises the issue of social comparison, where individuals assess their level of health based on the strategy of comparison used – whether they compare themselves to those who are better off or those who are worse off, potentially biasing self-rated health upwards or downwards (Beaument et al., 2004).

Measurement Invariance and Subjective Health

When conducting the test for measurement invariance, researchers examine the factor structure of latent constructs not only across groups but also over time (Cheung et al., 2000; Seddig et al., 2018). Only when a latent construct successfully passes the test for measurement invariance can latent mean differences be attributed to real differences between groups or time points, rather than being influenced by variations in the aforementioned contextual factors (Leitgöb et al., 2022).

The status of cross-cultural measurement invariance for subjective health measures remains unclear, with some authors affirming measurement invariance (Schulz, 2012), while others identify differences in factor structures based on cultural or ethnic background (Desouky et al., 2013; Fleishman et al., 2003; Lam et al., 2005). Longitudinal evidence for the invariance of subjective health measures is even more rare, although there is evidence for valid measures of subjective physical health over a period of up to four years (Cernat, 2015; Lynch et al., 2021).

Interest in the SF-12 scale as an instrument has been in both cross-cultural (Holz, 2022) and longitudinal contexts (O'Kelly et al., 2022; Teachman, 2011). However, most evidence for the measurement invariance of the SF-12 has come from separate investigations of the temporal and cultural/ethnic dimensions. To our knowledge, our study is the first to combine a cross-cultural and longitudinal examination of physical and mental health measurement. Based on our findings, we can provide evidence on whether the construction of additive indices or the application of the widely used scoring algorithm (Ware, 2007) leads to unbiased results in longitudinal and comparative studies.

Furthermore, we examine health differences between groups of origin and over time in models where measurement equivalence is partially ignored, in order to explore possible outcome bias due to violation of the invariance assumption. Although our case is limited to Germany, we believe that the results and issues addressed in this paper are transferable to other social and national contexts.

Data and Methods

Participants

We use secondary data from the German Socio-Economic Panel (GSOEP) (Liebig et al., 2021), a representative longitudinal survey of over 12,000 private households in Germany, conducted annually since 1984 by the German Institute for Economic Research (DIW). The survey modes used include CAPI, PAPI, CAWI and CASI, depending on the survey year (Deutsches Institut für Wirtschaftsforschung (DIW Berlin), 2023). Data from this panel is particularly well suited for Structural Equation Modelling, mainly due to its large sample size (more than 12,000 households), which increases the likelihood of detecting potential measurement biases (Meade & Lautenschlager, 2004). In addition, the panel is advantageous due to its deliberate oversampling of migrant respondents from (South) Eastern Europe and Southwest Asia (Herbert Brücker et al., 2014). Respondents were aged 17 and over. Health variables as repeated measures are available in the biennial survey waves of 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016 and 2018. In order to increase sample sizes for each migration group, waves 2002 and 2004 were excluded from the final sample, mitigating panel attrition concerns associated with a longer observation period.

SF-12

We use both the physical health scale and the mental health scale of the Short-Form 12-Item Health-Related Quality of Life Questionnaire (SF-12). The former is measured by six items: general health, limitations in climbing stairs and performing daily activities, presence of severe bodily pain in the past 4 weeks, limitations in performance due to physical health, and general limitations due to physical health (see Table 1 in the Supplementary Appendix for exact wording and scales). Mental health is measured by six items: frequency of feeling rushed and pressed for time, feeling down and gloomy, feeling calm and relaxed, feeling energetic, having achieved less than desired and doing tasks less thoroughly.

The debate over whether variables used for Health-Related Quality of Life (HRQoL) are reflective or formative indicators is critical (Testa et al., 2021). We argue for treating HRQoL indicators as reflective for the following reasons: firstly, the majority of items (7 out of 12) explicitly tie the health state of respondents as the cause of the health issues (for example: *"When you have to climb several flights of stairs on foot, does your health limit you greatly, somewhat or not at all".)* Secondly, we believe physical health issues cause pain and difficulty in climbing. For objective physical health problems, we argue these problems cause pain, not the reverse. Thirdly, the criterion for formative constructs, that a change in the

latent variable has low or no influence on indicators (Diamantopoulos et al., 2021; MacKenzie, 2003) does not apply; as subjective physical health declines, all indicators should tend to decline. Lastly, the widely-used SF-12, treated as reflective, consistently produced reliable results (Schulz, 2012; Kilbourne et al., 2008; Forero et al., 2018).

Migration Background

In our study, migrants are defined as respondents who were not born in the Federal Republic of Germany. Native Germans are identified when both the respondents and their parents were born in Germany. We do not consider indirect migration background or second generation migrants, where only one parent was born abroad or the respondent was born to foreign born parents in Germany, in this analysis. Additionally, we categorize migrants into European and Non-European groups based on the United Nations Statistics Division-Standard Country and Area Codes Classification (United Nations, 2013), utilizing the respondent's country of origin (country of birth) information.

For our focus on longitudinal effects, we only include cases with sufficient panel participation, excluding individuals with more than a total of 20 missing values across the 12 health indicators over all 7 waves. The final sample comprises data from waves 2006 to 2018, consisting of 8,922 cases. Among them, 8,427 are Native Germans (53.0% female, mean age=49.6 (sd=14.79) years), 485 are European migrants (57.4% female, mean age=49.5 (sd=14.51) years), and 164 are Non-European migrants (50.6% female, mean age=44.1 (sd=12.87) years).

In our sample, over 60% of European migrants predominantly originate from Eastern Europe (Poland, Russia, Czech Republic, Romania, Ukraine) and Southern Europe (Italy, Spain, Greece). Meanwhile, the majority (over 50%) of Non-European migrants are from Turkey.

Statistical Methods

The study aims to investigate the extent of measurement invariance in the SF-12 instrument across subgroups of European migrants, Non-European migrants, and native Germans over time, utilizing Multigroup Confirmatory Factor Analysis (MGCFA) within the framework of Structural Equation Modeling (SEM) (Kline, 2016). The procedure involves fitting a baseline model (configural model) where all factor loadings and intercepts are freely estimated across subgroups and waves (Model 0). Subsequent models progressively impose restrictions on factor loadings (metric invariance: Model 1 and Model 3) and intercepts (scalar invariance: Model 2 and Model 4) to be equal across subgroups and waves of the configural model. Measurement invariance is concluded when increasing constraints do not substan-

tially decrease model fit. Because in this analysis the focus lies on migrant health inequalities, multigroup invariance is tested before longitudinal invariance. If the construct does not pass, further invariance steps are unnecessary.

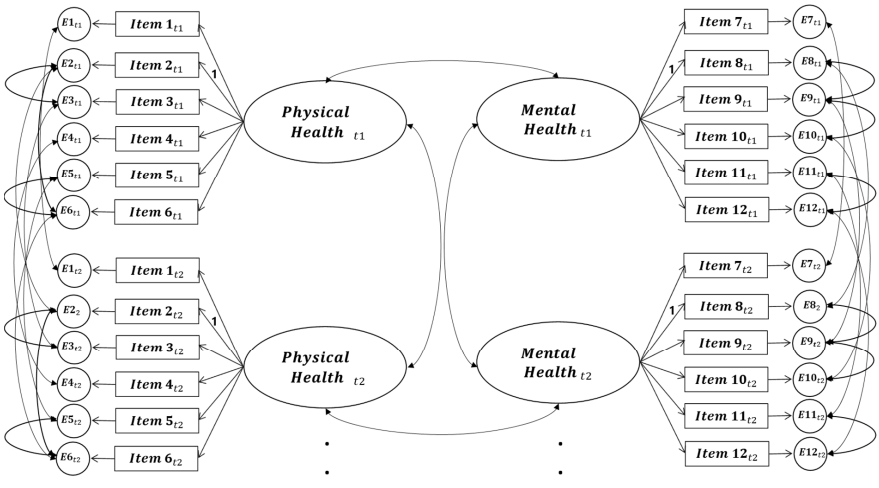
The criteria for establishing invariance include a lack of statistically significant increase in the model Chi-square value, a Comparative Fit Index (CFI) difference smaller than 0.01, a Root Mean Square Error of Approximation (RMSEA) difference smaller than 0.015 with overlapping 95% confidence intervals, and a Standardized Root Mean Square Residual (SRMR) difference smaller than 0.03 (Chen, 2007; Ploubidis et al., 2019). Single model fit criteria include a CFI above 0.95, and RMSEA and SRMR below 0.05 (Kline, 2016; Marsh et al., 2009).

Measurement invariance allows for meaningful comparison of latent factor means across groups and time without construct bias. This ensures that any observed differences in latent factor means (physical health and mental health) are attributable to real differences in the latent factors rather than variations in the properties of the dimensions (factor loadings and item intercepts) (Davidov et al. 2014; Mayerl 2016). Measurement error invariance testing is omitted due to the expected minimal impact on latent means (Joo & Kim, 2019).

Models 0 to 4 depict the primary invariance tests, wherein latent means are restricted to 0. Models 5, 6, and 7 illustrate the potential outcomes for unconstrained latent means in the absence of adequately established scalar invariance. The study calculates latent means adjusted for age and gender for each year by migration group (Model 5), using the Native German group in the first wave (year 2006) as a reference. In the context of SEM, by adjusted latent means we refer to the intercepts of the latent means after controlling for age and gender (both grand mean centered) in the regression (regression coefficients are set equal between origin groups). Potential consequences of insufficient invariance are explored, examining biased latent means due to non-equivalence of intercepts across groups (Model 6) or time (Model 7).

The analysis employs the Full Estimation Maximum Likelihood estimator (FIML) for its efficiency in handling missing values, conducted in RStudio (Version 2022.07.1) and lavaan (Version 06.-12).

Figure 1 illustrates the measurement model for physical and mental health, depicting indicators for each health construct. The model accounts for autocorrelation of error terms across survey years, autocorrelation for both health constructs, three contemporary error correlations per construct, and contemporary correlations between the latent factors physical and mental health. Item 2 in physical health and Item 8 in mental health serve as reference indicators with factor loadings set to 1.00. For brevity, the figure displays the model for the first two time points (t1 and t2), with subsequent waves up to 2018 following the same structure. E1, E2, etc., represent error terms or residuals of the indicators at specific time points. Additionally, contemporary correlations between the latent factors are included. Item



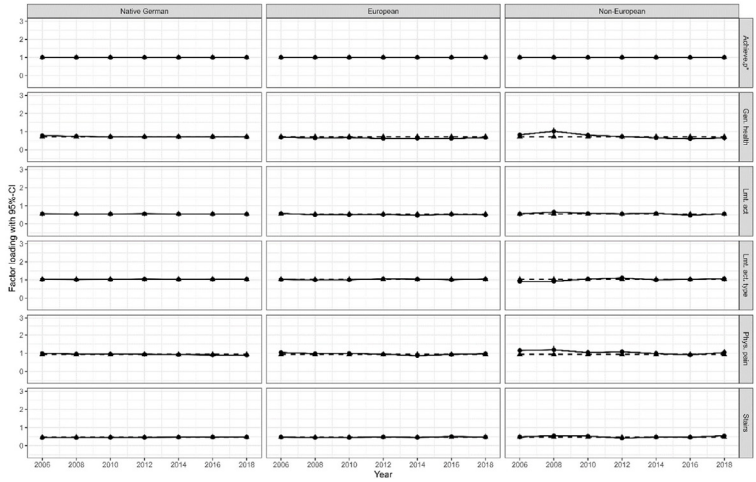
Note: Item wording and response scales can be found in Table 1 and in the Supplementary Material

Figure 1 Measurement model of the SF-12 physical and mental health component

wording details, as well as descriptive statistics can be found in the Supplementary Material (Table 1 and 2).

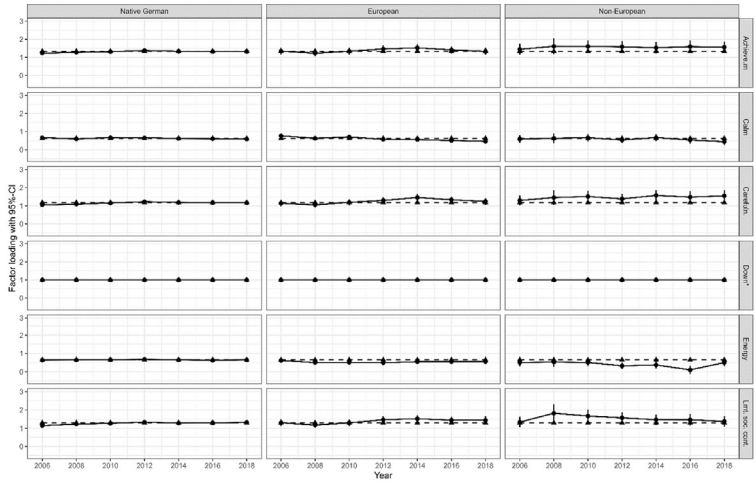
Results

Figures 2 and 3 depict unstandardized factor loadings over time for each migration group under the configural model (Model 0) and the full invariance model (Model 4). The straight line represents Model 0, while the dashed line shows factor loadings from the longitudinal and multigroup invariance model (Model 4). A closer alignment indicates a better fit. Results show that 'physical health' (Figure 2) remains consistent across survey years for each origin group, with minimal differences from the invariance model, as almost all factor loadings align and all confidence intervals overlap. In native Germans (Figure 3), 'mental health' exhibits no substantial differences between freely estimated factor loadings and metric invariance. However, Non-European migrants show more pronounced variations over time. In the Supplementary Material (Figure 1), standardized factor loadings are sufficiently high in physical health indicators over time, exceeding the 0.5 threshold. The 'mental health' indicator (Figure 2) shows weaker performance, especially in later survey waves (2012 to 2018).



Note: Achieve.p*=achieved less due to physical health; Gen.health=General health status; Lmt.act..= limited amount of activities due to physical health, Lmt.act.type=limited in type of activities due to physical health, Phys.pain=Physical pain; Stairs.=problems going up staris due to physical health; *Reference Indicator with factor loading set to 1.00; See Supplementary Material Table 1 for wording and scales

Figure 2 Unstandardized Factor loadings of physical health over time (Model 0 vs. Model 4)



Note: Achieve.m=achieved less due to mental health; Calm=felt calm; Carefuln.= work less thoroughly, Down*=felt down, Energy=felt energetic; Lmt.soc.cont.=limite social contatcs due to mental health; *Reference Indicator with factor loading set to 1.00; See Supplementary Material Table 1 for wording and scales

Figure 3 Unstandardized Factor loadings of mental health over time (Model 0 vs. Model 4)

Table 1 displays data fit measures for each step of invariance restriction. The 'x' in each row signifies parameters that were constrained to be equal and whether latent means were computed (in Models 0 to 4 latent means are constrained to 0). It is s worth noting that a sufficient model fit cannot be achieved without including three additional error correlations per construct (see Supplementary Material – Note to 4 for further explanation). At each invariance step, there is a notable rise in chi-square values. However, given that chi-square differences tend to be significant in larger sample sizes, closer scrutiny and detailed discussion are devoted to fit measures. Across all waves, both health constructs exhibit satisfactory fit indices in the configural model (Model 0) with RMSEA = 0.031, SRMR = 0.059, and CFI = 0.958. When factor loadings are restricted across groups (Model 1), the Chi-square value increases significantly, but other fit measures remain almost unchanged (RMSEA = 0.030, SRMR = 0.059, CFI = 0.958). The same holds for Model 2, where intercepts of indicators are set equal across origin groups, with minimal changes in fit indices except for the Chi-square value. In Model 3, setting factor loadings equal across waves results in no difference in RMSEA (0.030) and CFI (0.958), but an increase in SRMR by 0.001 (0.060). The final invariance step, constraining indicator intercepts over time (Model 4), leads to an RMSEA increase of 0.002 (0.032), an SRMR increase of 0.001 (0.061), and a CFI decrease of 0.006 (0.952).

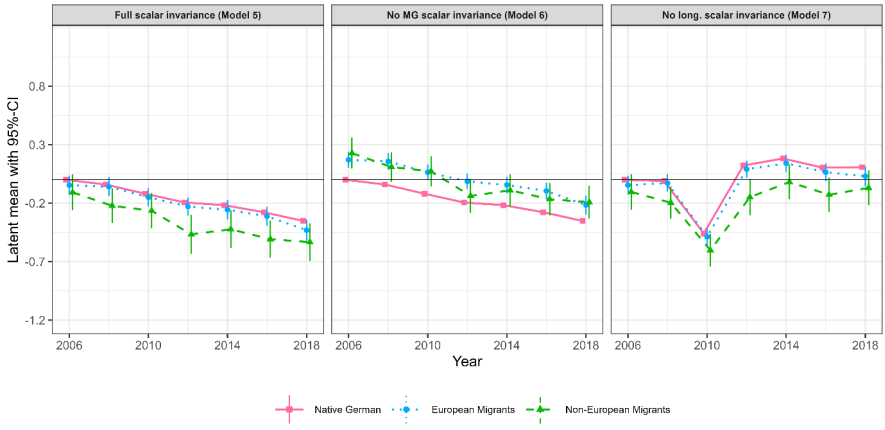
Models 6 and 7 do not establish full scalar invariance, complicating the estimation of latent means for comparing health measures between groups and over time. Comparing Model 5 (full scalar invariance) with Model 6 (no scalar invariance between groups) or Model 7 (no scalar invariance over time) allows us to assess potential outcome bias in health differences when scalar invariance is not fully specified (as in models 6 and 7).

In Figures 4 and 5, latent factor means of health constructs (controlled for gender and age) are presented, categorized by model restriction (Model 5 vs. Model 6 vs. Model 7). When comparing Model 5 and Model 6, differences in the trajectory of the latent construct 'physical health' for both migrant groups are evident. In Model 5 (full scalar invariance), European migrant health aligns with Native German health, while Non-European migrants consistently fall below both groups. In Model 6 (no multigroup scalar invariance), both migrant groups nearly follow the same trend, often lacking statistical significance compared to the reference group (Native Germans in the survey year 2006). Figure 5 illustrates that in Model 6, the trajectories for mental health almost align, indicating minimal negative slope. Full scalar invariance in Model 5 produces a similar trend as in the physical health trajectory, where European migrant health approximates Native German levels, and Non-European migrants consistently fall below, suggesting a slight decreasing tendency.

Table 1 Fit measures

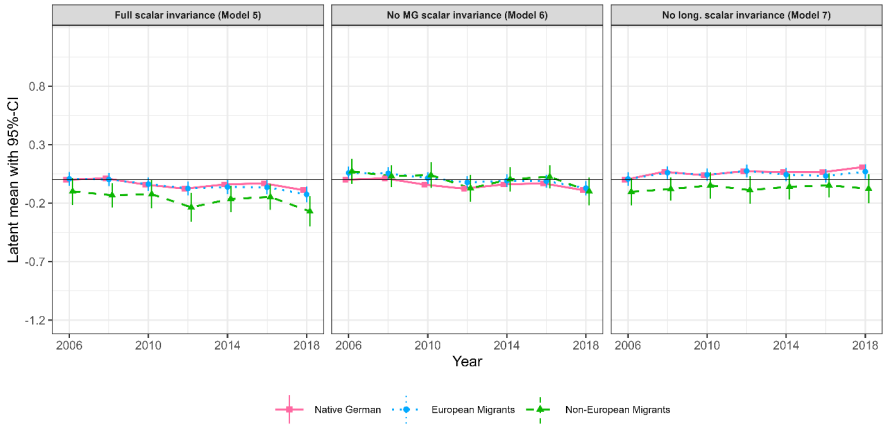
Model	Metric MG Invar.	Scalar MG. Invar.	Metric Longit. Invar.	Scalar Longit. Invar.	Latent Means	Chisq (Δ Chisq)	df (Δdf)	CFI	RMSEA (95%-CI)	SRMR
Testing for measurement invariance										
0						34297.22	9051	0.958	0.031 (0.030 - 0.031)	0.059
1	x					34580.80 (283.58***)	9191 (140)	0.958	0.030 (0.030 - 0.031)	0.059
2	x	x				34878.02 (297.22***)	9359 (168)	0.958	0.030 (0.030 - 0.031)	0.059
3	x	x	x			35273.93 (395.91***)	9419 (60)	0.957	0.030 (0.030 - 0.031)	0.060
4	x	x	x	x		38486.67 (3212.7***)	9491 (72)	0.952	0.032 (0.032 - 0.032)	0.061
Calculation of latent means in different invariance settings										
5	x	x	x	x	x	40017.79 (1531.1***)	9951 (460)	0.950	0.032 (0.032 - 0.032)	0.060
6	x		x	x	x	39878.92 (138.87***) [a]	9927 (24)	0.951	0.032 (0.030 - 0.032)	0.060
7	x	x	x		x	38488.25 (1529.5***) [b]	9883 (68)	0.953	0.031 (0.031 - 0.032)	0.061

Note: 'x' in each row indicates which parameters were restricted to be equal and if latent means were calculated.
MG.: Multigroup, Longit.: Longitudinal; Invar.: Invariance, Chisq: Chi-Square test value, df: degrees of freedom, RMSEA: Root Mean Square Error of Approximation, SRMR: Standardized Root Mean Squared Error, CFI: Comparative FIt Index;
Δ Chisq , Δ df, and Chisq significance difference tests always refer to values compared to the previous model.
[a] Model 6 is tested against Model 5; [b] Model 7 is tested against Model 5.
p-levels: $p \leq 0.000$: ****, $p \leq 0.001$: ***, $p \leq 0.01$: *



Note: Reference group: Native German in the year 2006 – effects controlled for age and gender (grand mean-centered)
MG: Multigroup; Longit.: Longitudinal; solid symbol: statistically significant to reference group with $p \leq 0.05$; empty symbol: statistically non-significant to reference group with $p > 0.05$

Figure 4 Latent means of physical health by scalar invariance restrictions



Note: Reference group: Native German in the year 2006 – effects controlled for age and gender (grand mean-centered)
MG: Multigroup; Longit.: Longitudinal; solid symbol: statistically significant to reference group with $p \leq 0.05$; empty symbol: statistically non-significant to reference group with $p > 0.05$

Figure 5 Latent means of mental health by scalar invariance restrictions

Further comparisons reveal differences in latent means between Model 5 and Model 7 (no longitudinal scalar invariance). By not setting intercepts equal across waves in both health constructs, the actual downward trend is not captured. Notably, in the physical component, there is a conspicuous abrupt decline in latent means in the year 2010 (Model 7).

Discussion

We can affirm that achieving acceptable metric and scalar measurement invariance is attainable for the latent constructs 'physical health' and 'mental health' of the SF-12 in a German panel survey across diverse groups and over the observation period, as per the established invariance criteria (Chen, 2007). Increasing restrictions on model parameters increases the deviation of the observed and expected matrices in the form of increasing Chi-square values. Nevertheless, the fit measures consistently signal satisfactory model performance (Kline, 2016; Marsh et al., 2009). It is important to highlight that achieving satisfactory data fit relies on incorporating additional error correlations. The improved data fit is presumed to result from factors such as question wording, position, and format.

Our findings align with the current literature (Ploubidis et al., 2019) and SF-12 research in Germany (Schulz, 2012). What sets our study apart is its contribution in integrating both longitudinal and cross-cultural dimensions of measurement invariance. While the invariance of the SF-12 has been examined in a more limited temporal context ($\leq$ 4 years) in previous studies (Cernat, 2015; Lynch et al., 2021) our research extends this examination, affirming the functionality of the SF-12 over a more extensive 12-year time span.

We noted a slightly heightened efficacy of the SF-12 survey for native Germans, with more consistent factor loadings over time, while other groups display more longitudinal variation, e.g. in the mental construct ('energy' in Figure 3)[1]. Despite literature highlighting cultural differences in self-rated health measures (Crockett et al., 2005; Desouky et al., 2013), our study aligns with global fit standards (Schulz, 2012). Limited German language proficiency may contribute to migrants showing disruptions in mental health factor loadings. Prior studies indicate that mental health indicators are prone to Different Item Functioning among ethnic groups (Crockett et al., 2005; Desouky et al., 2013; Fleishman et al., 2003), possibly stemming from diverse interpretations of mental illness (Crockett et al.,

1 We acknowledge that separate calculations for each group and a comparison of fit indices is needed to deliver an empirical test for differing functionality. We assume that a low level of variation of factor loadings over time is a sign for consistency of the construct and thus a sufficient but not necessary condition for functionality of a questionnaire.

2005; Roberts et al., 1992). Culturally distinct cognitive processes and response styles may also play a role. Investigating nuances like middle category or extreme responding is crucial in measurement invariance research (Weijters et al., 2008). However, merging respondents from different continents into the "Non-European" category may potentially lower the quality of correlational relationships between indicators and factors.

Moreover, we identified significant differences in latent factor means based on whether full scalar invariance between groups and/or time was specified. When the intercepts of the indicators are set equal only across waves, but not across origin groups (Model 5 vs. Model 6), the health of migrants is prone to substantial overestimation. In Model 6 (depicted in Figure 4), an initial physical health advantage of migrants over Native Germans endures over time, with European and Non-European migrants appearing almost indistinguishable. However, when intercepts are additionally set equal across groups (Model 5), the scenario changes markedly. Non-Europeans now exhibit a persistent health disadvantage over time, while the health trajectory for European migrants closely mirrors that of Native Germans. Beyond considerations related to survey response (Fleishman & Lawrence, 2003; Weijters et al., 2008), the potential overestimation of migrant health levels might be attributed to the positive selection of healthier individuals participating in large household surveys (Saß et al., 2015).

Latent means for the 'mental health' construct (refer to Figure 5) also vary depending on the invariance setting. In the 'softer' invariance model, Model 6, we observe minimal differences in latent means between migrants and native Germans, both over time and in terms of longitudinal trends. However, when implementing longitudinal multigroup scalar invariance (Model 5), the scenario changes, revealing that Non-European migrants consistently score below both Native Germans and European migrants. Disregarding longitudinal scalar invariance (Model 7) results in a sudden drop in all latent means of physical health in the year 2010. We attribute this to a potential mode effect, as the composition of survey modes became more reliant on Computer-Assisted Personal Interviews (CAPI) from 2010 onward (Deutsches Institut für Wirtschaftsforschung (DIW Berlin), 2023). The mental health trajectory of Native Germans and European migrants remains almost identical, displaying only a slight decreasing tendency.

This finding contradicts the cross-sectional results of Schulz (Schulz, 2012), where no significant mean differences between origin groups were identified. However, it aligns with the results of Fleishman et al. (2003), where adjustments for Different Item Functioning (DIF) reduced ethnic minority health advantages. Unlike the cross-sectional study by Fleishman et al. (2003), our study reveals changes in latent means in both dimensions of the SF-12 (physical and mental) after imposing invariance constraints. We attribute these differing findings primarily to our lon-

gitudinal approach and our focus on (first generation) migration status rather than ethnic minority status.

We conducted a robustness check by recalculating the entire invariance test using the Weighted Least Squares with Mean and Variance adjustment (WLSMV) estimator, applying the threshold invariance approach for ordered-categorical variables as suggested by Liu et al. (2017). Furthermore, we recalculated the model by using the Robust Maximum Likelihood estimator. The results from the threshold invariance and the robust analysis consistently supports our findings (see Supplementary Material). Despite this, for consistency with approaches in the literature (Schulz, 2012; Testa et al., 2021; Anagnostopoulos et al., 2009), we maintain the FIML estimation in our primary analysis.

In addition to our contributions, the analysis comes with certain limitations. We faced a trade-off between the number of waves and sample sizes across various origin groups. Given our specific focus on the consistency concept of 'health' over time, delving deeper into more specific regions of origin was unfeasible due to compromised sample sizes. Future studies could explore this by utilizing a reduced number of later waves from the GSOEP and delving into country-specific differences, as demonstrated in Schulz (2012), while adopting a longitudinal approach.

Another issue dependent on sample size that we could not address is the categorization of migrant groups into recent and non-recent migrants, a crucial element for analyzing the Healthy Migrant Effect. As highlighted in the introduction, chronic health conditions vary based on migration status. Whether the SF-12 yields reliable and valid results when considering different cultural groups over time and under varying chronic conditions (objective health measures) is a question that requires exploration in future research.

Language poses another challenge. While there is some information available about whether a translation of the GSOEP questionnaire was used, the topic itself is intricate. Depending on the survey year, demand and costs; various translations, translation devices, aids, or in-person interpreters were available for the interview (Liebau et al., 2015). There is no information on the language version concerning the questionnaire language at the beginning of our observation period (wave 2006). Drawing valid conclusions about the influence of language on factor structures between groups necessitates further research.

Conclusion

Utilizing seven waves spanning from 2006 to 2018 of the GSOEP and employing a Structural Equation Modelling approach, we examined the intercultural and longitudinal measurement invariance of the SF-12 in both its physical and mental health components. Our findings provide empirical evidence that both scales

achieve metric and scalar measurement invariance across native Germans without a migration background, European, and Non-European migrants over time. This finding supports the functionality of summative indices or the standard scoring algorithm (Ware, 2007). However, despite attesting measurement invariance, differences persist among these groups. It is crucial to note that the identification of measurement invariance does not imply that invariance steps can be overlooked in longitudinal multigroup analysis. Instead, we demonstrated that neglecting scalar invariance could lead easily to biased results in latent mean comparisons.

When the longitudinal latent mean difference between the investigated origin groups within a Structural Equation Modelling framework (using lavaan) is the estimand of interest, it is crucial that all models are constrained to full scalar invariance across groups *and* time[2]. Given that the values of latent means heavily rely on the intercept structure of the indicators, the health of migrants, especially Non-European migrants, is susceptible to overestimation if indicator intercepts are not equated.

Consequently, we advocate for the use of a Structural Equation Modelling approach when engaging in intercultural and longitudinal analyses of the SF-12. Special attention should be given to specifying metric and scalar invariance when the focus involves multigroup latent mean differences or health trajectories over time.

Note

The Online Appendix containing the results of robustness checks and analysis scripts can be retrieved from
https://doi.org/10.5281/zenodo.10521878

Statements and Declarations

No potential conflict of interest was reported by the authors.
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Ethical approval

Not Applicable.

Informed Consent

Not applicable.

Consent for publication

Not applicable.

2 The built-in invariance option of lavaan only equalizes the parameters of the grouping variable; if more time points are defined in the measurement model, these parameters have to be set equal manually.

Data availability

The data that support the findings of this study are available from German Institute for Economic Research (Deutsches Institut für Wirtschaftsforschung). Restrictions apply to the availability of these data, which were used under license for this study.

Contribution

Both authors contributed equally to this work and were involved in drafting the manuscript. Both authors read and approved the final manuscript.

References

Acevedo-Garcia, D., Bates, L. M., Osypuk, T. L., & McArdle, N. (2010). The effect of immigrant generation and duration on self-rated health among US adults 2003-2007. *Social Science & Medicine, 71 (*6), 1161–1172. DOI: 10.1016/j.socscimed.2010.05.034.

Anderson, C. A. (1999). Attributional Style, Depression, and Loneliness: A Cross-Cultural Comparison of American and Chinese Students. *Personality and Social Psychology Bulletin, 25(*4), 482–499. DOI: 10.1177/0146167299025004007.

Anagnostopoulos, F., Niakas, D., & Tountas, Y. (2009). Comparison between exploratory factor-analytic and SEM-based approaches to constructing SF-36 summary scores. *Quality of Life Research, 18,* 53-63.

Bakanauskas, A. P., Kondrotienė, E., & Puksas, A. (2020). The Theoretical Aspects of Attitude Formation Factors and Their Impact on Health Behaviour. *Management of Organizations: Systematic Research, 83* (1), 15–36. DOI: 10.1515/mosr-2020-0002.

Bates, M. S., Edwards, T. W., & Anderson, K. O. (1993). Ethnocultural influences on variation in chronic pain perception. *Pain, 52 (*1), 101–112. DOI: 10.1016/0304-3959(93)90120-E.

Beaumont, J. G., & Kenealy, P. M. (2004). Quality of life perceptions and social comparisons in healthy old age. *Ageing and Society, 24 (*5), 755–769. DOI: 10.1017/S0144686X04002399.

Campostrini, S., Carrozzi, G., Severoni, S., Masocco, M., & Salmaso, S. (2019). Migrant health in Italy: a better health status difficult to maintain-country of origin and assimilation effects studied from the Italian risk factor surveillance data. *Population Health Metrics, 17 (*1), 14. DOI: 10.1186/s12963-019-0194-8.

Cernat, A. (2015). Impact of mixed modes on measurement errors and estimates of change in panel data. *Survey Research Methods, 9 (*2), 83-99. DOI: 10.18148/srm/2015.v9i2.5851.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14 (*3), 464–504. DOI: 10.1080/10705510701301834.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-Cultural Psychology, 31 (*2), 187–212. DOI: 10.1177/0022022100031002003.

Crockett, L. J., Randall, B. A., Shen, Y.-L., Russell, S. T., & Driscoll, A. K. (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. *Journal of consulting and clinical psychology, 73 (*1), 47–58. DOI: 10.1037/0022-006X.73.1.47.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology, 40* (1), 55–75. DOI: 10.1146/annurev-soc-071913-043137.

Desouky, T. F., Mora, P. A., & Howell, E. A. (2013). Measurement invariance of the SF-12 across European-American, Latina, and African-American postpartum women. *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 22 (*5), 1135–1144. DOI: 10.1007/s11136-012-0232-5.

Deutsches Institut für Wirtschaftsforschung (DIW Berlin). (2023). Survey Concepts and Modes. Retrieved from https://companion.soep.de/Survey%20Design/Survey%20Concepts%20and%20Modes.html

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of marketing research, 38 (*2), 269–277.

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning? *Medical care,* III75-III86.

Herbert Brücker, Ingrid Tucci, Simone Bartsch, Martin Kroh, Parvati Trübswetter, & Jürgen Schupp. (2014). Neue Muster der Migration. *DIW Wochenbericht, 81 (*43), 1126–1135. Retrieved from http://hdl.handle.net/10419/104052.

Hergenç, G., Schulte, H., Assmann, G., & Eckardstein, A. von. (1999). Associations of obesity markers, insulin, and sex hormones with HDL-cholesterol levels in Turkish and German individuals. *Atherosclerosis, 145 (*1), 147–156. DOI: 10.1016/S0021-9150(99)00027-1.

Holz, M. (2022). Health inequalities in Germany: Differences in the 'Healthy migrant effect'of European, non-European and internal migrants. *Journal of Ethnic and Migration Studies, 48 (*11), 2620–2641.

Joo, S.-H., & Kim, E. S. (2019). Impact of error structure misspecification when testing measurement invariance and latent-factor mean difference using MIMIC and multiple-group confirmatory factor analysis. *Behavior research methods, 51 (*6), 2688–2699. DOI: 10.3758/s13428-018-1124-6.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health economics, 16 (*2), 163–178. DOI: 10.1002/hec.1134.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling (Fourth; T. D. Little, Ed.).* New York (UK): The Guilford Press.

Lam, C. L. K., Tse, E. Y. Y., & Gandek, B. (2005). Is the standard SF-12 health survey valid and equivalent for a Chinese population? *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 14 (*2), 539–547. DOI: 10.1007/s11136-004-0704-3.

Lee, F., Hallahan, M., & Herzog, T. (1996). Explaining Real-Life Events: How Culture and Domain Shape Attributions. *Personality and Social Psychology Bulletin, 22 (*7), 732–741. DOI: 10.1177/0146167296227007.

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., Roover, K. d., et al. (2022). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*. DOI: 10.1016/j.ssresearch.2022.102805.

Liebau, E., & Tucci, I. (2015). Migrations- und Integrationsforschung mit dem SOEP von 1984 bis 2012: Erhebung, Indikatoren und Potenziale. *Berlin: Deutsches Institut für Wirtschaftsforschung (DIW) (SOEP Survey Papers, 270).* Retrieved from https://www.econstor.eu/handle/10419/111916.

Liebig, S., Goebel, J., Schröder, C., Grabka, M., Richter, D., Schupp, J., et al. (2021). Sozio-oekonomisches Panel, Daten der Jahre 1984-2019 (SOEP-Core, v36, EU Edition). *Berlin: Kantar Deutschland GmbH.*

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological methods, 22* (3), 486.

Lynch, C. P., Cha, E. D. K., Mohan, S., Geoghegan, C. E., Jadczak, C. N., & Singh, K. (2021). Two-year validation and minimal clinically important difference of the Veterans RAND 12 Item Health Survey Physical Component Score in patients undergoing minimally invasive transforaminal lumbar interbody fusion. *Journal of neurosurgery. Spine*. DOI: 10.3171/2021.6.SPINE21231.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16 (*3), 439–476. DOI: 10.1080/10705510903008220.

Mayerl, J. (2016). Environmental concern in cross-national comparison: Methodological threats and measurement equivalence. In: *Green European: Routledge,* 210–232.

McDonald, J. T., & Kennedy, S. (2004). Insights into the 'healthy immigrant effect': health status and health service use of immigrants to Canada. *Social science & medicine, 59 (*8), 1613–1627. DOI: 10.1016/j.socscimed.2004.02.004.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo Study of Confirmatory Factor Analytic Tests of Measurement Equivalence/Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 11 (*1), 60–72. DOI: 10.1207/S15328007SEM1101_5.

Metzing, M., & Schacht, D. (2019). Gesundheitliche Situation der Bevölkerung mit Migrationshintergrund in Deutschland - Sonderauswertung für die Bundesintegrationsbeauftragte 2019. *SOEP Survey Papers.* Retrieved from http://hdl.handle.net/10419/196880.

O'Kelly, B., Vidal, L., Avramovic, G., Broughan, J., Connolly, S. P., Cotter, A. G., et al. (2022). Assessing the impact of COVID-19 at 1-year using the SF-12 questionnaire: Data from the Anticipate longitudinal cohort study. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases, 118,* 236–243. DOI: 10.1016/j.ijid.2022.03.013.

Ploubidis, G. B., McElroy, E., & Moreira, H. C. (2019). A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts. *Longitudinal and Life Course Studies, 10 (*4), 471–489. DOI: 10.1332/175795919X15683588979486.

Raza, Q., Nicolaou, M., Dijkshoorn, H., & Seidell, J. C. (2017). Comparison of general health status, myocardial infarction, obesity, diabetes, and fruit and vegetable intake between immigrant Pakistani population in the Netherlands and the local Amsterdam population. *Ethnicity & health, 22 (*6), 551–564. DOI: 10.1080/13557858.2016.1244741.

Roberts, R. E., & Sobhan, M. (1992). Symptoms of depression in adolescence: A comparison of Anglo, African, and Hispanic Americans. *Journal of Youth and Adolescence, 21 (*6), 639–651. DOI: 10.1007/BF01538736.

Sabbah, I., Drouby, N., Sabbah, S., Retel-Rude, N., & Mercier, M. (2003). Quality of life in rural and urban populations in Lebanon using SF-36 health survey. *Health and Quality of Life Outcomes, 1 (*1), 30. DOI: 10.1186/1477-7525-1-30.

Sara Rellstab, Marco Pecoraro, Alberto Holly, Philippe Wanner, & Karine Renard. (2016). The Migrant Health Gap and the Role of Labour Market Status: Evidence from Switzerland. *IRENE Working Paper.* Retrieved from http://hdl.handle.net/10419/191494

Saß, A.-C., Grüne, B., Brettschneider, A.-K., Rommel, A., Razum, O., & Ellert, U. (2015). Beteiligung von Menschen mit Migrationshintergrund an Gesundheitssurveys des Robert Koch-Instituts. *Bundesgesundheitsblatt, 58 (*6), 533–542. DOI: 10.1007/s00103-015-2146-1.

Schulz, M. (2012). Messartefakte bei der Erfassung der Gesundheit von Migranten in Deutschland: Zur interkulturellen Äquivalenz des SF-12-Fragebogen im Sozio-oekonomischen Panel (SOEP). *SOEPpapers on Multidisciplinary Panel Data Research*. Retrieved from http://hdl.handle.net/10419/59013.

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In: *Survey methods in multinational, multiregional, and multicultural contexts,* 175–190.

Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: concept and application with panel data. *Survey Research Methods, 12 (*1), 29-41. DOI: 10.18148/srm/2018.v12i1.7210.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* Jossey-Bass.

Teachman, J. (2011). Are veterans healthier? Military service and health at age 40 in the all-volunteer era. *Social Science Research, 40 (*1), 326–335. DOI: 10.1016/j.ssresearch.2010.04.009.

Testa, S., Di Cuonzo, D., Ritorto, G., Fanchini, L., Bustreo, S., Racca, P., & Rosato, R. (2021). Response shift in health-related quality of life measures in the presence of formative indicators. *Health and Quality of Life Outcomes, 19,* 1-11.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103 (*3), 299–314. DOI: 10.1037/0033-2909.103.3.299.

United Nations. (2013). United Nations Statistics Division-Standard Country and Area Codes Classifications.

Ware, J. E. (2007). *User's manual for the SF-12v2TM health survey.* Lincoln: QualityMetric Incorporated.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science,36* (3), 409–422. DOI: 10.1007/s11747-007-0077-6.

Weitoft, G. R., Gullberg, A., Hjern, A., & Rosén, M. (1999). Mortality statistics in immigrant research: method for adjusting underestimation of mortality. *International Journal of Epidemiology, 28* (4), 756–763. DOI: 10.1093/ije/28.4.756.

Wengler, A. (2011). The health status of first- and second-generation Turkish immigrants in Germany. *International journal of public health, 56 (*5), 493–501. DOI: 10.1007/s00038-011-0254-8.

methods, data, analyses | Vol. 18(1), 2024, pp. 79-108 DOI: 10.12758/mda.2023.06

Challenges in Assigning Panel Data With Cryptographic Self-generated Codes – Between Anonymity, Data Protection and Loss of Empirical Information

Christina Beckord

ehs University of Applied Sciences for Social Work, Education and Nursing, Dresden

Abstract

The assignment of questionnaires between the 13 survey waves in the panel study "Crime in the Modern City" (CrimoC) was done by matching self-generated codes. This method was challenging because the individual codes tend to be ambiguous, prone to errors and the resulting panel data can be biased. The individual data were merged over time using an error-tolerant matching process with manual handwriting comparison. Despite these problems, there is no alternative to the chosen method with regard to anonymity and data-protection. Until now, the self-generated codes of each new survey wave were matched against the codes of the last and second-last wave. Over the years, this led to an increasing discrepancy between the data originally collected and the data linked to the panel. For this reason, first in a pretest and later for the complete sample, the cases that had not yet been linked to the panel were subsequently matched with earlier waves. This panel consolidation proved to be very successful. A total of 3,589 original missing units were subsequently filled with case data. This paper describes the steps taken to optimize the quality of the panel data set and illustrates exemplarily on specified criteria which properties of the panel data set could be improved. Since the importance of panel studies is steadily increasing in social science research this paper is relevant for researchers who need to make matching decisions within panel studies. Assurance of anonymity can counteract panel attrition. Self-generated codes represent one possibility in this regard, and are discussed in terms of feasibility and effectiveness.

Keywords: panel data, missing unit, personal codes, assignment rates

Longitudinal and panel designs are useful for analyzing intra- and inter-individual changes. A major challenge in this context is the matching of individual data over time. If no data from a new survey time point can be assigned to a previous time point, this can have two causes: Either the person did not actually participate in the new survey (refusal) or he or she did participate but the data could not be linked to previous data. Both cases lead to missing data in the panel data set: the so-called wave nonresponse or missing units.

Probably the simplest and safest matching method is to use participants' plain names. This, however, has the decisive disadvantage that the participants cannot be assured of the anonymity of their information, which can lead to refusals to participate, especially when sensitive content is being surveyed, as in the example of juvenile delinquency used here. In addition, the initial population of the reported study "Crime in the Modern City" (CrimoC) consisted of pupils aged 13 on average who attended a school in the city of Duisburg in 2002 (see Bentrup, 2019). Thus, a data protection concept also had to be developed due to the young age and the associated necessary declaration of consent by the parents. Together with the State Commissioner for Data Protection and Freedom of Information of North Rhine-Westphalia, it was decided to use self-generated personal codes that would allow the individual data to be combined while guaranteeing anonymity. This procedure was chosen for two interrelated reasons: first, to grant respondents the anonymity of their answers, and second, to make any possibility of de-anonymization by third parties impossible, since, violations relevant to criminal law were inquired about. These individual codes are self-generated by each respondent through responses to 6 to 10 targeted questions on time-stable characteristics (Pöge, 2008: 60). To ensure good reproducibility, letters from own name or the name of close relatives are often used. The goal is to obtain combinations that are as unique as possible. Over a total of 13 survey waves, this procedure proved to be a stable allocation procedure for most participants. Nevertheless, at each point in time, it was not possible to link a certain proportion of participations to the panel dataset. For this reason, the missing units were composed of individuals who either did not participate or did participate, but the individual data could not be matched to the panel data set using the self-generated code. It is precisely in this last case that the described data optimization comes into play. The panel consolidation describes a procedure with which missing units are subsequently replaced by originally collected data. The question that arises after such a time-consuming and challenging process whether the new data situation represents an improvement over the original panel.

Direct correspondence to
Christina Beckord, ehs University of Applied Sciences for Social Work, Education and Nursing, Dresden
E-mail: christina.beckord@ehs-dresden.de

While there are possibilities to address missing values at the statistical level (Rubin, 1987; Reinecke & Weins, 2013; Kleinke, Reinecke & Weins, 2021), even these methods have their limits. For this reason, the stated goal should be to integrate as many cases into a panel dataset as possible. For example, it is not possible to impute outstanding events that are not influenced by any predictors. One such example are typical stages of life such as starting an own family. For a sufficient subgroup analysis with longitudinal data as much cases of the data collection as possible should be included in the panel data. In addition to certain subgroups, an existing bias in the linkage by certain characteristics, e.g., gender, also poses a difficulty in interpreting the results. But does the consolidation call into question the quality of the previous panel data set? For this purpose, the main variable "juvenile delinquency" is examined in more detail below. If there are no changes in this characteristic in longitudinal analysis, this would indicate that the new cases compared to the already matched cases are at random regarding the dependent variable.

All in all, the described panel consolidation is considered a success if dropout from relevant subgroups can be minimized, biases in the panel data set can be reduced, and at the same time the structure of main dependent variables (here: juvenile delinquency) do not change from the original data set.-

Therefore, this paper begins with a description of (1) the starting point – the original CrimoC-data, the application and limitations of the self-generated codes and (2) the performed optimization of matching cases within the existing 13-wave panel data. Furthermore, it is (3) defined when the panel consolidation is considered successful and (4) what improvement could be achieved by the newly connected cases.

The Starting Point: "Crime in the Modern City" (CrimoC)

Crime in the Modern City (CrimoC) is a prospective panel study that began in 2002, surveying 7th grade pupils from public schools in the German city of Duisburg. The self-report questionnaire had the goal of explaining and monitoring the emergence and development of deviant and delinquent styles of behavior throughout the phase of adolescence (Sedding & Reinecke, 2017; Reinecke et al., 2015). As possible causes of these phenomena, the study focuses not only on structural conditions and processes on the macro-level but also on the meso- and micro-level (e.g., social milieus, moral orientation, lifestyle, how spare time is spent, attitudes, norm orientations, social environment; detailed information about the study can be obtained from the webpage www.crimoc.org; Boers et al., 2010; Boers & Reinecke, 2019). Due to the satisfying re-interview rates, and the successful panel construc-

tion, the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) has extended its funding in three-year intervals up to now.

Data Collection

The longitudinal self-report panel design evoked three major challenges: (1) respondents' retrieval after age-related school leaving despite the assurance of anonymous answers, (2) the necessity of different data collection modes, and (3) the matching of individual data over time by simultaneous assurance of response anonymity.

At the beginning of the CrimoC-study, the researchers attempted to collect data from all 7th graders in all public schools of Duisburg, an industrial city in the Rhine-Ruhr-Area with a long tradition in coal mining. In Germany, there are five different types of schools that follow elementary school: the *Hauptschule*, a school with a lower level of education which ends after grade 9, the *Realschule*, a medium-level school which ends after grade 10, the *Gymnasium*, the highest educational level which ended for our cohort after grade 131, the *Gesamtschule*, a combination of Realschule and Gymnasium which enables more pupils to achieve a higher educational level, and the *Förderschule* where pupils with learning disabilities receive special support. Of all 56 schools of Duisburg, 16 refused to participate. The progress of data collection was adjusted to the age and life stage of the respondents. From age 13 to 20, the survey was conducted annually, and from age 20 to 30, every two years (figure 1, in detail, see Bentrup, 2019).

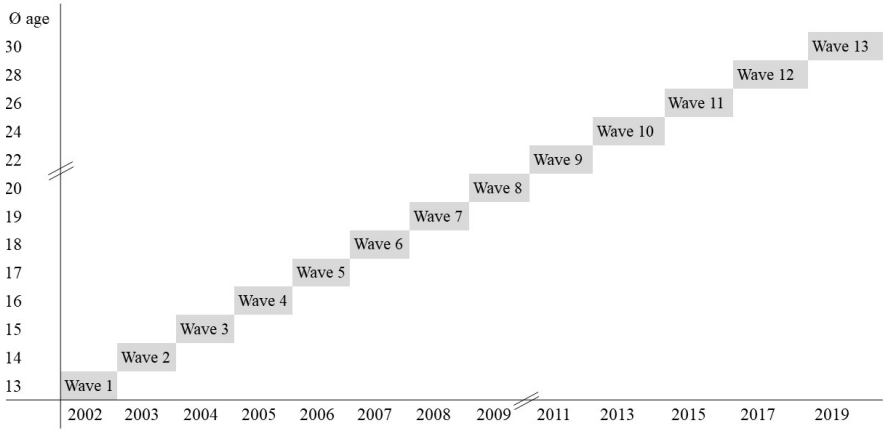


Figure 1 CrimoC survey design

1 Meanwhile the Gymnasium and Gesamtschule end after grade 12.

The first four waves took place in the school context with self-administered paper-pencil questionnaires, while the following waves of data collection were used for a stepwise change into postal mode. In order to contact the participants after leaving school, all respondents were asked to share their addresses (independent of the completed questionnaire). The resulting contact database was updated within each of the following data collections. If participants changed their residence, they had the possibility to communicate the new address to the project team via the project's webpage. Respondents who did not report their new residence were searched in local registers of residence. After every postal data collection, an additional personal contact phase was carried out for all contacts in the database who did not participate in the actual wave. This could be executed despite the assurance of anonymity because all participants filled out a separate address card to receive an incentive of 25 Euros for their participation. In this case, interviewers contacted respondents to motivate them to participate after all. If the respondents agreed, they were given a questionnaire by the interviewers (if necessary), which was to be completed without the presence of the interviewers and later collected again by the interviewers. In this way, 13 waves could be realized in 18 years.

The complex study design necessitates a closer look at the different datasets. First, one has to distinguish between different terms: the cross-sectional datasets for each time point t (CS_t), which include all individuals who filled out a questionnaire during a data collection wave. Second, there are the matched individual data – the 13-wave panel dataset that includes all cases with at least one match to another time point. The single cases it this panel dataset differ regarding the number of participation (independent of whether this missing unit is due to non-participation or not being matched to the dataset). The possible data range is between 2 and 13 points in time or in other words, the number of missing units varies from 0 to 11. The largest number of cases per time point is therefore obtained when all missing units are tolerated. This number of cases in the panel per time point (t) is referred to below as panel-cross-sectional data[2] (PCS_t). Additionally, there are the complete panel datasets, which contain only those respondents who have participated any time during the period of interest, and which could be successfully matched to the previous individual data ($P_{t1\text{-}t13}$). Fourth, one can use panel data sets with missing units, which include all cases with a tolerated number of missing units ($P_{txi,\ txj,\ldots,tX}$).

2 Even though strictly speaking it is the number of 2-wave panels from t to t+1 or t-1.

The study started in 2002 with a survey of the initial population of 3,411 7th-grade pupils in Duisburg. In the following years, the cross-sectional re-interviewing rates ranged between 85 and 92%[3].-

Previous Matching of Individual Data Over Time

In order to enable the questionnaires from the different survey waves to be assigned, individual codes were used which were requested via code sheets. In the course of the interview, each respondent filled out a code sheet containing five or - from the 2003 survey onward - six personal questions, the answer to each of which represented a specific letter or number and was to be noted down accordingly. The questions referred to unchangeable characteristics of the respondent or his environment (natural hair color, name of father, etc.). This letter-number combination finally formed the entire code. In each survey wave, the code was filled in by the participants at the beginning of the survey. Since the codes in each survey contained the same information, the codes filled out by the same person in the different waves should have to be identical.

The questions to create the code included:

Co001: The first letter of the father's first name

Co002: The first letter of the mother's first name

Co003: The first letter of your first name

Co004: The two-day digits of your own birthday

Co005: The last letter of the own hair color

Co006: The last letter of your own eye color

Since 2009 additional:

Co011: The last letter of own surname (in case of name change, the birth name)

Since the survey year 2003, the following questions have also been asked:

Co007: Survey participation in the previous year (yes/no)

Co008: Change of school in the past year (yes/no)

Co009: Not transferred in the past year (yes/no)

---

3 In detail: 2003 n = 3,392; 2004 n = 3,339; 2005 n = 3,243; 2006 n = 4,548; 2007 n = 3,336; 2008 n = 3,086; 2009 n = 3,090; 2011 n = 3,050; 2013 n = 2,850; 2015 n = 2,754, 2017 n= 2,778; 2019 n= 2,697. The data collection in the year 2006 was the most challenging one. Due to the school leave of respondents in the lower educational level schools and the compulsory school attendance for all adolescents up to age 18, the attempt was made to retrieve these school leavers in selected classes at vocational schools. A consequence was that the cross-sectional data includes additional cases of individuals who attended these classes but who did not participate before. These additional cases leave no impact on the panel-dataset because they could not be matched to previous cases.

Co004, Co007, Co008 and Co009 have been included in the code sheet since the year 2003. In addition, the design of the code sheet has been changed. In 2002, respondents had to provide their respective answers to the code questions in a box in handwriting; since 2003, all possible letters have been shown as answer options to be checked off (see appendix A).

Since the fifth wave of the survey (2006), Co008 has not been collected due to the end of the school career of most respondents. Co009 has been collected since 2006 only for those respondents who attended a Gymnasium or a Gesamtschule. Since the eighth wave of the survey (2009), only survey participation in the previous year (Co009) was asked. In addition to the six code questions and the supplementary questions, information on the respondent's gender and the (most recent) school attended was available for the questionnaire assignments.

The function of the code requires that the codes are a) unique, i.e., that the individual parameters have enough variance so that the codes can be uniquely identified (identification), b) that the participants answer the individual code questions exactly the same over time (replication), and c) that the queried characteristics are indeed time-invariant characteristics.

In 2002, the problem of identifying individual data over time was posed by multiple occurrences of the same complete codes. By adding one code question (the last letter of one's first name), the uniqueness of the code could be greatly increased. In 2002, there were 324 double occurrences of the complete code (7.9%), 18 triple occurrences (0.5%), and 5 quintuple occurrences (0.1%); in 2003, the six-digit code had only 32 double occurrences (0.9%) (cf. Pöge, 2007: 6; Pöge, 2008: 62). This figure remained between 2.0% in 2006 and 0.1% in 2009 across all subsequent survey waves. In principle, respondents were willing to fill in the code with an overwhelming majority (98.5% in 2005 to 99.6% in 2006).

However, the problem of replicating the individual codes remained. For this reason, an error-tolerant matching procedure was developed in which gradually more and more errors in the code were allowed (cf. Pöge, 2005: 66). To provide additional certainty about the matching, each potentially matching questionnaire from two points in time was subjected to a manual handwriting check.

The steps of the error-tolerant matching procedure are hierachical and allow more variation in the code with each step. Accordingly, the assignment rate decreases with each additional step (table 2). Each step consisted of two sub-steps to keep the number of reconciliations to be performed manageable: first, gender and school attended had to be compared in addition to the codes (with the number of errors allowed in each case). Moreover, students were matched on the basis of additional variables (Co007, Co008, Co009), which asked whether they had participated in the survey in the last year, as well as whether they had changed schools or stayed behind. Second, the additional conditions to be fulfilled were successively relaxed and in some cases omitted altogether. In this way, there were controls for

0-3 errors in the code and the release of the control variables, so the basic structure was a 4*2 pattern.

After each step, the matched questionnaires were then subjected to a manual handwriting check. This check was performed for several reasons: on the one hand, there was the possibility that individual codes were not unique (especially when tolerating errors) On the other hand, it was an additional control on the basis of the handwriting style and/ or similarities in the content in the questionnaire. Those pairs of questionnaires that had obviously been completed by the same person were removed from the data sets so that they were no longer available for the subsequent matching steps. Non-matching questionnaires remained in the data sets, possibly to be identified as matching in one of the next matching steps.

The 2007 and 2008 data collections will serve as an example of the chosen approach (table 1). The respective cross-sectional data comprised n=3.336 in 2007; n=3.086 in 2008 (Daniel & Erdmann, 2017: 8).[4]

It can be seen that the number of comparisons increases as the error tolerance increases, whereas the assignment rate decreases. A total of 4,407 potentially matching pairs of questionnaires were identified, of which 2,698 (61.2%) were found to be matches during the handwriting checks. In terms of the cross-sectional data set of 2008, this means that of the 3,086 cases available, 2,698 (87.4%) could be linked to the cross-sectional data of the previous wave.

In addition, controls were also conducted between survey waves that were not directly consecutive (figure 2). The first four survey waves were fully matched. For economic reasons, starting with the fifth survey wave in 2006, the codes of a cross-sectional data set, which had not yet been assigned to the panel after the matching with the immediately preceding wave described above, were compared with the unassigned codes from the penultimate wave.

Between these data, in a first step in which the condition of fully matching codes and fully matching additional variables were checked, 1,403 potentially matching pairs of questionnaires were identified. 1,343 (95.7%) of these were found to be matches in the subsequent handwriting checks. These were marked as matches and removed from the two cross-sections for further matching. The control steps shown in table 2 followed in order.

Since the matching was based on the cross-sectional data, these cases were matched to the existing panel data set in a next step. This again reduced the number of cases, so that in the previous example, the original PCS (oPCS) for 2008 included a total of 2,412 cases (Erdmann, 2021).

4 The original table was summarized to the 4*2 steps described above for illustrative purposes, even though a total of 10 steps were performed in the matching process to keep the size of each list to be compared manageable.

Table 1 Performed control steps 2007/2008

Step	Codevariables	Additional variables
S1	without errors	without errors
S2	without errors	no restriction
S3	one error	without errors
S4	one error	no restriction
S5	two errors	without errors
S6	two errors	only selected restricitons
S7	three errors	without errors
S8	three errors	only selected restrections

Table 2 Number of checks and matches

		Number of checks	Match		No match	
Errors	Step	n	n	%	n	%
Without errors	1	1,403	1,343	95.7	60	4.3
	2	584	506	86.6	78	13.4
One error	3	415	370	89.2	45	10.8
	4	371	258	69.5	113	30.5
Two errors	5	138	104	75.4	32	24.6
	6	1,190	89	7.5	1,101	92.5
Three errors	7	194	24	12.4	170	87.6
	8	112	4	3.6	108	96.4
Total	1-8	4,407	2,698	61.2	1,709	38.8

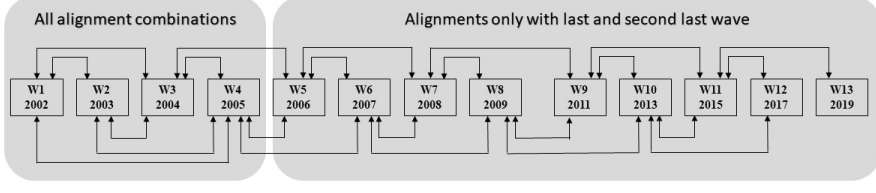


Figure 2 Matches performed as part of the original panel construction

Discrepancies Between Cross-sectional and Panel Data

Consequently, as was made clear in the previous section, there is a discrepancy in the number of cases between CS_t and $oPCS_t$. Table 3 shows the differences in the number of cases between the cross-sections and the associated panel cross-sections, as well as the differences between CS_t and $oPCS_t$. Two things become clear: The first four waves, which were fully matched, show the best assignment rate. The increased difference between CS_t and $oPCS_t$ in the first survey are due to the shorter code, the lack of additional questions, and the more difficult layout of the code query (see previous section). All other data collections show a much larger difference between CS_t and $oPCS_t$. Ideally, this drop out is at random, i.e., does not exhibit systematic failures.

In summary, it can be stated at this stage that since the sixth survey wave in 2006, between 21.8% and 38.1% of the participating individuals could not be assigned to the panel. However, since contact data are available from all individuals to ensure the postal survey, it should theoretically be possible to assign them to the panel data set.

In addition to the loss of cases, the linkage to the panel exhibits additional biases. In the earlier waves (w1-w4), these relate to the type of school. High school students are more strongly represented in the panel than in the cross-section. For all waves, there is a clear bias with respect to the gender of the respondents; female participants are significantly overrepresented in the panel data set (counts are in table 8).

Finally, due to the general loss of cases, some subgroups of special interest were significantly reduced. This reduction becomes more pronounced the later the point in time considered for the identification of a subgroup in the dataset (e.g. parenthood). For instance, in 2011, 168 of the respondents reported having at least one child of their own. Of these, however, only 106 were found in the oPCS of the year 2011. For this reason, a pretest in 2011 attempted to link additional parents to the panel by performing code matches for survey periods more than two time points apart. The developed method turned out to be surprisingly successful. Further matching increased the number of parents in the consolidated PCS_t ($cPCS_t$) by 48 cases to a total of 154, representing 92% of the parents in the CS. Due to its success, it was decided to apply this procedure to all cases of the full panel. The procedure is explained in the next section.

Table 3 Case numbers of data sets in cross sections, panel cross sections and their difference

	2002	2003	2004	2005	2006	2007	2008	2009	2011	2013	2015	2017	2019
CS_t	3,411	3,392	3,339	3,243	4,548	3,336	3,086	3,090	3,050	2,849	2,754	2,778	2,697
$oPCS_t$	2,750	3,132	3,177	3,195	3,032	2,587	2,412	2,304	2,100	1,912	1,812	1,760	1,670
$oDiff_t$	661	260	162	48	1,516	1,026	674	786	950	937	942	1,018	1,027
%	19.4	7.7	4.9	1.5	33.3	30.8	21.8	25.4	29.7	32.9	34.2	36.6	38.1

CS_t: cross-sectional data for time point t; $oPCS_t$: original panel data for time point t; $oDiff_t$: CS_t-$oPCS_t$; %: $oDiff_t/(CS_t/100)$

Improvement in Assignments – The Panel Consolidation

First of all, it should be summarized to which criteria the improvement of data quality should be determined. Four criteria are applied in this paper:

1. *Increase in the number of cases per time point in the panel*: the initial aim is to replace as much missing units as possible through additional matching with aubsequent allocations. This means that the consolidated PCS_t ($cPCS_t$) have a larger number of cases than the corresponding $oPCS_t$.
2. *Reduction of socio demographic bias:* since all cases that have not yet been allocated originate from the CrimoC-population, the overrepresentation of females should decrease within the consolidation, since more data from male respondents should be matched.
3. *Improving the number of cases of relevant subgroups:* As parents are an important subgroup for a follow-up project, the difference in the number of cases between cross-sectional and panel data should be reduced.
4. *No changes in the structure of the dependent variable:* the longitudinal structure of the main dependent variable (juvenile delinquency) should not change significantly. Otherwise this would be an indicator for a relevant bias in the previous panel construction and related to this in the interpretation of previous results.

The Consolidation Procedure

The procedure was analogous to the original panel construction. First, the cases of the cross sections were selected that could so far not be matched to the panel ($oDiff_t$ in table 3). For each of the 29 potential additional checks listed in table 4, the cases of the oPCSt were selected that so far have a missing unit for the wave of interest. For example, the 2007 cross-section was reduced to those cases that were not previously part of the 2007 panel cross-section. For the matching with the 2004 panel wave, the 2004 panel cross-section was reduced by the cases that already had a link to 2007. For the resulting two partial data sets, SQL queries were run in *Access* to identify identical codes or, in the context of the error-tolerant procedure, the corresponding potential matches.

Because these subsamples were considerably smaller than had been the case in previous panel checks, matching was performed in two steps: Step 1 included all cases with identical codes for each match, and the additional variables were not equated. This corresponds to S2 of the original panel controls (table 1).

Table 4 Potential for panel consolidation

Survey year	Already performed checks	Potential further checks
2006	2005, 2004	2003
2007	2006, 2005	2004, 2003
2008	2007, 2006	2005,2004, 2003
2009	2008, 2007	2006, 2005, 2004, 2003
2011	2009, 2008	2007, 2006, 2005[1)]
2013	2011, 2009	2008, 2007, 2006 2005
2015	2013, 2011	2009, 2008, 2007, 2006
2017	2015, 2013	2011, 2009, 2008, 2007
2019	2017, 2015	2013, 2011, 2009, 2008

1) In the first comparisons, it turned out that the complete comparison of waves 1 to 5 already performed meant that further checks in these waves for later points in time were not very successful. For this reason, an additional comparison with 2004 was not performed in 2011.

In the second step, one error was tolerated in the code, and the additional variables remained unrestricted (S4 of the original panel controls). Further checks were deliberately omitted because manual handwriting comparison, which became of increasing importance especially for assignments with more than one tolerated error, becomes increasingly difficult over a greater temporal distance.

The number of reconciliations is summarized for each survey wave in Table 5; a detailed list of all reconciliations per survey wave can be found in appendix B. A total of 7,068 potential matches were checked, of which 3,589 (50.78%) resulted in new matches in the existing panel dataset. It is important to note here that the aim was not to link new cases to the dataset, but to fill gaps (in the form of missing units) through subsequent checks, i.e., the total number of cases before and after panel consolidation is identical at 4,076 cases (last row table 5). The table also illustrates that the number of matches, as well as the assignments found, increased with distance from the starting point of the study, the fully controlled five-wave panel. Appendix C illustrates two typical cases of the consolidated complete panel data set. A detailed documentation of the occurring errors by code question does not exist, as the queries since 2003 have been carried out and documented by number of errors, but not broken down by code question.

Table 5 Panel consolidation checks and matches

Aligned wave	Number of checks	New matches	Matches (%)	$oPCS_t$ (n)	$cPCS_t$ (n)	Increase (%)
t_5 2006	169	5	2.96	3,032	3,037	0.16
t_6 2007	428	123	28.74	2,587	2,710	4.75
t_7 2008	665	333	50.08	2,412	2,745	13.81
t_8 2009	815	459	56.32	2,304	2,763	19.92
t_9 2011	775	468	60.39	1,812	2,324	28.26
t_{10} 2013	976	480	49.12	1,912	2,392	25.10
t_{11} 2015	1,115	512	45.92	1,812	2,324	28.26
t_{12} 2017	1,058	597	56.43	1,760	2,357	33.92
t_{13} 2019	1,067	612	57.36	1,670	2,282	36.65
total	7,068	3,589	50.78	4,076	4,076	100.00

$oPCS_t$= original panel cross-sectional data set; $cPCS_t$= consolidated panel cross-sectional data set; Matches (%): new matches/ (number of checks/100); Increase (%) = $cPCS_t$/ ($oPCS_t$/100).

The 3,589 new matches are distributed among 1,071 participants, for whom one missing unit could be filled in 259 cases, two in 195 cases, three in 149 cases, four in 169 cases, five in 102 cases, six in 98 cases, seven in 73 cases, and eight original missing units could be replaced in 26 cases.

The greatest improvement was achieved for panel data sets with four to six missing units. Here, panel consolidation increased the number of cases by more than 500. But also the panel data sets with fewer missing units could be increased considerably. The 79 closed gaps for the continuous panel (first row table 6) are astonishing because, actually, comparisons were always carried out between three consecutive survey dates. Thus, a complete control was available for these cases. This may be due to three reasons: 1) in the handwriting control, a case was originally declared as non-matching but now declared as a match; 2) in the handwriting control, a questionnaire could not be found; 3) more than one gap was closed for some cases, so that there may be an increase in the number of cases for the continuous panel. The first possibility applies to 14 of the 79 new cases in the continuous panel dataset, and the second reason is crucial for 65 of the 79 cases: two missing units were filled with data for five of the cases, three missing units for one case, four missing units for nine cases, five missing units for 15 cases, seven missing units for 13 cases, and eight missing units for 16 cases. These cases were randomly tested for plausibility of assignment.

Table 6 Number of cases of original and consolidated panel data set by missing units

	oPCS		cPCS		Increase
Missing units	n	%	n	%	n
0	735	18.0	814	20.0	79
0-1	1,230	30.2	1,404	34.4	174
0-2	1,542	37.8	1,834	45.0	292
0-3	1,749	42.9	2,161	53.0	412
0-4	1,965	48.2	2,466	60.5	501
0-5	2,143	52.6	2,647	64.9	504
0-6	2,316	56.8	2,835	69.6	519
0-7	2,497	61.3	2,983	73.2	486
0-8	2,815	69.0	3,145	77.2	330
0-9	3,163	77.6	3,376	82.8	213
0-10	3,550	87.1	3,629	89.0	79
0-11	4,062	99.7	4,063	99.7	1
0-12*	4,076	100.0	4,076	100.0	0

% oPCS= n_{oPSC}/(4,076/100); % cPCS= n_{cPCS}/(4,076/100).
* 12 missing units are 14 (oPC_t) and 13 (cPC_t) cases, respectively, which were assigned to another time point, but the second case was classified as not qualitatively usable.

With regard to the first criterion for the improvement of data quality in the panel dataset - i*ncrease in the number of cases per time point in the panel* - it can be summarized that the number of cases in the cPCSt increased significantly compared to the oPCSt at all points in time. The later the time of the survey and thus the more additional comparisons were possible, the more missing units could be filled with empirical information.

Improvements in Content Due to the New Assignments

Following the encouraging results of the panel consolidation, the question arises as to its significance for the data structure. Based on the cross-sectional data, the quality of the assignments before and after panel consolidation can be assessed in terms of content to examine the quality criteria 2 to 4.

Examination of the Quality Criteria at the Content Level

Reduction of socio demographic bias: Table 7 illustrates the gender differences between the CS and oPCS. Within the oPCS, all time points are characterized by a higher proportion of female participants. If the panel consolidation meets the quality criterion, the difference between the proportion of females between the consolidated panel and the cross-sectional data should be smaller than between the original panel dataset and the cross-sectional data (cDiff % < oDiff %). Although the proportion is still higher than in the cross-sectional data all points in time of the consolidated panel meet this criterion.

Table 7 Gender differences between cross-sectional and panel data before and after panel consolidation

		Gender (% female)								
		17	18	19	20	22	24	26	28	30
Data		t5	t6	t7	t8	t9	t10	t11	t12	t13
CS		50.2	53.0	53.0	53.2	53.2	54.3	54.5	54.3	54.1
oPCS		54.3	56.8	56.6	57.8	59.1	60.9	58.6	61.8	62.3
oDiff	%	4.1	3.8	6.6	4.6	5.9	5.6	4.1	6.5	8.2
cPCS		54.2	56.4	54.9	54.8	56.0	57.5	58.1	57.1	58.0
cDiff.	%	4.0	3.4	4.9	1.6	2.8	3.2	3.6	2.8	3.9

oDiff. %= $\%oPCS_t$;- $\%CS_t$; cDiff %= $\%cPCS_t$ - $\%CS_t$

Improving the number of cases of relevant subgroups: The development of parents in the CrimoC-data is displayed in table 8. The number from the respective cross-section data serves as the reference category. The number of parents from the original panel and the consolidated panel are compared with this. The criterion is considered fulfilled if the proportion of parents in the consolidated data set is higher than that of the original data set.

Table 8 Development cases parents between cross-sectional and panel data before and after panel consolidation

		Number of parents				
Data		22 t9	24 t10	26 t11	28 t12	30 t13
CS	n	168	286	490	732	1.004
oPCS	n	106	153	260	392	540
% CS		63.1	53.5	53.1	53.6	53.8
cPCS	n	154	214	386	590	823
% CS		91.7	74.8	78.8	80.6	82.0

% CS= Percentage of cases in relation to the cross-section.

The original panel data set includes only about half of the parents from the cross-sectional data at four of the five points in time shown in the table. Through panel consolidation, the proportion of parents could be drastically increased to 75-82%. In figures, this means, for example, that in t11 126 parents could be subsequently matched, in t13 even 283. Criterion 3 is thus fulfilled.

No changes in the structure of the dependent variable: In the present criminological study, the extent of delinquent behavior is of particular importance. This can be operationalized in two different ways per survey time: A sum index of the annual prevalence rates over the queried 15 offenses (Have you committed the offense in the last 12 months?). This so-called *versatility score* thus has a range of values from 0 to 15. 0 means that an individual has committed none of the offenses, 15 means that an individual has committed all of the offenses queried, while the values in between indicate the respective number of types of offense committed. Strictly speaking, this score measures the number of different types of offense committed. The second possibility is a sum index of the *incidence rates* for each survey time. The incidence corresponds to the frequency of offenses committed within the last 12 months.

However, this sum score is very susceptible to extreme values. For this reason, criminology usually uses the versatility score for complex models, which has proven to be a comparable, less distributionally skewed alternative to the incidence rates (Sweeten, 2012). For both scores, mean values can be found for the different survey waves in table 9. As can be seen, these two variables do not deviate significantly from each other between the two panel data sets, with the mean values of the incidence rates showing somewhat greater deviations than the versatility scores.

Table 9 Versatility scores and incidence rates per time point before and after panel consolidation[5]

Data	Versatility score per time point (and average age)								
	17 t5	18 t6	19 t7	20 t8	22 t9	24 t10	26 t11	28 t12	30 t13
CS_t	0.48	0.27	0.15	0.13	0.10	0.08	0.07	0.06	0.04
oPCS	0.44	0.25	0.14	0.11	0.08	0.06	0.06	0.05	0.04
cPCS	0.44	0.24	0.15	0.12	0.09	0.08	0.06	0.06	0.04
	Incidence rates per time point (and average age)								
CS_t	4.67	4.40	2.50	1.80	0.74	0.57	0.38	0.32	0.31
oPCS	4.82	4.57	2.09	1.52	0.62	0.31	0.28	0.32	0.22
cPCS	4.82	4.51	2.10	1.42	0.66	0.50	0.34	0.33	0.28

Overall, the descriptive results of both panel data sets appear comparable. On the content level, both data sets lead to the same results. Criterion 4 seems to be fulfilled but in longitudinal criminological research, the development of juvenile delinquency is often described using complex trajectories. These are mostly based on *Latent Class Growth Analyses* (LCGA) or on *Growth Mixture Models* (GMM) (Nagin & Land 1993; Vermunt & Magidson, 2004; Muthén, 2004). Using the previously reported versatility score, LCGAs will be calculated for the original and the consolidated panel for two different age periods. Missing values were accounted for using the *full information maximum likelihood estimator* (FIML). In order to check the comparability of both data sets (original versus consolidated panel), two LCGAs are calculated. The first covers age 13 to 19, thus also including the first four waves that were not affected by the consolidation. All cases with a maximum of one missing participation were included in this analysis (original n= 1,907; consolidated n= 2,051). Since the description of the consolidation could show that more missing units could be filled with data at later points in time, another model will be calculated for age 20-30 and up to two missing participations will be tolerated here (original n= 1,865; consolidated n= 2,419). Since the comparability of the results is the focus of this paper, the detailed description of the modelling is omitted (the necessary information can be found in appendix D). Instead, the class solutions found for the original and the consolidated panel are cross-tabulated. The

5 The tables are always described only from the 5th wave onwards, since the first five waves were already fully matched against each other as part of the original panel construction.

quality criterion is still considered to be fulfilled if the class solutions found for the individual cases do not deviate significantly from each other.

At the beginning, all data sets were tested to determine which distributional assumption best fits the data. Due to the fact that a large number of respondents indicated that they had not committed any crime, the versatility score shows many zeros. It was found that the *negative binomial distribution* assumption best fit the highly right-skewed data. A zero-inflated model did not lead to a substantial improvement of model fit.

For age 13 to 19, both models reach a five-class solution. As expected, the model fit values are higher for the consolidated data set due to the higher number of cases.

The five classes found describe typical developmental patterns of delinquent behavior during youth (table 10). The class of *non-offenders* is characterized by the reporting of no or only very isolated offenses. The *Adolescent limited* class shows higher mean versatility scores in early adolescence but commits fewer and fewer offenses with increasing age. The *early desistance* class shows high delinquency scores at the start of adolescence that steadily decrease with age. Compared to the other groups, the *late onset* group shows its highest delinquency levels later, at age 16. The *persistent* class shows the highest burden of delinquency across all waves, although a decline toward young adulthood is also observed for this group. These patterns are found in both the original and consolidated panel data sets. The proportion of cases attributed to a particular class varies only marginally by a maximum of one percent between the data sets, i.e., the consolidated data set can be considered comparable at the content level even in the case of the LCGA for the juveniles.

Based on the variance and co-variance structure of both data sets the latent classes are estimated quite similar. This is reflected in the fact that in Table 11 the diagonal of the crosstab has the highest numbers. 1,093 of the total of 1,096 non-offenders in the original classification are also assigned to this class in the consolidated data set. In total, only 66 of the original 1,907 cases (=3.4%) were assigned to a different class within the consolidated data set, which indicates a stable class solution.

But what happens in the later waves under the acceptance of more missing units? For this purpose, 1,865 cases of the original panel data set and 2,419 cases of the consolidated panel for the age group 20-30 years were conducted with a maximum of two missing units. Both data sets differ by more than 500 cases.

Table 10 Comparison of the versatility score mean values for each class and age for the original and consolidated panel

	Age						
Class	13	14	15	16	17	18	19
Non-offenders							
Original (57%, n=1,096)	0.07	0.06	0.05	0.04	0.03	0.03	0.02
Consolidated (58%, n=1189)	0.07	0.06	0.05	0.04	0.03	0.03	0.02
Adolescent limited							
Original (15%, n=280)	0.60	0.77	0.57	0.24	0.06	0.01	0.00
Consolidated (12%, n=247)	0.61	0.82	0.59	0.23	0.05	0.01	0.00
Early desistance							
Original (10%, n=198)	2.47	2.46	1.83	1.02	0.42	0.13	0.03
Consolidated (12%, n=241)	2.27	2.29	1.74	1.00	0.43	0.14	0.04
Late onset							
Original (11%, n=213)	0.27	0.56	0.89	1.10	1.03	0.75	0.41
Consolidated (12%, n=241)	0.23	0.50	0.82	1.03	0.98	0.71	0.39
Persistant							
Original (6%, n=120)	3.13	3.74	3.85	3.41	2.60	1.70	0.96
Consolidated (6%, n=133)	3.14	3.81	3.94	3.48	2.63	1.69	0.93

n and % based on the most likely latent class membership

Table 11 Cross-tabulation class solution original and consolidated panel age 13 to 19

	Consolidated classification*					
Original Classification*	Non-offenders	Adolescent limited	Early desistance	Late onset	Persistant	total
Non-offenders	1,093	0	0	3	0	1,096
Adol. limited	24	229	12	15	0	280
Early des.	0	0	197	0	1	198
Late onset	0	0	9	202	2	213
Persistant	0	0	0	0	120	120
Not matched	*72*	*18*	*23*	*21*	*10*	*144*
total	1,189	247	241	241	133	2,051

* n based on the most likely class membership, $\chi^2 = .00068$, $p < .001$

Table 12 Comparison of the versatility score mean values for each class and age for the original and consolidated panel age 20 to 30

	Age					
Class	20	22	24	26	28	30
Non-offenders						
Original (88%, n=1,634)	0.02	0.01	0.01	0.00	0.00	0.00
Consolidated (88%, n=2,130)	0.02	0.01	0.01	0.01	0.01	0.00
Adult onset						
Original (9%, n=165)	0.14	0.16	0.18	0.19	0.20	0.20
Consolidated (8%, n=183)	0.14	0.20	0.25	0.28	0.30	0.28
Late desistance						
Original (2%, n=45)	0.87	0.61	0.30	0.11	0.03	0.00
Consolidated (3%, n=75)	0.93	0.66	0.34	0.12	0.03	0.01
Persistent						
Original (1%, n=21)	1.39	1.52	1.50	1.34	1.09	0.79
Consolidated (1%, n=31)	2.05	1.93	1.70	1.40	1.07	0.77

n and % based on the most likely latent class membership

The class solution (table 12) consists of the *non-offenders*, (individuals who, compared to their peers, do not start committing offenses until adulthood (*adult onset*), individuals who do not stop committing offenses in adolescence but in young adulthood (*late desistance*)), and the *persistent offenders*, who commit a comparatively large number of offenses even in adulthood. The percentages of participants in the groups are comparable. Overall, less delinquency was reported for this age range.

The final cross-tabulation of both most likely class memberships leads to a stable class solution, as for adolescence (table 13). Only 42 cases of the original classification were assigned to other classes, the number of cases of the diagonal shows the highest values.

Overall, a satisfactory stability and thus comparability of the data sets with respect to the analysis of developmental trajectories can thus be observed.

Table 13 Cross-tabulation class solution original and consolidated panel age 20 to 30

	Consolidated classification*				
Original Classification*	Non-offenders	Adult onset	Late desistance	Persistant	total
Non-offenders	1,634	0	0	0	1,634
Adult onset	35	127	3	0	165
Late des.	0	0	45	0	45
Persistant	0	2	2	17	21
Not matched	*461*	*54*	*25*	*14*	*554*
total	2,130	183	75	31	2,419

* n based on the most likely class membership, $\chi^2 = .00046$, $p<.001$

Discussion

In this paper, the difficulties of missing units in the construction of panel data with self-generated individual codes in the context of anonymous surveys were discussed. Self-generated codes offer the advantage of assuring anonymity to survey participants. At the same time, they have the disadvantage that they only work if the respondents generate the code identically at all times. If no current code of a new case can be assigned to a case in the data set during the panel construction, a missing unit is created. For time and economic reasons, the previous comparisons of the reported 13-wave panel in the past, except for the first four waves, only took place between a current survey and the two previous surveys. It was shown that although this procedure resulted in a usable panel data set, there were still numerous cases that could not previously be assigned to the panel. With the help of so-called panel consolidation, a procedure in which additional comparisons were made with surveys conducted further apart in time, the quality of the previous data was to be increased. Four criteria were used to assess the quality of the consolidated data set: The number of additional cases or the number of reduced missing units, the reduction of socio-demographic bias, improvement of relevant subgroups and stability of the dependent variable (juvenile delinquency).

Panel consolidation allowed 3,589 missing units in the data set to be replaced with empirical data. This is accompanied by a considerable increase in the number of cases in possible subdata sets. This increase is smaller for data sets without acceptance of missing units, but is greater if missing units are also tolerated in the consolidated data set (table 6). It was also shown that the gender bias could be

reduced across all time points (table 7), and that the cases of the subgroup of parents can be increased enormously (table 8; for example, the number of parents in the 2019 panel cross-section could be increased from 540 to 823 (+34.4%)).

In order to be able to classify the analyses carried out so far on the basis of the original panel and their interpretation in comparison to the consolidated data, the central dependent variable was examined as the last criterion for assessing the impact of the consolidation. It was assumed that there was no systematic bias due to the original panel construction if the consolidation data showed comparable results with regard to this variable.

Both, the descriptive analysis and the longitudinal modelling of LCGAs, lead to the result that both data sets do not differ significantly with regard to the outcome for the dependent variable *juvenile delinquency.* However, the panel consolidation could reduce existing biases and optimize the starting point for subgroup analysis.

The limits of self-generated codes are clearly to be named in their susceptibility to error. Some respondents do not answer identically over time, even to questions on selected, time-stable characteristics. Therefore, it is necessary to design the procedure to be error-tolerant.

Overall, however, this code procedure represents a method of guaranteeing anonymity that is comprehensible to participants.

However, this does not mean that panel consolidation was not necessary. Although the process was very time-consuming and personnel-intensive, numerous missing units could be replaced by empirical information. This automatically also means that data imputation techniques can fall back on a more secure basis. Furthermore, panel consolidation helps to increase the number of cases for subgroup analyses.

References

Bentrup, C. (2020a). The Dual Trajectory Approach. Detecting Developmental Behavioural Overlaps in longitudinal and intergenerational research. *Quality & Quantity* 54(1): 43-65. doi: 10.1007/s11135-019-00934-1

Bentrup, C. (2020b). Gewaltsame Erziehung und ihre Folgen im Altersverlauf. *Monatsschrift für Kriminologie und Strafrechtsreform,* 103(2): 97-120. doi: 10.1515/mks-2020-2042

Bentrup, C. (2019). Untersuchungsdesign und Stichproben der Duisburger Kriminalitätsbefragung. In K. Boers, & J. Reinecke (eds.), *Delinquenz im Altersverlauf. Erkenntnisse der Langzeitstudie Kriminalität in der modernen Stadt* (pp. 95-120). Münster, New York: Waxmann.

Bentrup, C. (2018). First Results of Cross-Generational (Dis-)Similarities Between Three CrimoC Generations. The Relationship Between Experienced Violent Parenting Practice, Delinquency and Own Parenting Style. In V. Eichelsheim, & S. van de Weijer (eds.), *Intergenerational Continuity of Criminal and Antisocial Behaviour. An International Overview of Studies (pp. 235-259).* London & New York: Routledge.

Boers, K., & Reinecke, J. (eds.) (2019). *Delinquenz im Altersverlauf. Erkenntnisse der Langzeitstudie Kriminalität in der modernen Stadt*. Münster; New York: Waxmann.

Boers, K., Reinecke, J., Mariotti, L., & Seddig, D. (2010). Explaining the Development of Adolescent Violent Delinquency. *European Journal of Criminology*, 7(6), 499-520. doi: 10.1177/1477370810376572

Daniel, A., & Erdmann, A. (2017). *Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2013, Zehn-Wellen-Panel. Schriftenreihe: Jugendkriminalität in der modernen Stadt - Methoden Nr. 23,* Münster, Bielefeld.

Erdmann, A. (2021). *Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2019 – Dreizehn-Wellen-Panel. Schriftenreihe Kriminalität in der modernen Stadt – Methoden, Heft 27.* Münster, Bielefeld.

Kleinke, K.; Reinecke, J.; Salfrán, D., & Spiess, M. (2020). *Applied Multiple Imputation. Advantages, Pitfalls, New Developments and Applications in R*. Wiesbaden: Springer VS.

Kleinke, K., Reinecke, J., & Weins, C. (2021). The Development of Delinquency During Adolescence: A Comparison of Missing Data Techniques Revisited. *Quality & Quantity* 55(3), 877-895. doi: 10.1007/s11135-020-01030-5

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of Quantitative Methodology for the Social Sciences (pp. 345–368).* Newbury Park, CA: Sage Publications.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology,* 31, 327–362. doi: 10.1111/j.1745-9125.1993.tb01133.x

Pöge, A. (2008). Persönliche Codes „reloaded“. *Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung*, 2 (1), 59-70.

Pöge, A. (2007). *Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002-2005, Vier-Wellen-Panel.* Schriftenreihe: Jugendkriminalität in der modernen Stadt- Methoden Nr. 13. Münster, Bielefeld.

Pöge, A. (2005). Persönliche Codes bei Längsschnittstudien. Ein Erfahrungsbericht. *ZA-Information*, 56, 50-69.

Reinecke, J.; Meyer, M., & Boers, K. (2015). Stage-Sequential Growth Mixture Modeling of Criminological Panel Data. In M. Stemmler, A. von Eye, & W. Wiedermann (eds.), *Dependent Data in Social Science Research (pp. 67-89).* Wiesbaden: Springer VS.

Reinecke, J., & Weins, C. (2013). The development of delinquency during adolescence: a comparison of missing data techniques. *Quality & Quantity*, *47*(6), 3319-3334. doi: 10.1007/s11135-012-9721-4

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Seddig, D., & Reinecke, J. (2017). Exploration and Explanation of Adolescent Self-Reported Delinquency Trajectories in the Crimoc Study. In A. Blokland, & V. van der Geest (eds.), *The Routledge International Handbook of Life-Course Criminology (pp. 159-178).* London: Taylor & Francis.

Sweeten, G. (2012). Scaling criminal offending. *Journal of Quantitative Criminology*, *28*(3), 533-557. doi: 10.1007/s10940-011-9160-8

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 2, 549-553.

Appendix

A The query for creating the individual code

Wenn du eine der Fragen überhaupt nicht beantworten kannst, kreuze bitte kein Feld an!

Hier nun die sechs Fragen zur Erstellung deines persönlichen Codes:

1 Bitte kreuze den **ersten** Buchstaben des Vornamens deines **Vaters** (oder einer Person, die für dich einem Vater am nächsten kommt) an. (z. B. Anton, Bernd, Hans-Peter usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

2 Bitte kreuze den **ersten** Buchstaben des Vornamens deiner **Mutter** (oder einer Person, die für dich einer Mutter am nächsten kommt) an. (z. B. Anna, Beate, Jutta, Maria, usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

3 Bitte kreuze den **ersten** Buchstaben deines **Vornamens** an (z. B. Michael, Thomas, Ute usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

4 Bitte kreuze den **Tag** deines **Geburtsdatums** an (z.B. Geburtstag am 7. Januar = 7, am 12. Mai = 12, am 31. Oktober = 31)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

5 Bitte kreuze den **letzten** Buchstaben deiner natürlichen **Haarfarbe** an. (z. B. braun, Glatze, schwarz, usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

6 Bitte kreuze den **letzten** Buchstaben deiner **Augenfarbe** an. (z. B. braun, grün, grau, usw.)

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
p	q	r	s	t	u	v	w	x	y	z	ä	ö	ü	ß

Hast du im letzten Jahr an der Befragung teilgenommen? ☐ ja ☐ nein
Hast du im letzten Jahr die Schule gewechselt? ☐ ja ☐ nein
Bist du im letzten Jahr sitzen geblieben? ☐ ja ☐ nein

B All alignments, matches and new case counts of the panel cross-sections

Alignment	Number of checks	New matches	Exhaustion in %	oPCS	cPCS	%
2006 with 2005	169	5	2.96	3,032	3,037	
2006 total	**169**	**5**	**2.96**	**3,032**	**3,037**	**+0.16**
2007 with 2004	290	73	25.17			
2007 with 2003	138	50	36.23			
2007 total	**428**	**123**	**28.74**	**2,587**	**2,710**	**+4.75**
2008 with 2005	349	190	54.44			
2008 with 2004	202	99	49.01			
2008 with 2003	114	44	38.60			
2008 total	**665**	**333**	**50.08**	**2,412**	**2,745**	**+13.81**
2009 with 2006	323	203	62.85			
2009 with 2005	236	125	52.97			
2009 with 2004	158	93	58.86			
2009 with 2003	98	38	38.78			
2009 total	**815**	**459**	**56.32**	**2,304**	**2,763**	**+19.92**
2011 with 2007	300	203	67.67			
2011 with 2006	220	116	52.73			
2011 with 2005	255	149	58.43			
2011 total	**775**	**468**	**60.39**	**1,812**	**2,324**	**+28.26**
2013 with 2008	448	309	68.97			
2013 with 2007	180	95	52.78			
2013 with 2006	170	46	27.06			
2013 with 2005	178	30	16.85			
2013 total	**976**	**480**	**49.12**	**1,912**	**2,392**	**+25.10**
2015 with 2009	524	263	50.19			
2015 with 2008	273	137	50.18			
2015 with 2007	164	68	41.46			
2015 with 2006	154	44	28.57			
2015 total	**1,115**	**512**	**45.92**	**1,812**	**2,324**	**+28.26**
2017 with 2011	488	325	66.60			
2017 with 2009	304	155	50.99			
2017 with 2008	153	69	45.10			
2017 with 2007	113	48	42.48			
2017 total	**1,058**	**597**	**56.43**	**1,760**	**2,357**	**+33.92**
2019 with 2013	489	352	71.98			
2019 with 2011	244	121	49.59			
2019 with 2009	213	95	44.60			
2019 with 2008	121	44	36.36			
2019 total	**1,067**	**612**	**57.36**	**1,670**	**2,282**	**+36.65**
total	**7,068**	**3,589**	**50.78**	**4,076**	**4,076**	**+0.00**

C Two examples of post-hoc matching of units

	Code	Participation last year	Gender	Citizenship	Education	New match
First example of eight new matches over time						
w2	HRS2NU	yes	male	German	low level	
w3	HRS2NU	yes	male	German	low level	
w4	HRS2NU	yes	male	German	low level	
w5	HRS2**D**U	yes	male	German	low level	
w6	HRS2NU	yes	male	German	low level	yes
w7	HRS2**D**U	yes	male	German	low level	yes
w8	HRS2**D**UE	-	male	German	low level	yes
w9	HRS2**D**UE	yes	male	German	low level	yes
w10	HRS2**D**UE	yes	male	German	low level	yes
w11	HRS2N**N**E	yes	male	German	low level	yes
w12	HRS2**B**UW	yes	male	German	low level	yes
w13	HRS2**B**UE	yes	male	German	low level	yes
Second example of one new match						
w2	ENB10NN	yes	male	Turkish	high level	
w3	ENB10NN	yes	male	Turkish	high level	
w4	ENB10NN	yes	male	Turkish	high level	
w5	ENB10NN	yes	male	Turkish	high level	
w6	ENB10NN	yes	male	Turkish	high level	
w7	ENB10NN	yes	male	Turkish	high level	
w8	ENB10NNK	yes	male	Turkish	high level	yes
w9	ENB10NNK	yes	male	Turkish	high level	
w10	ENB10NNK	yes	male	Turkish	high level	
w11	ENB10NNK	yes	male	Turkish	high level	
w12	ENB10NNK	yes	male	Turkish	high level	
w13	ENB10NNK	yes	male	Turkish	high level	

The first example reflects a case that was present from w1 to w5 without missing units in the panel data set before the panel consolidation. It can be seen that up to this point, this case only had an error in the code in w5. During the consolidation process, eight units were added to this individual data set. In all cases the code fit

within the error tolerance and also other visible indicators (such as the similarity of the school name) allowed the conclusion that the newly linked units are the same person. The errors in the code are quite easy to justify. It concerns Co005 (the last letter of the own hair color). Numerous respondents had a problem with the change of the query of the first letter (Co001-Co003) to the last letter. If the respondent now had the hair color "dark brown," the error could be explained with the choice of the first letter. If in addition in w12 and w13 only "brown" was meant by the respondent, this error could also be explained. The second case is an example of an individual data set that had only one missing unit until the panel consolidation, which was closed by the additional matching. In this case, a questionnaire could not be found or it could have been subjectively decided during the handwriting check that the questionnaire from the eighth wave should not be linked to w7 or w6.

D Results LCGAs

Original Panel (age 13-19, n=1,907)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	19,006	19,0840	19,039	1,945.57	0.00
3	18,714	18,814	18,757	290.59	0.03
4	18,541	18,663	18,593	175.57	0.00
5	18,498	18,643	18,560	48.79	0.01
6	18,482	18,649	18,554	23.32	0.21

Consolidated Panel (age 13-19, n=2,051)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	20,599	20,678	20,634	2,041.12	0.00
3	20,262	20,363	20,306	334.21	0.03
4	20,067	20,194	20,124	193,98	0.00
5	20,028	20,174	20,092	48.23	0.01
6	20,000	20,168	20,073	35.22	0.19

Original Panel (age 20-30, n= 1,865)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	4,216	4,288	4,246	479.96	0.00
3	4,164	4,256	4,204	61.69	0.00
4	4,154	4,270	4,203	18.21	0.03
5	4,157	4,296	4,216	4.39	0.47

Consolidated Panel (age 20-30, n= 2,419)

Number of classes	AIC	BIC	Adj. BIC	LMR-LRT	p
2	5,904	5,979	5,938	665.24	0.00
3	5,832	5,930	5,876	77.81	0.00
4	5,811	5,933	5,866	27.84	0.03
5	5,811	5,956	5,876	7.77	0.33

methods, data, analyses | Vol. 18(1), 2024, pp. 109-138 DOI: 10.12758/mda.2023.01

Continuous Time Modeling with Criminological Panel Data: An Application to the Longitudinal Association between Victimization and Offending

Jost Reinecke[1], *Anke Erdmann*[2] *& Manuel Voelkle*[3]

[1] *Bielefeld University, Faculty of Sociology*

[2] *Federal Criminal Police Office, Wiesbaden*

[3] *Humboldt-Universität zu Berlin, Faculty of Life Sciences*

Abstract

Background: Criminological research shows that there is nearly always a strong and positive association between delinquency and being a victim of crime. This so-called victim-offender overlap is one of the most consistent and best documented findings in criminology. However, examinations using longitudinal panel data are rather scarce. Previous analyses based on latent growth and cross-lagged panel models showed that the developments of victimization and offending are parallel processes that expose similar stability and mutual influence over the period of adolescence and early adulthood (Erdmann & Reinecke, 2018).

Objectives: The present study examines the relationship between victimization and offending over the phase of adolescence and emerging adulthood. The focus is on the application of continuous time dynamic modeling and on comparing results using data from the criminological panel study *Crime in the Modern City*. For the present analyses, seven consecutive panel waves are used that contain information about German adolescents from the age of 14 to 20 years.

Approach: The relationship between victimization and offending is analyzed by continuous time structural equation modeling using the R package `ctsem` (Driver & Voelkle, 2018, 2021). In addition to the unconditional models, relevant predictors (gender, routine activities) are considered in the conditional models. Methododological and substantive aspects of continuous time dynamic modeling are highlighted in the discussion of the results.

Keywords: continuous time modeling, panel analysis, R, ctsem, juvenile delinquency, longitudinal data

Various dynamic specifications of longitudinal models based on structural equations are recently discussed in the methodological literature (Asparouhov & Muthén, 2020; Zyphur et al., 2019a, 2019b; Hamaker et al., 2018; Usami et al., 2019; Montfort et al., 2018). One direction of the discussion is based on potentially misleading findings and interpretations of the classical *cross-lagged panel model* (CLPM, cf. Kessler & Greenberg, 1981; McArdle & Nesselroade, 2014; Rogosa 1979, 1980) regarding the presence, predominance and sign of causal influences. As pointed out by Hamaker et al. (2015), the main critical point of the CLPM is the failure to separate the within-person and the between-person level in the presence of time-invariant trait differences (see also Usami et al., 2019). These arguments are driven by the multilevel structure of the data in panel designs with repeated measurements of the same persons under study. To cope with these major critiques, it has been proposed by Hamaker to extend the CLPM by random intercepts referring to stable between-persons trait differences in the measurements (*random intercept cross-lagged panel model*, RI-CLPM).

The second direction of the discussion is due to the underlying assumption of discrete time points in all major panel models including the CLPM. For example, Voelkle et al. (2012) argue that parameter estimates of the CLPM depend on the length of the time interval between measurement occasions and that this information is not considered in the estimation of the parameters. The authors recommend to model autoregressive processes with stochastic differential equation models using a continuous time approach (*continuous time structural equation model*, CTSEM), which estimate and visualize the continuous time parameters. They also show the derivation of discrete time parameters from these models for specific time intervals of interest. Further explanations and discussions are given in Oud et al. (2018) and Ryan et al. (2018).

This paper intends to provide an application of the CTSEM and to compare model restrictions and model results based on data from a criminological panel study which focuses on the development of delinquency from adolescence to early adulthood. The dynamic relationship between victimization and offending over a certain age period (14 to 20 years) will be the substantive focus of the present analyses. They are based on previous results from cross-lagged panel and growth curve models as well as mixture models considering unobserved heterogeneity in the development of offending and victimization (Erdmann & Reinecke, 2018, 2021).

Erdmann & Reinecke (2018) explored developmental processes of victimization and offending using data from the criminological panel study *Crime in the Modern City* (CrimoC) and found evidence that both processes peak at the age of 14 with a subsequent decrease over the phase of adolescence. Both victimiza-

Direct correspondence to
Jost Reinecke, Bielefeld University, Faculty of Sociology
E-mail: jost.reinecke@uni-bielefeld.de

tion and offending are highly parallel and positively related processes throughout the juvenile life course. Using the CLPM, positive effects from victimization on offending as well as from offending to victimization could be detected. In addition, the results show a tendency that at a younger age, victimization rather predicts later offending because the highest cross-lagged effects are detected between 14 and 16 years of age (Erdmann & Reinecke, 2018: 336).

Upon these findings, Erdmann & Reinecke (2021) explored interindividual differences in the development of victimization and offending and, accordingly, distinct patterns of trajectories are detected via specification of growth mixture models (e.g., Muthén, 2004). Three groups of offender development (high-level offenders, adolescence-limited offenders, and nonoffenders) and two groups of victimization development (nonvictims and decreasing victims) were identified. Examining the intersection between these trajectories provided more profound insights into the overlap between victimization and offending. The association between the particular group memberships showed that juveniles who exhibit a high level of delinquency over the phase of adolescence are usually in a trajectory of elevated victimization.

The present analyses will consider these previous findings and attempt to overcome restrictions regarding the longitudinal analyses with discrete time points. It has been shown in the literature (e.g., Voelkle et al. 2012) that estimates of autoregressive and cross-lagged parameters of the CLPM are highly dependent on the length of the time interval between the measurements. Under a continuous time framework, like the CTSEM, these dependencies will vanish. Furthermore, discrete time parameters for any time interval can be calculated from the continuous time estimates.

First, we will briefly discuss the continuous time approach as well as the implementation of the continuous time structural equation model in R. After a brief introduction of the panel data and the measurements, the results of the continuous time models will be discussed. Finally, a detailed discussion about advantages and disadvantages of longitudinal modeling in continuous time are provided.

Continuous Time Structural Equation Modeling

In contrast to most panel models, including the CLPM or the RI-CLPM, time is treated as a continuous variable in continuous time modeling. This allows a clear distinction between the oftentimes continuous nature of the constructs under consideration (e.g., vicitimization and offending) and the always discrete occasions at which the measurements take place (e.g., seven panel waves). Practically speaking, treating time as a continuous variable makes the approach independent of the assumption of equidistantly spaced measurement occasions, permits the compari-

son of parameter estimates across studies with different time intervals, and allows researchers a detailed study of temporal dynamics. A comprehensive introduction to continuous time modeling is beyond the scope of this article, but is provided, for example, by Voelkle et al. (2012). For a recent overview of continuous time models in the social and behavioral sciences, see van Montfort et al. (2018).

Mathematically, the basic idea of continuous time modeling is to predict change over an infinitesimally small time interval, more precisely, to predict the derivative of a vector of variables of interest η(t) with respect to time t (i.e., $\frac{d\eta(t)}{dt}$ by η(t) and possibly other variables. This is formalized in the stochastic differential Equation 1:

$$d\eta(t) = \big(A\eta(t) + b + M\chi(t)\big)dt + GdW(t) \tag{1}$$

Matrix A represents the so-called drift matrix, with auto-effects on the diagonal and cross-effects on the off-diagonals, characterising the temporal relationships of the processes. Vector b denotes the intercepts. Matrix M represents the effects of time dependent predictors χ(t) on the processes η(t). W(t) denotes the so-called Wiener process, a random-walk in continuous time.

Lower triangular matrix G represents the effect of the stochastic error term on the change in η(t), with Q = GG′ being the variance-covariance matrix of the diffusion process.

Vector η(t) can be directly observed or latent with the following measurement model equation:

$$y(t) = \tau + \Lambda\eta(t) + \epsilon(t) \tag{2}$$

In Equation 2, the vector y(t) represents the manifest variables, τ represents the vector of manifest intercepts, the matrix Λ contains the factor loadings, and ε is the vector of residuals with error covariance matrix Θ.

To connect the continuous time Equation 1 to the discrete time measurement occasions, the equation is solved for an initial time point and the observed (i.e., discrete) time intervals between measurement occasions in a given study. This is illustrated in Equation 3, where the stars (∗) denote that the discrete time parameters are constrained to the solution of the differential Equation 1. Importantly, because Equation 1 is a comparatively simple linear differential equation, an analytical solution exists and the constraints are well-known (e.g., Oud & Jansen, 2000). For this reason we refrain from reiterating them here, but limit ourselves to referencing the existing literature and the R-package `ctsem` (Driver et al., 2017; Driver & Voelkle, 2018, 2021) that implements these constraints and that will be used later on for the empirical analyses:

$$\eta_u = A^*_{\Delta t_u}\eta_{u-1} + b^*_{\Delta t_u} + M\mathrm{x}_u + \zeta^*_u \tag{3}$$

Note that in contrast to Equation 1, we introduce u as a new symbol in Equation 1 to denote the discrete time measurement occasion u, with U being the set of all measurement occasions. Thus, Δt_u denotes the continuous time interval between two discrete measurement occasions η_u and $\eta_u = -1$.

As described in detail by Driver & Voelkle (2018), parameters in Equation 1 and Equation 3 can differ across individuals. These differences may be explained by time-invariant predictors. In the following, we use the symbol β to denote the vector of effects of time-invariant predictors z.

Continuous time models that can be formulated in terms of Equation 1, 2 and 3 can be conveniently specified and estimated by the R-package `ctsem` (Driver et al., 2017; Driver & Voelkle, 2018, 2021). The initial version of the R package `ctsem` interfaces to OpenMx (Neale et al., 2016) to estimate CTSEM for wide-format panel and time series data based on a full information maximum-likelihood approach. This initial version is now implemented in the R package `ctsemOMX` (Driver et al., 2017). Current versions of the R package `ctsem` provide estimation options for maximum likelihood and Bayesian models, interfacing to Stan (Carpenter et al., 2017). For the latter, panel and time series data has to be provided in long-format.

Data Basis and Measurements

Data

The data used for this methodological examination stem from the research project *Crime in the Modern City* (CrimoC, e.g., Boers et al., 2010; Seddig & Reinecke, 2017; Boers & Reinecke, 2019).[1] The project is funded by the German Research Foundation (DFG) and aims at explaining and monitoring the emergence, development, and desistance of delinquent behaviour throughout adolescence and emerging adulthood. For this purpose, both cross-sectional and longitudinal data on deviant and criminal behaviour as well as on individual characteristics (e.g., values, family characteristics, activities with friends) were collected. The overall project started in the year 2000 with interviews among several cohorts of students in the German cities Münster (started 2000), Bocholt (started 2001) and Duisburg (started 2002). Yet, only the youngest cohort of 7th-graders (13 years old on average in 2002) in Duisburg was followed up to form a long-term panel data set. In 2019, the 13th and last wave of the project was conducted.

The data collection process was initially realized as self-administered paper-and-pencil interviews in school during class supervised by trained interviewers. As

1 Detailed information on the conceptual framework and the design of the study can be obtained from www.crimoc.org.

the students became older and successively started leaving school, their address information were retrieved and the interview mode was gradually changed to postal mode and an optional, subsequent face-to-face mode (for a comprehensive overview, see for example Bentrup, 2007, 2009). The first eight waves of the study (2002 to 2009, age 13 to 20) were conducted annually. When the data collection process was changed to postal mode, the efforts and field time increased accordingly. As a consequence, data were collected biennially after 2009 (five waves from 2011 to 2019, age 22 to 30).

The main objective of the CrimoC-study is to examine the development of delinquent behaviour over the life-course, thus, the according data were retrieved at every wave. Information on victimization, however, were not obtained beyond the panel wave in 2009. Also, the first wave from 2002 cannot be included in analyses that target victimization, because data on victimization experience was only retrieved for certain school types and not for the entire student sample at that time point. Consequently, the panel waves for studying the dynamic relationship between victimization and offending are restricted to the age period from 14 to 20 years. From a criminological point of view, this section of the life-course is well suited for analyzing the association between victimization and offending, because it covers the phase where onset, peak, and emerging desistance of delinquency are most prominent among German juveniles (for details, see Erdmann & Reinecke, 2021).

In summary, the data set employed in the following analyses contains seven waves of data collected annually between 2003 and 2009 (see Figure 1). Respondents who participated at least five out of seven times are included in the analysis (n = 2679) to reduce bias compared to the same panel data set which restricts respondents to those who participated in all seven panel waves (n = 1488).[2]

2 Because of the German data protection law, registered postal adresses could not be used to link the data of the particular panel waves. Instead, individual codes derived from time-stable characteristics (e.g., first letter of prename, day of birth, first letter of mother's prename) were retrieved in each panel wave and used to match the panel data. It has been shown that a sufficient replication of the personal code (i.e., errors in replicating the code were allowed) is associated with gender, delinquency rates, and education. If the analysis would be restricted to those respondents with complete data over all seven panel waves (i.e., continuous participation and sufficient replication of the code), females, respondents with low delinquency/victimization rates, and people with higher education would be overrepresented. Allowing missing participations reduces this bias. Even respondents that did not participate (or who failed to replicate their individual code sufficiently) in two subsequent waves are considered in the seven wave panel data under study.

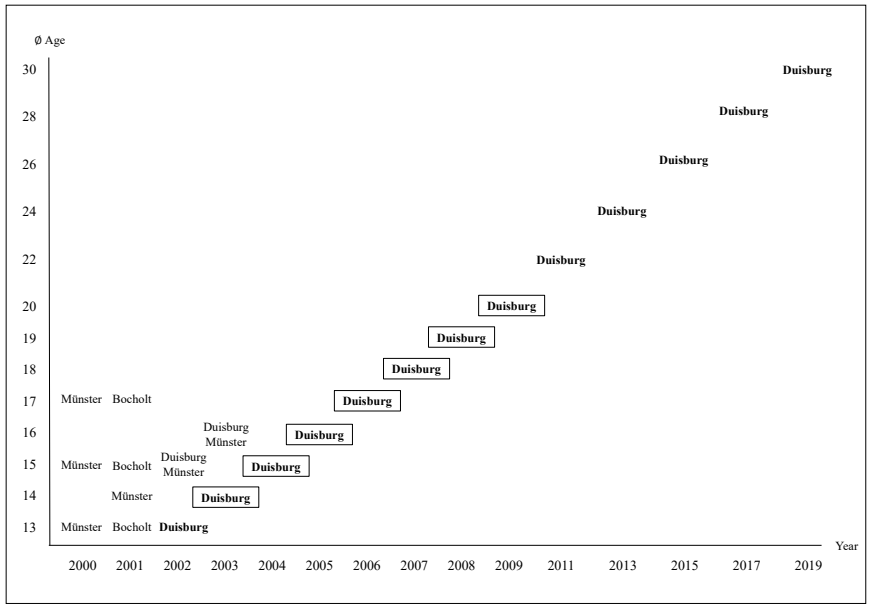


Figure 1 Design of the CrimoC-Study

Measurements

For the CTSEM, measurements of violent victimization and general offending are used as time dependent variables. Also, criminologically relevant predictors of victimization and offending – such as gender and activities with peers – are included in the later conditional models.

Violent Victimization. The present analysis considers violent victimization, which is measured via three violent offenses: *robbery (with threat of violence)*, *assault with a weapon*, and *assault without a weapon*. For each offense, participants were asked whether they have experienced this type of victimization within the last year preceding the interview. If yes, they were additionally asked how often they experienced this particular type of victimization. These annual incidences were summed over the three offenses for each wave (i.e., at every age under study)[3]. Hence, the variable reflects the intensity of violent victimization at a certain age. To be in line with previous longitudinal examinations of victimization (e.g., Higgins, Jennings, Tewksbury, & Gibson, 2009; Peterson, Taylor, & Esbensen, 2004;

3 Missing values were allowed for single items. If an item had a missing value, it was treated as zero in the sum. If all three items had missing values, the sum was also coded as missing.

Schreck, Stewart, & Fisher, 2006), the incidence was capped at the value of 12 and all values beyond were aggregated into one category. Thus, the highest category reflects at least 12 victimizations within a year which means at least once a month on average.

General Offending. The measurement of offending consists of 15 offenses covering a broad range of delinquent behaviour. It includes violence *robbery including threat of violence*, *violent bag snatching*, *assault with a weapon*, *assault without a weapon*), property offenses (*shoplifting*, *burglary*, *theft of bicycles*, *theft of cars*, *theft out of cars*, *theft out of a vending machine*, *fencing*, *other theft*), and criminal damage offenses (*graffiti*, *scratching*, *property damage*). The construction of the offending measurement was conducted equivalent to victimization: The annual incidences of the single offenses were added allowing missing values and all values of 12 and higher were combined into one category. Accordingly, the measurement reflects the intensity of offending at the considered time points, that is, at a certain age.

Gender. Gender is one of the most prominent predictors of offending and victimization. Independent of the panel waves (i.e., independent of age), it is expected that males have consistently higher incidence rates of offending and victimization compared to females. Hence, it is included as a time-invariant measurement in the CTSEM to explore possible gender effects. The measurement is binary and contains the two categories male and female.[4]

Routine Activities. The measurements describing the activities are derived from the lifestyle-routine activity approach, which is a combined framework based on routine activity theory (Cohen & Felson, 1979) and lifestyle-exposure theory (Garofalo, 1987; Hindelang et al., 1978). This approach is one of the most prominent theoretical concepts for investigating the association between victimization and offending and has been considered in numerous studies (e.g., Cho & Lee, 2018; Engström, 2018; Mustaine & Tewksbury, 2000; Plass & Carmody, 2005; Pyrooz, Moule, & Decker, 2014; Schreck, Stewart, & Osgood, 2008).

In general, the theory assumes that daily activities regulate the risk of committing criminal acts or – when transferred to victimization - the risk of becoming a victim of crime. A key element, that was later introduced by Osgood et al. (1996), is the distinction between *structured* and *unstructured activities*, also called structured or unstructured socializing. This differentiation states that participation in activities that take place an organized, monitored setting decreases the chances of deviance compared to unstructured and unsupervised activities. Among juveniles, particularly unstructured activities with peers are considered risk factors for crime because delinquent acts are perceived more easier and rewarding when friends are present and authority figures are absent.

---

4 The questionnaire does not differentiate between sex and gender. Thus, the information on sex provided by the respondents is designated as gender.

Similar mechanisms apply to victimization risk. On the one hand, structured activites reduce victimization risk due to, for example, amplified supervision, social control, and a more protective environment. Unstructured activities, on the other hand, entail a higher risk of victimization because, for example, people are more frequently placed in close proximity to motivated offenders or more exposed to hazardous situations.

In the later conditional models, two indicators derived from the lifestyle-routine activity framework are included in the analysis as time-invariant predictors of the dynamic relationship between victimization and offending. Specifically, we consider activities with peers that reflect *unstructured* and *structured socializing.* For measuring the activities, the respondents were asked how much certain statements apply to their friend group, each on a five-point Likert scale where higher values represent a higher frequency of the considered activity. The activities were measured at every age under study. Correlations show that the activities are mostly stable over the considered age span[5], thus, we averaged the values over the seven-year-period to obtain a single, time-invariant indicator as also practiced in previous studies (Erdmann & Reinecke, 2021; Labouvie, Pandina, & Johnson, 2016; Mulford et al., 2018).

For the unstructured activity, we use an indicator labeled as *partying* which consists of the two highly correlated items *alcohol consumption* ("When we are together, we drink a lot of alcohol.") and *going out* ("We visit bars, discotheques, or concerts together."), both measured on a five-point Likert scale from "does not apply" to "fully applies". Alcohol use has shown to be a consistent predictor of both delinquency and victimization (Engström, 2018; Felson & Staff, 2010; Mustaine & Tewksbury, 2000). Also, going to parties has regularly been considered an unstructured activity (Osgood et al., 1996). Thus, the indicator *partying* is suspected to facilitate both offending and victimization.

As a structured activity, we consider *studying* ("We study together for (vocational) school", measured on a five-point Likert scale from "does not apply" to "fully applies"). We expect this activity to have a mitigating effect on crime and victimization. This anticipation is based on the theoretical presumption that spending time in structured activites leaves less time available to conduct crime on the one hand (Osgood et al., 1996) and reduces exposure to potential offenders on the other hand.

5 The frequency of an activity at a certain age correlates strongly with the frequency of the same activity one year later (r between 0.44 and 0.66).

Descriptive Results

Table 1 shows the annual incidence rate of victimization and offending (mean and variance) for every age using the seven-wave panel data. At the age of 14, the mean incidence of victimization has an average frequency of 0.61 which drops down to 0.11 at the age of 20.

The variance of victimization decreases from 4.05 to 0.54 because the amount of zeros (i.e., no victimization) increases. A very similar development holds for general offending. The mean annual incidence is higher for offending than for victimization partly due to the higher number of different offenses included. At the age of 14, the mean incidence of offending has an average frequency of two offenses (2.05). At 20 years of age, the mean incidence drops down to 0.35. Also the variance decreases from 14.96 to 2.80 due to the increasing amount of zeros (i.e., no offenses). As expected, incidence rates of victimization and offending decrease throughout the phase of adolescence reflecting a parallel process of development.

Table 2 shows the descriptive results for the independent variables gender and peer activities. The panel data contains a somewhat higher percentage of females compared to males. The averaged distributions of the peer activities reflect a balanced activity pattern.

Table 1 Descriptive Results for Violent Victimization and General Offending

	Victimization				Offending			
Age	N	Mean	Var.	% Zero	n	Mean	Var.	% Zero
14	2201	0.61	4.05	83.2	2208	2.05	14.96	66.4
15	2406	0.47	2.89	85.0	2422	1.98	15.22	69.2
16	2563	0.40	2.45	87.1	2568	1.54	12.31	74.5
17	2480	0.36	2.25	88.7	2484	1.19	9.94	80.3
18	2435	0.23	1.53	92.0	2436	0.71	5.84	86.5
19	2455	0.15	0.84	94.0	2457	0.47	3.70	90.3
20	2436	0.11	0.54	95.1	2439	0.35	2.80	92.6

Note. Mean and variance of incidences for violent victimization and general offending, percentages of zero for each panel wave. Results are based on seven-wave panel, n = 2679, maximum of two missing wave information, full information maximum likelihood for estimating means and variances, and values rounded to two decimal digits.

Table 2 Descriptive Results for Gender and Peer Activity Variables

Gender	n	Proportion	
female	1469	0.55	
Male	1207	0.45	
Peer Activities	n	Mean	Var.
Partying	2600	2.66	0.89
Studying	2598	2.71	0.88

Note. Proportions of gender, means and variances for peer activities, based on seven-wave panel, n = 2679, maximum of two missing wave information, full information maximum likelihood for estimating means and variances, and values rounded to two decimal digits.

Model Specifications and Results

Unconditional and Conditional Model Specifications

According to the general specification of the CTSEM (Equation 1) the unconditional model contains the time-variant variables offending (off) and victimization (vict):[6]

$$d\begin{bmatrix} off \\ vict \end{bmatrix}(t) = \left(\begin{bmatrix} a\_off & a\_off\_vict \\ a\_vict\_off & a\_vict \end{bmatrix}\begin{bmatrix} off \\ vict \end{bmatrix}(t) + \begin{bmatrix} k\_cint1 \\ k\_cint2 \end{bmatrix}\right) dt +$$

$$cholsdcor\left\{\begin{bmatrix} q\_off & 0 \\ q\_vict\_off & q\_vict \end{bmatrix}\right\} d\begin{bmatrix} W_1 \\ W_2 \end{bmatrix}(t)$$

with initial latent state

$$\begin{bmatrix} off \\ vict \end{bmatrix}(t_0) \sim N\left(\begin{bmatrix} T0m\_off \\ T0m\_vict \end{bmatrix} covsdcor\left\{\begin{bmatrix} T0var\_off & 0 \\ T0var\_vict\_off & T0var\_vict \end{bmatrix}\right\}\right)$$

The CTM contains four parameters in drift matrix *A*, two parameters in vector *b* (intercepts) and three parameters in matrix *Q* for the diffusion process. Five parameters (two means, two variances and one covariance) are estimated for the initial latent state of the process. In previous analyses with latent growth and cross-lagged panel models Erdmann and Reinecke (2018) showed that the developments

6 *cholsdcor* converts lower triangular matrix of standard deviation and unconstrained correlation to Cholesky factor covariance, see Driver & Voelkle (2018: 11). *covsdcor* is the transposed cross product of *cholsdcor* which renders the stationary covariance matrix.

of offending and victimization are highly parallel processes that reflect similar stability and mutual influence over the time of adolescence. Therefore, it is intended to explore how the autoeffects differ between offending and victimization and how large would be the particular crosseffects. Absence of a particular crosseffect can be tested by restricting the particular parameter in the off-diagonal of drift matrix *A* to zero. A test of equal crosseffects would show that both processes are influencing each other with the same strength. Results of the different model specifications are shown and discussed in the next section.

The measurement part of the CTM (Equation 2) contains factor loadings (λ) which are fixed to 1.0. Measurement error variances (ε) and manifest intercepts (τ) are fixed to zero. For any different model specification regarding the elements of matrix *A*, there are no parameter to be estimated for the measurement model (cf. Voelkle et al., 2012).

As defined in the section Continuous Time Structural Equation Modeling, vector *b* contains the effects of the time independent predictors *z* (i.e., gender and indicators of routine activities) on the parameters of interest. Below, vector β is shown for the predictor gender:[7]

$$\beta = \begin{bmatrix} b\_T0m\_off \\ b\_T0m\_vict \\ b\_cint1 \\ b\_cint2 \\ b\_a\_off \\ b\_a\_off,vict \\ b\_a\_vict\_off \\ b\_a\_vict \end{bmatrix} [Gender]$$

The first two parameters are the effects of gender on the means of offending and victimization at the initial time point followed by the two parameters indicating the effects of gender on the intercepts of offending and victimization. The last four parameters consider the effects of gender on the parameters of the drift matrix *A*. For example, the parameter $b_{a_{off}}$ is the regression of gender on the autoeffect of offending.

The same specification of vector β was used for the variables of routine activities (partying and studying). Because of the complexity of the conditional CTM, the influence of the time independent predictors are considered in separate analyses.

7 Note, that not *all* possible parameters are included in vector β. For example, effects on the parameters of the diffusion matrix *Q*could be added. This was not done, because no theoretical reasons exist to justify these specifications.

Model Results

According to the propositions above, all models are estimated with the R package `ctsem` (Driver & Voelkle, 2018, 2021) using data from the seven panel waves of the CrimoC study. Maximum likelihood estimation procedure is used for the particular model estimation, prior information is not specified.[8]

Unconditional Models

Table 3 gives an overview about the log-likelihoods and the information criteria AIC and BIC (Kuha, 2004) for the estimated unconditional models.[9]

Model A contains the measurements of offending and victimization with full specification of the drift matrix. Model B restricts the crosseffect from offending to victimization to zero, Model C alternatively restricts the crosseffect from victimization to offending to zero. Therefore, both restricted unconditional Models B and C have the same number of parameters and one parameter less than Model A. Alternatively, Model D considers the restriction of equal crosseffects ($a_{off,vict} = a_{vict,off}$). Comparing the AIC across the four model variants, Model A has the lowest value. Comparing the BIC, Model D has the lowest value. But the difference of the BIC values between Model A and D is quite small. Therefore, Model A is chosen and will be described in more detail.

Table 3 Log-Likelihood and Information Criteria for Unconditional CTMs

Unconditional CTMs	Par.	- log (L)	AIC	BIC
Model A (unrestricted)	21	-70032.27	140106.5	140230.30
Model B (restricted) (Off→Vict = 0)	20	-70088.71	140217.4	140335.28
Model C (restricted) *(Vict→Off = 0)*	20	-70046.71	140133.4	140251.28
Model D (restricted) (Off→Vict = Vict→Off)	20	-70035.99	140112.0	140229.84

8 CTMs are estimated using the command `ctStanFit` with longformat data. Driver & Voelkle (2021: 894) recommend for the maximum likelihood approach to set the argument `nopriors=TRUE` in the command `ctStanFit` to disable the priors.

9 In the current version of the R package `ctsem`, only the AIC is provided. In addition, the BIC was calculated.

Table 4 Parameter Estimates of the Unconditional CTMs

	Model A		Model B		Model C		Model D	
Parameter	Estimate	SD	Estimate	SD	Estimate	SD	Estimate	SD
Drift Matrix (A)								
$a_{off,\ off}$	-1.061	0.040	-1.085	0.035	-0.976	0.029	-1.013	0.030
$a_{off,vict}$	0.569	0.111	0.280	0.093	-	-	0.285	0.034
$a_{vict,off}$	0.286	0.034	-	-	0.246	0.030	0.285	0.034
$a_{vict,vict}$	-2.598	0.140	-2.242	0.095	-2.595	0.136	-2.619	0.142
Intercepts (b)								
k_{off}	0.772	0.053	0.895	0.055	0.853	0.049	0.806	0.050
k_{vict}	0.416	0.048	0.632	0.043	0.464	0.049	0.426	0.048
Diffusion Matrix (Q)								
q_{off}	15.039	0.324	15.359	0.333	14.799	0.307	14.869	0.295
q_{vict}	7.546	0.366	6.906	0.268	7.564	0.363	7.603	0.373
$q_{off,vict}$	0.276	0.228	1.773	0.158	1.113	0.147	0.615	0.255
Initial Occasion								
$T0m_{off}$	2.223	0.084	2.213	0.082	2.218	0.082	2.222	0.084
$T0m_{vict}$	0.658	0.043	0.653	0.043	0.647	0.043	0.653	0.041
$T0var_{off}$	16.222	0.482	16.115	0.512	16.214	0.508	16.187	0.492
$T0var_{vict}$	4.153	0.129	4.150	0.126	4.126	0.123	4.136	0.122
$T0var_{off,vict}$	2.713	0.200	2.656	0.184	2.630	0.224	2.657	0.207

SD = Standard Deviation

According to the diagonal elements of the drift matrix (Model A, cf. Table 4), both processes are approaching an equillibrium in the future. The process is faster for victimization compared to offending ($|-2.598| > |-1.061|$). The off-diagonal elements of the drift matrix show that the impact of victimization on offending is stronger than the impact of offending on victimization ($|0.569| > |0.286|$).

The corresponding discrete time parameters can be computed at any arbitrary point in time. For time interval $\Delta t = 1$ autoregressive and cross-lagged discrete time parameters are calculated as follows:[10]

$$A(\Delta t) = e^{A\Delta t} = e^{\begin{pmatrix} -1.061 & 0.569 \\ 0.286 & -2.598 \end{pmatrix} \cdot 1} = \begin{pmatrix} 0.364 & 0.103 \\ 0.051 & 0.086 \end{pmatrix}$$

10 In the R package `ctsem` the argument `ctStanContinuousPars` can be used to calculate the discrete time parameters (Driver & Voelkle, 2021).

The transformation is unique under the condition that the eigenvalues of $A(\Delta t)$are real and have eigenvalues between zero and one (Kuiper & Ryan, 2018).

The calculated autoregressive discrete-time parameters show that victimization is less stable compared to offending ($|0.086| < |0.364|$). The cross-lagged parameters show that the impact of victimization on offending is stronger than the impact of offending on victimization ($|0.103| > |0.052|$). Restricting the latter cross-lagged effect to zero does not lead to an overall model improvement (cf. Model B in Table 3).

Furthermore, the R package `ctsem` allows a visual inspection of the development of victimization and offending over time and the relationship between both processes. Based on the estimates of the drift matrix (cf. Model A in Table 4), the autoeffect plots are shown in Figure 2. Estimates of the autoeffects are obtained by sampling from the subjects data. The red line shows the autoeffect process of offending, the turquoise line shows the same process of victimization. Over the time-interval on a range from 0 to 5 the processes are approaching asymptotically zero. As the curves show this is faster for victimization compared to offending meaning more changes for offending occur in future time periods. Since there are stable and stationary processes, there is an equilibrium to which the processes will return.

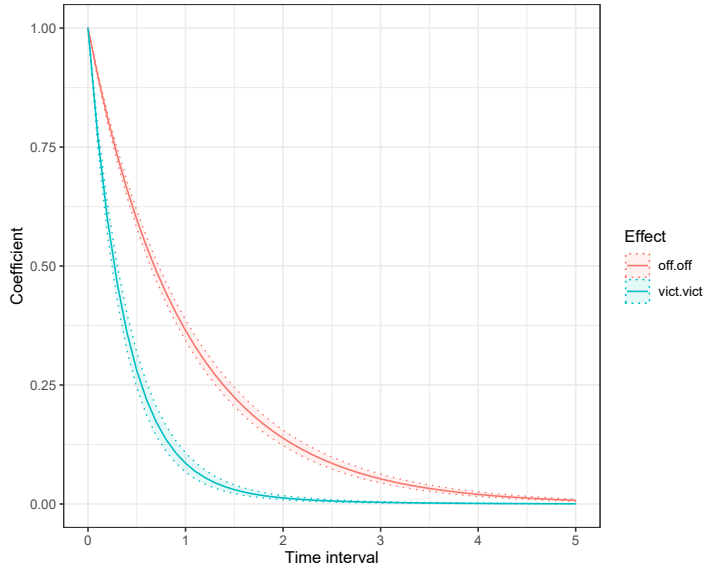


Figure 2 Autoeffect Plot of Victimization and Offending

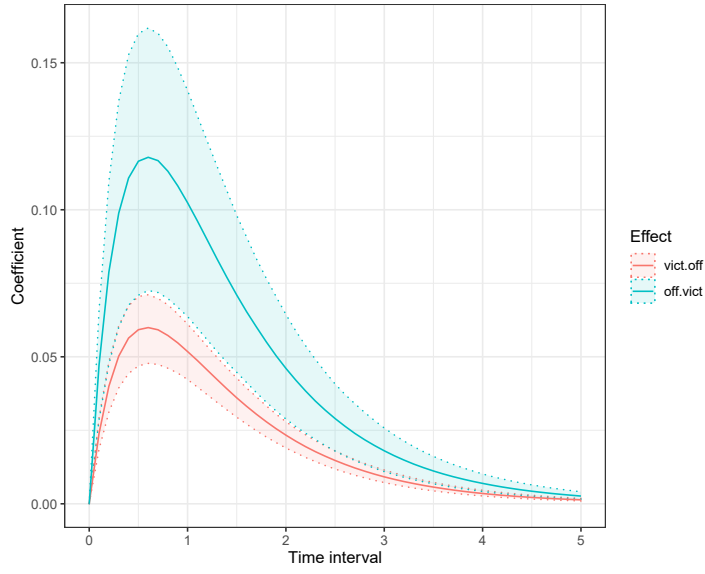


Figure 3 Crosseffect Plot of Victimization and Offending

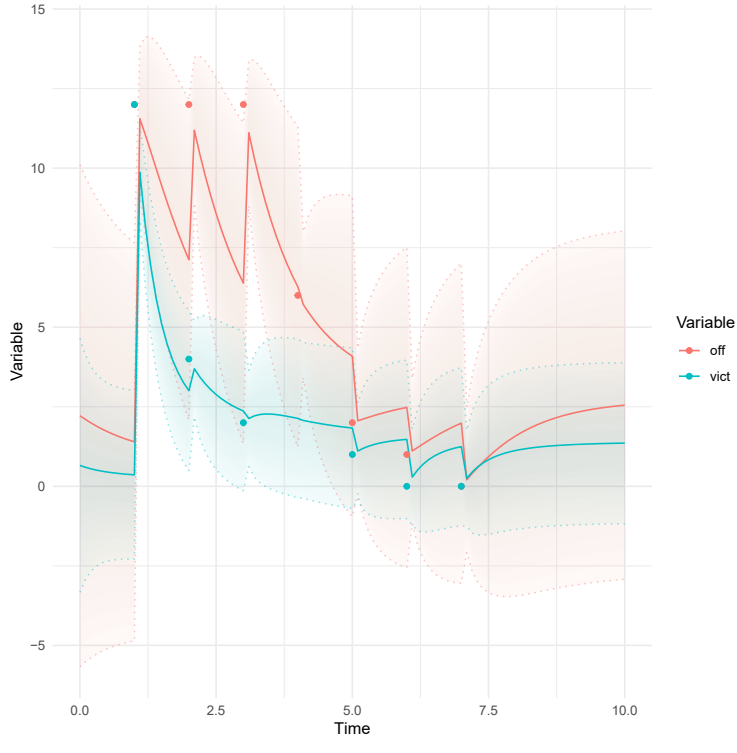


Figure 4 Prediction Plots for Subject 5

Crosseffects of offending and victimization are shown in Figure 3. For short time intervals the impact of victimization on offending (turquoise line) is larger than the impact of offending on victimization (red line). When the time interval becomes very large, the relationships dampens out.

For individual level analysis, Figure 4 shows observed and predicted scores for a particular person (Subject 5) obtained from the Kalman filter (Driver & Voelkle, 2018).[11]

The red line (offending) reflects higher score estimates on offending during adolescence (first panel waves) and lower scores later on. The green line reflects also some high scores on victimization but lower compared to offending. After the last panel wave, the Kalman filter estimates extrapolated future values.

Conditional Models

Table 5 summarizes the conditional models with gender and routine activities (partying and studying) as time independent predictors. Because of the model complexity, each time independent predictor is considered separately for the particular CTM. In the baseline versions of the conditional CTMs, eight regression parameters are estimated (cf. vector β in Section *Unconditional and Conditional Model Specifications*). Some of these regression estimates are low and not significant.

In the restricted versions of the conditional models, the non-significant regression parameters are restricted to zero for reasons of parsimony. When comparing the particular baseline models with the restricted ones (e.g., Model E with Model F for the time independent predictor gender), the information criteria AIC and BIC of the restricted models have always slightly lower values. In the following paragraphs, the results of the conditional models are discussed with emphasis towards the influences of the particular time independent predictor.[12]

Gender. Effects (vector β) of gender on the model parameters are summarized in Table 6 (Models E and F). Two parameters are restricted to zero in Model F: The effects of gender on the intercept of offending and victimization (b_{koff}, b_{kvict}).

11 The Kalman Filter produces subject specific estimates of the process variables based on all prior and current observations. It provides also a prediction of the future system state based on past estimations. In the R package `ctsem` predicted scores can be computed via the argument `ctKalman` (Driver & Voelkle, 2021).

12 In the conditional models the parameter estimates of the drift matrix, of the intercepts, of the diffusion matrix, and of the initial occasions do not differ substantially compared to the ones obtained by the unconditional models (cf. Table 4). Therefore, these estimates are not reported again.

Table 5 Log-Likelihood and Information Criteria for Conditional CTMs

Model variants	Par.	- log (L)	AIC	BIC
Model E (Gender)	29	-69089.98	138238.0	138408.9
Model F (Gender)	27	-69091.24	138236.5	138398.0
Model G (Party)	29	-66550.92	133159.8	133330.7
Model H (Party)	28	-66551.49	133159.0	133324.0
Model I (Study)	29	-67229.62	134517.2	134688.1
Model J (Study)	27	-67230.59	134515.2	134674.3

Table 6 Parameter Estimates of the Effects of Gender

	Model E			Model F		
Parameter	Estimate	SD	z	Estimate	SD	z
Effects on Drift Matrix (A)						
$b_{a_{off,off}}$	1.133	0.060	18.80	1.130	0.066	17.08
$b_{a_{off,vict}}$	-0.495	0.171	-2.89	-0.412	0.149	-2.77
$b_{a_{vict,off}}$	-0.222	0.050	-4.43	-0.238	0.052	-4.61
$b_{a_{vict,vict}}$	2.123	0.147	14.39	2.095	0.193	10.88
Effects on Intercepts (b)						
$b_{k_{off}}$	0.112	0.085	1.33	-	-	-
$b_{k_{vict}}$	-0.102	0.094	-1.09	-	-	-
Effects on Initial Occasion						
$b_{T0m_{off}}$	0.905	0.166	5.45	0.871	0.172	5.06
$b_{T0m_{vict}}$	0.475	0.084	5.66	0.474	0.091	5.19

SD = Standard Deviation; z = z-value

Positive values of the regression estimates indicate higher values for males, negative values indicate higher values for females. In Model E and Model F regression coefficients for the elements of drift matrix A and means of the initial occasion are similar. Gender differences are higher for the autoeffect of victimization (Model E: 2.12, Model F: 2.10) compared to the autoeffect of offending (Model E/F: 1.13). The opposite gender difference can be observed for the crosseffects. Gender differences are higher for the crosseffect of victimization to offending (Model E: -0.50; Model F: -0.41) compared to the reversed crosseffect (Model E: -0.22; Model

F: -0.24). Higher gender differences are observed for the initial mean of offending (Model E: 0.91; Model F: 0.87) compared to the ones for the initial mean of victimization (Model E: 0.48; Model F: 0.47).

Figure 5 shows how the expectations for individuals parameter values change as a function of the particular value of the time independent predictor gender (Driver & Voelkle, 2021: 898). Four discrete time parameters of matrix A (dtDrift) based on the estimates of Model F are included in the graph. For males (-axis value of one) the likelihood of change for the autoregressions of offending (red line in the graph) and victimization (violet line in the graph) is higher compared to females (-axis value of zero). Gender differences are to be expected much higher for offending compared to victimization. The likelihood of change for both crossregressions (blue line: effect of victimization to offending; green line: effect of offending to victimization) is similar but on different levels. Note, that the effects of gender on both intercepts are restricted to zero in Model F and therefore not included in the graph (cf. Table 6).

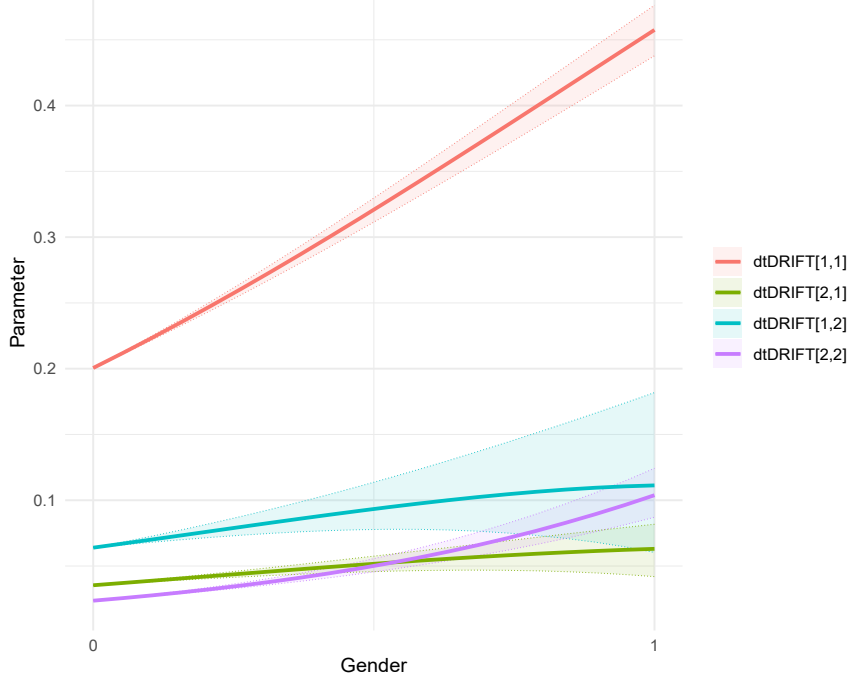


Figure 5 Expected Parameter Values as a Function of Gender. The four lines correspond to the four elements of the drift matrix. To ease interpretation, the discrete time parameters (dtDRIFT) for a time interval of 1 are presented.

Table 7 Parameter Estimates of the Effects of Partying

	Model G			Model H		
Parameter	Estimate	SD	*z*	Estimate	SD	*z*
Effects on Drift Matrix (A)						
$b_{a_{off,off}}$	1.520	0.040	37.67	1.531	0.061	25.00
$b_{a_{off,vict}}$	-0.511	0.099	-5.17	-0.492	0.099	-4.99
$b_{a_{vict,off}}$	-0.062	0.031	-2.00	-0.075	0.032	-2.30
$b_{a_{vict,vict}}$	2.221	0.103	21.62	2.248	0.203	11.06
Effects on Intercepts (b)						
$b_{k_{off}}$	0.069	0.061	1.13	-	-	-
$b_{k_{vict}}$	-0.155	0.069	-2.25	-0.141	0.066	-2.15
Effects on Initial Occasion						
$b_{T0m_{off}}$	1.246	0.086	14.56	1.228	0.020	14.76
$b_{T0m_{vict}}$	0.292	0.050	5.85	0.289	0.040	12.23

SD = Standard Deviation; *z* = *z*-value

Routine Activity: Partying Effects (vector β) of the routine activity partying on the model parameters of offending and victimization are summarized in Table 7 (Models G and H).

One parameter is restricted to zero in Model H: The effect of partying on the intercept of offending ($b_{k_{off}}$).

Regarding Models G and H, the regression estimates are positive for the diagonal elements of Matrix A (Model G: 1.52 and 2.22; Model H: 1.53 and 2.25): With more party activities, the autoeffects of offending and victimization increase. For the crosseffects, the regressions of partying are both negative in the particular models (Model G: -0.51 and -0.06; Model H: -0.49 and -0.08). The intercept of offending will be slightly higher for persons with higher party activities (Model G: 0.07) but these estimate turns to be not significant and is restricted to zero in Model H. The regression of the intercept of victimization on partying remains significant (Model G: -0.16; Model H: -0.14) meaning that this intercept will be lower for persons with higher party activities. For the means of the initial occasion of offending and victimization, positive and significant regressions of partying can be observed (for offending in Model G: 1.25 and in Model H: 1.23; for victimization in Models G and H: 0.29). At the beginning of the developmental process persons with more party activities are likely to have more offending and victimization experiences.

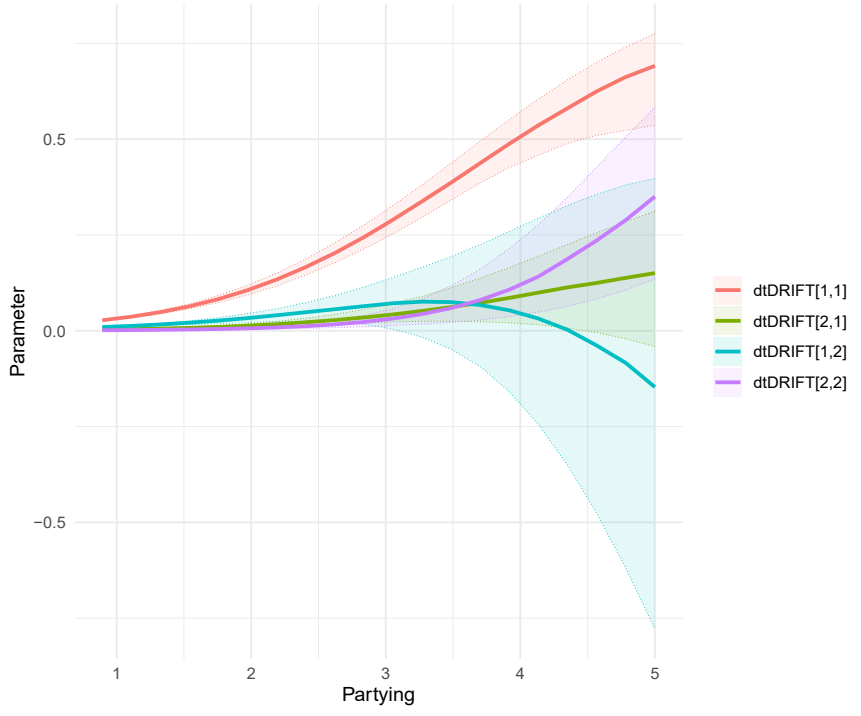


Figure 6 Expected Parameter Values as a Function of Partying. The four lines correspond to the four elements of the drift matrix. To ease interpretation, the discrete time parameters (dtDRIFT) for a time interval of 1 are presented.

Figure 6 shows how the expectations for individuals parameter values change as a function of the time independent predictor partying. It has a large effect on the autoeffect of offending (red line) compared to the autoeffect of victimization (violet line). With increasing party activities, it is likely that the developmental process of offending and victimization will change more often. That means that with extreme high numbers of party activities the developmental process of offending is likely to change (red line). The expected values of partying on the crosseffect between offending and victimization (green and blue line) are somewhat lower. For the crosseffect of offending on victimization a dampening effect can be observed (blue line).

Table 8 Parameter Estimates of the Regression on Studying

	Model I			Model J		
Parameter	Estimate	SD	*z*	Estimate	SD	*z*
Effects on Drift Matrix (A)						
$b_{a_{off,off}}$	-0.228	0.014	-16.92	-0.230	0.013	-17.85
$b_{a_{off,vict}}$	0.587	0.166	3.53	0.520	0.217	2.40
$b_{a_{vict,off}}$	0.181	0.052	3.46	0.197	0.043	4.62
$b_{a_{vict,vict}}$	-0.908	0.150	-6.04	-0.879	0.150	-5.85
Effects on Intercepts (b)						
$b_{k_{off}}$	-0.058	0.057	-1.03	-	-	-
$b_{k_{vict}}$	0.098	0.078	-1.25	-	-	-
Effects on Initial Occasion						
$b_{T0m_{off}}$	-1.004	0.084	-11.97	-0.989	0.086	-11.50
$b_{T0m_{vict}}$	-0.210	0.045	-4.66	-0.209	0.046	-4.55

SD = Standard Deviation; *z* = *z*-value

Routine Activity: Studying Effects of routine activity studying (vector β) on model parameters are summarized in Table 8 (Models I and J). Two parameters are restricted to zero in Model J: The regression of the intercept of offending (k_{off}) and victimization (k_{vict}) on partying.

Regarding Models I and J the regression estimates are negative for the diagonal elements of Matrix A (Model I: -0.23 and -0.91; Model J: -0.23 and -0.88): With more activities to study, the autoeffects of offending and victimization decrease. For the crosseffects, the regressions of studying are both positive in the particular models (Model I: 0.58 and 0.18; Model J: 0.52 and 0.20). The influences on the intercepts of offending and victimization are not significant and the parameters are restricted to zero in Model J. For the means of the initial occasion of offending and victimization, negative and significant regressions of studying can be observed (for offending in Model I: -1.00 and in Model J: -0.99; for victimization in Models I and H: -0.21). At the beginning of the developmental process persons with more study activities are likely to have less offending and victimization experiences.

Figure 7 shows how the expectations for individuals parameter values change as a function of the time independent predictor studying. Similar to partying, it has a large impact on the autoeffect of offending (red line) compared to the autoef-

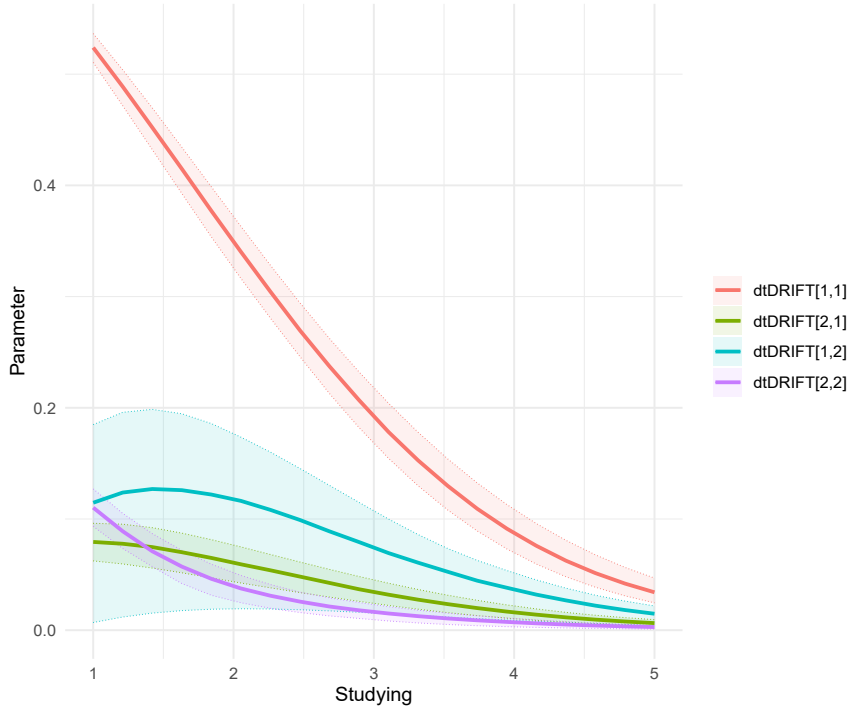


Figure 7 Expected Parameter Values as a Function of Studying. The four lines correspond to the four elements of the drift matrix. To ease interpretation, the discrete time parameters (dtDRIFT) for a time interval of 1 are presented.

fect of victimization (violet line). But the direction of the expected values is completely opposite in comparison to partying (cf. Figure 6). With increasing activities to study, it is likely that the developmental process of offending will change less. This means that with low study activities, the developmental process of offending is likely to change (red line). In principle, the expected value change for victimization goes into the same direction but on a much lower level. The expected values of studying on the particular crosseffects between offending and victimization (green and blue line) are positive. The values reflect a dampening effect of studying.

5 Discussion

Several advantages of stochastic differential equation models for the social and behavioral sciences have been adressed and discussed in the statistical literature for decades. Foremost is the use of time for the modeling process. Discrete-time methods are often used although the underlying longitudinal processes require models based on continuous time. In their editorial introduction to a special issue on continuous time modeling of panel data, Oud and Singer (2008, p. 1) remark that the use of discrete-time models might work as long as the time interval in the data is small (e.g., time-series data). But in the social and behavioral sciences panel data with far less measurement frequencies than observations are more common. It has been shown that in widely used cross-lagged panel models the results are inherently bound to the time intervals of the panel data (e.g., Delsing & Oud, 2008; Voelkle et al., 2012). More and more large-scale panel studies employ different time intervals due to substantive reasons or financial restrictions and researchers have to cope with such designs when analyzing the data.

Continuous time models on the basis of stochastic differential equations can overcome limitations of standard autoregressive models like the cross-lagged panel model. We have briefly shown the relationship between estimated parameters of the continuous-time model (auto- and crosseffects in the drift matrix A) and the corresponding discrete-time parameters in the autoregressive cross-lagged matrix $A^*_{\Delta t_u}$ (cf. Equation 1 and 3). Discrete and continuous time parameters are directly available during estimation and it is possible to transform the parameters of an estimated continuous time model to the discrete time parameters for any time interval.

Continuous time models are implemented in the R package `ctsem` which has been used here to study the long-term relationship between victimization and offending during the age of adolescence. Unconditional as well as conditional models are estimated. The parameters of the unconditional models show that the process of victimization is less stable compared to offending while the impact of victimization on offending is stronger than the impact of offending on victimization. The particular crossregression plot shows that this impact holds for the phase of early adolescence (14 to 16 years of age) but tends to diminish later (Figure 3).

Gender as well as unstructured and structured routine activities (partying and studying) are used as time independent predictors in the conditional models. Gender differences are higher for the autoregression of victimization compared to the autoregression of offending. In both cases, males would have larger negative values in the diagonal of the drift matrix meaning that the process is more unstable and refers to a larger amount of activities. Individual parameter change is more likely for males compared to females (Figure 5). A similar picture can be observed for the unstructered routine activity *partying*. The more party activities are observed the higher is the instability of the developmental process of offending and victimiza-

tion. The tendency for individual parameter value change is increasing (Figure 6). For the structured routine activity *studying*, the opposite result is gained from the model estimates. With more study activities the developmental process of offending and victimization is becoming more stable.

The tendency for individual parameter value change is constantly decreasing (Figure 7).

These results support previous findings that the risk for males to be in a group of victimized high-level offender is much higher compared to females (cf. Erdmann & Reinecke, 2021). In addition, group activities like meeting with friends, partying and hanging out with friends also increased the risk to be a victimized high-level offender whereas studying with friends has a decreasing impact.

Like in the previous publications of Erdmann and Reinecke (2018, 2021), the panel data of the CrimoC-study used here contains seven panel waves limited to persons who participated at minimum in five out seven waves (n=2679). We also tested the models using all persons for the particular time interval between 2003 and 2009 (cf. Figure 1) including those who participated less than five times (n=4076). No substantive differences in model parameters compared to the reported ones could be detected.

Of course, the application of continuous time models with criminological panel data has some limitations. The dependent variables offending and victimization are summed indices of annual incidences. So, we treated both variables without specifying a measurement model (cf. Equation 2). Furthermore, the dependent variables are treated as continuous measurements although they are based on count data (number of incidences per year). For count measurements other link function for estimating a CTSEM should be used like the Poisson or the negative binomial model (Hilbe, 2011). But unfortunately, these link functions are not yet implemented in the R package `ctsem` (but see Hecht et al., 2019).

We explored the impact of the time independent predictors one by one instead of using them simultaneously in a single conditional CTSEM. This was done for substantive reasons as well as to reduce the model complexity, but does not consider potential dependencies among the predictors.

Although fully Bayesian approaches are implemented in the current version of the R package `ctsem` (Driver & Voelkle, 2018), we restricted ourselves to maximum likelihood estimation, respectively maximum a posteriori estimates. This was done for reasons of computation time. Comparing our approach and the empirical results to a fully Bayesian analysis with Hamiltonian Monte Carlo sampling as implemented in Stan (Carpenter et al., 2017) would be an interesting future research direction.

References

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27(2)*, 275-297. https://doi.org/10.1080/10705511.2019.1626733

Bentrup, C. (2007). Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2006. Schriftenreihe: Jugendkriminalität in der modernen Stadt - Methoden: Nr. 12. Bielefeld, Münster.

Bentrup, C. (2009). Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2007. Schriftenreihe: Jugendkriminalität in der modernen Stadt - Methoden: Nr. 15. Bielefeld, Münster.

Boers, K., Reinecke, J., Seddig, D., & Mariotti, L. (2010). Explaining the development of adolescent violent delinquency. *European Journal of Criminology*, *7(6)*, 499-520. https://doi.org/10.1177/1477370810376572

Boers, K., & Reinecke, J. (2019). *Delinquenz im Altersverlauf. Erkenntnisse der Langzeitstudie Kriminalität in der modernen Stadt*. Münster: Waxmann.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76(1)*, 1-32. https://doi.org/10.18637/jss.v076.i01

Cho, S., & Lee, J. M. (2018). Explaining physical, verbal, and social bullying among bullies, victims of bullying, and bully-victims: Assessing the integrated approach between social control and lifestyles-routine activities theories. *Children and Youth Services Review*, *91*, 372-382. https://doi.org/10.1016/j.childyouth.2018.06.018

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, *44*, 588-608.

Delsing, M. J. M., & Oud, J. H. L. (2008). Analyzing reciprocal relationships by means of the continuous-time autoregressive latent trajectory model. *Statistica Neerlandica*, *62(1)*, 58-82.

Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software*, *77(5)*, 1-35. https://doi.org/10.18637/jss.v077.i05

Driver, C. C., & Voelkle, M. C. (2018). Hierarchical bayesian continuous time dynamic modeling. *Psychological Methods*, *23(4)*, 774-799.

Driver, C. C., & Voelkle, M. C. (2021). Hierarchical continuous time modeling. In J. Rauthmann (Ed.), *The handbook of personality dynamics and processes* (pp. 887-908). London: Elsevier. https://doi.org/10.1016/B978-0-12-813995-0.00034-0

Engström, A. (2018). Associations between Risky Lifestyles and Involvement in Violent Crime during Adolescence. *Victims & Offenders*, *1(2)*, 1-23. https://doi.org/10.1080/15564886.2018.1503984

Erdmann, A., & Reinecke, J. (2018). Youth violence in Germany: Examining the victim-offender overlap during the transition from adolescence to early adulthood. *Criminal Justice Review*, *43(3)*, 325-344. https://doi.org/10.1177/0734016818761529

Erdmann, A., & Reinecke, J. (2021). What influences the victimization of high-level offenders? A dual trajectory analysis of the victim-offender overlap from the perspective of routine activities with peer groups. *Journal of Interpersonal Violence*, *36(17-18)*, 9317-9343.

Felson, R. B., & Staff, J. (2010). The effects of alcohol intoxication on violent versus other offending. *Criminal Justice and Behavior*, *37*, 1343-1360.

Garofalo, J. (1987). Reassessing the lifestyle model of criminal victimization. In M. R. Gottfredson & T. Hirschi (Eds.), *Positive Criminology* (pp. 23-42). Newbury Park: SAGE Publ.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20(1)*, 102-116.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research*. Online first. https://doi.org/10.1080/00273171.2018.1446819

Hecht, M., Hardt, K., Driver, C. C., & Voelkle, M. C. (2019). Bayesian continuous-time Rasch models. *Psychological Methods*, *24(4)*, 516-537. https://doi.org/10.1037/met0000205

Higgins, G. E., Jennings, W. G., Tewksbury, R., & Gibson, C. L. (2009). Exploring the link between low self-control and violent victimization trajectories in adolescents. *Criminal Justice and Behavior, 36*, 1070-1084.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). New York: Cambridge University Press.

Hindelang, M. J., Gottfredson, M. R., & Garofalo, J. (1978). *Victims of personal crime. An empirical foundation for a theory of personal victimization*. Cambridge, MA: Ballinger.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis. Models of quantitative change*. New York, NY: Academic Press.

Kuha, J. (2004). AIC and BIC. Comparisons of assumptions and performance. *Sociological Methods & Research*, *33(2)*, 188-229. https://doi.org/10.1177/0049124103262065

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

Labouvie, E. W., Pandina, R. J., & Johnson, V. (2016). Developmental trajectories of substance use in adolescence: Differences and predictors. *International Journal of Behavioral Development*, *14*, 305-328.

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, D.C: American Psychological Association.

Montfort, K. van, Oud, J. H. L., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. Cham: Springer International.

Mulford, C. F., Blachman-Demner, D. R., Pitzer, L., Schubert, C. A., Piquero, A. R., & Mulvey, E. P. (2018). Victim offender overlap: Dual trajectory examination of victimization and offending among young felony offenders over seven years. *Victims & Offenders*, *13*, 1-27.

Mustaine, E. E., & Tewksbury, R. (2000). Comparing the lifestyles of victims, offenders, and victim-offenders: A routine activity theory assessment of similarities and differences for criminal incident participants. *Sociological Focus*, *33*, 339-362. https://doi.org/10.1080/00380237.2000.10571174

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 345-368). Thousand Oaks, Calif: SAGE.

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

Osgood, D. W., Wilson, J. K., O'Malley, P. M., Bachman, J. G., & Johnston, L. D. (1996). Routine Activities and Individual Deviant Behavior. *American Sociological Review*, *61(4)*, 635-655. https://doi.org/10.2307/2096397

Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, *65(2)*, 199-215. https://doi.org/10.1007/BF02294374

Oud, J. H. L., & Singer, H. (2008). Continuous time modeling of panel data: SEM versus filter techniques. *Statistica Neerlandica*, *62*, 4-28.

Oud, J. H. L., Voelkle, M. C., & Driver, C. C. (2018). SEM based CARMA time series modeling for arbitrary N. *Multivariate Behavioral Research*, *53(1)*, 36-56.

Peterson, D., Taylor, T. J., & Esbensen, F.-A. (2004). Gang membership and violent victimization. *Justice Quarterly*, *21*, 793-816.

Plass, P. S., & Carmody, D. C. (2005). Routine activities of delinquent and non-delinquent victims of violent crime. *American Journal of Criminal Justice*, *29(2)*, 235-245.

Pyrooz, D. C., Moule, R. K., & Decker, S. H. (2014). The contribution of gang membership to the victim-offender overlap. *Journal of Research in Crime and Delinquency*, *51(3)*, 315-348. https://doi.org/10.1177/0022427813516128

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 263-302). New York: Academic Press.

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

Ryan, O., Kuiper, R. M., & Hamaker, E. L. (2018). A continuous-time approach to intensive longitudinal data: What, why, and how? In K. van Montfort, J. H. L. Oud & M. C. Voelkle, M. C. (Eds.), *Continuous Time Modeling in the Behavioral and Related Sciences* (pp. 27-54). Cham: Springer International.

Schreck, C. J., Stewart, E. A., & Fisher, B. S. (2006). Self-control, victimization, and their influence on risky lifestyles: A longitudinal analysis using panel data. *Journal of Quantitative Criminology*, *22*, 319-340.

Schreck, C. J., Stewart, E. A., & Osgood, D. W. (2008). A reappraisal of the overlap of violent offenders and victims. *Criminology*, *46(4)*, 871-905. https://doi.org/10.1111-/j.1745-9125.2008.00127.x

Seddig, D., & Reinecke, J. (2017). Exploration and explanation of adolescent self-reported delinquency trajectories in the CrimoC study. In A. Blokland & V. van der Geest (Eds.), *The Routledge international handbook of life-course criminology* (pp. 159-178). New York, NY: Routledge.

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, *24(5)*, 637-657. https://doi.org/10.1037/met0000210

Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17(2)*, 176-192.

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2019a). From data to

causes I: Building a general cross-lagged panel model (GCLM). *Organizational Research Methods*, 1-37.

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2019b). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods*, 1-29.

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript via www.mda.gesis.org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - be submitted as MS Word documents.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
  - should be anonymized ("blinded") for review.
- Please also send us an abstract of your paper (approx. 300 words), a front page with a brief biographical note (no longer than 250 words as supplementary file), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - pdf
  - jpg (uncompressed, high quality)
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- Please number the headings of your article. Doing so will make the work of the layout editor easier.

- If your text includes formulas we would like to ask you to upload your text also as a PDF, additional to the Word document.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.
- Response rates in research papers or research notes, where population surveys are analyzed, should be calculated according to AAPOR standard definitions.
- Please follow the APA guideline when structuring and formating your work.

When preparing in-text references and the list of references please also follow the APA guidelines:

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)