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

Julia M. Rohrer[1] and Kou Murayama[2,3]

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 for 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.

Daily diary studies, experience sampling, mobile sensing: Technological innovations have made it much easier for psychologists to collect intensive 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 in 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”

[1] Department of Psychology, Leipzig University, Germany. julia.rohrer@uni-leipzig.de.
[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany.
[3] Research Institute, Kochi University of Technology, Japan

This research was partly supported by JSPS KAKENHI (Grant Number 18K18696 to KM) and Jacobs Foundation Advanced Research Fellowship (to KM).

We thank Henrik Andersen, Ruben Arslan, Tobias Debatin, 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.

(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 effects (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). Within- and between-person effects can be statistically independent (i.e., they can take on different values, or even opposite signs), and it is the latter within-person effects 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 et al., 2021; Lüdtke & Robitzsch, 2021; Voelkle et al., 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 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. With this manuscript, we aim to provide some general conceptual guidance for researchers who want to make full use of the potential of

longitudinal data, and provide entry points into the more technical literature on causal inference.

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 which reflects a “typical” question that researchers may aim to answer with within-person data: does talkativeness, an aspect of 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 (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). 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, $Y^{a=1}$ and $Y^{a=0}$, 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 $Y^{a=0}$ and $Y^{a=1}$ 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 observe multiple people for a single day, it can become possible to estimate the *average* of their individual causal effects. Consider the possibility of a randomized experiment in which we assign a large number of people to either spending the day being talkative or spending the day untalkative. For the sake of the argument, let’s pretend we had two pills that manipulated 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.

Prior to the manipulation, 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 interchangeable 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 interchangeable 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 treatment group, the respective potential outcomes are realized (3). The assumption to ensure the equivalence between (2) and (3) is called consistency, which we will mention 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.[4] Going beyond experimental data, average causal effects can also be estimated with the help of non-experimental between-person data (or with the help of natural experiments). Such endeavors require strong additional assumptions that may vary in their plausibility (for accessible introductions, see e.g. Elwert, 2013; Hernán & Robins, 2010; Pearl et al., 2016; Rohrer, 2018; Rosenbaum, 2017).

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 function of (1) non-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 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

[4] Unless we additionally assume that the causal effect is the same for each individual (e.g., West & Thoemmes, 2010).

within-person associations. 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 socio-economic status, stable personality traits). In contrast, within-person associations cannot be explained away by third variables that are stable for the duration of the data collection. This means that it can also be theoretically (albeit unrealistic in many cases, in particular in the presence of stabilty due to traits or auto-regressive effects) 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).

Considering (2), another reason why within- and between-person associations may diverge are 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 (Lüdtke & Robitzsch, 2021). We will return to the issue of time frame when discussing design parameters.

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 again confounding. As discussed above, within-person associations are not affected by time-invariant confounding factors. 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, and 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 interchangeable—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 at day 1, as this may already capture the confounding influence of social events. However, this entirely 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. We bring up the issue of the “correct” time lag in our discussion on how to make the most of within-person data.

In short, causal inferences on the basis of within-person data 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 not affected by time-invariant confounders. Within-person data thus have potential to control for various types of between-person confounding, including unobservable confounders. One of the simplest ways that attempt to achieve this is so-called fixed effects models (Hamaker & Muthén, 2020; Imai & Kim, 2019; McNeish & Kelley, 2019). Conceptually, the fixed effect approach corresponds to the procedure in which we mean-center the variables of interest (i.e., subtract person-specific averages) before examining the effects of interest (e.g., Hamaker & Muthén, 2020; McNeish & Kelley, 2019). The fixed effects approach or mean centering cannot perfectly control for time-invariant confounders (e.g., it cannot control for unobserved

time-invariant confounders having time variant effects on observed variables). But it successfully identifies contemporaneous causal effects under assumptions that have been spelled out in great clarity by Imai and Kim (2019). In particular, in the standard fixed-effects model, we have to assume that there are no lagged causal dynamics: past happiness does not affect current talkativeness, past talkativeness does not affect current happiness.

Psychologists are, of course, often interested in precisely estimating 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, cross-lagged panel models, fail at the task of controlling between-subject confounding despite their use of longitudinal data (Hamaker et al., 2015). This has prompted researchers to use a modification of the model to separate stable trait effects (random intercepts) to estimate cross-lagged effects (i.e., random-intercept cross-lagged model and its variants (Hamaker et al., 2015). Random-intercept cross-lagged model has become very popular in psychology, but very similar models have already been proposed in other fields (e.g., dynamic panel model; for mathematical equivalence, see Andersen, 2020).

However, whether these random intercepts are sufficient to adjust for time-invariant confounders still depends on additional assumptions — like fixed-effects approach, the model cannot control for unobserved time-invariant confounders having time-variant effects (see also Lüdtke & Robitzsch, 2021). There is no magic to control for all time-invariant confounders at once without assessing them. The model is also unable to identify contemporaneous causal effects (i.e., we cannot identify both lagged and contemporaneous causal effects). In short, there is great potential to unravel causal effects in longitudinal data, but 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.

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. Usami et al. (2019) provide a discussion of the causal assumptions underlying popular models. Lüdtke and Robitzsch (2021) contrast the standard cross-lagged panel model and the random-intercept cross-lagged panel model with respect to 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. 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 (Robins et al., 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. 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.

Aside from causal identification, 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); there has been some methodological development 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.”

Setting the Analysis Goal

An approach that we believe to be more productive puts the substantive question first. 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 taken more slowly. 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. 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.[5]

A Side-Note on the Definition of Causal Effects in the Absence of Magic Pills

Earlier, we alluded to magic pills that fix an individual’s talkativeness at a given level. Such pills do not exist, and drugs that do affect talkativeness can have many “side” effects (e.g., potential direct effects on well-being). This simple example reflects a major complication to causal inference that we had put aside until now: at least in some cases, it may be impossible to come up with a well-defined causal effect even if the estimand can be verbalized and appears reasonable.

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*.

But 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 our hypothetical pill). 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. This violates an assumption called consistency (formally, this assumption is 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

[5] For example, Lüdtke and Robitzsch (2021) clarify how the inclusion of the random intercept changes the theoretical estimand of the cross-lagged panel model, and may create a mismatch with *some* research questions.

data cannot be well-defined (but see also Pearl, 2018). Yet body weight as an independent variable seems much more concrete than something like “talkativeness” or “subjective well-being” (Rohrer & Lucas, 2020). As highlighted by Hernán and Robins (2020, p. 34), trying to precisely define the causal effect of interest may alter the original question, and likely lead to narrower estimands. It may also sometimes lead us to discard certain variables as potential causes to begin with.

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 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, or 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 assessed. 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.[6] But if we decide that we want to investigate a relatively well-defined specific

[6] A further complication when it comes to self-report data is that the time point of assessment does not necessarily correspond to the time point of the assessed state. For example, when asked about their happiness, individuals may think back and aggregate over some time period. This, of course, makes it even harder to unravel the full underlying causal dynamics.

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.[7]

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 which (if any) 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.

Shouldn’t We Just Do Description Instead?

Having read the article 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 (although that term may be used in an ambiguous manner). Indeed, many longitudinal analyses claim to be descriptive in nature. However, this may partly be a strategic move to avoid the heightened scrutiny that may result 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.

Furthermore, it is not exactly easy to apply the standard longitudinal models for descriptive purposes. The literature is filled with many related “state-of-the-art” longitudinal models such as the random-intercept cross-lagged panel model (Hamaker et al., 2015), the

[7] 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.

stable trait/auto-regressive trait/state model (Kenny & Zautra, 1995), the autoregressive latent trajectory model (Bollen & Curran, 2004) and the dual-change score model (McArdle & Hamagami, 2001); many of which can be specified in multiple ways (e.g., with or without residual variables). In principle, with these models, we can elegantly address the underlying (causal!) within-person dynamics *if* we correctly specify the model that precisely reflects reality. In reality, we do not know which model generated the 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). In any case, trying to uncover the complete 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.

In comparison to attempting to describe longitudinal data with a large number of different models, we argue that it may actually be *easier* to estimate a causal effect of interest. 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 analysis goal: recovery of a specific causal effect (as opposed to uncovery of the whole causal dynamics). 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 tackle 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

Alvarez-Vargas, D., Braithwaite, D. W., Lortie-Forgues, H., Moore, M. M., Castro, M., Wan, S., Martin, E. A., & Bailey, D. H. (2020). *Hedges, Mottes, and Baileys: Causally Ambiguous Statistical Language can Increase Perceived Study Quality and Policy Relevance*. https://doi.org/10.31234/osf.io/nkf96

Andersen, H. K. (2020). *Equivalent approaches to dealing with unobserved heterogeneity in cross-lagged panel models?* https://doi.org/10.31235/osf.io/hcnfg

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

Cattell, R. B. (1952). The three basic factor-analytic research designs—their interrelations and derivatives. *Psychological Bulletin*, *49*(5), 499–520.

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

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). https://doi.org/10.18637/jss.v077.i05

Elwert, F., & Pfeffer, F. T. (2019). The Future Strikes Back: Using Future Treatments to Detect and Reduce Hidden Bias. *Sociological Methods & Research*, 0049124119875958.

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, *59*, 100785.

Foster, E. M. (2010a). Causal inference and developmental psychology. *Developmental Psychology*, *46*(6), 1454–1480.

Foster, E. M. (2010b). The U-shaped relationship between complexity and usefulness: a commentary [Review of *The U-shaped relationship between complexity and usefulness:*

a commentary]. *Developmental Psychology*, *46*(6), 1760–1766.

Galambos, N. L., Krahn, H. J., Johnson, M. D., & Lachman, M. E. (2020). The U Shape of Happiness Across the Life Course: Expanding the Discussion. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *15*(4), 898–912.

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.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *15*(5), 1243–1255.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2005). Statistical Modeling of the Individual: Rationale and Application of Multivariate Stationary Time Series Analysis. *Multivariate Behavioral Research*, *40*(2), 207–233.

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.

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.

Haslbeck, J. M. B., & Ryan, O. (2021). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, 1–32.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What If*. CRC Boca Raton, FL; https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf

Hernán, M. A., & Taubman, S. L. (2008). Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity* ,

32 Suppl 3, S8–S14.

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

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. In *https://psyarxiv.com › ...https://psyarxiv.com › ...* https://doi.org/10.31234/osf.io/d8f2z

Lüdtke, O., & Robitzsch, A. (2020). Commentary Regarding the Section “Modelling the Effectiveness of Teaching Quality.” *Zeitschrift Für Pädagogik Beiheft*, *1*, 210–223.

Lüdtke, O., & Robitzsch, A. (2021). A Critique of the Random Intercept Cross-Lagged Panel Model. In *https://psyarxiv.com › ...https://psyarxiv.com › ...* https://doi.org/10.31234/osf.io/6f85c

Lundberg, I., Johnson, R., & Stewart, B. (2020). What is your estimand? Defining the target quantity connects statistical evidence to theory. In *SocArXiv*. https://doi.org/10.31235/osf.io/ba67n

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

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*(1), 20–35.

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218.

Molenaar, P. C. M., & Campbell, C. G. (2009). The New Person-Specific Paradigm in

Psychology. *Current Directions in Psychological Science*, *18*(2), 112–117.

Murayama, K., Goetz, T., Malmberg, L.-E., Pekrun, R., Tanaka, A., & Martin, A. J. (2017). Within-person analysis in educational psychology: Importance and illustrations. *British Journal of Educational Psychology Monograph Series II: Psychological Aspects of Education--Current Trends: The Role of Competence Beliefs in Teaching and Learning*, 71–87.

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., & Lucas, R. E. (2020). *Causal Effects of Well-Being on Health: It's Complicated*. https://doi.org/10.31234/osf.io/wgbe4

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 1745691620966795.

Schmiedek, F., & Neubauer, A. B. (2020). Experiments in the Wild: Introducing the Within-Person Encouragement Design. *Multivariate Behavioral Research*, *55*(2), 256–276.

Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: Interindividual, intraindividual, and multivariate time-series analyses. *Journal of Personality and Social Psychology*, *64*(6), 1010–1028.

Thoemmes, F., & Ong, A. D. (2016). A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. *Emerging Adulthood*, *4*(1), 40–59.

Usami, S. (2020). Within-Person Variability Score-Based Causal Inference: A Two-Step Semiparametric Estimation for Joint Effects of Time-Varying Treatments. In *arXiv [stat.ME]*. arXiv. http://arxiv.org/abs/2007.03973

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, *24*(5), 637–657.

VanderWeele, T. J., Mathur, M. B., & Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *35*(3), 437–466.

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a Unified Framework for the Study of Between-Person and Within-Person Structures: Building a Bridge Between Two Research Paradigms. *Multivariate Behavioral Research*, *49*(3), 193–213.

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.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, *15*(1), 18–37.

Williamson, T., & Ravani, P. (2017). Marginal structural models in clinical research: when and how to use them? *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association*, *32*(suppl_2), ii84–ii90.

Yarkoni, T. (2020). The generalizability crisis. *The Behavioral and Brain Sciences*, 1–37.

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. (2020). From Data to Causes I: Building A General Cross-Lagged Panel Model (GCLM). *Organizational Research Methods*, *23*(4), 651–687.

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. (2020). From Data to Causes II: Comparing Approaches to Panel Data Analysis. *Organizational Research Methods*, *23*(4), 688–716.