It's Time To Abandon the Cross-Lagged Panel Model for Causal Analyses

Richard E. Lucas[1]

[1] Department of Psychology, Michigan State University

Abstract

CLPM

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

It's Time To Abandon the Cross-Lagged Panel Model for Causal Analyses

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.

The CLPM improves on simpler cross-sectional analyses by controlling for contemporaneous associations between the predictor and outcome. Presumably, confounding factors should be reflected in this association, which would mean that any additional cross-lagged associations 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. 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. They described the RI-CLPM as a multilevel model that separates between-person effects from within-person effects. As others have noted (e.g., Lüdtke & Robitzsch, 2021), this critique of the cross-lagged panel model has already been cited frequently and has 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. Critics of the RI-CLPM (e.g., Lüdtke & Robitzsch, 2021; Orth, Clark, Donnellan, & Robins, 2021) have argued that sometimes

researchers are actually interested in the between-person effects that a classic CLPM tests and that the choice of model should depend on one's theories about the underlying process. The goal of this paper is to examine these critiques, focusing first on the accuracy of the critical interpretation of the RI-CLPM, followed by simulations that demonstrate the problems with the CLPM and the utility of its alternatives. These simulations show that spurious cross-lagged associations are common and the likelihood of finding such spurious effects can reach 100% in many realistic scenarios. I conclude that there is no situation where the CLPM is preferable to the RI-CLPM and the CLPM should probably be abandoned as an approach for examining causal processes in longitudinal data.

A Note About Models and Terminology

Before I address the critiques of the RI-CLPM, it is necessary to clarify the terminology that I will use when describing the components of the models. The CLPM, the RI-CLPM, and a slightly more complex model—the bivariate Stable Trait, Autoregressive Trait, State (STARTS) model (Kenny & Zautra, 1995, 2001)—are presented in Panels A, B, and C of Figure 1. The common feature across all three models is that they include 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. The developers and critics of the RI-CLPM both refer to the autoregressive part of the model as the "within-person" part[1], but for reasons discussed below, I will follow Kenny and Zautra (2001) and refer to this as the "autoregressive" part. Similarly, I will rely on the STARTS terminology when describing the other components of the models.

The only difference between the CLPM and the RI-CLPM is that the RI-CLPM includes a random-intercept (labeled "Stable Trait" in the figure, according the STARTS terminology) that accounts for "time-invariant, trait-like stability" (Hamaker et al., 2015, p. 104). Thus, the CLPM is nested within the RI-CLPM; the CLPM is equivalent to the

[1] As do others, see, e.g., Curran, Howard, Bainter, Lane, and McGinley (2014).

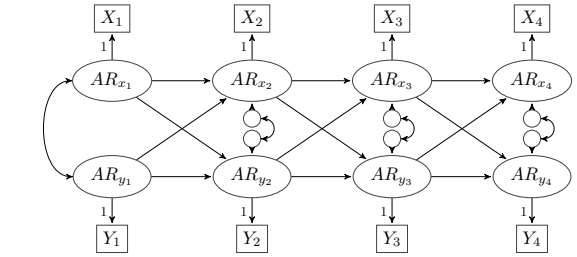
RI-CLPM with the random-intercept (or stable-trait) variance constrained to 0.

Notice that neither the CLPM nor the RI-CLPM include any measurement-error variance for the indicators. For the CLPM, 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). For the RI-CLPM, the observed variables are determined by the random intercept and the corresponding wave-specific latent variable from the autoregressive part of the model.

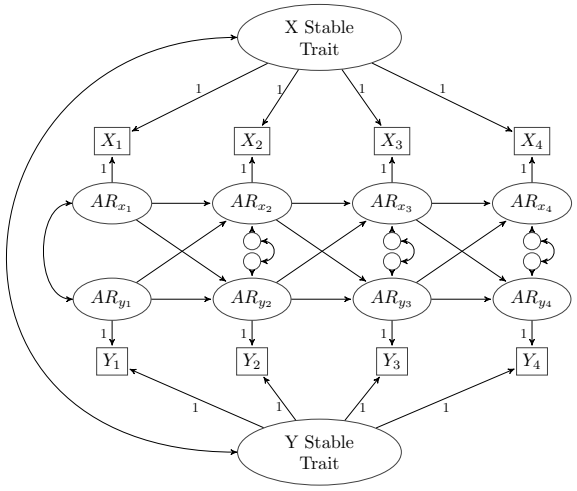
As can be seen in Figure 1, the only difference between the RI-CLPM and the STARTS model is the inclusion of a wave-specific "state" component (labeled s_t in the figure), which reflects variance in an observed variable that is perfectly "state-like" and unique to that occasion. Note that this state 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, but simpler models like the RI-CLPM have often been preferred because the STARTS requires more waves of data than the RI-CLPM and often has estimation problems (e.g., Cole, Martin, & Steiger, 2005; Orth et al., 2021).

Recently, Usami, Murayama, and Hamaker (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. For instance, an alternative model—the Latent Curve Model with Structured Residuals (Curran et al., 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 model, then all the models covered in this paper would be

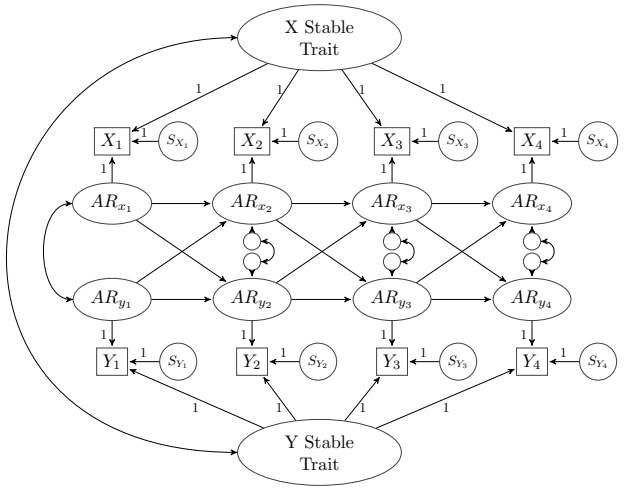
misspecified and could lead to biased estimates.



Panel A: CLPM



Panel B: RI-CLPM



Panel C: STARTS

Figure 1. Diagram of the three models used in this paper.

The Ambiguous Nature of "Between" Versus "Within"

In presenting the RI-CLPM, Hamaker et al. (2015) emphasized that a strength was its ability to separate between-person effects from within-person effects. Critics of the RI-CLPM focus on this distinction when explaining their concerns with the model. The terms "within-person" and "between-person," however, are ambiguous. As Usami (2021) recently noted, these terms are used differently in different contexts. I posit that this ambiguity and inconsistent usage has led to incorrect interpretations of the RI-CLPM and its alternatives.

When describing the RI-CLPM, Orth et al. (2021) argued 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). They go on to note that "in many fields researchers are also interested in gaining information about the consequences of between-person differences" (p. 1014). Similarly, in their critique of the RI-CLPM, Lüdtke and Robitzsch (2021) 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). Thus, these critics' central objection to the RI-CLPM is that it isolates within-person effects when sometimes that is not desirable[2]. But what is a within-person effect?

When answering this question, both Lüdtke and Robitzsch (2021) and Orth et al. (2021) are precise but inconsistent. For instance, Orth et al. (2021) initially state that "in the RI-CLPM, a cross-lagged effect indicates whether a within-person deviation from the trait level of one construct has a prospective effect on change in the within-person deviation from the trait level of the other construct" (p. 1014). This emphasis on "deviations from the

[2] Though Lüdtke and Robitzsch (2021) did also articulate some additional concerns about the ability of the RI-CLPM to adequately account for unobserved confounders.

trait level" reflects an accurate interpretation of the "within-person" or autoregressive part of the RI-CLPM. Similarly, Lüdtke and Robitzsch (2021) initially describe the within-person parts of the RI-CLPM as "deviations from between-person parts" of the model (p. 2), which is also correct. However, both authors later restate these description in ways that are either less precise or wrong. Importantly, the specific reasons both authors provide for preferring the CLPM to the RI-CLPM follow directly from the incorrect version of their description.

For instance, Orth et al. (2021) rephrase their original statement about "deviations from the trait level" to say that a within-person effect (in the context of their substantive example—a test of the causal effect of self-esteem on depression) means that "When individuals have lower self-esteem than usual, they will experience a subsequent increase in depression." But what is this "usual" level from which people deviate? Can we interpret the random intercept from the RI-CLPM as a person's "usual level" or their individual mean (as Lüdtke & Robitzsch, 2021, did)? And what exactly is left after scores are deviated from this random intercept?

Both Orth et al. (2021) appear to (incorrectly) assume that by including a random intercept, the RI-CLPM removes all between-person variance from the cross-lagged part of the model. For example, Orth et al. (2021) state this explicitly: They 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). But this statement reflects a fundamental misunderstanding of the RI-CLPM, one that results from the ambiguous use of the terms "between" and "within." Later on the same page, Orth et al. state that "The RI-CLPM includes [an] unrealistic assumption, specifically that the between-person variance in constructs is perfectly stable" (p. 1026). But again, this is wrong: The authors who proposed the RI-CLPM simply restricted the term "between-persons" to refer the variance that is perfectly stable over time. This is an issue of terminology. The RI-CLPM does not

assume that all between-persons variance is perfectly stable, it simply *defines* "between-persons" variance as the variance that *is* perfectly stable. There is clearly between-person variance (broadly defined) left in the cross-lagged part of the model.

An Example

To demonstrate, take the following example, where data are generated from a true autoregressive model with cross-lagged paths. In other words, the data-generating model looks like Panel A of Figure 1. For comparison with the substantive discussion by Orth et al., let's assume that the predictor is self-esteem and the outcome is depression. Specifically, in this example, I generated data for 10 waves of self-esteem and depression data, with no random intercept, starting variance of 1 for self-esteem and depression, stabilities of .5 for each, true cross-lagged paths of .50 from self-esteem to depression and .00 from depression to self-esteem. I also specified a Wave 1 correlation between self-esteem and depression of -.5. The code for the function to generate the data is available here and included in the appendix.

According to Orth et al. (2021), significant cross-lagged paths from self-esteem to depression in this model can be interpreted to mean 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." In other words, this model links between-person differences in self-esteem at Time 1 to change in depression from Time 1 to Time 2. They argue that this is precisely what many researchers would want to estimate in many common situations.

They also argue that when you test a model that includes random intercepts, the interpretation of the cross-lagged paths change. They state that a significant cross-lagged path in the context of the RI-CLPM means that "When individuals have lower self-esteem than usual, they will experience a subsequent increase in depression" (p. 1014). The problem is that what is captured by the "stable trait" in the RI-CLPM model is potentially (and

frequently) different than a person's "average level" or "usual level." These two things can be very different, both conceptually and empirically. This is because the stable trait does not incorporate all between-person variance; it only includes variance that is perfectly stable across all waves. So what a substantial cross-lagged path really reflects in a RI-CLPM context is that "When individuals have lower self-esteem *than what would be predicted from their levels on the stable trait*, they will experience a subsequent increase in depression."

This distinction may sound subtle, but it is important. For one thing, even when clear "between-person" differences exist in the data—for instance, in the data I generated—if one tries to fit the RI-CLPM to data without a stable-trait component, the estimate for the variance of the random intercept will be zero. If the RI-CLPM simply took what between-person variance exists and "relegated" it to a stable-trait component that reflected a person's long-term average, you would always be able to find a random intercept as long as variance in person means existed.

Moreover, although one will not always find random-intercept variance when testing the RI-CLPM on data with real between-person differences, it is always possible to do what Orth et al. (2021) and Lüdtke and Robitzsch (2021) claim the RI-CLPM does, which is to ipsatize the predictor and outcome variables by deviating them from a person's mean and then computing cross-lagged associations using this mean-deviated data. This procedure truly does separate all between-person variance from within; there would be no variance left in person-level means after ipsatizing in this way. To demonstrate and compare ipsatizing to fitting an RI-CLPM model, I subtracted each person's mean self-esteem score and mean depression score in each wave and then reran the CLPM using these ipsatized data. For comparison, I also ran the RI-CLPM on the original, undeviated data[3]. The first three pairs

[3] Note that the CLPM model will often result in nonpositive-definite matrices with the mean-deviated data. I chose a random seed that generates data where all models converge, but this may not be true for other randomly generated data. In addition, as expected, fitting the RI-CLPM to data with a true autoregressive structure resulted in negative variances (because the true variance of the random-intercept is zero, so estimates can drop below). The estimates for the RI-CLPM model in the table come from a solution with

of columns from Table 1 show the relevant estimates for the original CLPM and these two alternatives.

---

negative variances and are provided simply to show that they are close to the true parameters, even when the overall model solution is inadmissable.

Table 1

Comparison of Estimates for Cross-Lagged Effects

	CLPM		Deviated CLPM		RI-CLPM		RI-CLPM with Trait	
	est	se	est	se	est	se	est	se
Stability of Self-Esteem	0.50	0.00	0.36	0.00	0.50	0.00	0.50	0.00
Stability of Depression	0.50	0.00	0.38	0.00	0.50	0.00	0.50	0.00
Self-Esteem Predicted by Depression	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00
Depression Predicted by Self-Esteem	-0.50	0.00	-0.49	0.00	-0.50	0.00	-0.50	0.00
Variance of Self-Esteem AR Trait	1.01	0.01	0.87	0.01	1.01	0.02	1.02	0.02
Variance of Depression AR Trait	0.99	0.01	0.79	0.01	0.99	0.02	0.99	0.02
Covariance Between AR Components	-0.49	0.01	-0.37	0.01	-0.49	0.01	-0.50	0.01
Variance of Self-Esteem Stable Trait					0.00	0.00	0.49	0.01
Variance of Depression Stable Trait					0.00	0.01	0.51	0.01
Covariance of Stable Traits					0.00	0.01	-0.40	0.01

Note. The estimates for the CLPM columns reflect the CLPM fit to the original data. The estimates in the Deviated ClPM columns reflect the CLPM fit to the mean-deviated data. The estimates for the RI-CLPM column reflect the RI-CLPM fit to the original data. The estimates for the RI-CLPM column reflect the RI-CLPM fit to the original data with the stable trait variance added.

As can be seen in this table, the estimates from the mean-deviated CLPM model are quite different both from the CLPM and the RI-CLPM fit to the original (undeviated) data. In contrast, the estimates from the RI-CLPM are almost identical to those from the CLPM. Thus, the RI-CLPM is not equivalent to a CLPM with ipsatized (mean-deviated) data. Again, the cross-lagged part of the RI-CLPM reflects deviations from the latent stable trait, and if there is no variance in this stable trait, then the RI-CLPM is equivalent to the CLPM. The fact that the CLPM is a reduced version of the RI-CLPM where the random-intercept variance is set to 0 is, of course, obvious just by looking at the models. The point of this example is to clarify what that means for the interpretation of these cross-lagged paths.

A potential response to this example would be to acknowledge that using the RI-CLPM will not artificially create random-intercept variance when it does not exist, while still arguing that when stable between-person variance does exist, the model somehow distorts the cross-lagged paths. But this is also not correct. Imagine that we take the same data from above, but now we add some stable-trait variance. Let's assume for this example that the added stable-trait variance is just shared method variance. Specifically, we can generate data representing a method factor for self-esteem and one for depression. In generating these data, I set the variance of these stable-trait components to .50, and I assumed that the stable traits are pretty strongly (and negatively, given the direction of the items) correlated (though this doesn't really matter for this specific example). We can then just add the generated method-variance scores to the original data.

If we fit the RI-CLPM to the combined data, you get exactly what you'd expect: The estimates for the random intercept simply reflect the method variance we've added. These results are shown in the final columns of Table 1. The variance of each estimated random intercept is about .50 and the covariance is about -.40 (for a correlation of -.80). These random intercepts don't capture any of the between-person differences in the original data (which are substantial) because the original individual differences are not *perfectly* stable

across waves. In other words, the model does not, as Orth et al. suggested "relegate all between-person differences to the random intercept," it simply pulls out those between-person differences that are completely stable over time.

More importantly, Table 1 shows that the estimates for the original autoregressive, cross-lagged part of the model are the same as what you get when you run the CLPM on the original data. The variances, stabilities, and cross-lagged paths are the same (as they should be). Contrary to Orth et al., the cross-lagged part of the model still links the same between-person variance in self-esteem at Time 1 to between-person variance in depression at Time 2. The interpretation of this part of the model is exactly the same when no stable-trait variance exists as it is when there is stable-trait variance, but that stable-trait variance is modeled using the RI-CLPM. Specifically, in the RI-CLPM, the cross-lagged effects still assess whether those who have high self-esteem relative to others at Time 1 report larger changes in depression over time. The difference is that these effects have now been adjusted for stable differences in measurement error; they have certainly not been "ipsatized" in any way.

Of course, in real data, we typically do not know whether the stable-trait variance that exists reflects something that is clearly not theoretically interesting (as is true of method variance). My point is that adding some stable trait variance (in this case, method variance) and then modeling this stable variance using the RI-CLPM doesn't suddenly make the cross-lagged part any more "within-persons" than it was before. Indeed, I am pretty sure that what Orth et al. (2021) would call a between-persons model—the CLPM fit to the original data I generated—would actually be described by Hamaker et al. (2015) as a within-persons model, not because of what the model is doing, but simply because there is no stable-trait variance. Again, this is an issue of terminology.

An additional response that the critics of the RI-CLPM could make is that there is some important reason to avoid separating purely stable between-person variance from the

between-person variance that remains in the autoregressive part of an RI-CPLM or STARTS model when testing causal associations with longitudinal data. However, this is *not* the argument that these critics made. They explicitly state that their objection to the RI-CLPM was that it removed all between person differences from the autoregressive/cross-lagged part of the model, which resulted in cross-lagged paths that capture only associations between ipsatized scores (which reflect deviations from people's usual level) for the predictor and outcome. Perhaps there is a reason why both types of individual differences (those captured by the stable trait and those that remain in the autoregressive part of the model) would be expected to be associated with wave-to-wave changes, but these reasons were not put forth.

It is noteworthy that neither Lüdtke and Robitzsch (2021) nor Orth et al. (2021) described a data-generating model that would result in biased estimates or incorrect conclusions if analyzed incorrectly using the RI-CLPM. As I demonstrated above, generating data from the model that they claim to prefer—the CLPM—results in correct estimates (despite 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 that would also lead to incorrect estimates when analyzed using the RI-CLPM[4].

The Source of Confusion

I think the source of the confusion about the RI-CLPM is that people use the terms "between" and "within" in different ways in different contexts (Usami, 2021). Many people (at least in my field) are used to thinking of the differences between within-person effects and between-person effects in the context of multilevel modeling. We are often warned that when

[4] 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 models that do result in data that, when analyzed using the RI-CLPM, lead to incorrect conclusions (including the more complex models described by Usami et al. (2019)). My point is that the critics of the RI-CLPM have not provided a model that matches the processes that they describe and also results in incorrect estimates when modeled using the RI-CLPM.

testing multilevel models, if we are not careful about how we enter variables into the model, what may look like within-person effects (e.g., the "Level-1" effects in repeated-measures data analyzed in a multilevel modeling framework) can actually reflect a mix of between- and within-person associations. The recommended solution is often to person-center the predictor (e.g., Enders & Tofighi, 2007), where each observation now reflects a deviation from a person's mean. When centering this way, the Level-1 part of these models completely separates within-person variance from between—all between-person variance has been removed from this part of the model (at least in the predictor).

So when methodologists talk about separating "within" from "between," there are at least two possible meanings to this distinction. In the first (used by Curran et al., 2014; Hamaker et al., 2015, and others to describe models with a stable-trait component), "between-person" variance and effects refer to those that are perfectly stable over time, and "within-person" variance and effects include everything that is not perfectly stable. In the second meaning, "within-person" variance and effects are those that involve ipsatized scores from which all "between-person" variance (broadly defined) has been removed. These two meanings are not equivalent.

Why does this matter? Critics of the RI-CLPM appear to interpret the within/between distinction in the "ipsatizing" way, and they derive their concerns about the model from this interpretation. They explicitly state that their primary reason for avoiding the RI-CLPM is that it focuses so narrowly on within-person deviations from a person's "usual level". For instance, as previously noted, Lüdtke and Robitzsch (2021) 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." But if the first part of that statement is incorrect, then the second would not follow. Similarly, all of the limitations of the RI-CLPM and all of

the reasons for preferring the CLPM to the RI-CLPM discussed by Orth et al. (2021, p. 1026) are based on an incorrect description of what the cross-lagged part of the RI-CLPM really does. Orth et al. conclude that "the RI-CLPM does not allow testing what many researchers ... are interested in: the prospective between-person effect" (p. 1026). I think the very term "prospective between-person effect" reflects the confusion about the meaning of the terms "between" and "within" that I noted; but to the extent that such a thing exists, the RI-CLPM captures it just as well as the CLPM.

Beyond "Within" and "Between"

Given the ambiguity of the terms "within" and "between" in the context of models like the CLPM and its alternatives, it can be helpful discuss benefits of the RI-CLPM over the CLPM without referring to this distinction[5]. Specifically, the RI-CLPM is useful because it tests an extremely plausible alternative explanation of the underlying pattern of correlations that is being modeled when the CLPM is used. It is important to consider this alternative explanation when testing the CLPM, regardless of what theoretical model the researcher prefers.

The logic of the CLPM is very similar to the logic of any other regression model where we assess whether one variable predicts another after controlling for relevant confounds. When we test whether Time 1 X predicts Time 2 Y after controlling for Time 1 Y, we 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 X) is not a perfect measure of what we're trying to account for, then it is possible—indeed, quite easy—to find spurious "incremental validity" effects. This, I think, is

[5] Though ultimately, I think that Hamaker et al.'s (2015) framing is a more precise way of saying the same thing; my point is that we don't necessarily need to the within/between framing to understand the problems with the CLPM.

a simpler way of thinking about the strengths of the RI-CLPM relative to the CLPM.

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

The problem with the CLPM is that is easy—in fact, extremely easy—to find spurious cross-lagged associations under conditions that are extremely likely in the typical situations where the CLPM is used. Hamaker et al. (2015) 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 simply 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 guaranteed to find spurious cross-lagged effects.

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 choose to study over time have a STARTS-like 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 and true state-like influences. Starting with this assumption, it is then possible to test how variation in these factors affect the estimated cross-lagged paths when the CLPM is used. A Shiny app is available where variations of this starting model can be specified and the effects on cross-lagged paths can be tested: (shinyapps.io/rlucas11/clpm)[https://shinyapps.io/rlucas11/clpm]. 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 "reliability" of the measures (defining reliability as the proportion of variance due to stable-trait and autoregressive-trait variance) across three levels: .5, .7, and .9. Even the lowest level is not unrealistic, given that the state component includes both measurement error and reliable state variance. 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 six sample sizes: 25, 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. 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 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 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 readers can interpret these results either in comparison to a baseline error rate of 5% or 10% depending on how these multiple comparisons are considered.

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.

The proportion of simulations that resulted in at least one significant cross-lagged effect are presented in Figure 2. The X-axis shows results for different sample sizes. The Y-axis reflects the percentage of runs in which a significant (spurious) cross-lagged path was found. Ideally, this would be close to 10%, which would be the error rate without taking multiple comparisons into account[6]. The columns reflect variation in the reliability 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) is reported in Table 2. So, when can you find spurious effect? Here are a few common situations.

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 2, where the reliability is a very high .9. 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 large ($N = 1{,}000$). Even when correlations are more moderate (e.g., $r = .5$), however, these error rates approach 50%

[6] More precisely, the error rate based on chance alone should be .0975.

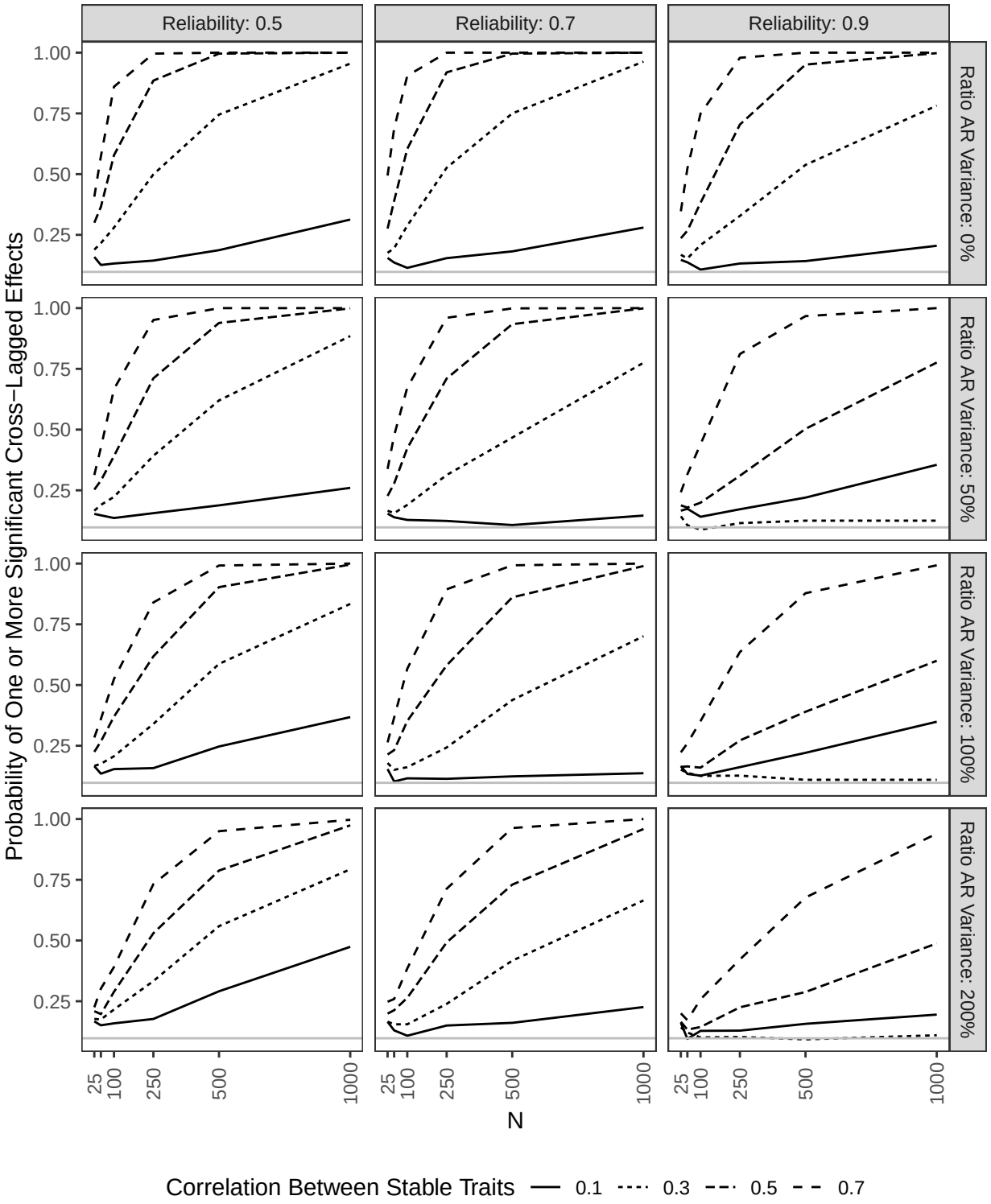


Figure 2. Simulation results for two-wave CLPM. Columns reflect different reliabilities. Rows reflect different ratios of autoregressive to stable-trait variance. Lines reflect different correlations between stable-trait components. Grey line reflects expected number of significant effects due to chance (assuming a critical p-value of .05).

Table 2

Average Estimated Cross-Lagged Paths In Each Simulation Condition

		Reliability		
Stable Trait r	AR Variance Ratio	0.5	0.7	0.9
0.10	0%	0.03	0.02	0.01
	50%	0.03	0.01	-0.02
	100%	0.03	0.01	-0.03
	200%	0.04	0.02	-0.02
0.30	0%	0.08	0.06	0.03
	50%	0.07	0.05	0.01
	100%	0.07	0.05	0.01
	200%	0.07	0.05	0.01
0.50	0%	0.13	0.12	0.06
	50%	0.11	0.10	0.05
	100%	0.10	0.09	0.04
	200%	0.09	0.08	0.04
0.70	0%	0.20	0.20	0.11
	50%	0.16	0.16	0.10
	100%	0.14	0.14	0.09
	200%	0.12	0.11	0.07

in large samples.

Interestingly, error rates are not monotonically associated with the size of the correlation when reliability is high. 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 reliability is high, the correlation between the stable traits is low, and there is substantial amounts 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 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[7]. 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, 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 correlations are almost guaranteed.

When Measures Have Error or Reliable Occasion-Specific Variance. The simulations described above focus on situations where reliability is very high. When there is measurement error or reliable state variance (as is very likely), this effect gets worse—potentially *much* worse. Consider the panel in the first row and the first column of Figure 2. In this case, reliability is set to .5 and there is no autoregressive variance. Error rates are very high, approaching 100% with large samples, even when the stable-trait correlation is just .3. Even samples of 100 can result in spurious cross-lagged effects approximately 60% of the time when stable traits are correlated .5.

---

[7] This can be understood by using tracing rules. If we randomly generate data for 10,000 participants from the data-generating model just described, the correlation between *X1* and *X2* and between *Y1* and *Y2* are both around .75. The correlation between *X1* and *Y1* would be about .25, and the correlation between *X1* and *Y2* 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 *X1* and *Y1* of .25. These values would imply an observed correlation of $.77 * .25 = .19$ between *X1* and *Y2*, 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.

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. Because X and Y are 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 2) is unrealistic, as there is sure to be some form of autoregressive structure to most variables we 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's 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 relatively high reliability (e.g., .8 for this simulation), 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 Y1 does not control for enough. The result is a spurious cross-lagged path.

In addition, measurement error also affects estimates from the RI-CLPM. If we specify a data-generating model 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 0.06, which would be easily detectable with moderate to large sample sizes. Note, this limitation of the RI-CLPM is not an argument *for* the CLPM (though it is an argument for using the STARTS, when possible).

One response to the above simulations is to suggest that we 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 these results in exactly the same way as random measurement error. Unfortunately, we don't know how common this reliable state component is in real data, though we have at least some evidence that it can exist and be large enough to be meaningful (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 reasonbly 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).

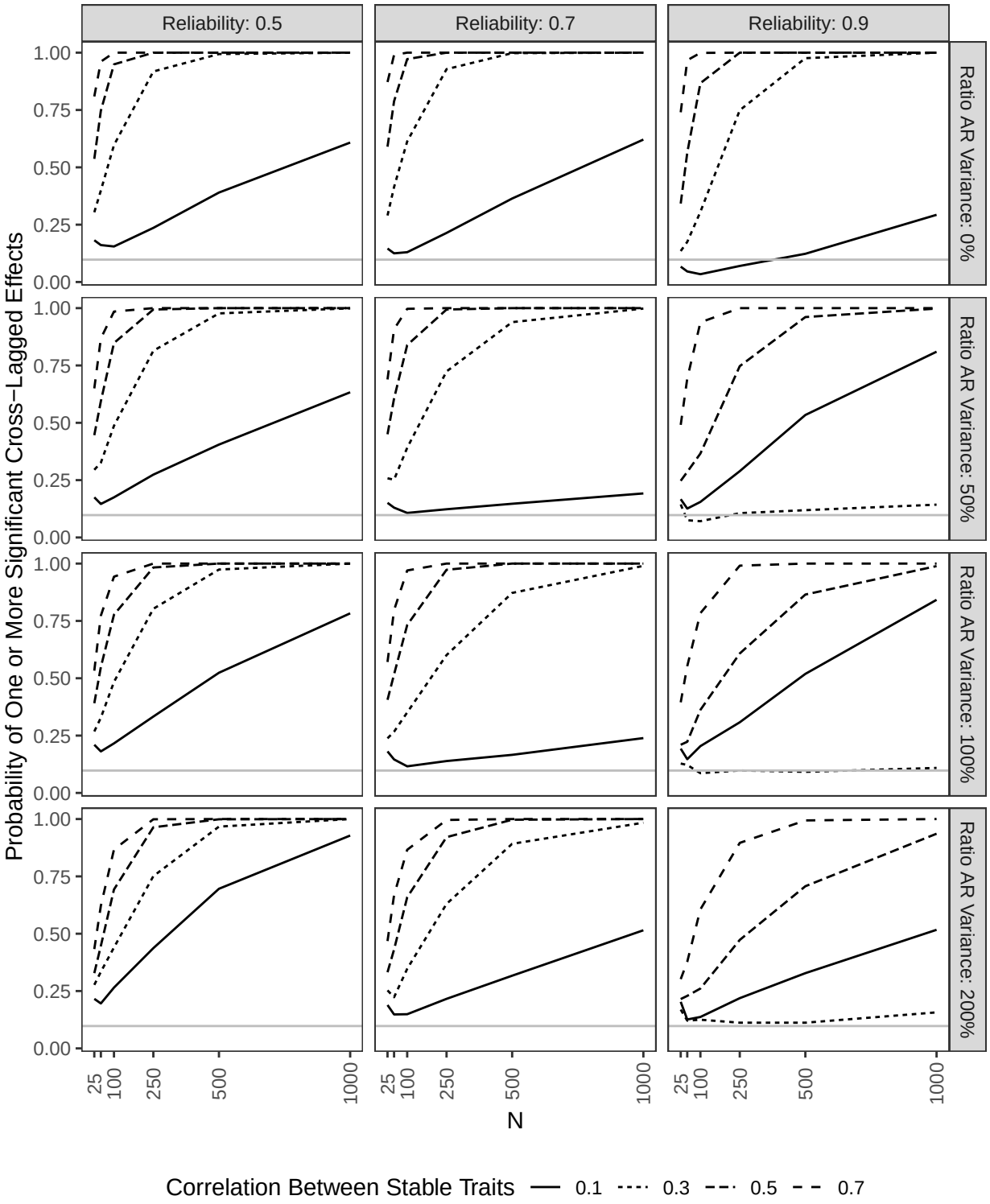


Figure 3. Simulation results for five-wave CLPM. Columns reflect different reliabilities. Rows reflect different ratios of autoregressive to stable-trait variance. Lines reflect different correlations between stable-trait components. Grey line reflects expected number of significant effects due to chance (assuming a critical p-value of .05).

Figure 3 shows a set of simulations that are similar to those reported in Figure 2, 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, reliabilities of .7, and a 1:1 ratio of stable-trait to autoregressive variance, the error rate increases from 58.10% to 97.30% when moving from a two-wave study to a five-wave study. So features that are generally desirable—large sample sizes and multiple waves of assessment—increase the likelihood of finding spurious cross-lagged paths.

The RI-CLPM Is Not More Conservative Than the CLPM

The examples above focused almost entirely 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 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, reliability was set to be 1 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 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, where as 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 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 RI-CLPM is actually *more likely* to find them than is the CLPM. 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 underestimate of the cross-lagged paths in most situations. If applied researchers are observing a pattern where cross-lagged paths routinely emerge when the CLPM is used and these paths disappear when the RI-CLPM is applied, then this would suggest that the effects themselves are likely spurious. Given the existence of stable-trait variance, true effects are more detectable with the RI-CLPM than the CLPM; if there is no stable-trait variance, the RI-CLPM simply reduces to the CLPM, and thus is equally likely to detect those effects.

One caveat 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

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	0.20	0.24	0.30
0.10	1.00	0.26	0.29	0.33	0.38
0.10	2.00	0.33	0.37	0.38	0.43
0.30	0.50	0.15	0.16	0.20	0.26
0.30	1.00	0.22	0.25	0.28	0.35
0.30	2.00	0.31	0.32	0.37	0.42
0.50	0.50	0.11	0.13	0.16	0.23
0.50	1.00	0.18	0.21	0.25	0.31
0.50	2.00	0.25	0.29	0.33	0.39
0.70	0.50	0.07	0.09	0.12	0.17
0.70	1.00	0.12	0.15	0.19	0.25
0.70	2.00	0.20	0.23	0.28	0.34

Note. AR r = Correlation between autoregressive components; AR Ratio = Ratio of autoregressive variance to stable-trait variance. Reliability is set to 1 for all simulations.

underused in the literature because of frequent estimation problems, recent methodological advances in Bayesian modeling have helped address these concerns (Lüdtke, Robitzsch, Link to external site, & Wagner, 2018).

Finally, most of the simulations described above set the variance components, stabilities, and reliabilities to be equal across the X and Y variables. Yet researchers often have predictions about which variables have causal priority. If these structural features differ—for instance if the measures of X are more reliable than the measures of Y, or the

autoregressive part of X is more stable than the autoregressive part of Y—then this can lead to evidence that one variable has causal priority over the other. I won't go into detail about these effects as others have discussed them and because the Shiny App is available to play around with. But researchers often report support for the causal priority of one variable using data from a CLPM, and subtle differences in these structural and psychometric properties can lead to important differences.

Conclusion

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." This sounds like advice with which no one could argue. But if there is a plausible alternative model that describes the data as well as (or better than) the preferred model, then much more work is needed to defend that selection.

As an obvious example, if the true causal process linking self-esteem to depression was that changes in self-esteem instantaneously caused a corresponding change in depression (and there were 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, because that would be conditioning on a collider. Yet few would find a cross-sectional correlation between self-esteem and depression to be compelling evidence for a causal effect of self-esteem precisely because there are so many plausible alternative models.

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. And if there is stable trait variance, it is quite plausible that two constructs correlate at the stable-trait level. The simulations describe above show that it is quite easy to find cross-lagged effects in the presence of stable traits, especially correlated stable traits, and especially when there is measurement error. The RI-CLPM and STARTS models provide a way to test this compelling and plausible alternative model.

References

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., 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.

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

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., Link to external site, this link will open in a new window, & 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

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

Usami, S. (2021). Within-Person Variability Score-Based Causal Inference: A Two-Step Estimation for Joint Effects of Time-Varying Treatments. *arXiv:2007.03973 [Stat]*. Retrieved from https://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. https://doi.org/gf4fqx

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