These are not the effects you are looking for: Causality and the within-/between-person distinction in longitudinal data analysis

Julia M. Rohrer

Department of Psychology, Leipzig University

Kou Murayama

Hector Research Institute of Education Sciences and Psychology, University of Tübingen & Research Institute, Kochi University of Technology

This manuscript has been accepted for publication in *Advances in Methods and Practices in Psychological Science.*

Author Note

Correspondence concerning this manuscript should be addressed to julia.rohrer@uni-leipzig.de

JMR and KM jointly conceptualized the manuscript. JM wrote the initial draft, both JM and KM reviewed and edited the resulting manuscript.

We thank Henrik Andersen, Ruben Arslan, Tobias Debatin, Ellen Hamaker, Oliver Lüdtke, Brent Roberts, Anne Scheel, Stefan Schmukle, Felix Thoemmes, Satoshi Usami and Manuel Voelkle for their feedback on a draft of this manuscript. We also thank Hayley Jach for proofreading the manuscript.

Abstract

In psychological science, researchers often pay particular attention to the distinction between within- and between-person relationships in longitudinal data analysis. Here, we aim to clarify the relationship between the within- and between-person distinction and causal inference, and show that the distinction is informative but does not play a decisive role in causal inference. Our main points are threefold. First, within-person data are not necessary for causal inference; for example, between-person experiments can inform us about (average) causal effects. Second, within-person data are not sufficient for causal inference; for example, time-varying confounders can lead to spurious within-person associations. Finally, despite not being sufficient, within-person data can be tremendously helpful for causal inference. We provide pointers to help readers navigate the more technical literature on longitudinal models, and conclude with a call for more conceptual clarity: Instead of letting statistical models dictate which substantive questions we ask, we should start with well-defined theoretical estimands which in turn determine both study design and data analysis.

These are not the Effects You are Looking for: Causality and the Within-/Between-Person Distinction in Longitudinal Data Analysis

Daily diary studies, experience sampling, mobile sensing: Technological innovations have made it much easier for psychologists to collect longitudinal data from multiple participants. Accordingly, the number of studies making use of such data has increased steadily (e.g., Hamaker & Wichers, 2017), relevant statistical models have gained prominence, and interest in psychology as an idiographic science has been rekindled (Molenaar, 2004). That is not to say that the idea of assessing a person multiple times is a new one—"occasions" constitute one of the three axes of Cattell's well-known "data cube" (the other two being "persons" and "variables", Cattell, 1952)—but empirical research is finally catching up with a dimension that has always been considered conceptually important.

With the increased amount of longitudinal data available, recent studies have paid particular attention to "disentangling" within- and between-person associations (e.g., Curran & Bauer, 2011; Hamaker et al., 2005; Voelkle et al., 2014). For example, a positive association between talkativeness and subjective well-being may exist on the between-person level—people who are (on average) more talkative than others are (on average) happier than others—or it may exist on the within-person level—people who are more talkative today (than they usually are) are happier today (than they usually are). Between- and within-person associations can be statistically independent (i.e., they can take on different values, or even opposite signs), and it is the latter within-person associations that psychologists often deem more interesting as they are meant to inform us about "within-person processes" (e.g., Molenaar & Campbell, 2009). In line with traditions of the field (Grosz et al., 2020), however, the psychological literature on within-person data often shies away from explicitly interpreting such processes as causal effects; and

only recently have authors tried to explicitly bridge the gap with the causal inference literature (e.g., Gische, West, & Voelkle, 2021; Lüdtke & Robitzsch, 2022; Voelkle, Gische, Driver, & Lindenberger, 2018). However, just because the term "causal" was not used does not mean that it was not implied all along. For example, Curran and Bauer (2011) invoke the following "within-person process": when an individual engages in effective coping, this mitigates the effects of stress for them. The most plausible reading of this "process" is that effective coping has a *causal* effect on various relevant outcomes for the individual.

In some parts of the psychological literature, concerns about the within/between-person distinction have taken center stage. Here, we are going to argue that we should change gears and put causal inference upfront when planning to collect and analyze longitudinal data. The within/between distinction only plays an instrumental role in this endeavor, and we make three points to clarify its substantive utility. First, it is not necessary to investigate within-person associations to identify causal effects—other designs can do so, too. Second, it is not sufficient to investigate within-subject associations to identify causal effects—confounding can still be an issue. Third, while longitudinal within-person data are neither necessary nor sufficient for causal inference, they still can be tremendously helpful. They can aid causal identification, allowing us to relax some assumptions; they can inform us about interindividual differences in causal effects; and they can give us a more dynamic view of how effects unfold over time. We conclude with some recommendations for how to approach longitudinal data analysis from a causal inference perspective. With this manuscript, we aim to provide some general, non-technical guidance for researchers who may have some initial experience with longitudinal data analysis but are less familiar with the causal inference literature, who would like to understand how those two topics

are connected, and who are looking for entry points into the more technical literature on causal inference and statistical models of longitudinal data.

Within-Person Data Are Not Necessary for Causal Inference

In the potential outcomes framework (Holland, 1986; Rubin, 1974), individual causal effects are defined as differences in potential values of an outcome variable (Y) under different treatments (A). We will start with an example that reflects a "typical" question that researchers may aim to answer with observational within-person data: does talkativeness, a behavior associated with the personality trait extraversion, increase subjective well-being? As a starting point for a formalization of this effect, we focus on a single unit (for our purposes, an individual), measured without error, and we consider the case of a binary independent variable to simplify matters (see West & Thoemmes, 2010 for a more comprehensive introduction). Here, Y may be an individual's subjective well-being at the end of today, and A may refer to the treatment of spending the day being talkative (A = 1) as opposed to spending the day being untalkative (A = 0; we will return to the actual realization of such a treatment below). If the person was talkative, $Y^{a=1}$ would be observed; if the person was untalkative, $Y^{a=0}$ would be observed. These two values ($Y^{a=1}$, $Y^{a=0}$) are the individual's potential outcomes. The individual's causal effect of being talkative today on their subjective well-being at the end of the day is defined by the contrast between the two, $Y^{a=1} - Y^{a=0}$. This individual causal effect is unobservable. Today, the individual will have *either* been talkative or untalkative, so only one of the individual potential outcomes ($Y^{a=0}$ or $Y^{a=1}$) will be realized as Y and become observable. How could we possibly recover it from between-person data?

In virtually all circumstances, we cannot. With between-person data, the *individual* causal effect is out of reach. However, if we collect data from multiple people for a single day, it can

become possible to estimate the *average* of their individual causal effects. This works best in randomized experiments, and most psychological researchers will be aware of the special status of this research design. Nonetheless, it is worth spelling out the details to clarify some terminology and crucial assumptions.

Consider the possibility of a randomized experiment in which we assign a large number of people to either spend the day being talkative or spending the day untalkative. For the sake of the argument, let's pretend there was a psychological intervention that manipulates individuals' talkativeness in a highly reliable and targeted manner for the duration of a single day, rendering them either talkative or untalkative to an exactly pre-specified degree, without any unintended side effects.

Prior to this hypothetical intervention, individuals' "natural" talkativeness may be correlated with the potential outcomes. For example, people who are more talkative might be those who are happier *generally*, meaning that both of their potential outcomes are higher. But after the randomized treatment, the assigned talkativeness will not be correlated with the individuals' potential outcomes $Y^{a=0}$ and $Y^{a=1}$; this means that the two treatment groups are exchangeable with respect to their potential outcomes. Thus,

$$E[Y^{a=1} - Y^{a=0}] = E[Y^{a=1}] - E[Y^{a=0}] \quad (1)$$

$$= E[Y^{a=1}|A = 1] - E[Y^{a=0}|A = 0] \quad (2)$$

$$= E[Y|A = 1] - E[Y|A = 0]. \quad (3)$$

The expected value of the individual level causal effect across individuals is the arithmetic mean; the expected value of a difference is equivalent to the difference between the expected values (1). As the groups are exchangeable with respect to their potential outcome, their

expected potential outcomes will not systematically vary; we can thus substitute the expected potential outcomes across all individuals with the expected potential outcomes in the respective groups (2). In the two different groups, the respective potential outcomes are realized (3). The assumption to ensure the equivalence between (2) and (3) is called consistency, and we will discuss potential violations later.

Thus, by taking the difference between the group means of the outcome, we recover the *average* of the individual causal effects, which is often called the average treatment effect (ATE). This fundamental property is what renders randomization such a valuable tool. And it clearly demonstrates that between-person data can inform us about within-person processes—albeit only in aggregate.[1]

Going beyond experimental data, average causal effects can also, at least in theory, be estimated with the help of non-experimental between-person data (including natural experiments). Such attempts require strong additional assumptions (for accessible introductions, see e.g., Elwert, 2013; Hernán & Robins, 2010; Pearl et al., 2016; Rohrer, 2018; Rosenbaum, 2017) which may range in their plausibility from defensible to untenable, depending on the application. How does this mesh with the often-emphasized fact that between- and within-person associations are statistically independent (e.g., Schmitz & Skinner, 1993)? Even though these associations are sometimes also labeled between- and within-person *effects*, they do not refer to causal quantities. The between- and within-person associations are a combination of (1) non-

[1] If we additionally assume that the causal effect is precisely the same for every individual, then this is a non-issue because the average causal effect will equal every single individual level causal effect (e.g., West & Thoemmes, 2010). However, many psychologists would probably object to this assumption.

causal associations between the two variables of interest (e.g., associations induced by confounders) and (2) causal associations (induced by effects flowing either way).

Considering (1), one of the reasons why between- and within-person associations may diverge is that some confounders—time-invariant factors, such as stable socio-demographic variables, but also stable personality traits—will only affect the between-person association but not within-person associations. This is because within-person association are calculated based on time-by-time fluctuations within a person, and time-invariant factors do not have any variance within persons.[2] In practice, this means that between-person associations (e.g., between talkativeness and happiness) can often plausibly be explained away by time-invariant third variables (e.g., gender, age, childhood socioeconomic status, stable personality traits). In contrast, within-person associations cannot be explained away by third variables that are stable and have constant effects for the duration of the data collection.[3]

Considering (2), another reason why within- and between-person associations may diverge is scenarios in which causal associations between the variables of interest cannot be captured by within-person associations. There may be a lack of within-person variability in the independent variable of interest over the course of the study; or the causal effects may unfold over a time frame longer than the duration of the study. We will return to the issue of time frame when discussing design parameters.

[2] Sometimes people define within-person associations as computed by mean-centered data of multiple persons (see section "Within-Person Data Can Be Very Helpful for Causal Inference" and Box 1). With this definition, within-person associations can suffer from time-invariant confounders that have varying effects over time (Usami et al., 2019). In this scenario, within-person associations are protected *only* from time-invariant confounders with constant effects, but not from time-invariant confounders with time-varying effects.

[3] This means that it can also be, in theory, possible to recover average within-person associations from between-person associations when the relevant time-invariant confounders are adjusted for, just like average causal effects can potentially be recovered from between-person data (Murayama et al., 2017); but again very strong assumptions are necessary.

Within-Person Data Are Not Sufficient for Causal Inference

Now we have seen that within-person data are not *necessary* for causal inference. But they are not *sufficient* either: longitudinal data on their own do not justify causal inferences. The reason for this is, once again, confounding. As discussed above, within-person associations are not affected by time-invariant confounding factors with constant effects. However, they can still be influenced by time-varying confounding factors.

Let us assume we had intensive within-person data of individuals' talkativeness and subjective well-being. This allows us to compare days on which they were talkative to days on which they were untalkative. But talkativeness was not randomized, so it is possible that the treatment (being talkative vs. untalkative) is correlated with the potential outcomes (potential well-being that day).

For example, social events (such as dates or parties) may both affect talkativeness as well as happiness. Let us first consider the "contemporaneous" association between the reported level of talkativeness on a given day and happiness at the end of that day. This association will be confounded because talkative days are not exchangeable—they are days on which more social events happened, and those events alone may be sufficient to make one happier.

Next, let us consider the lagged association between the reported talkativeness on a given day (day 1) and happiness at the end of the next day (day 2). We might think that confounding by social events is no longer an issue when we adjust for happiness on day 1, as this may already capture the confounding influence of social events. However, this depends on the time course over which the causal effects of social events unfold. If they immediately induce talkativeness (small talk at the party) but affect happiness more slowly over multiple days (the warm, ongoing

glow after having reconnected with old friends), then they will end up confounding the lagged associations as well. And in this particular substantive example, the lagged effect may not be particularly informative, to begin with—if we assume that the affective benefits of talkativeness are reaped immediately, we would expect them to be mostly captured in happiness on the same day rather than the next day.

In short, causal inferences on the basis of within-person data still rest on the assumption that all (time-varying) confounders have been appropriately adjusted for.

Within-Person Data Can Be Very Helpful for Causal Inference

As we have explained, both between- and within-person data require strong assumptions to warrant causal inference. However, the use of within-person data allows us to relax certain assumptions. Thus, although we still need to think clearly about the remaining assumptions, within-person data can aid causal inference.

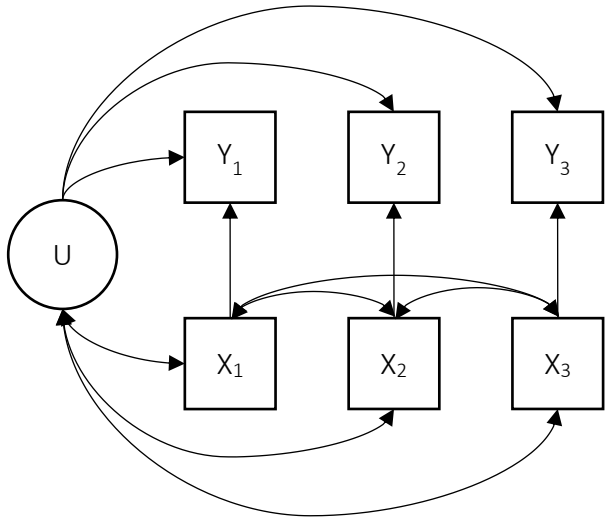
As stated above, within-person associations are in principle not affected by time-invariant confounders. Within-person data thus have the potential to control for various types of between-person confounding, including *unobservable* confounders. One of the simplest ways to do this is through so-called fixed effects models (Hamaker & Muthén, 2020; Imai & Kim, 2019; McNeish & Kelley, 2019). While fixed effects models are not very commonly used in psychological research, the approach is conceptually equivalent to the procedure in which variables are mean-centered within persons before entering them into multilevel models (e.g., Hamaker & Muthén, 2020; McNeish & Kelley, 2019). In Box 1, we go into detail about the assumptions under which the fixed effects model can identify contemporaneous causal effects. Importantly, this model focuses on the contemporaneous effect of X on Y, and not on their broader causal dynamics.

Box 1: The Fixed Effects Model

The fixed effects approach (or alternatively within-person mean centering) can control for unobserved time-invariant confounders whose effects do not change over time. For example, when considering the effects of talkativeness on happiness, extraversion (a stable personality trait) may be such a confounder: extraverted individuals are habitually more talkative, but extraverted individuals may also simply be dispositionally happier. Figure 1, which has been adapted from (Hamaker & Muthén, 2020, p. 367) shows the causal model underlying the standard fixed effects model. Note that this model focuses only on the (contemporaneous) effects of X on Y, and X is treated as exogenous (i.e., the model does not impose constraints on the causal relationship between X variables and U).

Figure 1

Causal Graph Underlying the Fixed Effects Model



When does this model successfully identify the effects of X on Y? As indicated by the *absence* of certain arrows in model 1, we have to assume that there are no lagged causal dynamics, for example: past happiness does not affect current talkativeness (no cross-lagged paths from Y to X), past talkativeness does not affect current happiness (no cross-lagged paths from X to Y), past happiness does not affect current happiness (no auto-regressive paths among the Y). Such dynamics will bias estimates, although the standard model can be modified to partially relax assumptions (Imai and Kim, 2019).

Another common scenario in which a fixed effects model is biased occurs if people vary in their change over time (i.e., heterogeneous slopes) and if these differences are related to the effect of interest; this can be addressed by another modification of the model (Rüttenauer & Ludwig, 2020). Lastly, we have to assume that any time-varying confounder has been included in the model, which is also the case for the following models (Box 2, Box 3).

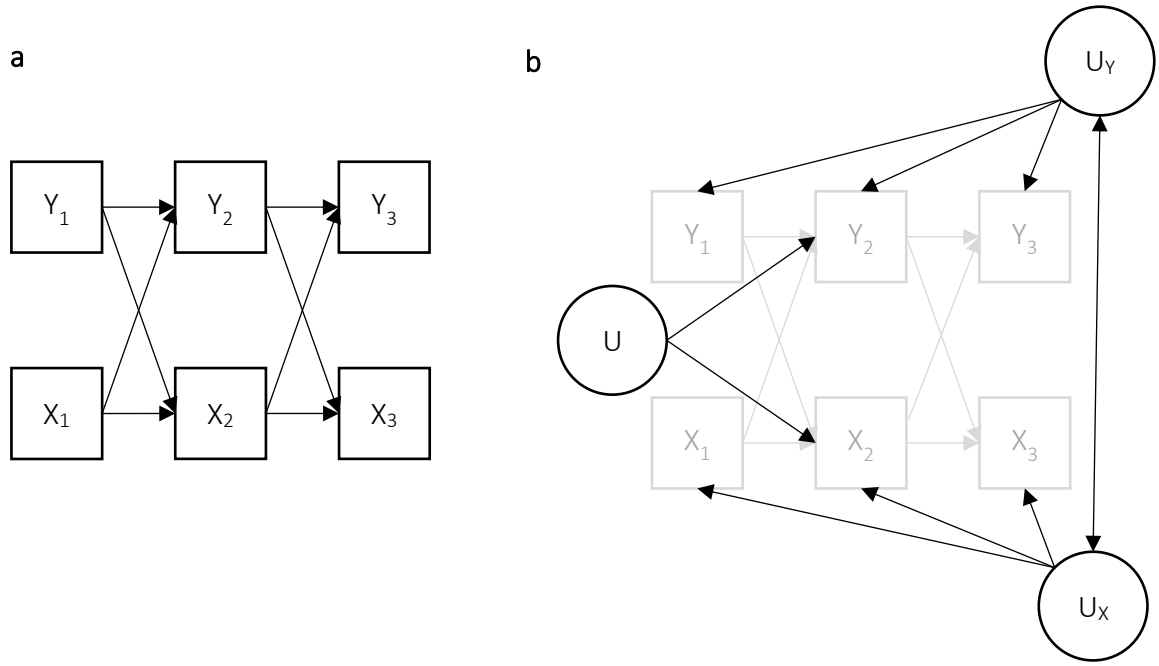
The fixed effects model only considers within-person changes over time. In our example, the resulting effect estimate would only be informed by those people who actually do experience some changes in their talkativeness over the course of the study.

Psychologists are, of course, often interested in estimating precisely these causal dynamics, which may explain why models including reciprocal effects are so much more popular in the psychological literature. Can we identify such reciprocal causal effects in longitudinal data? Recent literature has suggested that the most widely-used form of such models, the cross-lagged panel model (Box 2), does not sufficiently control for between-subject confounding despite its use of longitudinal data (Hamaker et al., 2015).

Box 2: The Cross-Lagged Panel Model

Figure 2

(a) Causal Graph Underlying the Cross-lagged Panel Model. (b) Different Scenarios Involving Unobserved Confounders.



The cross-lagged panel model (Figure 2) has different aims than the fixed effects model. First, it aims to identify *lagged* effects (not the contemporaneous effects examined in fixed effects model). Second, it usually aims to identify *reciprocal* effects (i.e., X influences Y *and* Y influences X). The model addresses so-called Granger causality (Granger, 1969), but Granger causality is a form of prediction—and such prediction only implies causation when certain assumptions are met. For example, the model provides biased causal estimates when there are contemporaneous causal effects (e.g., current happiness affects current talkativeness). This highlights trade-offs when trying to simultaneously consider contemporaneous and lagged effects which are also discussed by Imai and Kim (2019).

The cross-lagged panel model can partly account for unobserved confounders. For example, when we are interested in the causal effect of X_2 on Y_3, the existence of the unobserved confounder U, which only has a temporal effect, is unproblematic (Figure 2, Panel b): the

confounding goes through Y_2 which is included in the model and is thus statistically accounted for. However, the model fails if the constructs are trait-like in nature. The existence of more stable effects U_X and U_Y (Figure 2, Panel b) would be problematic since they directly open a confounding path between X_2 and Y_3. Thus, the resulting estimates of the cross-lagged paths would be biased.

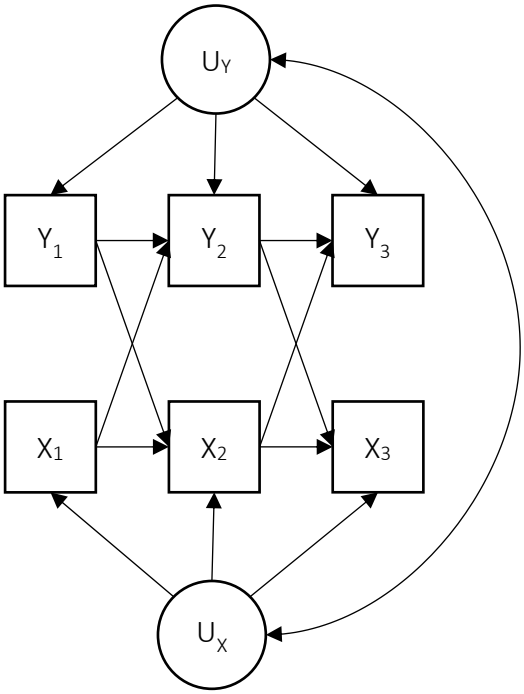
The shortcomings of the cross-lagged panel model have prompted researchers to use modifications of the model to separate such between-subject confounding from cross-lagged effects. In psychology, the random-intercept cross-lagged panel model (Hamaker, Kuiper, & Grasman, 2015) has become the most popular choice, but similar models have been proposed in other fields. In Box 3, we highlight one of such models, the dynamic panel model, which is a combination of the fixed effects model with the cross-lagged panel model.

Importantly, whether these modified approaches are sufficient to adjust for all time-invariant confounders still depends on additional assumptions about the precise nature of the confounding (e.g., Lüdtke & Robitzsch, 2022; Murayama & Gfrörer, 2022). Furthermore, models are usually unable to identify both contemporaneous and lagged effects simultaneously. This highlights that there is no “one-size-fits-all” procedure to enable causal inference. Instead, we need to be very clear about the type of causal effects we want to examine (e.g., lagged effect vs. contemporaneous effects), and we need to carefully evaluate the underlying assumptions. However, even if those assumptions may be deemed unrealistic, observational longitudinal data combined with an appropriate model may often provide answers that are “less wrong” (i.e., potentially less biased) than answers provided by observational cross-sectional data, all else being equal.

Box 3: The Dynamic Panel Model

Figure 3

Causal Graph Underlying the Dynamic Panel Model



Dynamic panel models exist in several different versions, but we focus on the model depicted in Figure 3. Note that Figure 3 omitted some details for the purpose of simplicity, but more detailed practical tutorial of the model can be found in Dishop and DeShon (2021). This dynamic panel model has the same goal as the cross-lagged panel model—it aims to identify lagged reciprocal causal effects. Like the fixed-effects model, it takes into account (constant) effects of time-invariant confounders; like the cross-lagged panel model, it allows for reciprocal lagged dynamics. While the model can control confounders that cross-lagged panel model cannot (i.e., time-invariant confounders), important assumptions of the more basic models (Box 1, Box 2) still apply: Like in the fixed-effects model, we need to make

assumptions about the type of time-invariant confounders (i.e., no heterogeneous slopes). And like in the cross-lagged panel model, the existence of contemporaneous causal effects would bias our estimates of the cross-lagged causal effects.

Helpful Further Readings

Maybe due to psychology's fraught relationship with causality (Grosz et al., 2020), the literature on the many models discussed in our field—such as varieties of change-score models, cross-lagged models, and latent curve models—is unfortunately not always transparent with respect to the assumptions under which these models can successfully identify causal effects. However, more recently, researchers have tried to bridge the gap between longitudinal data modeling in psychology and causal inference.

Gische, West and Voelkle (2021) introduce graphical causal models for researchers familiar with structural equation modeling and the cross-lagged panel design; and Voelkle et al. (2018) provide a more general discussion of the role of time for understanding psychological mechanisms, which they quite explicitly describe as a sequence of causal effects. Usami et al. (2019) provide a discussion of the causal assumptions underlying popular models. Both Andersen (2021) and Lütdke and Robitzsch (2022) discuss different classes of longitudinal models with respects to the conditions under which they recover the (cross-lagged) causal effects of interests, and conditions under which they will result in equivalent results. Lastly, Zyphur et al. (2020; 2020) develop a comprehensive general cross-lagged panel model as a generic approach to translate assumptions into a statistical model.

Of course, other fields have also tackled the issue of causal inference with longitudinal data. For example, the sociologists Elwert and Pfeffer (2019) developed an approach that uses future values of the independent variable to detect and reduce omitted variable bias. In epidemiology, the particularly promising approach of marginal structural models (Cole & Hernán, 2008; Robins, Hernán, & Brumback, 2000) has been developed. These models implement a multi-step estimation procedure to control for time-varying confounding variables (Williamson & Ravani, 2017). The promise of such models for causal inference in psychology, however, has not yet been well recognized (Lüdtke & Robitzsch, 2020; Usami, 2020). Thoemmes and Ong (2016) provide an introduction to marginal structural models in combination with inverse probability weighting as a means for third-variable adjustment in longitudinal data, including annotated SPSS and R code for psychologists. The tutorial by Bray et al. (2006) showcases an implementation in SAS and highlights how this method can, unlike other common methods, successfully adjust for time-varying confounders. Lastly, VanderWeele, Mathur and Chan (2020) have developed a comprehensive template for so-called outcome-wide longitudinal designs, in which the goal is to identify the causal effects of an independent variable on a number of outcome variables, and longitudinal data is leveraged to reduce concerns about confounding.

Additional Advantages of Longitudinal Data

Aside from causal identification in the narrow sense (getting rid of confounding), which often focuses on average effects, longitudinal data may also enhance causal inference for other reasons. First, longitudinal data can improve our understanding of how causal effects unfold over time (Voelkle et al., 2018). Second, they may provide the means to actually estimate individual-level causal effects. Causal effects may vary between individuals, and we can take into account such between-person variability of causal effects with longitudinal data. The optimal approach to

identify such effects are experiments in which we observe individuals repeatedly in different experimental conditions (see Fine Point 2.1, Hernán & Robins, 2020, p. 16), and there have been some recent methodological developments to obtain a better causal estimate with this type of design (Schmiedek & Neubauer, 2020). Doing this with observational longitudinal data once again requires more and stronger assumptions, and this is an important avenue for future methodological work.

Making the Most of our Within-Person Data

In recent years, we have observed considerable enthusiasm for the within-person approach in psychology, with various advanced statistical models proposed. In line with this enthusiasm, we believe that within-person data is a promising way to advance causal inference. Yet we also feel like its promises have led people to put the technological and methodological cart before the conceptual horse. Researchers may decide to collect within-person data with an ESM study because it is the innovative thing to do right now; they may decide to apply certain statistical models because they appear novel and highly sophisticated. Journals may further implicitly reinforce this style of research when they automatically dismiss studies that are "merely cross-sectional" or do not employ "sophisticated statistical modeling." An approach that we believe to be more productive, and which we describe in the following, puts the substantive question first. While this may sound trivial—of course, the substantive question should be the starting point of any empirical investigation—debates such as the one surrounding the age trajectory of happiness (Kratz & Brüderl, 2021) and the interpretation of the "Many Analysts" project (Auspurg & Brüderl, 2021; Silberzahn et al., 2018) highlight how arguments often focus on statistical aspects when in fact researchers do not even agree about which substantive question is being addressed.

Setting the Analysis Goal

Researchers should start by explicitly spelling out the theoretical estimand of interest in precise terms that exist outside of any statistical model (Lundberg et al., 2020). At this point, it may become clearer whether the research question targets causal quantities or not—but even non-causal endeavors, such as "description of developmental trajectories", require conceptual clarity. This estimand, in combination with the additional assumptions researchers are (or are not) willing to make, determines which research design is appropriate, be it experimental or non-experimental, cross-sectional or longitudinal, needing many time points or not.

What does such a well-defined estimand look like? Psychologists like to make claims about broad concepts (Yarkoni, 2020) and to address broad research questions ("what is the interplay between talkativeness and happiness?"). However, from a causal inference perspective, things need to be broken down and taken more slowly (see also Rohrer, Hünermund, Arslan, & Elson, 2022). For example, a more tractable research question may concern the effect of being continuously talkative (as opposed to continuously untalkative) for a certain defined amount of time on well-being immediately after the episode. Formalization, for example with the help of the potential outcomes model, makes it explicit that causal effects are defined by contrasts of specific treatments on specific outcomes. Treatments may be time-varying—there are many different sequences of talkativeness and untalkativeness that one could contrast to learn something about the effects of talkativeness on happiness (see Hernán & Robins, 2020, Chapter 19 for an introduction to time-varying treatments)—and outcomes can be evaluated at different points in time.

Thus, there is no such thing as "the" effect of talkativeness on happiness; there are many different possible theoretical estimands that may all be deemed informative with respects to the

overarching question of whether and how talkativeness affects happiness. This nuance may get lost if we simply apply an out-of-the-box model, which may target a different estimand than the one we actually have in mind. Furthermore, insufficient clarity about estimands may lead to ostensible contradictions between empirical studies which in reality target different estimands.

Once we try to be more precise about the causal effects we have in mind, we may run into deeper issues. Theories in psychology are often quite underspecified—a topic that was addressed in a recent special issue of the journal *Perspectives on Psychological Science* (Volume 16 Issue 4, July 2021)—and may thus only make vague predictions about which causal effects are to be expected. But even if theories were more precise, causal effects involving psychological variables pose conceptual obstacles. While these are neither unique nor central to the within-/between-distinction and thus outside of the focus of the present article, we provide more details in Box 4 for interested readers.

Box 4: Hypothetical Interventions, Real World Complications

Earlier, we alluded to a hypothetical intervention that fixes an individual's talkativeness at a given level, without any side effects. Such an intervention does not exist—no psychological intervention will achieve precisely the desired level of talkativeness for everyone. Furthermore, the intervention may affect all sorts of other variables, and some of these may in turn explain any effect of the intervention on well-being. In general, psychological variables as causes pose challenges as interventions targeting them are often "fat-handed" (Eronen, 2020), meaning that they will affect multiple variables simultaneously. This not only constrains experimentation but also makes it hard to pin down which hypothetical states of the world we have in mind when estimating causal effects on the basis of

observational data. For example, we likely do not encounter many situations in which individuals' talkativeness could have varied *while all other psychological variables were held constant*.

And there is a second concern that goes beyond this. The effects of talkativeness may depend on how talkativeness was induced (e.g., by an individual's genetic disposition, by a specific situation, or by a psychological intervention). In such a scenario, does it even make sense to talk about effects of talkativeness *per se*? An example from a different field of research may illustrate the matter more clearly. Does obesity shorten life? If we take a particular individual, there may be many different ways to intervene on their body weight. For example, we may put them on a specific diet or a specific exercise regime, or we may chop off a body part. Any of these will affect body weight, but how this change in body weight subsequently affects mortality may vary between interventions. Formally speaking, this violates the assumption of consistency (i.e., the assumption needed to ensure the equivalence of equations (2) and (3) described earlier), and Hernán and Taubman (2008) go so far as to state the effects of BMI on mortality in observational data cannot be well-defined. In contrast, the effects of specific interventions on BMI can be well-defined, and can at least potentially be recovered from observational data. In line with this, Hernán and Robins (2016) argued that we should conceptualize (hypothetical) target trials to precisely define which causal effects we are interested in.

However, this "interventionist" account of causality has been challenged. Pearl (2018) champions a structural account of causality in which causal relations exist independently of hypothetical interventions. In this account, consistency is not an assumption but a theorem.

Actual interventions may often have side effects that need to be reckoned with, but these do not render causal effects inconsistent.[4]

We do not aim to solve this philosophical debate, but would like to highlight that degree of concreteness in variables/constructs renders causal inference more or less challenging, and psychological variables tend to be less concrete. "Obesity" as an independent variable is much more concrete than concepts like "talkativeness" or "subjective well-being" (Rohrer & Lucas, 2020). Assuming an interventionist account of causality, problems may arise because we cannot even come up with hypothetical targeted interventions, or because consistency fails (the effects of talkativeness may vary depending on whether talkativeness is induced through drugs or through verbal encouragement), resulting in ill-defined causal effects. Assuming a structural account of causality, problems may arise because we lack knowledge of the structure of the causal web linking psychological variables (such as talkativeness, extraversion and other personality traits). In this case, effects are not necessarily ill-defined, but estimating them may still be virtually impossible given the current state of knowledge.

Identification Strategy and Design Parameters

Once we have settled on an estimand, we can start thinking about appropriate identification strategies. Accessible articles provide some guidance on this step (Foster, 2010a; Grosz et al., 2020). Considerations may include whether or not a sufficiently targeted

[4] To render matters even more complex, there are more than two sides in this debate about the nature of causal inference. For example, Krieger and Smith (2016) highlight the limits of both counterfactuals thinking (championed by Hernán) and of directed acyclic graphs (as championed by Pearl) and propose a broader and more flexible framework of "inference to the best explanation."

intervention is available to manipulate talkativeness (but see Eronen, 2020) and thus whether an experiment is plausible, whether a suitable natural experiment may exist (e.g., a situation that affects talkativeness in a plausibly random manner), and which time-invariant or time-varying confounders are deemed relevant.

If within-person data turn out to be a productive way forward—for example, because time-invariant confounders are deemed particularly relevant, and because we can assume that there is relevant within-person variability in the independent variable of interest—the causal angle can clarify specific design parameters. Consideration of potential time-varying confounders tells us what needs to be measured. Consideration of the precise definition of the causal effect of interest tells us which time lag between assessments is sensible.

Discussions of the appropriate time lag in psychology often focus on attempts to uncover the true underlying dynamic system (Haslbeck & Ryan, 2021) which is, of course, unknown. Hence, one might conclude that the narrowest possible sampling is desirable, as it still allows one to estimate effects with a wider lag (for example, in the crudest case, one might just drop the measurement points in between). In practice, when it comes to time lags, pragmatic concerns need to be considered as well. High-frequency sampling can overburden participants, and there is a very real possibility that the assessment interferes with the causal system of interest. Self-reporting positive affect one hundred times a day may influence mood; filling out a personality questionnaire over and over again may change the way individuals answer the items. Thus, the smallest possible time lag is not always advisable. But if we decide that we want to investigate a relatively well-defined specific causal effect, such as “the *immediate* effect of picking up one’s smartphone on well-being”, or “the effect of cumulative smartphone usage over the course of a

day on well-being on the next day", the research question already implies how data needs to be collected.[5]

Statistical Estimation

If the theoretical estimand is set, the aim of the statistical analysis is to provide an actual empirical estimate. We have already extensively referred to the longitudinal modeling literature above and will thus just briefly emphasize a central concern. Psychological researchers have often relied on out-of-the-box longitudinal models such as cross-lagged panel models, which could, in principle, be applied to any pair of variables. Such default solutions have multiple shortcomings. First, it should be noted that currently, the psychological literature on within-person associations is not well integrated with the causal inference literature—thus, for at least some of the out-of-the-box solutions, it is unclear or at least intransparent which causal effect is targeted by the analysis. Second, the causal webs linking different sets of variables can look very different, and a model that is not tailored to the specific underlying causal web cannot recover the causal effects of interest. Third, many published implementations of these models pay less attention to including measured (time-invariant and time-varying) confounders, further limiting the chances that the causal effects of interest will be recovered.

Interpreting Model Results

Lastly, once the model has been estimated—how do we interpret it? In the psychological literature, model coefficients associated with particular paths are often treated as the relevant analysis output. But even if we are lucky and our coefficients correctly identify the causal

[5] Continuous time modeling (e.g., Driver et al., 2017) can be a useful tool to identify an underlying continuous process regardless of the particular time lags, and this information may be useful to determine an optimal lag for a new research design.

processes of interest, they may still not provide a straightforward answer to our (causal) research question. For example, the model-implied effects of X at a given point in time on Y at a given point in time may vary between individuals, either because of systematic heterogeneity (e.g., interactions) or because of non-linearity (see Rohrer & Arslan, 2021 for a discussion of why in non-linear models, "everything interacts"). Furthermore, in a cross-lagged panel model, the effect of a predictor on the outcome at a later time point can be the sum of both direct and indirect effects (Lüdtke & Robitzsch, 2022), so that multiple coefficients need to be added up when contrasting different states of the world.

Here again, causal thinking can clarify how to summarize effects in complex models in an interpretable manner. We can once again consider a hypothetical intervention and use the model to predict how it would affect individuals' outcomes at a point in time we consider relevant. Gische, West, and Voelkle (2021) demonstrate how to work with such hypothetical interventions in the context of cross-lagged panel models and how to calculate both average and person-specific effects.

The general framework for using models to make predictions about the effects of various interventions are so-called marginal effects. Marginal effects have received comparatively little attention in psychology outside of methods journals; they are more common in, for example, sociology (e.g., Mize, Doan, & Long, 2019), possibly because the statistical software Stata popular in that field makes calculating them quite easy (Williams, 2012). However, there are now packages available that allow researchers to calculate marginal effects in R based on structural equation models (Mayer, 2019; Mayer, Dietzfelbinger, Rosseel, & Steyer, 2016), as well as based on a vast number of other model classes including multilevel models (Arel-Bundock, 2022; Lenth, 2022). As far as we know, no comprehensive primer to marginal effects

for psychologists has been published to date, but recent blog posts try to provide a gentle introduction (Heiss, 2022; Rohrer, 2022).

Shouldn't We Just Do Description Instead?

Having read so far, readers may feel that inferring causal effects from (non-experimental) within-person data is an overwhelming task and may instead prefer a "descriptive" approach. Indeed, many longitudinal analyses claim to be descriptive in nature, although that term may be used in an ambiguous manner. This may partly be a strategic move to avoid the heightened scrutiny that results from overtly causal claims (Alvarez-Vargas et al., 2020; Grosz et al., 2020). We do believe that descriptive research is currently undervalued in psychology (see e.g., Scheel et al., 2020). But many models in psychology are too complex to produce good descriptions (Foster, 2010b), and this holds true for longitudinal models in which the explanation for how coefficients behave quickly turn opaque.

An actually informative descriptive analysis should involve much more *basic* description than we routinely encounter in studies analysing longitudinal data. For example, a fruitful first step to describe associations in longitudinal data may consist of a fixed effects model in its most basic specification, or the equivalent multi-level model with within-person centering. This tells us the strength of the contemporaneous association between the variables, after removing stable between-person differences in the level of the predictor and the outcome. With sufficient data points per individual, we may even simply calculate the bivariate association for every individual in isolation. As we discussed earlier, any association in such analyses may still be confounded and thus does not necessarily provide a convincing causal estimate, but at least we have narrowed down the range of confounders while sticking with estimates that still have a straightforward descriptive interpretation. Reporting results from such analyses before moving

on to more complex models also mirrors established practices in cross-sectional studies, in which authors routinely report bivariate correlations before moving on to more complex regression analysis.

In contrast, it is not easy to apply the "standard" longitudinal models for descriptive purposes. The literature is filled with many related "state-of-the-art" longitudinal models (e.g., the autoregressive latent trajectory model, Bollen & Curran, 2004; the random-intercept cross-lagged panel model, Hamaker et al., 2015; the stable trait/auto-regressive trait/state model, Kenny & Zautra, 1995; the dual-change score model, McArdle & Hamagami, 2001) and these in turn can usually be specified in multiple ways. Assuming we chose the right model that correctly reflects the data-generating mechanism, we could elegantly capture the underlying causal within-person dynamics. But in reality, we do not know which model generated the observed data, and presented with a daunting number of different models and little guidance which nuances matter and which don't, we may resort to the standard that is accepted in the field—and that might not be optimal, as we know from the story of cross-lagged panel models (Hamaker et al., 2015). Trying to uncover the complete causal dynamics of a system is a more ambitious task than identifying a specific causal effect; once we fail to uncover the complete dynamics, the interpretation of any specific component of the model becomes questionable.

Putting causal inference upfront, we are still confronted with a challenging task, but one that is potentially more tractable because there is at least a clear circumscribed analysis goal: recovery of a specific causal effect. This also opens the possibility to use available experimental evidence as a benchmark to evaluate our observational longitudinal analysis—if a longitudinal model implies certain effects that contradict existing evidence from intervention studies targeting similar cause and effect, this can at least be taken as a warning sign (see Wan, Brick, Alvarez-

Vargas, & Bailey, 2021 for an implementation of this logic). The choice of the data-analytic model only matters insofar as it should map onto the assumptions about the underlying causal web that we are willing to make. These assumptions will often be strong and potentially unrealistic—there is no free lunch—but at least we are actually tackling the question of interest.

Consider, for example, the debate surrounding the age trajectory of happiness (Galambos et al., 2020). This actually seems to be one of the easier questions one could answer with longitudinal data, yet it has spawned a bloated literature and lots of confusion about how to specify the model. If we tackle the problem from a causal inference perspective, as demonstrated by Kratz and Brüderl (2021), it becomes clear that some analytic decisions are just wrong (e.g., statistical adjustment for mediators, which only makes sense if we are trying to address a *different* research question), whereas others hinge on additional assumptions (e.g., about the existence and shape of period and cohort effects). This does not mean that the debate is automatically settled, but at least we can pinpoint where exactly analysts disagree and how to make progress on the research question.

We believe that a better understanding of causal inference, and how it can be enhanced with the help of within-person data, has the potential to clarify other debates in psychological research as well, resulting in an overall improvement of the quality of our inferences.

References

Andersen, H. K. (2021). Equivalent approaches to dealing with unobserved heterogeneity in cross-lagged panel models? Investigating the benefits and drawbacks of the latent curve model with structured residuals and the random intercept cross-lagged panel model. *Psychological Methods*. doi:10.1037/met0000285

Arel-Bundock, V. (2022). marginaleffects: Marginal Effects, Marginal Means, Predictions, and Contrasts. https://vincentarelbundock.github.io/marginaleffects/

Auspurg, K., & Brüderl, J. (2021). Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project. *Socius*, *7*, 23780231211024420.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive Latent Trajectory (ALT) Models A Synthesis of Two Traditions. *Sociological Methods & Research*, *32*(3), 336–383.

Bray, B. C., Almirall, D., Zimmerman, R. S., Lynam, D., & Murphy, S. A. (2006). Assessing the total effect of time-varying predictors in prevention research. *Prevention Science: The Official Journal of the Society for Prevention Research*, *7*(1), 1–17.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*(6), 656–664.

Dishop, C. R., & DeShon, R. P. (2021). A tutorial on Bollen and Brand's approach to modeling dynamics while attending to dynamic panel bias. *Psychological Methods.* https://doi.org/10.1037/met0000333

Gische, C., West, S. G., & Voelkle, M. C. (2021). Forecasting Causal Effects of Interventions versus Predicting Future Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 475–492.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica: Journal of the Econometric Society*, *37*(3), 424–438.

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

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates

to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379.

Heiss, A. (2022, June 30). Marginalia: A guide to figuring out what the heck marginal effects, marginal slopes, average marginal effects, marginal effects at the mean, and all these other marginal things are. Retrieved May 20, 2022, from https://www.andrewheiss.com/blog/2022/05/20/marginalia/

Hernán, M. A., & Robins, J. M. (2016). Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*, *183*(8), 758–764.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data?: Unit fixed effects models for causal inference. *American Journal of Political Science*, *63*(2), 467–490.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*(1), 52–59.

Kratz, F., & Brüderl, J. (2021). The Age Trajectory of Happiness. doi:10.31234/osf.io/d8f2z

Lenth, R. V. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. https://CRAN.R-project.org/package=emmeans

Lüdtke, O., & Robitzsch, A. (2022). A Comparison of Different Approaches for Estimating Cross-Lagged Effects from a Causal Inference Perspective. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–20.

Mayer, A. (2019). Causal effects based on latent variable models. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, *15*(S1), 15–28.

Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR Approach for Analyzing Average and Conditional Effects. *Multivariate Behavioral Research*, *51*(2–3),

374–391.

McArdle, J. J., & Hamagami, F. (2001). *Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data*. Retrieved from https://psycnet.apa.org/record/2001-01077-005

Mize, T. D., Doan, L., & Long, J. S. (2019). A General Framework for Comparing Predictions and Marginal Effects across Models. *Sociological Methodology*, *49*(1), 152–189.

Murayama, K., & Gfrörer, T. (2022, October 17). Thinking clearly about time-invariant confounders in cross-lagged panel models: A guide for model choice from causal inference perspective. https://doi.org/10.31234/osf.io/bt9xr

Pearl, J. (2018). Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, *6*(2), 20182001.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* , *11*(5), 550–560.

Rohrer, J. M. (2022, May 27). ✨ Unleash your inner stats sparkle ✨ with this very non-technical introduction to marginal effects. Retrieved June 30, 2022, from The 100% CI website: http://www.the100.ci/2022/05/27/%e2%9c%a8-unleash-your-inner-stats-sparkle-%e2%9c%a8-with-this-very-non-technical-introduction-to-marginal-effects/

Rohrer, J. M., & Arslan, R. C. (2021). Precise Answers to Vague Questions: Issues With Interactions. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007370.

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a Lot to Process! Pitfalls of Popular Path Models. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221095828.

Rüttenauer, T., & Ludwig, V. (2020). Fixed Effects Individual Slopes: Accounting and Testing for Heterogeneous Effects in Panel Data or Other Multilevel Models. *Sociological Methods & Research*, 0049124120926211.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356.

Voelkle, M. C., Gische, C., Driver, C. C., & Lindenberger, U. (2018). The Role of Time in the Quest for Understanding Psychological Mechanisms. *Multivariate Behavioral Research*, *53*(6), 782–805.

Wan, S., Brick, T. R., Alvarez-Vargas, D., & Bailey, D. H. (2021). *Toward a Causally Informative Fit Index of Longitudinal Models: A Within-Study Design Approach*. doi:10.31234/osf.io/5cbkt

Williams, R. (2012). Using the Margins Command to Estimate and Interpret Adjusted Predictions and Marginal Effects. *The Stata Journal*, *12*(2), 308–331.