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

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

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 has made the CLPM a popular choice for the analysis of longitudinal data.

Hamaker, Kuiper, and Grasman (2015) pointed out that the CLPM does not adequately account for stable-trait-level confounds, and they proposed the random-intercept cross-lagged panel model (RI-CLMP) 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 (e.g., Lüdtke & Robitzsch, 2021) have noted, this critique of the cross-lagged panel model is 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. 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, 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 RI-CLPM with the random-intercept (or stable-trait) variance constrained to 0.

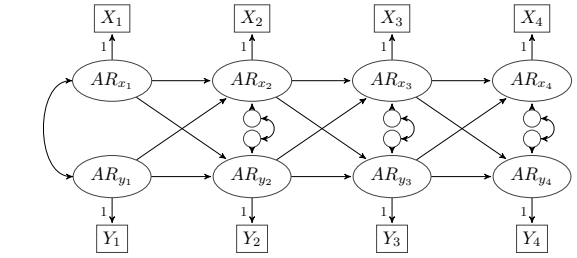
Notice that neither the CLPM nor the RI-CLPM include any measurement-error variance for the indicators. For the CLPM, this means that the latent variables from the autoregressive part of the model are equivalent to the observed variables (which is why it is also possible to draw an equivalent CLPM model 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 at 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., Orth et al., 2021).

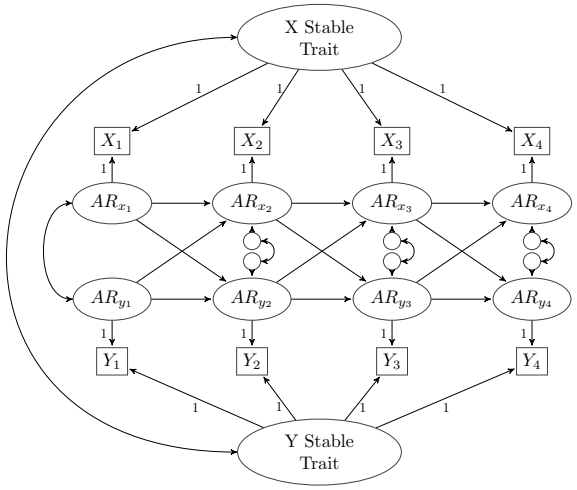
Recently, Usami, Murayama, and Hamaker (2019) clarified that the CLPM, RI-CLPM, STARTS and many other longitudinal models could be thought of as variations of an overarching "unified" model that captures many different forms of change. For instance, an alternative model—the Latent Curve Model with Structured Residuals—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.

The Ambiguous Nature of "Between" Versus "Within"

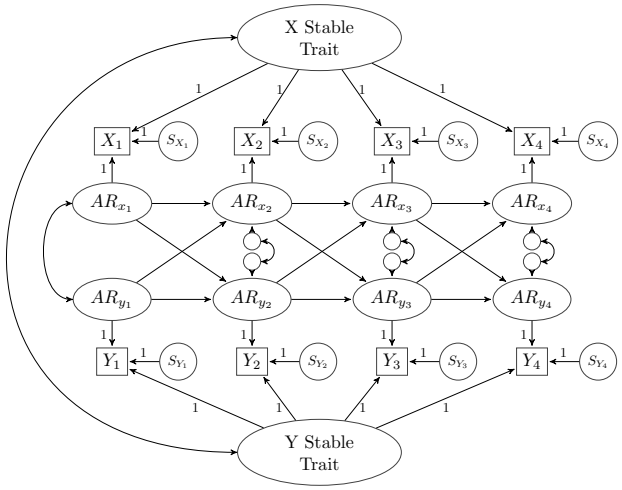
In presenting the RI-CLPM, Hamaker et al. (2015) emphasized that a strength was its ability to separate between-person effects from within-person effects. Critics of the RI-CLPM focus on this distinction when explaining their concerns with the model. The terms "within-person" and "between-person," however, 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.



Panel A: CLPM



Panel B: RI-CLPM



Panel C: STARTS

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

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[1]. 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 trait level" reflects an accurate interpretation of the "within-person" 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

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

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 individual mean (as Lüdtke & Robitzsch, 2021, did) or their "usual level?" 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 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 between-person differences in depression at 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) I think 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." I think that 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 some reasonably stable between-person variance exists.

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. To demonstrate, I subtracted each person's mean self-esteem score and mean depression score in each wave and then reran the CLPM using these data. For comparison, I also ran the RI-CLPM on the original, undeviated data [2]. The first three pairs of columns from Table 1 show the relevant estimates for the original CLPM and these two alternatives.

---

[2] 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 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.

My point is that adding some stable trait variance (in this case, just some 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.

One final response that I could imagine critics of the RI-CLPM making 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 the critics made. . . .

It is noteworthy that neither Lüdtke and Robitzsch (2021) nor Orth et al. (2021)

described a data-generating model that would result in data that would lead to 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[3].

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. 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 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 data (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.

The critics of the RI-CLPM often talk as if the RI-CLPM does this too. As noted earlier, Lüdtke and Robitzsch (2021) state in their abstract that the cross-lagged effect from

[3] 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.

the RI-CLPM "is typically less relevant for testing causal hypotheses with longitudinal data *because it only captures temporary fluctuations around the individual person means* (p. 1, emphasis mine). But, the RI-CLPM does not capture fluctuations around the person means, it captures deviations from the *perfectly stable trait*, which, as my example above demonstrates, can be a very different thing. This suggests to me that the critics of the RI-CLMP are incorrectly equating what the RI-CLPM does to what person-centering does in a more traditional multilevel model.

Why does this matter? Both of the main critics of the RI-CLPM have explicitly stated 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". As previously noted, Lüdtke and Robitzsch (2021) caution 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 does 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've been talking about; but to the extent that such a thing exists, the RI-CLPM captures it just as well as the CLPM.

Beyond "Within" and "Between"

Honestly, though, my concerns about the critiques of the RI-CLPM (and the defenses of the CLPM) emerged long before I had time to think through this issue, and I would still

have concerns about these critiques even if my interpretation of these "within-person" effects was wrong. That is because the appeal of the RI-CLPM, to me, did not rest on its strengths as a multilevel model that separated within-person effects (no matter how they are defined) from between. Instead, I saw value in the RI-CLPM because it tested an extremely plausible alternative explanation of the underlying pattern of correlations that is being modeled when the CLPM is used[4].

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

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

The problem with the CLPM is that is easy—in fact, incredibly 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 post simply builds on their simulations and tries to clarify when such spurious effects are likely to occur.

---

[4] 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.

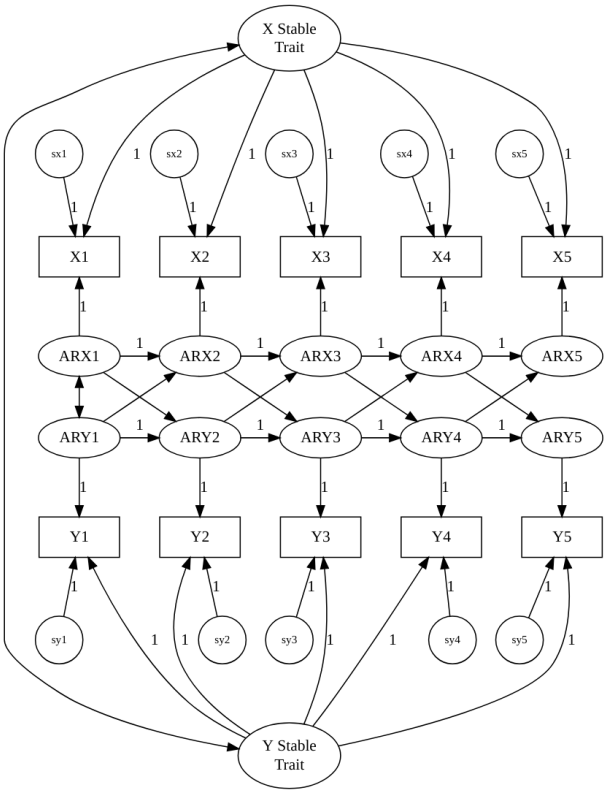


Figure 2. The STARTS Model. Residual variances for the autoregressive components are included but not shown in the figure.

I also created a Shiny app that allows the user to specify any STARTS-like structure (including the restricted RI-CLPM or CLPM). The app then fits the RI-CLPM and CLPM (and optionally the STARTS) to the data it simulates. The app is here, though I only use the free tier at shinyapps.io, so you can also download the app from my github page if it is not available due to capacity issues.

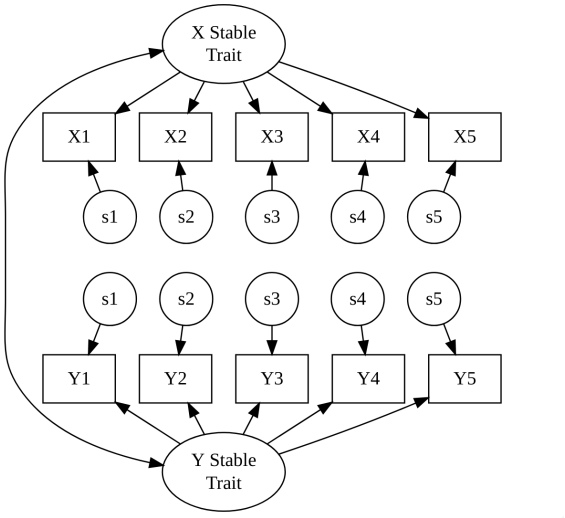
So, when can you find spurious effect? Let's go through a few common situations.

When The Constructs You're Measuring Are Correlated At The Stable-Trait Level. If you are measuring a construct that includes some amount of stable-trait variance, and the constructs are associated at the stable-trait level, it is possible to find spurious cross-lagged paths. To be sure, this factor alone—correlated stable traits—only produces spurious cross-lagged effects when the correlations between these

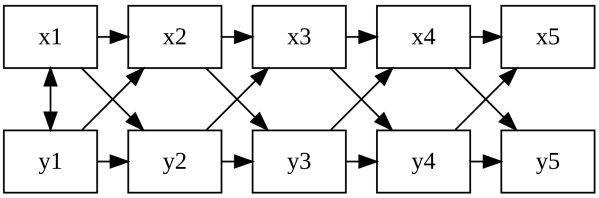
stable trait factors are quite high. For instance, if we simulate data, setting the variance of the X Stable Trait, the Y Stable Trait, the X Autoregressive Trait, and the Y Autoregressive Trait each to 1 (i.e., each contributes equally), you typically would not find a spurious cross-lagged correlation when the stable-trait correlation was .50, but you would reliably find spurious effects with correlations of .75 or higher. Note, however, that I did not specify any measurement error in this initial simulation.

When Your Measures Have Error. When there is measurement error (as is very likely), this effect gets worse—potentially *much* worse. Let's specify a very simple model where the measures at each wave result from two latent stable traits (X and Y) correlated at .50, plus measurement error, but no autoregressive variance (and therefore, no cross-lagged paths). In other words, this is a STARTS model with the autoregressive variance set to 0. Even when the reliability is high (e.g., .80), fitting the CLPM model results in estimated cross-lagged paths large enough to be detected even with samples of just 100 people. Increasing the correlation between the two stable traits or lowering the reliability increases the size of these estimates even further. And as Westfall and Yarkoni (2016) noted, larger sample sizes make it even easier to detect these spurious effects. So even when there are absolutely no within-person dynamics whatsoever, cross-lagged paths are not hard to find.

In other words, when the true data-generating model looks like this:



You will almost always find support for this:



That's a problem.

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) work because they do a better job accounting for this underlying association[5].

[5] One might argue that a model that just includes stable-trait variance and error is unrealistic, as there is sure to be some form of autoregressive structure to most variables we study. That's true, but the existence of

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 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 zero, and specify no cross-lagged paths, and even with reliabilities of .80, there will be enough power to detect spurious cross-lagged effects with just a couple hundred participants. Again, Westfall and Yarkoni's (2016) explanation can account for these results.

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). The reasons for this are the same as those discussed above, though slightly harder to wrap your head around in this more complicated case. 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).

When There Is Reliable Occasion-Specific Variance. 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. And 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). In any case, even the use of latent occasions in the CLPM can't solve this problem.

---

this stable trait causes problems for the CLPM even when all three sources of variability (stable trait, autoregressive trait, and state) exist. Play around with the Shiny App to see its effects.

Two Final Observations

The examples above focused almost entirely on cases where spurious effects were found when there were no cross-lagged paths in the data-generating model. However, failure to model associations between stable-trait components can also lead to the *underestimation* of real cross-lagged paths. For instance, we can use the Shiny App to specify a model with variances of 1 for the two stable traits and the two autoregressive components, and then specify stabilities of .5 for X and Y and correlations of .5 for both the random intercepts and the initial autoregressive components. 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. So the RI-CLPM is not necessarily more conservative than the CLPM when testing cross-lagged effects. Of course, everyone knows that an incorrect model will not provide accurate estimates; but my point is that most variables that psychologists study likely have at least some stable-trait-like variance, and ignoring it can have important consequences for the results that are obtained.

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.

Wrapping Up

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 possibilities.

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. 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 correlated stable traits (especially when there is measurement error), and the RI-CLPM and STARTS models provide a way to test this compelling and plausible alternative model.

References

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/10.1037/a0038889

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

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/10.1037/met0000210

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