Why The Cross-Lagged Panel Model is Almost Never the Right Choice

Richard E. Lucas[1]

[1] Department of Psychology, Michigan State University

Author Note

Correspondence concerning this article should be addressed to Richard E. Lucas, 316 Physics Rd., Michigan State University, East Lansing, MI 48823. E-mail: lucasri@msu.edu

Abstract

The cross-lagged panel model (CLPM) is a widely used technique for examining reciprocal causal processes using longitudinal data. Critics of the CLPM have noted that it fails to account for certain person-level confounds. Because of this, models that incorporate stable-trait components (such as the random intercept cross-lagged panel model [RI-CLPM] or the bivariate Stable Trait Autoregressive Trait [STARTS] model) have become popular alternatives. Debates about the merits of the CLPM have continued, however, with some researchers arguing that the CLPM is more appropriate than modern alternatives for examining common psychological questions. In this paper, I discuss the ways that these defenses of the CLPM fail to acknowledge widely known problems with the interpretation of analyses of multilevel data. I propose some possible sources of confusion regarding between- and within-person effects that these models estimate, and provide alternative ways of thinking about the problems with the CLPM. I then show in simulated data that with realistic assumptions, the CLPM is very likely to find spurious cross-lagged effects when they don't exist, while also underestimating them when they do. I argue that there are no situations where the CLPM is preferable to alternatives that incorporate information about stable traits (though there are, of course, research questions for which neither the CLPM nor alternatives that incorporate a stable trait are appropriate).

Keywords: cross-lagged panel model, longitudinal, structural equation modeling

Why The Cross-Lagged Panel Model is Almost Never the Right Choice

The cross-lagged panel model (CLPM) is a widely used technique for examining causal processes using longitudinal data. With at least two waves of data, it is possible to estimate the association between a predictor at Time 1 and an outcome at Time 2, controlling for a measure of the outcome at Time 1. With some assumptions, this association can be interpreted as a causal effect of the predictor on the outcome. The simplicity of the model along with its limited data requirements have made the CLPM a popular choice for the analysis of longitudinal data. For instance, Usami, Murayama, and Hamaker (2019) reviewed medical journal articles published between 2009 and 2019 and found 270 papers that used this methodological approach. A broader search of google scholar returned 3,910 papers that use the term "cross-lagged panel model" in the last 40 years.[1]

The CLPM improves on simpler cross-sectional analyses by controlling for contemporaneous associations between the predictor and outcome when predicting future scores on the outcome. Presumably, confounding factors should be reflected in this initial association, which would mean that any additional cross-lagged paths between the Time 1 predictor and the Time 2 outcome would reflect a causal effect of the former on the latter (again, with some assumptions). Hamaker, Kuiper, and Grasman (2015) pointed out, however, that the CLPM does not adequately account for stable-trait-level confounds, and they proposed the random-intercept cross-lagged panel model (RI-CLPM) as an alternative (also see Allison, 2009; Berry & Willoughby, 2017; Zyphur, Allison, et al., 2020). The RI-CLPM includes stable-trait variance components that reflect variance in the predictor and outcome that is stable across waves. Hamaker et al. showed that failure to account for these random intercepts and the associations between them can lead to incorrect conclusions about cross-lagged paths. As others have noted (e.g., Lüdtke & Robitzsch, 2021; Usami, 2020), this critique of the cross-lagged panel model has already been cited frequently and has

[1] As of July 20, 2022.

had an important impact on researchers who use longitudinal data.

Despite this impact, debates about the relative merits of the CLPM versus the RI-CLPM (and more complex alternatives) continue. Most notably, Orth, Clark, Donnellan, and Robins (2021) argued that sometimes researchers are actually interested in the effects that a classic CLPM tests and that the choice of model should depend on one's theories about the underlying process. Orth et al.'s paper has already been cited over 120 times even though it was only published one year ago at the time of this writing. Many of the citing papers justify their use of the CLPM based on the arguments that Orth et al. put forth. The goal of the current paper is to examine this defense of the CLPM, focusing first on the interpretation of models like the RI-CLPM that include a stable-trait component, followed by simulations that demonstrate the problems with the CLPM and the utility of its alternatives. These simulations show that when the CLPM is used, spurious cross-lagged associations are common and the likelihood of finding such spurious effects can reach 100% in many realistic scenarios. At the same time, the CLPM is also likely to underestimate cross-lagged associations when they do exist. I conclude that there is no situation where the CLPM is preferable to alternatives that model a stable trait and that the CLPM should be abandoned as an approach for examining causal processes in longitudinal data.

Ambiguity About Between- and Within-Person Effects

In their critique of the CLPM, Hamaker et al. (2015) described the RI-CLPM as a multilevel model that separates between-person associations from within-person associations. But what *is* a between-person association and how does it differ from a within-person association? Why is it important to separate these levels of analysis in the context of lagged effects? These questions are critical, as answers to them form the basis for some debates about the CLPM.

Certain aspects of the between/within distinction are clear and unambiguous. When

data are collected from multiple participants at a single point in time, there can only be between-person variance. All associations that can be observed in these data are necessarily between-person associations. For instance, in cross-sectional data, a negative correlation between self-esteem and depression can only be interpreted as a between-persons association: People who score high on measures of self-esteem tend to score low on measures of depression. If, on the other hand, just a single individual is assessed repeatedly over time, all variance is within-person variance and all associations would be within-person associations. For example, if a single person's self-esteem and depression were tracked over time, a negative correlation would reflect a within-person association: When self-esteem is high in that individual, feelings of depression tend to be low.

The potential for confusion arises, however, when data are collected from multiple people across multiple occasions. Such data include information both about how people differ from one another (between-person variance) and how each person changes over time (within-person variance). Describing effects and associations as "between" versus "within" becomes more challenging with these multilevel data. For instance, if feelings of depression were assessed multiple times over the course of a school semester and cross-sectional differences in self-esteem from the start of the semester predicted changes in depression over that period, would this interaction reflect a between-person association or one that is within-person? Although I believe that most methodologists would label this a between-person association (because individual differences in self-esteem are predicting individual differences in depression slopes), within-person data (each person's change over the course of the study) are used to estimate this between-person association. The example shows that the decision to label an association as "between" or "within" is not always linked in a straightforward way to the type of data that contribute to the effect.

This is important because Orth et al. (2021) rely heavily on the *description* of the cross-lagged paths in the RI-CLPM as a *within-person effect* in their defense of the CLPM.

They state that “a potential disadvantage of the proposed alternatives to the CLPM is that they estimate within-person prospective effects only, but not between-person prospective effects” (p. 1014) and that “in many fields researchers are also interested in gaining information about the consequences of between-person differences” (p. 1014). They go on to argue that “a limitation of the RI-CLPM is that it does not provide any information about the consequences of between-person differences. In the RI-CLPM, the between-person differences are relegated to the random intercept factors” (p. 1026). Later on the same page, they state that “The RI-CLPM includes [an] unrealistic assumption, specifically that the between-person variance is perfectly stable” (p. 1026). Orth et al. (2021) do acknowledge later on the same page that “some portion of the systematic between-person variance will be included in the residualized factors” (p. 1026). However, they argue that this discrepancy is a conceptual problem for the RI-CLPM: They state that “the cross-lagged effects in the RI-CLPM are not pure within-person effects but partially confounded with between-person variance” (p. 1026).

These statements reflect an apparent misunderstanding of the RI-CLPM and related models. Confusion about these issues likely results from the previously mentioned ambiguities regarding the terms “between-person” and “within-person” along with fundamental differences in the ways that different types of multilevel models separate between- and within-person effects. The RI-CLPM separates between-person associations from within-person associations, but it does not do so by “relegating all between-person differences to the random intercept.” Because Orth et al.’s (2021) defense of the CLPM rests on a flawed interpretation of the alternatives, their defense against these alternatives is not valid.

As Curran and Bauer (2011) noted, what is probably the most familiar way to separate between-person effects from within-person effects is the use of person-mean centering. For instance, in the context of multilevel modeling, researchers are often warned that if they are

not careful about how they enter variables into the model, what may look like within-person effects (e.g., the "Level-1" effects in repeated-measures data) can actually reflect a mix of between- and within-person associations. The recommended solution in this context is to person-center the predictor (e.g., Curran & Bauer, 2011; Enders & Tofighi, 2007), where each observation now reflects a deviation from a person's mean. When centering this way, the Level-1 part of the model tests whether occasion-specific deviations from a person's mean predict variability in the outcome.

In their own critique of the RI-CLPM, Lüdtke and Robitzsch (2021) imply that the RI-CLPM and related models accomplish the separation of between- and within-person associations in the exact same way as the multilevel modeling approach described by Curran and Bauer (2011) (for discussions about the similarity and differences between lagged models in the multilevel modeling and structural equation modeling contexts, see Hamaker and Muthén (2020), Falkenström, Solomonov, and Rubel (2022)). They cautioned that "researchers should be aware that within-person effects are based on person-mean centered (i.e., ipsatized) scores that only capture temporary fluctuations around individual person means" which would be "less appropriate for understanding the potential effects of causes that explain differences between persons" (p. 18). They do not clarify, however, that the "person mean" in their description is not the observed person mean calculated from the actual observations (as it would typically be in the traditional multilevel modeling context), but a latent mean that reflects only the variance that is perfectly stable over time. These "means" are—conceptually and empirically—very different things. This also means that what is left after adjusting for these means (the "ipsatized" scores) can also be very different depending on which "mean" is used.

To appreciate this difference, consider the example in Figure 1. The models shown in this figure represent a single variable, X, measured across three occasions. Panel A shows the data-generating process that I used for this example, which reflects a very simple

autoregressive model. In this example, the initial variance for X_1 was set to 1, and the wave-to-wave stability was set to .5. This simple autoregressive model links between-person differences at Time 1 to between-person differences at Time 2 through a stability coefficient. Note that it would be possible to extend this model to a traditional CLPM by adding an outcome variable at each wave and then testing the lagged paths from the predictor to the outcome and the outcome to the predictor.

With these data, it is possible (and meaningful) to consider what each person's mean would be across these three occasions. A simple latent-trait model (like the one shown in Panel B) would capture the variance in these means. Note that in this simple latent-trait model, the variance in each wave is partitioned into variance in the person mean (the variance of the common latent trait, which is .48 in this example) and variance in the wave-specific deviations from that mean (.68, .42, and .64). The components of the model in the blue box could be considered the "within-person" part of this model, as these residuals reflect wave-specific deviations from the person mean. The partitioning of variance in this example is quite similar to the descriptions that Lüdtke and Robitzsch (2021) and Orth et al. (2021) provide for the variance partitioning in the RI-CLPM.

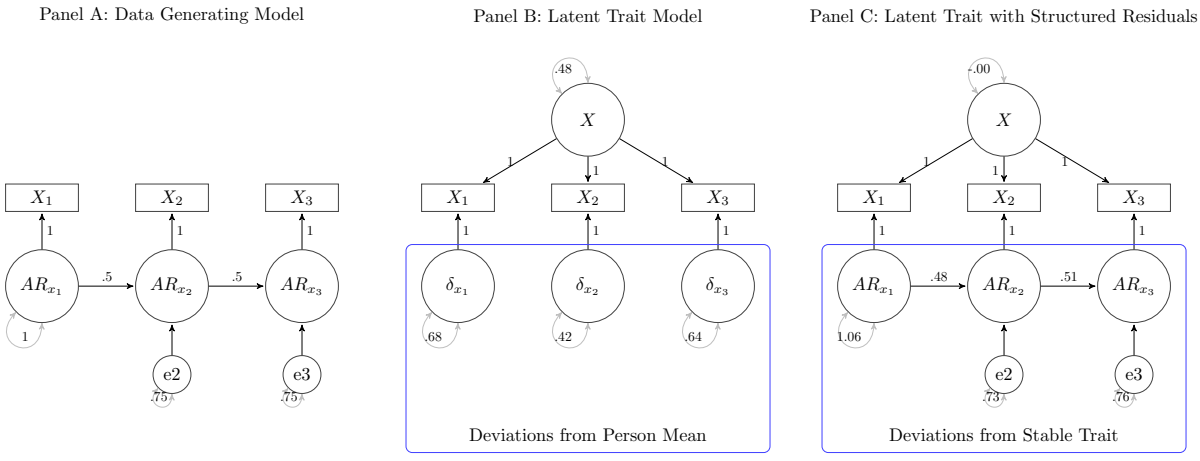


Figure 1. Different ways of conceptualizing between-person variance.

It is the third panel of this figure, however, that most closely represents (in a univariate setting) what the RI-CLPM and related models actually do. In this panel, the residuals have an autoregressive structure, where the residuals for each wave are predicted from the residuals of the wave prior (exactly as they are in the RI-CLPM). When residuals are structured in this way, they capture between-person variance that is somewhat—but usually not perfectly—stable over time. By structuring the residuals to allow for some wave-to-wave stability, the latent trait now includes only the variance that is *perfectly* stable over this time period. The residuals no longer represent deviations from the person mean, they represent deviations from a *perfectly stable trait*[2]. Importantly, although the parts of the model included in the blue box still represent the "within-person" model, these structured residuals now link meaningful between-person differences at Time 1 to between-person differences at Time 2. Indeed, in this specific example, because the data-generating process specifies that there is no perfectly stable variance, these "deviations" are identical to the original variables themselves. The parameter estimates from this model almost perfectly recover the values specified in the simple autoregressive data-generating process[3].

A comparison of the variance estimates across Panels B and C show that the latent trait that links the three observations captures something very different in a simple latent-trait model as compared to a model with structured residuals. However, one could describe either the latent trait from Panel B or the one from Panel C as reflecting the "between-person variance" in these assessments. The latent trait in Panel B captures between-person differences in mean levels of X during the observed period of assessment; the latent trait in Panel C captures between-person differences in *a hypothetically perfectly stable* trait. Moreover, one could describe the residuals in either model as reflecting the "within-person" part of the model even though they correspond to conceptually different

[2] Thanks to Kou Murayama for discussions that clarified this issue.

[3] Note that this model fit to data with no stable trait variance will often result in inadmissable solutions because with true latent-trait variance of 0, it is possible to get estimates that are negative.

things. The occasion-specific residuals in Panel C do not reflect deviations from the person mean, they reflect deviations from the perfectly stable trait.

Orth et al. criticize the RI-CLPM for assuming that all between-person differences are perfectly stable over time and for removing all between-person variance from the within-person associations, but those critiques are not correct. In the RI-CLPM and other related models, between-person variance is simply *defined* as the variance that is perfectly stable over time. This is an issue of terminology, not assumptions. Indeed, relying on the terminology of the RI-CLPM, Orth et al.'s preferred CLPM would be said to assume that *no between-person variance exists whatsoever*, an assumption that is rarely defensible in studies of psychological phenomena. Moreover, between-person variance (broadly defined as differences between individuals) is clearly included in the "within-person" part of the model, as Panel C of Figure 1 shows. Indeed, for variables for which there is no stable-trait variance, the estimates for the cross-lagged paths from the RI-CLPM will be identical to those from the CLPM because the RI-CLPM reduces to the CLPM in such cases. In these cases, what Orth et al. (2021) would describe as a "between-person prospective effect" (the cross-lagged path from X_1 to Y_2 is equivalent to what would be described as a "within-person effect" in the RI-CLPM. Thus, the first problem with the arguments presented by Orth et al. (2021) is that their primary critique of the RI-CLPM (and corresponding defense of the CLPM) rests on an incorrect intepretation of that model and its relation to the CLPM.

Beyond Between and Within

In the previous section, I discussed ambiguity in the definition of between and within-person effects, focusing on how these ambiguities might lead to misinterpretations of models like the RI-CLPM that include a stable trait component. In the next section, I first review previous explanations for why separating these levels of analysis is critical when examining lagged effects. I then propose an alternative way of thinking about the problems with the CLPM that does not rely on an understanding of the distinction between within-

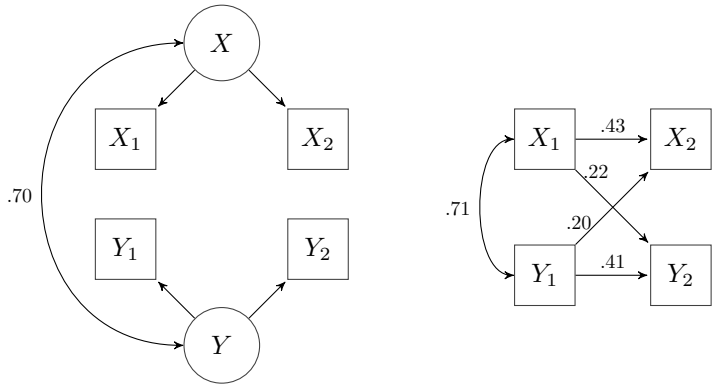
and between-person analyses.

As noted previously, data that have been collected from multiple people across multiple occasions include information both about how people differ from one another (between-person variance) and how each person changes over time (within-person variance). Methodologists have, for many decades, warned that failure to consider multilevel structures can lead to incorrect conclusions (see Curran & Bauer, 2011, for a review and discussion). It would be wrong, for instance, to draw conclusions about within-person associations from between-person data or to draw conclusions about between-person differences from within-person data because the association can be completely different at these different levels. In addition—and most importantly for debates about the merits of the CLPM—when data do have a multilevel structure, but this multilevel structure is not taken into account through appropriate analytic methods, the estimates obtained from analyses of these data reflect an uninterpretable mix of between and within-person effects (Raudenbush & Bryk, 2002).

The traditional CLPM is a classic example of an analysis that fails to separate between-person associations from within. It is quite easy to show that simple between-person differences can masquerade as within-person effects when the CLPM is used. As a simple demonstration, I generated two waves of data for two variables X and Y for 10,000 people. Panel A of Figure 2 shows the data-generating process, which is a very simple correlated-latent-trait model. I set the variance of X and Y to be 1, and the reliability of the indicators to be .5. X and Y are only associated at the between-person level ($r = .7$)[4]. In other words, X and Y are related only because people who tend to score high on X on average also tend to score high on Y average; X does not predict change in Y or vice versa and they have no unique associations within any particular wave. Panel B shows what

[4] Note that the terminological ambiguities from the previous section do not play a role in this simple two-wave example, so the meaning of "between-person" differences should be relatively straightforward and unambiguious in this simple case.

happens if we fit the CLPM to the generated data. As can be seen in this panel, there would be clear evidence for reciprocal associations between the two, even though there are no over-time associations between X and Y whatsoever. The CLPM simply cannot distinguish between associations that occur at the stable-trait level from those that incorporate some change over time.



Panel A: Data-Generating Model Panel B: CLPM Results

Figure 2. Spurious cross-lagged effects in data with only between-person associations. Panel A is the data-generating model; Panel B shows estimates from the CLPM fit to the generated data. Coefficients are unstandardized estimates. Residuals are not shown but are estimated.

Although concerns about the inappropriate handling of multilevel data emerge any time multilevel data are analyzed, they should be particularly salient when lagged effects are examined. This is because a goal of such lagged analyses is typically to clarify causal processes. Only the within-person part of these models, however, provides information about causal effects. If associations at the two levels are conflated, then drawing causal conclusions from these models would be inappropriate. In short, the primary benefit of the CLPM—its ability to provide evidence for causal effects—would be invalidated by the failure to isolate within-person effects.

Orth et al. (2021) acknowledge that they do indeed believe that the CLPM estimates causal effects. They state that in the context of their focal case study of self-esteem and depression, "the hypothesized causal effect" can be stated to be that: "when individuals have low self-esteem (relative to others), they will experience a subsequent rank-order increase in depression compared to individuals with high self-esteem" (p. 1014). Although this description of the cross-lagged path is technically correct, Orth et al. (2021) do not go on to explain how this association can be interpreted as a causal effect. No formal causal analysis is presented. Unfortunately, when stable-trait variance exists in the measures being analyzed, then this association is confounded. Referring again to Panel B of Figure 2, it is technically correct to say that the path from X_1 to Y_2 shows that those who score high on X_1 have a rank-order increase in Y from Wave 1 to Wave 2, but this is due solely to the confounding effect of the stable trait reflected in the latent X variable. The rank order on Y at Time 2 changes from Time 1 not because of any causal effect of X, but because the rank order at Y_1 imperfectly reflects the true rank order of Y.

Thus, cross-lagged paths in the CLPM can result from purely between-person associations, purely within-person effects, or some combination of the two (even combinations where the within and between-person associations are in the opposite direction). Rather than clarifying how the CLPM solves the uncontroversial interpretational issues that are inevitably involved with this type of analysis, Orth et al. (2021) sidestep this perennial analytic issue. If these authors believe that the CLPM somehow avoids the interpretational challenges inherent in all analyses of multilevel data, then they must do more than simply assert this to be true. Estimates from the CLPM are uninterpretable for all the same reasons that any analysis of multilevel data that fails to account for the multilevel structure are.

It can sometimes be difficult to think about how the complex nature of multilevel data can affect conclusions about underlying processes. Indeed, I worry that framing the discussion of the CLPM solely as an issue of multilevel structure has led to more confusion

than clarity (even though I agree that this conceptualization is technically correct and has led to practical analytic solutions to the problem). Figure 2 helps, however, by providing an alternative way of thinking about the problems with the CLPM and the benefits of alternative models that incorporate a stable-trait component. Models like the RI-CLPM are useful because they test an extremely plausible alternative explanation of the underlying pattern of correlations that is being modeled when the CLPM is used, an alternative explanation that has nothing to do with over-time associations.

The logic of the CLPM is very similar to the logic of any other regression model where researchers assess whether one variable predicts another after controlling for relevant confounds. When they test whether Time 1 X predicts Time 2 Y after controlling for Time 1 Y, they hope to capture whether there is something unique about X—something that cannot be explained by the concurrent association between X and Y—that helps us predict Y at a later time. But as Westfall and Yarkoni (2016) pointed out when discussing the difficulty of establishing incremental predictive validity of any kind, if the measure that we include as a control (i.e., Time 1 Y) is not a perfect measure of what researchers are trying to account for, then it is possible—indeed, quite easy—to find spurious "incremental validity" effects. Referring to Figure 2, Y_1 is an imperfect measure of the latent variable Y. Thus, controlling for Y_1 does not control for all of the association between X and Y, which means that X_1 will still have incremental predictive validity of Y_2 even after controlling for Y_1. Technically, Orth et al. (2021) are correct that the cross-lagged effect in Panel B means those who score high on X_1 would report a "rank-order increase" on Y over time, but this rank-order increase results from imperfect control of Y, not true change over time.

One might argue that concerns about inadequate control in the CLPM become moot as long as researchers avoid talking about causal effects. For instance, one might be tempted to interpret the cross-lagged path from X_1 to Y_2 purely descriptively: Using Orth et al.'s (2021) substantive example, it might seem reasonable to conclude from a significant

cross-lagged path that initial levels of self-esteem are associated with change in depression over time. However, even this more limited interpretation of this path is not warranted when stable-trait variance exists. Again, Figure 2 shows that the path from X_1 to Y_2 can emerge simply due to unmodeled stable-trait associations.

In summary, decades of methodological work show the importance of distinguishing between-person associations from within-person effects when data have a multilevel structure. Failing to do so results in uninterpretable estimates of the association between predictors and outcomes. The CLPM is not an exception to this widely discussed rule. In defending the CLPM, Orth et al. (2021) sidestep the issue of how a model that fails to distinguish between levels can lead to interpretable results; instead, they simply assert that the effects from the CLPM are meaningful. They claim to want to test a "between-person prospective effect" but do not define what a between-person prospective effect is and they offer no causal analysis that explains the meaning of such an effect. It is easy to show, however, that when the CLPM is used, it is possible to mistake purely between-person associations for over-time effects, which confirms the long-standing methodological warning about the failure to separate effects at different levels. Moreover, researchers do not even need to think about these issues in terms of multilevel models and the separation of between-person and within-person effects to appreciate the problems with the CLPM. The CLPM simply cannot rule out an extremely plausible alternative explanation for the underlying pattern of correlations. I now turn to a set of simulations that demonstrate just how bad this problem likely is.

It's Extremely Easy to Find Spurious Cross-Lagged Effects

The issues discussed in the previous sections show that hypothetically, it is possible to mistake purely between-person associations for over-time effects when the CLPM is used. But how likely are such spurious effects? Unfortunately, it is extremely easy to find spurious cross-lagged associations under conditions that are quite likely in the typical situations

where the CLPM is used. Hamaker et al. (2015) and Usami, Todo, and Murayama (2019) conducted simulations to show that the estimates from a cross-lagged panel model were often biased in realistic situations. I don't think they went far enough, though, in describing the practical implications of these simulations or showing just how likely spurious effects are in realistic situations. So the rest of this paper builds on their simulations and tries to clarify when such spurious effects are likely to occur. As I show, there are many realistic scenarios where researchers are almost guaranteed to find spurious cross-lagged effects.

A Note About Models

At this point, it is necessary to introduce the models used in these simulations and to clarify the terminology that I will use when describing the components of the models. As Falkenström et al. (2022) noted, "it is important to first reflect on the relevance of [a statistical analysis method] to the real-world processes a researcher attempts to model" and that "researchers familiar with the study subject can make educated guesses about the nature of this process" (p. 447). It has long been recognized that psychological variables often have features that are both "state-like" and "trait-like" (Hertzog & Nesselroade, 1987). In other words, these variables exhibit stability and change, and it is possible to think about different ways that constructs can stay the same or change over time.

For instance, Nesselroade (1991) noted that there are three types of latent factors that are frequently very useful for explaining variability in repeated measures of individual difference constructs—state factors, slowly changing "trait" factors, and a completely stable trait factors. State factors are the most fleeting, as they reflect variance that is unique to a single measurement occasion. These state factors can include random measurement error, but they also any reliable variance that does not carry over from one wave to the next. If a construct consisted solely of state variance, there would be no stability from wave to the next.

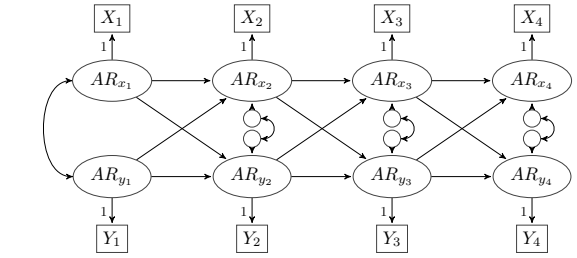
In contrast, stable-trait factors reflect variance that is perfectly stable across all waves

of assessment. If a construct consisted solely of stable trait variance, then wave-to-wave stability would be perfect, regardless of the interval between them. In between these two extremes are slowly changing trait factors where variance at one wave predicts variance at the next, but with less than perfect stability. Stability of this slowly changing trait factor declines with increasing interval length. Of course, these three components do not exhaust all possible patterns of stability and change Zyphur, Allison, et al. (2020), but they reflect reasonable assumptions about features that are likely to generalize to a wide range of psychological variables[5].

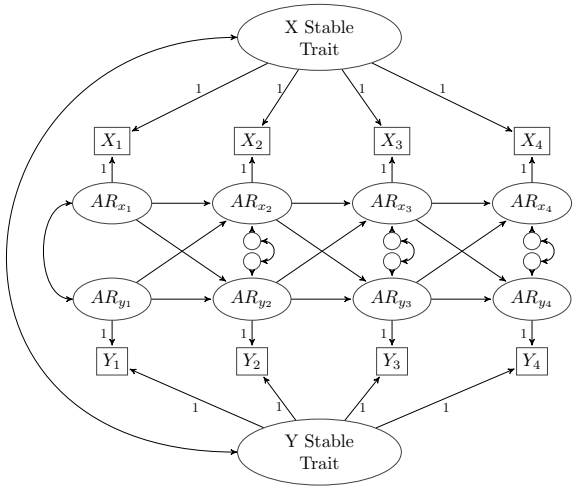
It is possible to frame the CLPM and its modern alternatives in the context of these sources of variance. This is especially helpful when considering what data to simulate. For instance, Panel A of Figure 3 shows a diagram of the model that has been the focus of this paper, the CLPM. This model includes one latent variable per wave for the predictor (X) and the outcome (Y), and these latent variables have an autoregressive structure with cross-lagged associations. Notice that the CLPM does not include any measurement error for the indicators. This means that the latent variables from the autoregressive part of the model are equivalent to the observed variables (which is why it is also possible to draw an equivalent CLPM model with only observed variables). The CLPM assumes that all variance is of the slowly-changing, autoregressive variety described by Nesselroade (1991).

Panel B of 3 shows the diagram of the RI-CLPM. The difference between the CLPM and the RI-CLPM is that the RI-CLPM includes a random intercept that accounts for "time-invariant, trait-like stability" (Hamaker et al., 2015, p. 104). The random intercept corresponds to purely stable trait variance and is thus labeled "Stable Trait" in the figure. Including this stable-trait component changes the meaning of the autoregressive part of the model. Whereas in the CLPM, the cross-lagged paths reflect associations between the X and Y variables over time, in the RI-CLPM, these paths reflect associations among wave-specific

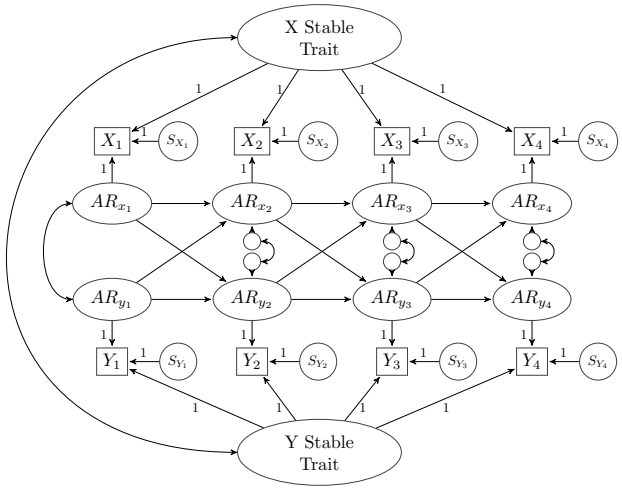
[5] One obvious addition would be mean-level change including linear growth.



Panel A: CLPM



Panel B: RI-CLPM



Panel C: STARTS

Figure 3. Diagram of the three models used in this paper.

deviations from a person's stable-trait level. This is what allows for the separation of between and within-person effects (Allison, 2009; Curran & Bauer, 2011; Hamaker et al., 2015). Note that the CLPM is nested within the RI-CLPM; the CLPM is equivalent to the RI-CLPM with the random-intercept (or stable-trait) variance constrained to 0. This also means that if one tries to fit the RI-CLPM to data with no stable-trait variance, the interpretation of the "within-person" or autoregressive part of the model will be identical to the interpretation of the CLPM.

The final model, presented in Panel C 3, is the bivariate Stable Trait, Autoregressive Trait, State (STARTS) model (Kenny & Zautra, 1995, 2001), which I have only briefly mentioned up to this point. The STARTS model differs from the RI-CLPM in that it includes includes a wave-specific "state" component (labeled s_t in the figure), which reflects variance in an observed variable that is perfectly "state-like" and unique to that occasion. This state component can include measurement error or any reliable variance that is unique to a single wave of assessment. The idea that some amount of pure state variance would exist in measures of psychological constructs is quite plausible (Fraley & Roberts, 2005), but simpler models like the RI-CLPM have often been preferred because the STARTS requires more waves of data than the RI-CLPM (four, to be precise) and often has estimation problems (e.g., Cole, Martin, & Steiger, 2005; Orth et al., 2021; Usami, Todo, et al., 2019).

Recently, Usami, Murayama, et al. (2019) clarified that the CLPM, RI-CLPM, STARTS and many other longitudinal models could be thought of as variations of an overarching "unified" model that captures many different forms of change (also see Zyphur, Allison, et al., 2020). For instance, an alternative model—the Latent Curve Model with Structured Residuals (Curran, Howard, Bainter, Lane, & McGinley, 2014)—can be thought of as an RI-CLPM with a random slope. Because debates about the utility of the CLPM have primarily focused on debates about the inclusion of the random intercept, I focus here only on the comparison of the CLPM to the RI-CLPM and the STARTS, as this comparison

highlights these debates most clearly. It is certainly true, however, that if the other forms of change included in the unified model were part of the actual data generating process, then all the models covered in this paper would be misspecified and could lead to biased estimates.

The Simulations

When considering what types of situations to simulate, I focus on realistic scenarios for the types of data to which the CLPM is likely to be applied. For instance, it is likely that most variables that psychologists (and other social and health scientists) choose to study over time have a longitudinal structure where stability declines with increasing interval length (reflecting an autoregressive structure), yet this decline approaches or reaches an asymptote where further increases in interval length are no longer associated with declines in stability (reflecting the influence of a stable trait). It is also likely that most measures of psychological constructs have some amount of pure state variability, which could reflect measurement error or true state-like influences.

To test the plausibility of these starting assumptions, I selected ten diverse variables that have been included in almost every wave of the long-running Household Income and Labour Dynamics in Australia (HILDA) panel study, which now spans 20 waves of assessment (Watson & Wooden, 2012). I intentionally selected variables from different domains and variables that might have different psychometric properties due to how easily observable they are (e.g., weight and income are likely to be measured with less systematic and random error than life satisfaction or social support). I then fit the most inclusive of the three models discussed (the STARTS model) to see how much variance each component accounted for. Results are shown in Table 1.

The first thing to note from this table is that the CLPM, which assumes a purely first-order autoregressive structure, would be misspecified when applied to any of these

Table 1

Variance components from the STARTS model for 10 variables in the HILDA. Table entries reflect the proportion of total variance accounted for by that component. All variables were assessed in each of the 20 waves except (), which were included in all but Wave 1.

Variable	Stable Trait	Autoregressive Trait	State
Life Satisfaction	0.36	0.30	0.34
Social Support	0.41	0.31	0.29
General Health	0.50	0.26	0.24
SF-36 Pain	0.24	0.40	0.36
Weight	0.58	0.39	0.03
Physical Activity	0.29	0.34	0.38
Pressed for Time	0.15	0.49	0.35
Household Income	0.24	0.37	0.39
Household Wages	0.38	0.54	0.09
Minutes Commuting	0.14	0.62	0.25

variables, as there are substantial stable-trait components for all ten variables[6]. Second, although the size of stable-trait variance component varies across the ten variables, it is often comparable in size and sometimes exceeds the estimates for the autoregressive component that is the focus of the CLPM. Finally, for almost all of the variables that were analyzed, the state component is also quite large, often accounting for one-quarter to one-third of the variance in these measures. These estimates can be used to evaluate the plausibility of the values that I chose for the simulation studies.

I used the simulations to test how variation in these factors affects the estimated cross-lagged paths when the CLPM is used. A Shiny app is available where variations of this data-generating model can be specified and the effects on cross-lagged paths can be tested:

[6] It is important to note that there are other possible data-generating processes that will lead to the appearance of stable trait variance, including autoregressive effects beyond the first order (Lüdtke & Robitzsch, 2021). However, omitting these higher-order autoregressive effects from a lagged model when they exist will likely have a similar effect on the cross-lagged paths as the omission of a stable trait factor, and thus, these possibilities are not discussed further as the subtle differences are beyond the scope of this paper.

http://shinyapps.org/apps/clpm/[7]. Readers can use this app to examine the specifications described in the text and to test alternatives.

Because the focus of this paper is on examining the effects of unmodeled stable trait variance, I set the variance of the stable trait component for the predictor and outcome to be 1 in the primary simulations (though occasionally, I do set stable-trait variance to zero to address specific questions). I then varied the ratio of autoregressive variance to stable-trait variance across four levels: 0, .5, 1, and 2. Similarly, I varied the ratio of non-state to state variance (which would reflect the reliability of the measures if state variance consisted only of measurement error) across three levels: .5, .7, and .9. The results in Table 1 show that these values correspond to what we might find in real data. Finally, I varied the size of the correlation between the stable traits across four levels from very weak to very strong: .1, .3, .5, and .7. I ran 1,000 simulations for each of five sample sizes: 50, 100, 250, 500, and 1,000). In all simulations, I set the correlation between the initial autoregressive variance components for the predictor and outcome to be .50 and the stability of the autoregressive components to be .50 (though, later, I discuss some modifications to this). I also set the correlations between state components to be 0. Most importantly, all true cross-lagged paths were set to be 0. Consistent with the canonical STARTS model, I included a stationarity constraint, so that variances, correlations, and stability coefficients are constrained to be equal over time. This constraint is not absolutely necessary, but it simplifies discussion of the estimated cross-lagged paths, as there is just one estimate per model.

After generating the data, I tested a simple two-wave CLPM, keeping track of the average size of the estimated cross-lagged paths and the number of cross-lagged paths that were significant at a level of .05. Note that researchers are often interested in determining which of the two variables in the model has a causal impact on the other rather than on

[7] The app and source code are also available on the corresponding OSF site: https://osf.io/4qukz/. In the "Shiny" component, download the "app.R" file and the "scripts" folder and then run it like any other Shiny app.

simply testing the effect of one predictor on an outcome. Thus, an effect of X on Y, Y on X, or both would often be interpreted as a "hit" in common applications of the CLPM. This means that error rates are typically elevated in the CLPM even without unmodeled stable-trait effects unless corrections for multiple comparisons are used. In these simulations, I report the percentage of runs that result in at least one significant cross-lagged effect (out of two tested), and these can be compared to a baseline error rate of approximately 10%, assuming multiple comparisons are ignored.

Finally, although I focus on the common two-wave CLPM design, it is important to note that more waves of data lead to increased power to detect smaller effects—even spurious effects. This means that spurious cross-lagged associations are more likely to be found with better, multi-wave designs. Thus, I will also present results from simulations with more waves of data after presenting the primary results. Code used to generate the data, test the models, and run the simulation are available here: https://osf.io/4qukz/. All analyses were run using R [Version 4.2.1; R Core Team (2021)][8].

Simulation Results

The proportion of simulations that resulted in at least one significant (spurious) cross-lagged effect in this initial simulation are presented in Figure 4. The X-axis shows results for different sample sizes. The Y-axis reflects the percentage of runs in which a significant cross-lagged path was found. The columns reflect variation in the ratio of state to non-state variance of the measures. The rows reflect variation in the ratio of autoregressive variance to stable-trait variance. The individual lines in each plot reflect different correlations between the two stable traits. The averaged estimates for the cross-lagged paths in each set of simulations (averaging across sample sizes, as this will not affect the estimated effect) are

[8] We, furthermore, used the R-packages *dplyr* (Version 1.0.9; Wickham, François, Henry, & Müller, 2021), *ggplot2* (Version 3.3.6; Wickham, 2016), *knitr* (Version 1.39; Xie, 2015), *lavaan* (Version 0.6.12; Rosseel, 2012), *mnormt* (Version 2.1.0; Azzalini & Genz, 2020), *papaja* (Version 0.1.1; Aust & Barth, 2020), and *rethinking* (Version 2.13; McElreath, 2020).

reported in Table 2. What do these simulations tell us about when spurious effects are likely?

When Constructs Have Some Stable-Trait Structure. If the measures include some amount of stable-trait variance–even if the stable traits are uncorrelated—it is likely that spurious cross-lagged paths will emerge. To be clear, this is most problematic when the stable traits are correlated and the correlation is quite high. However, error rates are elevated across most simulations. For instance, consider results in the third column of Figure 4, where most of the variance is non-state variance. Specifically, focus on the fourth row, where the ratio of autoregressive variance to stable-trait variance is 2:1. This panel reflects the least problematic set of values tested, and even here, error rates approach 100% when correlations between the stable traits are strong ($r = .70$) and sample sizes are moderately large ($N = 1{,}000$). Even when correlations are more moderate (e.g., $r = .5$), however, these error rates approach 50% in large samples.

Interestingly, error rates are not always monotonically associated with the size of the correlation when state variance is low. Consider the panels in Rows 2, 3, and 4 of Column 3. In these panels, where the ratio of autoregressive variance to stable-trait variance is .5 or higher, the error rates for the lowest correlation tested ($r = .1$, shown in the solid line) are actually higher than error rates for a higher stable-trait correlation of .3. A look at the actual estimates across simulations in Table 2 provides insight into why this is. This table shows that the average estimated cross-lagged path is actually negative when state variance is low, the correlation between the stable traits is low, and there is a substantial amount of autoregressive variance. These negative estimates emerge even though all associations among the latent components were specified to be positive.

This effect can be demonstrated even more clearly by simulating data with uncorrelated stable traits, an equal amount of autoregressive and stable-trait variance, and no state variance whatsoever (this simulation is not shown in the figure). In this case, the estimated cross-lagged paths will be approximately -.07. This is due to the fact that by

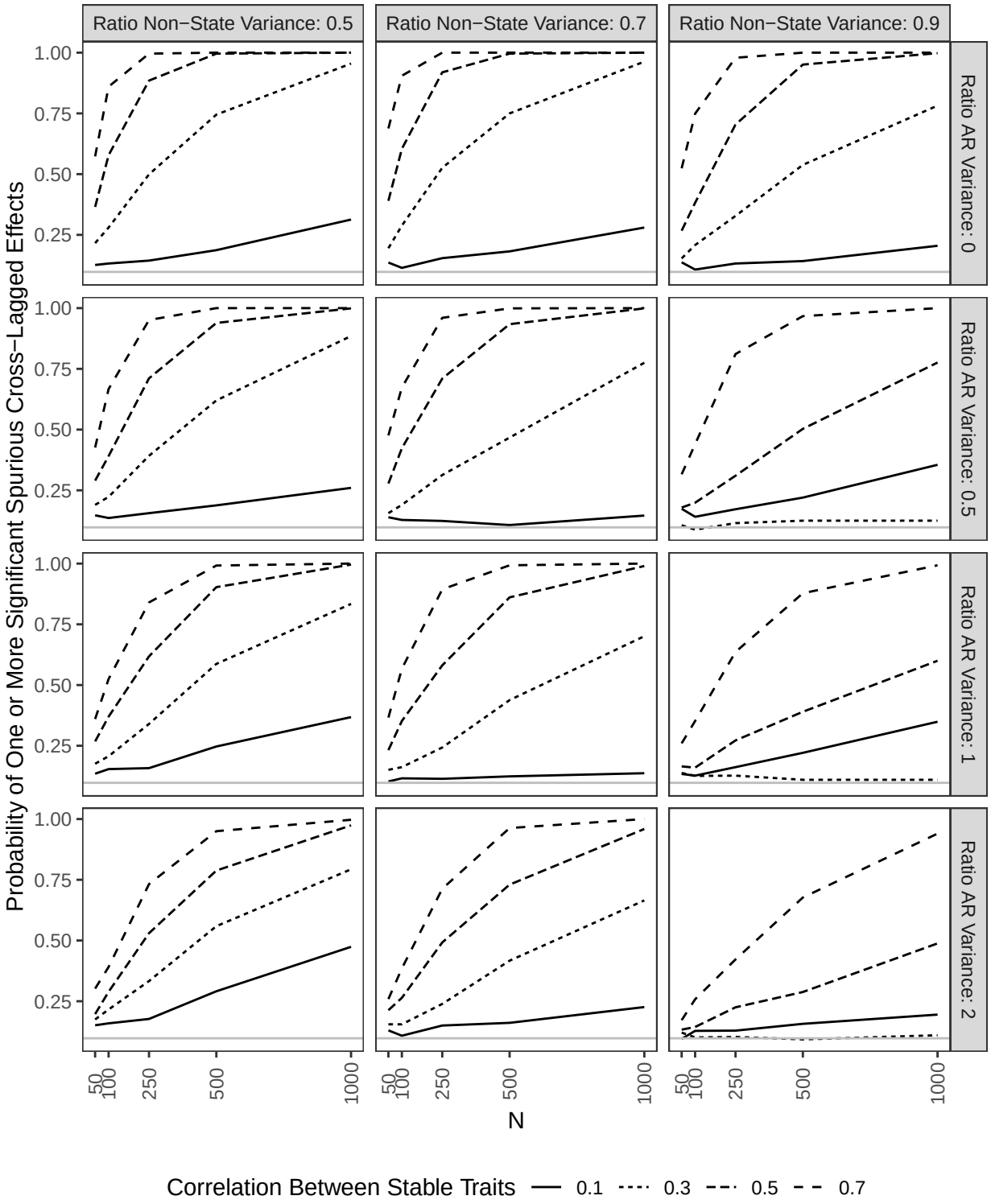


Figure 4. Simulation results for two-wave CLPM. Columns reflect different ratios of non-state to total variance. Rows reflect different ratios of autoregressive to stable-trait variance. Lines reflect different correlations between stable-trait components. Grey line reflects expected number of significant effects due to chance (assuming a critical p-value of .05).

Table 2

Average Estimated Cross-Lagged Paths In Each Simulation Condition

		Non-State Ratio		
Stable Trait r	AR Ratio	0.5	0.7	0.9
0.10	0.00	0.03	0.02	0.01
	0.50	0.03	0.01	-0.02
	1.00	0.03	0.01	-0.03
	2.00	0.04	0.02	-0.02
0.30	0.00	0.08	0.06	0.03
	0.50	0.07	0.05	0.01
	1.00	0.07	0.05	0.01
	2.00	0.07	0.05	0.01
0.50	0.00	0.13	0.12	0.06
	0.50	0.11	0.10	0.05
	1.00	0.10	0.09	0.04
	2.00	0.09	0.08	0.04
0.70	0.00	0.20	0.20	0.11
	0.50	0.16	0.16	0.10
	1.00	0.14	0.14	0.09
	2.00	0.12	0.11	0.07

failing to account for the stable trait, the model overestimates the stability of X and Y, which means that the observed correlation between X at Time 1 and Y at Time 2 is lower than what would be expected based on the initial correlation between X and Y at Time 1 and the stability over time[9]. These simulations show that when the variables being examined have a trait-like structure, this can lead to spurious cross-lagged effects, even when the stable trait variance is not correlated. When correlations at the stable-trait level are strong,

[9] This can be understood by using tracing rules. Randomly generating data for 10,000 participants from the data-generating model just described, the correlation between X_1 and X_2 and between Y_1 and Y_2 are both around .75. The correlation between X_1 and Y_1 would be about .25, and the correlation between X_1 and Y_2 would be about .12. Fitting a CLPM to these data results in estimated stabilities for X and Y of approximately .77, and a correlation between X_1 and Y_1 of .25. These values would imply an observed correlation of $.77 * .25 = .19$ between X_1 and Y_2, which is greater than the actual correlation of .12. This discrepancy between the predicted and observed correlations results in the negative estimates for the cross-lagged paths.

however, the effects of ignoring the stable-trait structure can be substantial. In some realistic scenarios (e.g., moderate correlations between stable traits, reliabilities of .70, and sample sizes over 100), significant spurious cross-lagged paths are almost guaranteed.

It is also worth highlighting that the size of the spurious effects shown in Table 2 are consistent with the size of cross-lagged effects typically found in the literature. For instance, Orth et al. (2022) sampled from papers that used the CLPM to determine how large these effects typically are. In their analysis, the 25th, 50th, and 75th percentiles for cross-lagged effects were .03, .07, and .12. Even the largest of these values is similar to the spurious effects found in realistic scenarios from the simulations[10].

When Measures Have Error or Reliable Occasion-Specific Variance. The simulations described above focus on situations where the ratio of non-state to total variance is very high, or in other words, when state variance is low. When there is a lot of state variance, including either reliable state variance or even just measurement error, this effect gets worse—potentially *much* worse. Consider the panel in the first row and the first column of Figure 4. In this case, the ratio of non-state to total variance is set to .5 and there is no autoregressive variance. Note that values in this range are not unrealistic, because estimated this ratio is reduced both by the existence of measurement error and reliable occasion-specific variance. In this scenario, error rates are very high, approaching 100% with large samples, even when the stable-trait correlation is just .3. Samples of 100 can result in spurious cross-lagged effects approximately 60% of the time when stable traits are correlated .5. Even in samples as small as 50, error rates exceed 25% in many situations.

This outcome is actually quite easy to understand. Indeed, we don't really need simulations at all to predict it. This result is a simple consequence of the issues that Westfall and Yarkoni (2016) discussed and those that I highlighted in Figure 2. Because X and Y are

[10] The values in the table are average unstandardized coeficients, but given the way that the model is specified, they are mostly equivalent to the standardized estimates aggregated by Orth et al. (2022).

measured with error at each occasion, controlling for Time 1 Y when predicting Time 2 Y from Time 1 X does not fully account for the true association between X and Y. There will still be a residual association between Time 1 X and Time 2 Y, which can be accounted for by the freed cross-lagged path in the CLPM. The RI-CLPM (and the STARTS) are useful because they do a better job accounting for this underlying association than the CLPM.

One might argue that a model that just includes stable-trait variance and error (which is true of all simulations in the first row of Figure 4) is unrealistic, as there is sure to be some form of autoregressive structure to most variables psychologists study. That is true, but as the other rows of the figure show, the existence of this stable trait causes problems for the CLPM even when all three sources of variability (stable trait, autoregressive trait, and state) exist.

At this point, it is important to highlight the fact that at least some of these effects are due more to the existence of measurement error (or reliable state variance) than to the existence of the stable trait. For instance, we could simulate data with an autoregressive structure, set the variance of the stable trait components to be 0, and specify no cross-lagged paths. Even with a relatively high ratio of non-state to total variance (e.g., .8 for this simulation, which is not shown in the figure), the average estimated cross-lagged paths would be 0.05 and spurious effects would be found 35% of the time in a two-wave design with samples of 500 participants. Again, Westfall and Yarkoni's (2016) explanation can account for these results: The existence of measurement error or state variance in the observed measures of Y means that controlling for Y_1 does not control for enough. The result is a spurious cross-lagged path.

It is also important to note that measurement error and reliable state variance also affects estimates from the RI-CLPM. If we specify a data-generating process that includes all three sources of variance (stable trait, autoregressive trait, and state/measurement error), but no cross-lagged paths, the CLPM will find substantial cross-lagged effects, but so will

the RI-CLPM (at least if the autoregressive components of X and Y are correlated). To demonstrate, I simulated data with the following characteristics. The X and Y stable traits had variances of 1 and a correlation of .5 and X and Y autoregressive traits had a variance of 1 and a starting correlation of .5 with stability coefficients of .5. The average estimated cross-lagged path was 0.07, which would be easily detectable with moderate to large sample sizes.

To examine this issue more systematically and to compare the likelihood of finding spurious effects when using the RI-CLPM to the likelihood when using the CLPM, I repeated the primary simulation using the RI-CLPM. Because the estimates of the cross-lagged paths are not affected by the size of the correlation between the stable trait components when the RI-CLPM is used, instead of varying the correlation between stable traits, I varied the correlation between the initial wave autoregressive components. In addition, the RI-CLPM requires three waves of data instead of the two that I used in the initial simulation. The results are shown in Figure 5.

As can be seen in this figure, when there is measurement error variance or reliable state variance, then there is a chance for spurious cross-lagged paths even when the RI-CLPM is used. To be sure, these effects are much less likely than with the CLPM: Error rates typically only exceeded 25% with large samples and very strong correlations between the autoregressive traits, whereas they often approached 100% with the CLPM. Note, this limitation of the RI-CLPM is not an argument *for* the CLPM (though it is an argument for using the STARTS or other more complicated models, when possible).

One response to the above simulations is to suggest that researchers simply need to use very reliable measures or perhaps model latent variables at each occasion instead of relying on observed variables with less than perfect reliability. This will certainly help, but it is important to remember that the "state" component in the STARTS model includes measurement error *and* reliable occasion-specific variance. Reliable state variance will affect

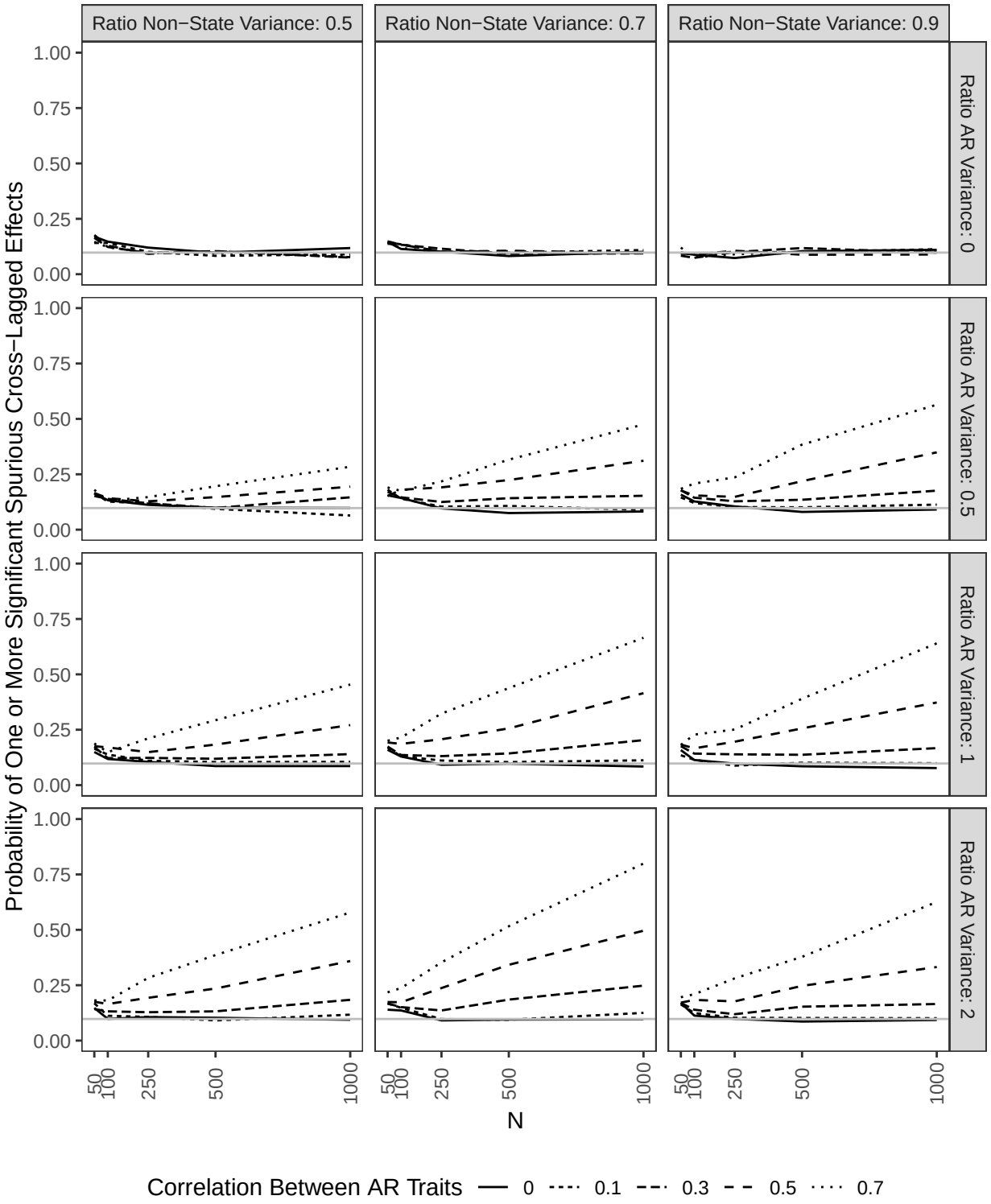


Figure 5. Simulation results for three-wave RI-CLPM. Columns reflect different ratios of non-state to total variance. Rows reflect different ratios of autoregressive to stable-trait variance. Lines reflect different correlations between autoregressive-trait components. Grey line reflects expected number of significant effects due to chance (assuming a critical p-value of .05).

these results in exactly the same way as random measurement error. Unfortunately, researchers don't know how common this reliable state component is in real data, though there is at least some evidence that it can exist and be large enough to be meaningful (Anusic, Lucas, & Donnellan, 2012; Lucas & Donnellan, 2012). Thus, even the use of latent occasions in the CLPM can't solve this problem.

When There Are Many Assessment Waves. Although the CLPM is often used with just two waves of assessment, it can also be used with more complex data. Indeed, a general rule for longitudinal data is that more waves are better than fewer, and in situations where stationarity could reasonably be expected, including more waves and imposing equality constraints should lead to more precise estimates of cross-lagged paths. When estimating true effects, this has the benefit of increasing power. When spurious effects would be expected, however, the use of more waves will also increase the probability of those spurious effects being significant (again, see Westfall & Yarkoni, 2016, for a discussion of how factors that improve power can increase the ability to find spurious effects).

Figure 6 shows a set of simulations that are similar to those reported in Figure 4, but this time using five waves of data and the CLPM with equality constraints across waves. When comparing these two figures, the effect of increasing the number of waves is immediately apparent: Error rates increase considerably. For instance, in the very realistic scenario of an N of 250, a correlation between stable traits of .5, non-state variance ratio of .7, and a 1:1 ratio of stable-trait to autoregressive variance, the error rate increases from % to 97% when moving from a two-wave study to a five-wave study. With five waves of data, error rates often exceed 50%, even in samples as small as 50. So features that are generally desirable—large sample sizes and multiple waves of assessment—increase the likelihood of finding spurious cross-lagged paths.

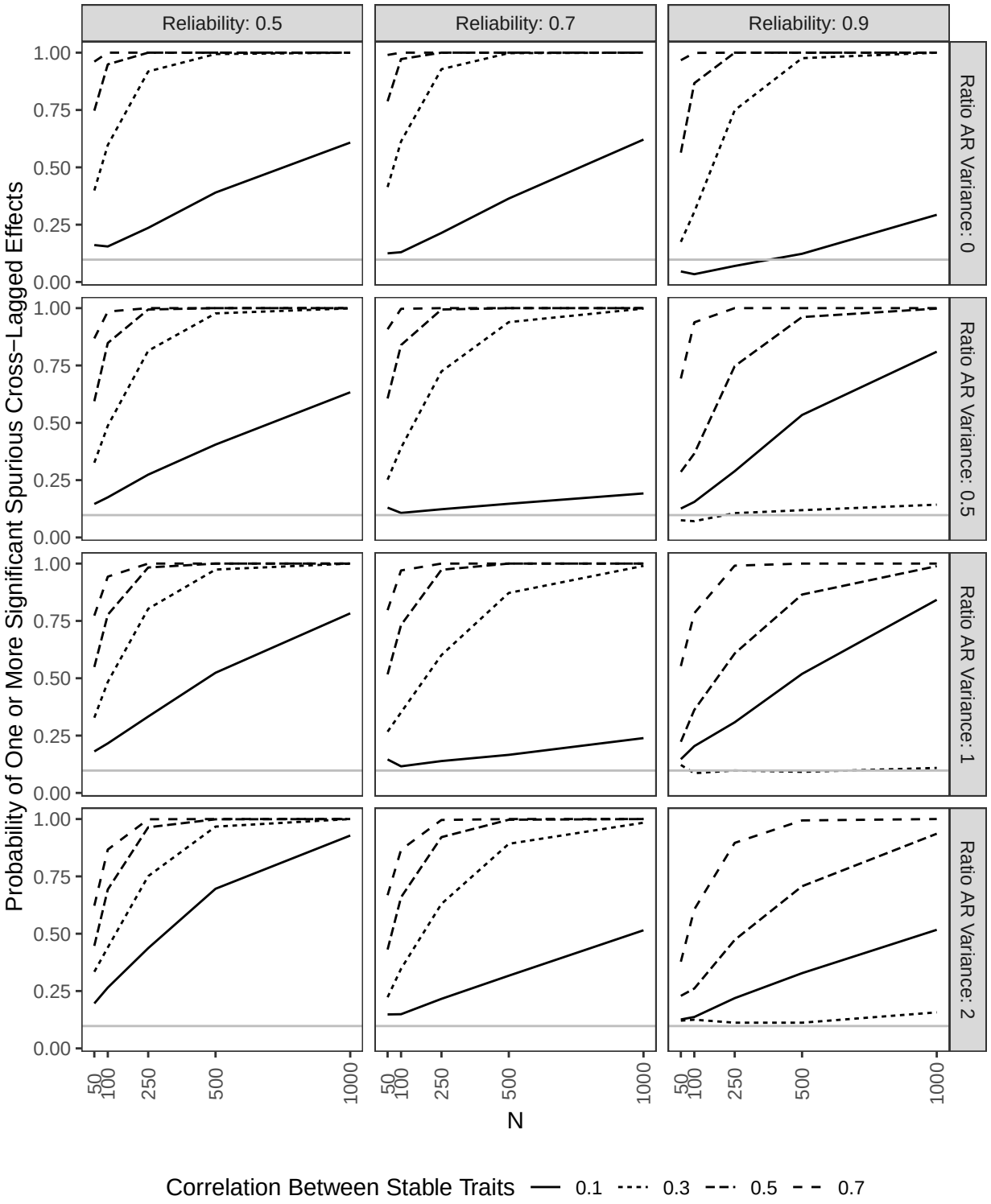


Figure 6. Simulation results for five-wave CLPM. Columns reflect different reliabilities. Rows reflect different ratios of autoregressive to stable-trait variance. Lines reflect different correlations between stable-trait components. Grey line reflects expected number of significant effects due to chance (assuming a critical p-value of .05).

The RI-CLPM Is Not Conservative

The examples above focused on cases where there were no true cross-lagged associations in the data-generating model. The simulations showed that spurious paths are often very likely to be found. This pattern matches the intuition that the RI-CLPM is more conservative than the CLPM (Lüdtke & Robitzsch, 2021). However, failure to model associations between stable-trait components can also lead to the *underestimation* of real cross-lagged paths. For instance, consider a situation where the measures are perfectly reliable (and there is no reliable state variance) and the stable trait and autoregressive trait contribute equally (in this particular case, I also specified the correlations among the stable trait and autoregressive traits to be .5). If we simulate data with cross lagged paths of .5 from X to Y and .2 from Y to X, the RI-CLPM reproduces these effects perfectly. However, even with no measurement error, the estimates from the CLPM are half the size that they should be, approximately .25 and .10.

The precise manner in which estimates will be affected depends on the size of these variance components and the correlations between them. Table 3 shows the results from a separate simulation that examines these effects. Specifically, because the parameter estimates were the focus (rather than the frequency of errors), I followed Lüdtke and Robitzsch (2021) and generated just one set of 10,000 responses for each of 48 combinations. I varied the correlation between the stable traits across four levels: .1, .3, .5, and .7. Similarly, I varied the correlation between the autoregressive traits across the same four levels. I also varied the ratio of autoregressive to trait variance across three levels: .5, 1, and 2. For this example, state variance was set to be 0 and the stability of the autoregressive components were set to .5. The table only shows results for one cross-lagged path, for which the true value is .50.

First consider the example just discussed. Looking at the column where the correlation between the stable traits is .5, and the row where the correlation between the autoregressive traits is .5 and the ratio of autoregressive variance to stable-trait variance is 1, the true

Table 3

Average Estimated Cross-Lagged Path In Each Simulation Condition When True Value = .5.

		Stable Trait Correlation			
AR r	AR Ratio	0.1	0.3	0.5	0.7
0.10	0.50	0.17	0.19	0.24	0.30
0.10	1.00	0.25	0.27	0.33	0.37
0.10	2.00	0.33	0.36	0.38	0.43
0.30	0.50	0.15	0.17	0.20	0.27
0.30	1.00	0.22	0.26	0.29	0.34
0.30	2.00	0.30	0.33	0.36	0.41
0.50	0.50	0.12	0.13	0.17	0.23
0.50	1.00	0.17	0.21	0.25	0.31
0.50	2.00	0.26	0.29	0.34	0.39
0.70	0.50	0.07	0.09	0.12	0.16
0.70	1.00	0.13	0.15	0.19	0.24
0.70	2.00	0.20	0.23	0.28	0.33

Note. AR r = Correlation between autoregressive components; AR Ratio = Ratio of autoregressive variance to stable-trait variance. State variance is set to 0 for all simulations.

cross-lagged path of .5 is estimated to be .25. One can then move up and down that column or across that row to see the effects of the other factors on this underestimation. For instance, looking at the values in the rows immediately above and below this value shows that the underestimation of the cross-lagged paths is greater when there is more stable trait variance than when there is less. The true cross-lagged path of .50 is estimated to be .16 when there is twice as much stable trait variance as autoregressive variance, whereas it is estimated to be .33 (still an underestimate, but not as bad), when there is twice as much autoregressive variance as stable-trait variance.

Moving across the same row shows how this estimate is affected by variation in the correlation between stable traits. As can be seen, the estimate for a true cross-lagged association of .50 declines from .31 when the correlation between the stable traits is a high .70 to .25 when the correlation is .50, to .21 when the correlation is .30, to .18 when the correlation is 0.1. Again, stable trait variance affects estimates of cross-lagged paths even when the stable traits are weakly correlated or uncorrelated.

Finally, moving across groups of rows (e.g., Rows 1 through 3 compared to Rows 4 through 6) shows the effect of the correlation between autoregressive components. In this case, the estimate of the cross-lagged parameter is *negatively* associated with the size of the correlation between autoregressive components, declining from .33 when the correlation between autoregressive components is .10 to .19 when the correlation is .70 (again, for the example where the stable-trait correlation is .5 and there is an equal amount of autoregressive and stable-trait variance).

These simulations show that the RI-CLPM is not more conservative than the CLPM when testing cross-lagged effects. Indeed, when true cross-lagged associations exist, the CLPM is likely to underestimate them when there is stable-trait variance. Again, this pattern is actually easily predictable just by considering tracing rules for structural equation models. When stable-trait variance exists, the stability of the observed variables is overestimated in a CLPM, resulting in a corresponding underestimation of the cross-lagged paths in most situations. If researchers observe a pattern where cross-lagged paths routinely emerge when the CLPM is used but these paths disappear when the RI-CLPM is applied, then this would suggest that the effects themselves are likely spurious.

When considering the implications of this finding, there are two things to keep in mind. First, although the CLPM typically underestimates the true cross-lagged paths, this may not result in decreased power to detect effects relative to alternative models like the RI-CLPM. This is partly due to the fact that models that separate between-person effects from within

frequently have less bias, but at the cost of efficiency; the standard errors in such models are often greater than models that do not correctly separate levels (see Allison, 2009, for an explanation; see Usami, Todo, et al., 2019 for similar examples). Second, it is difficult to consider how to think about power in the context of the CLPM, where it is so easy to find spurious effects. Indeed, I ran simulations similar to those described earlier, but with true cross-lagged effects. However, when simulating data where one cross-lagged path was zero and the other was greater than zero, estimated "power" to detect both the true effect and the spurious one were quite high.

The second thing to remember is that the above simulations were conducted specifying that the measures are perfectly reliable and that there is no occasion-specific state variance. This is unlikely in practice. Indeed, when state variance is included, the estimates from the RI-CLPM are also biased. The precise way that state variance impacts estimates is a quite complicated function of all the factors included in the previous simulations, the state factor, and the direction of the correlations and true cross-lagged paths. Because of this complexity, these simulations are beyond the scope of this paper, though the Shiny app is provided for readers to examine the effect of different combinations on estimated cross-lagged paths. Importantly, the STARTS model is appropriate for modeling data that includes stable-trait, autoregressive-trait, and state variance. Although the STARTS model has been somewhat underused in the literature because of frequent estimation problems, recent methodological advances in Bayesian modeling have helped address these concerns (Lüdtke, Robitzsch, & Wagner, 2018).

Moving Forward with the CLPM

After describing the various approaches available to model longitudinal data, Orth et al. (2021) made the following recommendation: "Before selecting a model, researchers should carefully consider the psychological or developmental process they would like to examine in their research, and then select a model that best estimates that process." There are two

problems, however, when using this guidance to advocate for the CLPM. First, the CLPM is explicitly and inarguably a model of constructs that change over time. Specifically, the CLPM is used to estimate reciprocal associations between two (or more) variables *that have a first-order autoregressive structure*. The model is fundamentally misspecified when used to model variables that have a stable-trait structure. Of course, it is a truism to say that "all models are wrong," but it is rare to select a model that is known to be wrong in such a fundamental way, especially when a more appropriate model exists. The mathematics of the model, along with the simulations provided in this and other papers show that this clear and inarguable misspecification can frequently lead to spurious effects in realistic scenarios. So, if researchers' careful consideration of the psychological and developmental processes under examination leaves open the possibility that the variables have some stable-trait structure, then the CLPM will always be the wrong choice.

Even if we ignore the misspecification involved when applying the CLPM to data with a stable-trait structure, there is an additional problem with Orth et al.'s (2021) guidance. If there is a plausible alternative model that describes the data as well as (or better than) the preferred model, then much additional work is needed to defend that selection. As an obvious example (using the substantive question that motivated Orth et al.'s analysis), if the true causal process linking self-esteem to depression is that changes in self-esteem instantaneously cause a corresponding change in depression (and there are no confounding factors), then that causal effect would be perfectly captured by the cross-sectional correlation between the two variables. Indeed, it would actually be problematic to rely on the cross-lagged association between self-esteem and depression controlling for earlier levels of depression as an estimate of the causal effect. Yet few would find a cross-sectional correlation between self-esteem and depression to be compelling evidence for a causal effect of self-esteem even if a researcher's preferred model only predicted such cross-sectional associations. This is because there are so many plausible alternative models that explain that cross-sectional effect. Unfortunately, the situation is no better with longitudinal data

tested using the CLPM. The CLPM can never rule out the plausible alternative explanation that cross-lagged paths are due to simple between-person differences.

When researchers acquire data from multiple people on multiple occasions, those data have a multilevel structure; this point is uncontroversial. It is also uncontroversial to note that when multilevel data are analyzed using analytic approaches that do not separate these levels, the estimates from those models reflect an uninterpretable mix of between and within-person effects. This concern is especially problematic in the analysis of lagged effects precisely only one level of analysis (the within-person level) can inform causal conclusions. Perhaps there is some reason why this long-standing and well-documented concern about the analysis of multilevel data does not apply in the case of the CLPM, but Orth et al. (2021) did not articulate such a reason. Indeed, they did not acknowledge this problem or discuss how the CLPM solves it. Moreover, it is quite easy to show that the CLPM simply cannot distinguish between the "between-person prospective effects" Orth et al. (2021) claim to want to test and simple between-person differences.

It is also noteworthy that Orth et al. (2021) did not propose a data-generating process that would lead to correct results when modeled using the CLPM but that would result in biased estimates or incorrect conclusions if modeled using the RI-CLPM. It is easy to show that generating data from the model that they say should sometimes be preferred—the CLPM—results in correct estimates (despite the possibility of inadmissable solutions due to negative estimated variances) when analyzed using the RI-CLPM. I believe that it is not possible to specify a data-generating model that corresponds to the processes that the critics describe and that is appropriately modeled using the CLPM but that would lead to incorrect estimates when analyzed using the RI-CLPM or STARTS [11].

[11] Lüdtke and Robitzsch (2021) did show that the RI-CLPM cannot successfully control for all types of confounds, but this is an issue that is distinct from the question of whether we can simply recover the structure of these variables over time. Of course, it is possible to specify data-generating processes that do result in data that, when analyzed using the RI-CLPM, lead to incorrect conclusions (including the more complex models described by Usami, Murayama, et al., 2019). My point is that the critics of the RI-CLPM

In addition to these conceptual reasons for abandoning the CLPM, the simulations reported in this paper show that relying on the CLPM when the variables in the model have some stable trait variance leads to dramatically inflated error rates (often reaching 100% in realistic scenarios, even with small to moderate sample sizes). The constructs that psychologists study very rarely have a purely autoregressive structure. At some point, the long-term stabilities of most constructs are stronger than would be suggested by the short-term stabilities and the length of time that has elapsed alone. This is likely due to the fact that many constructs have at least some stable-trait variance that is maintained over time (even if that stable-trait variance reflects something like response styles or other method factors). And if there is stable trait variance, it is quite plausible that two constructs correlate at the stable-trait level. The RI-CLPM and STARTS models provide a way to test this compelling and plausible model and to adjust for the problematic effects of these stable-trait components when testing reciprocal effects.

So, if using the CLPM results in dramatically elevated error rates for cross-lagged paths when they do not exist, while also underestimating estimates of these associations when they do exist (again, sometimes dramatically in realistic scenarios), should we ever rely on the CLPM for causal analyses? My suggestion in the title of this paper that the CLPM be abandoned was not intended to be attention-grabbing hyperbole; it simply reflects the difficulty identifying any situation where the CLPM would be preferable to alternatives[12]. When at least three waves of data are available, the RI-CLPM can be used. If no stable trait variance exists, the RI-CLPM will simply reduce to a CLPM.

To be sure, there are certain situations when alternatives to the CLPM cannot be used, most notably in the very common situation where only two waves of assessment are available.

---

have not provided a model that matches the processes that they describe, that results in correct estimates when modeled using the CLPM, but also results in incorrect estimates when modeled using the RI-CLPM.

[12] Indeed, although they do not state this quite so emphatically, Zyphur, Voelkle, et al. (2020) also argued that the use of the standard CLPM be abandoned. They stated that "In sum, classic cross-lagged SEMs and MLMs should be avoided when seeking to make causal inferences in panel data" (p. 700).

However, as Figure 2 showed, there are too many plausible alternative models that can lead to the same set of six correlations among two variables at two time points to draw any conclusions about which of those models is correct. Many have suggested that two-wave designs barely count as "longitudinal" (e.g., Ployhart & MacKenzie, 2014; Rogosa, 1995), and these authors' reasons for this claim are quite salient when considering the application of the CLPM to such data. There is simply not enough information available in these data to distinguish among multiple competing models. In discussing reciprocal associations using cross-lagged analyses, Rogosa (1995) noted that there is a "hierarchy of research questions about longitudinal data [that] might start with describing how a single attribute—say aggression—changes over time. A next step would be questions about individual differences in change of aggression over time, especially correlates of change in aggression. Only after such questions are well understood does it seem reasonable to address a question about feedback or reciprocal effects, such as how change in aggression relates to change in exposure to TV violence or, does TV violence cause aggressive behavior?" (p. 34). Many researchers have noted that this first step—describing how a construct changes over time—is not possible with only two waves of data (e.g., Fraley & Roberts, 2005; Ployhart & MacKenzie, 2014; Rogosa, 1995). Because we cannot understand the longitudinal structure of the variables with two waves of data, the estimates from a two-wave CLPM are uninterpretable.

All researchers want to obtain information about the phenomena they study as efficiently as possible. The widely used CLPM is a simple analysis that can be applied in many situation with very little data. Unfortunately, this simple model is not up to the task of clarifying causal processes in longitudinal data. By failing to separate between-person and within-person levels, the CLPM cannot distinguish over-time effects from simple between-person associations. Simulations show that the CLPM results in extremely elevated error rates when stable-trait (or state) variance exists; spurious associations are very likely in many different realistic scenarios. This confirms what methodologists have highlighted for many years: By failing to account for the multilevel structure of longitudinal data, models

like the CLPM result in uninterpretable estimates. Berry and Willoughby (2017) suggested that it was time to rethink the CLPM, which they described as a workhorse of developmental research. I concur that the introduction of useful alternatives like the RI-CLPM and STARTS, when combined with the demonstrable problems with the CLPM, show that it is time for this workhorse to be retired.

Disclosures

Author Contributions

Richard E. Lucas was responsible for all contributions, including conceptualization, methodology, formal analysis, and writing.

Conflicts of Interest

The author declares that there were no conflicts of interest with respect to the authorship or the publication of this article.

Prior Versions

A preprint of this paper was posted on the PsyArXiv preprint server: https://psyarxiv.com/pkec7/ .

Data Usage

This paper uses unit record data from Household, Income and Labour Dynamics in Australia Survey [HILDA] conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this paper, however, are those of the author[s] and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. DOI: ####

References

Allison, P. D. (2009). *Fixed effects regression models*. Sage.

Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Dependability of Personality, Life Satisfaction, and Affect in Short-Term Longitudinal Data. *Journal of Personality*, *80*(1), 33–58. https://doi.org/10.1111/j.1467-6494.2011.00714.x

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Azzalini, A., & Genz, A. (2020). *The R package `mnormt`: The multivariate normal and t distributions (version 2.0.2)*. Retrieved from http://azzalini.stat.unipd.it/SW/Pkg-mnormt/

Berry, D., & Willoughby, M. T. (2017). On the Practical Interpretability of Cross-Lagged Panel Models: Rethinking a Developmental Workhorse. *Child Development*, *88*(4), 1186–1206. https://doi.org/gbf8jt

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and Conceptual Problems With Longitudinal Trait-State Models: Introducing a Trait-State-Occasion Model. *Psychological Methods*, *10*(1), 3–20. https://doi.org/10.1037/1082-989x.10.1.3

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*(1), 583–619. https://doi.org/d54m6f

Curran, P. J., Howard, A. L., Bainter, S. A., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person and within-person components of individual change over time: A latent curve model with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*(5), 879–894. https://doi.org/10.1037/a0035297

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121.

Falkenström, F., Solomonov, N., & Rubel, J. A. (2022). How to model and interpret cross-lagged effects in psychotherapy mechanisms of change research: A comparison of multilevel and structural equation models. *Journal of Consulting and Clinical Psychology*, *90*(5), 446–458. https://doi.org/10.1037/ccp0000727

Fraley, R. C., & Roberts, B. W. (2005). Patterns of Continuity: A Dynamic Model for Conceptualizing the Stability of Individual Differences in Psychological Constructs Across the Life Course. *Psychological Review*, *112*(1), 60–74. https://doi.org/10.1037/0033-295X.112.1.60

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. https://doi.org/f67cvh

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. https://doi.org/ggbk25

Hertzog, C., & Nesselroade, J. R. (1987). Beyond Autoregressive Models: Some Implications of the Trait-State Distinction for the Structural Modeling of Developmental Change. *Child Development*, *58*(1), 93–109. https://doi.org/10.2307/1130294

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*(1), 52–59. https://doi.org/10.1037/0022-006X.63.1.52

Kenny, D. A., & Zautra, A. (2001). The trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New Methods for the Analysis of Change* (pp. 243–263). Washington, DC: American Psychological Association.

Lucas, R. E., & Donnellan, M. B. (2012). Estimating the Reliability of Single-Item Life Satisfaction Measures: Results from Four National Panel Studies. *Social Indicators Research*, *3*, 323–331.

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

Lüdtke, O., Robitzsch, A., & Wagner, J. (2018). More stable estimation of the STARTS model: A Bayesian approach using Markov chain Monte Carlo techniques. *Psychological Methods*, *23*(3), 570–593. https://doi.org/gd86g5

McElreath, R. (2020). *Rethinking: Statistical rethinking book package*.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10099-006

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034. https://doi.org/gg7zfw

Orth, U., Meier, L., L., Buhler, J. L., Dapp, L. C., Krauss, S., Messerli, D., & Robins, R. W. (2022). Effect Size Guidelines for Cross-Lagged Effects. *Psychological Methods*.

Ployhart, R. E., & MacKenzie, W. I. (2014). Two Waves of Measurement Do Not a Longitudinal Study Make. In C. E. Lance & R. J. Vandenberg (Eds.), *More Statistical and Methodological Myths and Urban Legends* (pp. 85–99). Routledge.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models* (Second). Thousand Oaks, CA: Sage.

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (Ed.), *The Analysis of Change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum Associates.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from https://www.jstatsoft.org/v48/i02/

Usami, S. (2020). On the Differences between General Cross-Lagged Panel Model and Random-Intercept Cross-Lagged Panel Model: Interpretation of Cross-Lagged Parameters and Model Choice. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–14. https://doi.org/gj6znt

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal

models to examine reciprocal relations. *Psychological Methods*, *24*(5), 637–657. https://doi.org/gf4fqx

Usami, S., Todo, N., & Murayama, K. (2019). Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models. *PLOS ONE*, *14*(9), e0209133. https://doi.org/10.1371/journal.pone.0209133

Watson, N., & Wooden, M. (2012). The HILDA Survey: A Case Study in the Design and Development of a Successful Household Panel Study. *Longitudinal and Life Course Studies*, *3*(3), 369–381.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., … Diener, E. (2020). From Data to Causes I: Building A General Cross-Lagged Panel Model (GCLM). *Organizational Research Methods*, *23*(4), 651–687. https://doi.org/gf8rt5

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., … Diener, E. (2020). From Data to Causes II: Comparing Approaches to Panel Data Analysis. *Organizational Research Methods*, *23*(4), 688–716. https://doi.org/gf8rt7