CHILD DEVELOPMENT

Child Development, July/August 2017, Volume 88, Number 4, Pages 1186–1206

On the Practical Interpretability of Cross-Lagged Panel Models: Rethinking a Developmental Workhorse

Daniel Berry
University of Minnesota

Michael T. Willoughby
RTI International

Reciprocal feedback processes between experience and development are central to contemporary developmental theory. Autoregressive cross-lagged panel (ARCL) models represent a common analytic approach intended to test such dynamics. The authors demonstrate that—despite the ARCL model's intuitive appeal—it typically (a) fails to align with the theoretical processes that it is intended to test and (b) yields estimates that are difficult to interpret meaningfully. Specifically, using a Monte Carlo simulation and two empirical examples concerning the reciprocal relation between spanking and child aggression, it is shown that the cross-lagged estimates derived from the ARCL model reflect a weighted—and typically uninterpretable—amalgam of between- and within-person associations. The authors highlight one readily implemented respecification that better addresses these multiple levels of inference.

The idea of dynamic, reciprocal feedback processes between experience and development is central to most contemporary developmental models. Whether the cycles are vicious, virtuous, or benign, developmental continuity and change for most complex traits are presumed to be driven by self-organizing transactions between individual and context over time. Thus, it makes good sense that our field has historically been attracted to statistical models that appear to estimate such dynamics. One of the most well-known and oft-used models is the autoregressive cross-lagged panel model (ARCL; e.g., Figure 1). Indeed, a quick *Google Scholar* search for the term "cross-lagged" in only the journals *Child Development* and *Developmental Psychology* (2010–2015) yielded well over 200 hits (as of June 29, 2015).

Introduced by Campbell (1963), refined by Kenny (1973; Kenny & Harackiewicz, 1979), and made practical through the advent of structural equation modeling (SEM; Jöreskog, 1973), the ARCL model represents a simultaneous equation that provides estimates of the respective autoregressive relations (i.e., rank order stability) of two (or more) variables that unfold over time (i.e., Figure 1; B_1, B_2). Additionally, and often of most substantive interest, are the time-lagged regressions of Y_{it} on X_{it-1} and X_{it} on Y_{it-1} (i.e., Figure 1; B_3, B_4). The cross-lagged parameters are typically interpreted as the between-person effect of X_{it-1} on Y_{it}, controlling for Y_{it-1} (and vice versa). Or, in simple terms, the estimated average difference in Y_t that is associated with one's *own* value of X_{t-1}, compared with a unit difference in *someone else's* X_{t-1} value, despite sharing the same lagged value Y_{it-1}. The respective structural parameters (e.g., autoregressive and cross-lagged estimates) and the respective residual (co)variances are often constrained to be equal over time, though this need not be the case.

Not without controversy (e.g., Duncan, 1969; Hertzog & Nesselroade, 2003; Kenny & Harackiewicz, 1979; Rogosa, 1980), the comparative statistical significance and sometimes comparative size of the respective standardized effects of X on Y and Y on X are typically used as evidence of lead–lag or bidirectional relations between the variables over time. We refer the reader to

This research was supported by a grant from the National Institute of Child Health and Human Development (1PO1HD39667 and 2PO1HD039667). Cofunding was provided by the National Institute of Drug Abuse, the NIH Office of Minority Health, the NIH Office of the Director, the National Center on Minority Health and Health Disparities, and the Office of Behavioral and Social Sciences Research. We offer our sincere gratitude to all of the families and children who participated in this research as well as the investigators of the *Family Life Project*, the *Fragile Families and Child Well-Being Study*, and the authors of the Lansford et al. (2011) article cited herein. Their thoughtful design and stewardship of these landmark studies made the present article possible.

Correspondence concerning this article should be addressed to Daniel Berry, Institute of Child Development University of Minnesota, Twin Cities 51 E River Road Minneapolis, MN 55455. Electronic mail may be sent to dberry@umn.edu.

 0009-3920/2017/8804-0016
DOI: 10.1111/cdev.12660

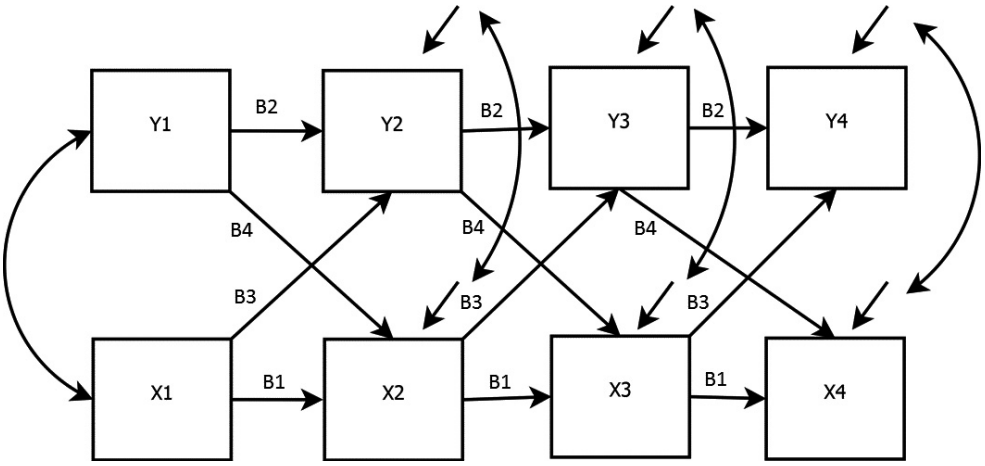

Figure 1. A prototypical autoregressive cross-lagged panel model.

Hertzog and Nesselroade (2003) for a discussion of the merits and problems of these practices and interpretations. They are not the focus of the present study.

Rather, our aim is to consider two (deceptively) straight-forward questions: (a) What do ARCL models really tell us? and (b) Does this square with the developmental process we are trying to understand? Building on recent discussions regarding the importance of disaggregating within- and between-person relations (e.g., Curran & Bauer, 2011; Curran, Lee, Howard, Lane, & MacCallum, 2012; Hoffman, 2015; Hoffman & Stawski, 2009), we propose that, despite their intuitive appeal and ubiquity in the child development literature, ARCL models typically give rise to estimates that are difficult (or impossible) to interpret meaningfully. That's the bad news.

The good news is that there are now a number of tools available to help us better tailor our statistical models to developmental processes we aim to test. We highlight a variant of Bollen and Curran's (2005) autoregressive latent trajectory (ALT) model as one such tool. In addition to allowing one to simultaneously model multiple developmental processes, we find that ALT-like specifications are in most cases *required* to make much sense of more typical ARCL cross-lagged estimates. We adopt the ALT model specifically because it marries two methodological approaches that are likely familiar to developmentalists—latent growth curve models and ARCL models. It is, of course, one tool of many innovative approaches (e.g., latent difference score, McArdle, 2001; state–trait, Kenny & Zautra, 2001; latent differential equations, Boker & Laurenceau, 2007; state–space, Molenaar, 2003).

We illustrate our discussion using a simulation study and empirical examples of the well-studied link between parental spanking and child aggression (e.g., Gershoff, 2013; Gershoff, Lansford, Sexton, Davis-Kean, & Sameroff, 2012; Gromoske & Maguire-Jack, 2012; Lansford et al., 2011; Lee, Altschul, & Gershoff, 2013; Strassberg, Dodge, Pettit, & Bates, 1994). We choose the potential bidirectional effects of spanking and aggression as an example because the link between spanking and aggression—a relation often established using ARCL models—has become accepted as somewhat of a truism. To be clear, we agree with Gershoff (2013): Physical discipline is a violation of children's rights. This is reason enough to "stop hitting our children" (p. 133). We, nonetheless, see it as a useful example, given the seeming intuitiveness of the (causal) inference and generalist appeal.

Levels of Analysis: Theoretical Models and Statistical Models

There is often a mismatch between developmental theory and the statistical models we specify to test developmental theory (see Curran & Bauer, 2011; Hoffman & Stawski, 2009; Molenaar & Newell, 2010). At its heart, developmental theory is typically concerned with intraindividual or within-person variability—how phenotypes change or remain stable *within* individuals over time. It is certainly reasonable that, on average, *compared to other children who are spanked less*, children who are spanked more might tend to evince higher levels of aggression. This might be true even after we hold other

important things, such as their levels of early aggression statistically constant across children. One could make a similarly reasonable case for the relation in the opposite direction—compared to less aggressive children, on average, more aggressive children tend to be spanked more. These are the bidirectional between-person inferences that are typically ascribed to the ARCL model.

It is, however, somewhat difficult to devise a developmental *theory* underlying the between-person relation between spanking and aggression. Rather, developmental theory is largely a within-person endeavor—it is about change. For example, from an (oversimplified) social-information processing perspective (e.g., Crick & Dodge, 1994): (1) spanking, (a) models aggression as a means to solve disagreements and (b) initiates the child's expectation of aggression; (2) over time, this becomes a representation on the child's mind; as such, (3) the child is more likely to interpret benign social gestures as acts of aggression, and (4) respond by aggressing, which (5) elicits aggression in her counterpart, (6) confirms the initial representational biases, and thus (7) exacerbates aggressive tendencies. Reading this, our guess is that most readers picture *an individual* child in the context of a series of social exchanges that unfold over time, rather than, say, a system of temporally ordered between-person distributions in which children's rank orders shift conditionally over time (i.e., between-person relations).

Indeed, with regard to developmental theory, it is difficult to conceptualize *most* developmental process from a between-person orientation. This is not to say that individual differences don't matter. They do. Some aspects are inherently more trait-like, showing virtually no intraindividual change (e.g., DNA, sex, race). Other aspects vary between people by definition. For example, one's developmental history of spanking over time is likely explained by both systematic components that vary only between people (e.g., mean levels, growth rates), as well as time-specific within-person deviations from these more systematic trends. Empirically, (main-effect) between-person differences can only explain developmental differences that emerge between people. Yet, with rare exception, in terms of theory, these between-person individual differences most likely manifest by having downward (or bidirectional) impacts on *within-person* processes (e.g., one or more of the seven steps listed above).

Between-person statistical models are seemingly based on the implicit assumption that between-person relations are aggregated representations of within-person developmental processes (i.e., "ecological fallacy"; Robinson, 2009; see Curran & Bauer, 2011). For instance, the between-person observation that children who are spanked more tend to be more aggressive than children who are spanked less seems plausible because it is thought to reflect the aggregate of n within-person instances in which increased spanking led to increased aggression.

This alignment between the within- and between-person effects is certainly possible. However, it need not be the case. In fact, there are often good reasons to expect the magnitude and/or direction of these inferences to differ. The association between exercise and blood pressure is a common example (e.g., Curran & Bauer, 2011; Hoffman, 2015). On average, people who exercise more tend to have lower blood pressure than do those who exercise less (i.e., between-persons). However, it is also true that, on average, blood pressure tends to increase when one is exercising more intensely than his or her normal level of intensity (i.e., within-persons). Both are accurate statements about the relation between exercise and hemodynamic functioning, yet the direction of the effect varies as a function of the level of analysis. They differ, in part, because they represent substantive differences in the exercise construct. Being *a person* who exercises a good deal on average—perhaps, a between-person proxy for ongoing, incremental within-person effects on cardiac functioning—leads to better overall cardiac health and lower blood pressure, relative to *a person* who exercises less. Yet, *at any given time*, the cardiovascular exertion associated with increasing one's workout intensity is associated with real-time blood pressure increases required to circulate oxygenated blood throughout the body.

It is not terribly difficult to think of other similar substantive examples. In fact, both substantively and methodologically it seems that differences in the magnitude—if not direction—of between- and within-person relations will be the rule rather than the exception (Hoffman & Stawski, 2009; Snijders & Bosker, 2012). At the most basic level, between- and within-person estimates typically lay on different scales (Hoffman & Stawski, 2009), carry distinct levels of statistical power (Bolger & Laurenceau, 2013), and thus tend to (quite reasonably) yield different regression parameters and different standard errors. That alone is probably enough to caution one against expecting identical effects at different levels of variation.

Perhaps more provocatively, we argue that failing to account for these multiple levels of analysis in the context of time-varying covariates—a weakness we find in the typical ARCL specification—will often yield cross-lagged estimates that are essentially uninterpretable (see also Hamaker, Kuiper, & Grasman, 2015). The importance of disaggregating within- and between-person effects is not a new realization (e.g., Cronbach & Webb, 1975; Firebaugh, 1978). However, it has been primarily discussed in the context of multilevel modeling (e.g., Bryk & Raudenbush, 1992; Curran & Bauer, 2011; Hoffman, 2015; Snijders & Bosker, 2012). With some recent exceptions (e.g., Allison, 2009; Bollen & Brand, 2010; Curran, Howard, Bainter, Lane, & McGinley, 2014; Curran et al., 2012; Hamaker et al., 2015; Voelkle, Brose, Schmiedek, & Lindenberger, 2014), these issues have been conspicuously absent from the applied SEM and developmental literatures, despite having the same interpretative implications.

What can go wrong? To set the stage, we first borrow some concepts from multilevel modeling (see Bryk & Raudenbush, 1992; Hoffman, 2015; Singer & Willett, 2003). A key idea behind longitudinal data is that there is something systematic about an individual that provides information about where we might expect his or her value of a given outcome to be at a given point in time. For instance, knowing about one's *average* level of aggression over time will likely allow us to do a better job of guessing his or her level of aggression at any given time than, say, guessing blindly without this information. Thus, one's average aggression level provides information about *dependencies* in the aggression data that occur within-individuals over time.

Accounting for such dependencies is the motivation underlying multilevel modeling. On one hand, modeling these dependencies is critical to estimating accurate standard errors for our parameters of substantive interest (e.g., regression coefficients). On the other, these dependencies also typically carry substantive meaning—it is *meaningful* that some children's average aggression levels are greater than the average aggression levels of other children. Using a common two-level representation, this idea is represented by Equations 1a–b, where γ_{00} represents the population grand mean level of aggression, and ζ_{0i} represents individual *i*'s average level aggression over time. Rather than estimate each individual's actual aggression mean, we make some assumptions about how the entire population's aggression means are distributed around the grand mean (e.g., mean of zero, normally distributed) and estimate this between-person variance (σ_0^2). Each individual has his or her "*own* mean," distributed around the grand mean (Hoffman, 2015, p. 84).

Of course, a person's mean does not tell the whole story. At any given point in time, his or her aggression level will rarely fall directly on his or her mean level. Rather, over time, each individual's time-specific aggression level (i.e., ε_{ij}) will be distributed around his or her *own* aggression mean. Again, we typically make some assumptions about this distribution (e.g., mean of 0, normally distributed, homogenous over time) and estimate the variance, σ_ε^2, rather than each individual's time-specific deviation from his or her own mean. As this variation occurs around one's own mean, it represents *within-person* variation.

Level 1:

$$\mathrm{AGG}_{ij} = \pi_{0i} + \varepsilon_{ij} \tag{1a}$$

Level 2:

$$\pi_{0i} = \gamma_{00} + \zeta_{0i} \tag{1b}$$

Some of this residual within-person variation will reflect nonsystematic, truly random error. However, presumably, there are also substantively meaningful reasons for why one's time-specific aggression level deviates from his or her own mean level. For many developmental processes, this will include the possibility that different children change at different rates—indeed, such differences in growth rates represent another *dependency* that needs to be accounted for in the data. This is reflected in Equations 2a–b, where each individual's time-specific aggression level is a function of (a) the population average intercept (γ_{00}) and growth rate (γ_{10}), (b) the deviation of his or her *own* intercept (ζ_{0i}) and *own* growth rate (ζ_{1i}) from the respective population averages, and (c) the time-specific bit of aggression that is left unexplained (ε_{ij}). As above, we typically invoke assumptions and model these person-specific intercepts (σ_0^2) and slopes (σ_1^2) as variances and covariances (σ_{01}; i.e., the association between intercept and slope). Of course, the between-child variability in children's intercepts and growth rates are substantively meaningful. Indeed, these are the growth trajectories that we often attempt to predict with between-child predictor variables.

Level 1:

$$\mathrm{AGG}_{ij} = \pi_{0i} + \pi_{1i}\mathrm{Time}_{ij} + \varepsilon_{ij} \quad (2a)$$

Level 2:

$$\pi_{0i} = \gamma_{00} + \zeta_{0i} \quad (2b)$$

$$\pi_{1i} = \gamma_{10} + \zeta_{1i} \quad (2c)$$

Beyond "time," we are also often interested in whether within-person deviations in our outcome are predicted by within-person variation in our substantive predictors. An obvious example here for aggression is spanking. Spanking can be measured longitudinally and treated as a time-varying predictor. As such, our instinct might be to center the time-varying spanking variable on some meaningful constant, like the grand mean, add it to the Level 1 model (i.e., $\mathrm{Spank}_{ij} - \overline{\mathrm{Spank}}$; Equations 3a–d), and fit the model to the data.

Level 1:

$$\mathrm{AGG}_{ij} = \pi_{0i} + \pi_{1i}\mathrm{Time}_{ij} + \pi_{2i}(\mathrm{Spank}_{ij} - \overline{\mathrm{Spank}}) + \varepsilon_{ij} \quad (3a)$$

Level 2:

$$\pi_{0i} = \gamma_{00} + \zeta_{0i} \quad (3b)$$

$$\pi_{1i} = \gamma_{10} + \zeta_{1i} \quad (3c)$$

$$\pi_{2i} = \gamma_{20} \quad (3d)$$

In turn, our statistical program will provide an estimate of the spanking effect, γ_{20}. But what would the estimated spanking effect, γ_{20}, really mean? It is actually pretty hard to say. Like aggression, longitudinal spanking data likely carry multiple levels of variability. The amount a child is spanked at any given point time is likely explained by (a) systematic processes—such as his or her overall trajectory (e.g., intercept, growth rate) of spanking over time—which vary *between* children—and (b) his or her *own* time-specific deviations from his or her *own* spanking trajectory—which vary *within* children. By virtue of not disaggregating the within-child spanking variation from the between-child spanking variation, we have implicitly forced this mixed-bag of spanking variation to be captured by a single parameter. To use Hoffman's (2015) wonderfully descriptive terminology, this "smushes" the between- and within-person effects together as a single estimate. Technically, this "smushed effect" represents a blend of within- and between-person spanking effects, weighted as a function of their respective reliabilities (Bryk & Raudenbush, 1992; Snijders & Bosker, 2012). This weighted mix of between- and within-individual relations is sometimes referred to as a "convergence" effect (Bryk & Raudenbush, 1992). As per Bryk and Raudenbush (1992), a "convergence" effect with balanced data is approximated as:

$$B_{\mathrm{convergence}} \approx \frac{\frac{B_{\mathrm{WP}}}{SE^2_{\mathrm{WP}}} + \frac{B_{\mathrm{BP}}}{SE^2_{\mathrm{BP}}}}{\frac{1}{SE^2_{\mathrm{WP}}} + \frac{1}{SE^2_{\mathrm{BP}}}}, \quad (4)$$

where B represents an unstandardized regression coefficient, SE represents the standard error of the estimate, and the WP and BP subscripts refer to within-person and total between-person effects of spanking on aggression, respectively. The "convergence" name stems from the fact that the only time these estimates are readily interpretable is when the effects are identical or "converge" across the two levels of analysis. If the convergence assumption is met, then the "smushed" model represented in Equations 3a–d will be the most efficient estimate of the spanking effect (Bryk & Raudenbush, 1992). In practice, though, convergence is rare (Hoffman, 2015; Snijders & Bosker, 2012). Indeed, as Hoffman (2015) notes, on the few occasions in which convergence appears to hold, the similarity of the between- and within-person effects can change quickly as other predictors are added to the model. In short, the assumption of convergence implied by failing to disaggregate the two levels of inference: (a) rarely holds, and thus (b) typically yields uninterpretable estimates (Hoffman, 2015; Snijders & Bosker, 2012).

Another troubling possibility is that one ends up with what appears to be a within-person effect, that is, in reality, due completely to *between*-person differences in the predictor. For instance, the unmodeled between-child spanking variation (i.e., child means/intercepts/growth rates over time) "sneaks in" and masquerades as a within-child effect, when the within-child effect would not be present were the spanking effect "fixed" to be *within* children. This is actually where "fixed-effect" regression common in econometrics gets its name (Allison, 2009). When variation is fixed to an individual, the estimated spanking effect is based on only within-person variation. This disaggregation makes the

spanking effect interpretable because it disaggregates the two levels of inference. In this example, the child-fixed effect would mean that, on average, within-child temporal shifts in spanking from his or her own typical spanking level over time are associated with temporal shifts in aggression. "Typical" here can refer to within-person deviations from one's own *mean* level of spanking over time. However, as we show below, there are often times when one may want to consider within-child deviations from his or her own overall longitudinal trajectory of the time-varying predictor.

In either case, the two levels of analyses clearly carry different substantive interpretations. Being *spanked more relative to other children* means something different than being *spanked more relative to one's own typical amount of spanking*. These different levels of inference also carry different strengths and weaknesses. For example, when the spanking effect is fixed within persons, each individual conceptually serves as his or her own control group. As such, all observed and unobserved time-invariant variables are accounted for in the model because one is estimating the effect based only on within-person variation—the person compared to him/herself. This is a methodological strength of fixed-effect estimates. In contrast, between-child inferences maintain the advantage of being able to consider aspects that differ systematically between children—for instance, temporally stable or "trait-like" aspects of spanking. However, with between-person relations, considerable effort must be taken to rule out biases due to unobserved confounds.

Collectively, the take-home messages from this digression into multilevel modeling (MLM) are that between-person and within-person relations need to be disaggregated because they (a) carry different substantive meanings, (b) have different strengths and weaknesses with respect to internal validity, and (c) typically yield uninterpretable and/or biased estimates, if they are not disaggregated.

What does any of this have to do with ARCL models? Regardless of whether one uses MLM or SEM to model relations between two (or more) longitudinal variables over time, the fact remains that longitudinal data (such as spanking and aggression) often carry both within- and between-person variation. Given its remarkable flexibility, SEM affords a variety of ways to disaggregate these two levels of inference. Unfortunately, they are wholly absent from the typical ARCL specification. As such, the cross-lagged estimates derived from the typical ARCL specification will typically represent an amalgam of between- and within-person relations that are interpretable only under the dubious—though easily testable—assumption of convergence.

To show how this is the case, we integrate work from latent growth curve modeling (Curran et al., 2012, 2014) and fixed-effect regression (Allison, 2009; Bollen & Brand, 2010). First, we draw on the now well-known observation that MLMs used to model change, such as those introduced above, can be readily respecified as latent growth models in the context of SEM (Chou, Bentler, & Pentz, 1998; Singer & Willett, 2003; Willett & Sayer, 1994). For instance, take the longitudinal variables shown in the path model displayed in Figure 2. The rectangles represent hypothetical longitudinal aggression data—interval-scaled level of aggression for a given month, measured across four consecutive months. As indicated by the latent factor and loading constraints of one, this represents a random intercept specification. It is identical to the MLM specification above in Equation 1 (provided that the within-person residual variances are constrained to be equal over time). We assume that children's true mean aggression levels over this period are independent and vary randomly around a population grand mean. We identify this latent mean by constraining the measurement intercepts to zero. For simplicity, we also assume that there is no systematic growth in aggression nor are there between-person differences in children's aggression growth rates over this short span. However, these assumptions are readily testable (see below). Collectively, this model implies that there is something systematic about one's mean level of aggression that is

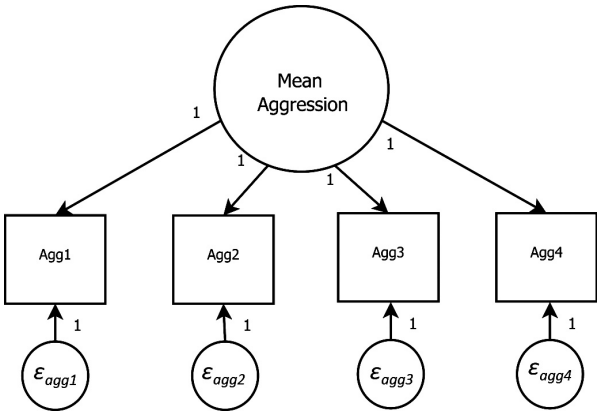

Figure 2. A prototypical random intercepts model, disaggregating between-person variation in mean levels of aggression around an estimated population average mean (i.e., latent variance and mean, respectively), and each individual's time-specific deviation in aggression from his or her own mean level of aggression (i.e., ε_{aggt}).

informative about one's aggression level in any particular month during this period.

Irrespective of what is consistent about an individual over time, there will also typically be time-specific deviations from these individual typical levels—the within-person residual variation (ε_{aggt}, Figure 2). These time-specific deviations from one's own mean comprise meaningful differences—on some occasions one is actually spanked *more than usual* (Hoffman, 2015). If the time-varying variables are observed as they are here (i.e., not latent at each time point), the within-person residuals will also carry measurement error. To align with the typical assumptions of MLM, we constrain these residuals to be homogeneous over time.

Notably, here we model these residuals as "structured residuals" (see Curran et al., 2014). That is, we create new latent variables to conceptually absorb the within-person residual variation at each time point and hold it as a distinct entity. We identify the model by constraining the residual variances of the observed time-varying aggression measure to zero and the factor loadings to one. As such, we have simply moved the within-person residual variances from the observed variables to their respective latent representations. The benefits of this specification will become apparent below. For our immediate purposes, though, it is worth noting that this specification will yield identical estimates as an otherwise identical random intercepts model fitted without structured residuals in SEM or a "standard" random intercepts model in the MLM. They are all functionally identical.

Second, we introduce our time-varying predictor—spanking (Figure 3a). To align with the discussion of the ARCL models to follow, we lag spanking. Otherwise, spanking is kept on its raw, uncentered scale. Our aim is to estimate the lagged relation between spanking and subsequent aggression, as shown by the single-headed arrows representing simultaneous regressions. For simplicity, we first consider the relation in one direction and constrain them to be identical over time. Also, for simplicity, we first assume no autoregressive relations and that neither spanking nor aggression show any systematic growth over time. We add these complexities to our hypothetical model momentarily. Notably, this model (3a) makes the following assumptions: (a) there is nothing *systematic* about an individual that gives any order to one's time-specific levels of spanking and (b) there is no connection between time-varying spanking and children's mean aggression levels.

The first assumption is not terribly plausible. For the most part, at least one's own mean level of spanking will typically be informative about level of spanking at any given point in time. By excluding this systematic between-person representation of spanking, one likely misses out on a substantively meaningful between-person relation between systematic aspects of spanking and aggression. If one excludes this systematic between-person variation *and* fails to provide direct connections between time-varying spanking and the random aggression intercept, then this forces the between-person spanking effect to blend with the within-person spanking effect. As such, this leads to a weighted composite of the two effects (i.e., a "convergence" effect). In fact, this is a key assumption of multilevel modeling—the predictors at one level need be uncorrelated with the random effects of another (Bryk & Raudenbush, 1992).

This assumption represents a fundamental difference between multilevel modeling and SEM (Allison, 2009; Bollen & Brand, 2010; Curran et al., 2012). SEM allows a straight-forward test of this assumption—simply test the constraint that the respective covariances between time-varying spanking and the random aggression intercept are jointly zero (Bollen & Brand, 2010; Curran et al., 2012; Figure 3b). This is functionally equivalent to a Hausman test (Bollen & Brand, 2010; Hausman, 1978) seen commonly in econometrics. It is also functionally equivalent to testing a *contextual* effect as in the multilevel modeling literature (see Curran et al., 2012; Snijders & Bosker, 2012). The contextual effect is the *incremental* between-person effect, *net of* the within-person effect (Bryk & Raudenbush, 1992; Snijders & Bosker, 2012) or, more simply, the difference between the between- and within-person effects.

In MLM, tests of the convergence assumption are readily estimated by including each individual's mean level of the time-varying covariate (e.g., mean spanking) as a Level 2 predictor—thus, disaggregating the two levels of the effects by controlling statistically for the other. This is the definition of a *contextual effect*. For example, in simple terms, the contextual effect means that *for all children who were spanked at the same level on a given occasion* (i.e., holding absolute occasion-specific level constant), the contextual effect is the *additional* bump in aggression that would be expected, on average, due to a unit difference in children's *mean* levels of spanking over time (i.e., one's typical level). Adjusting for it removes the between-person effect from within-person effect—thus, "fixing" the effect to be within-person.

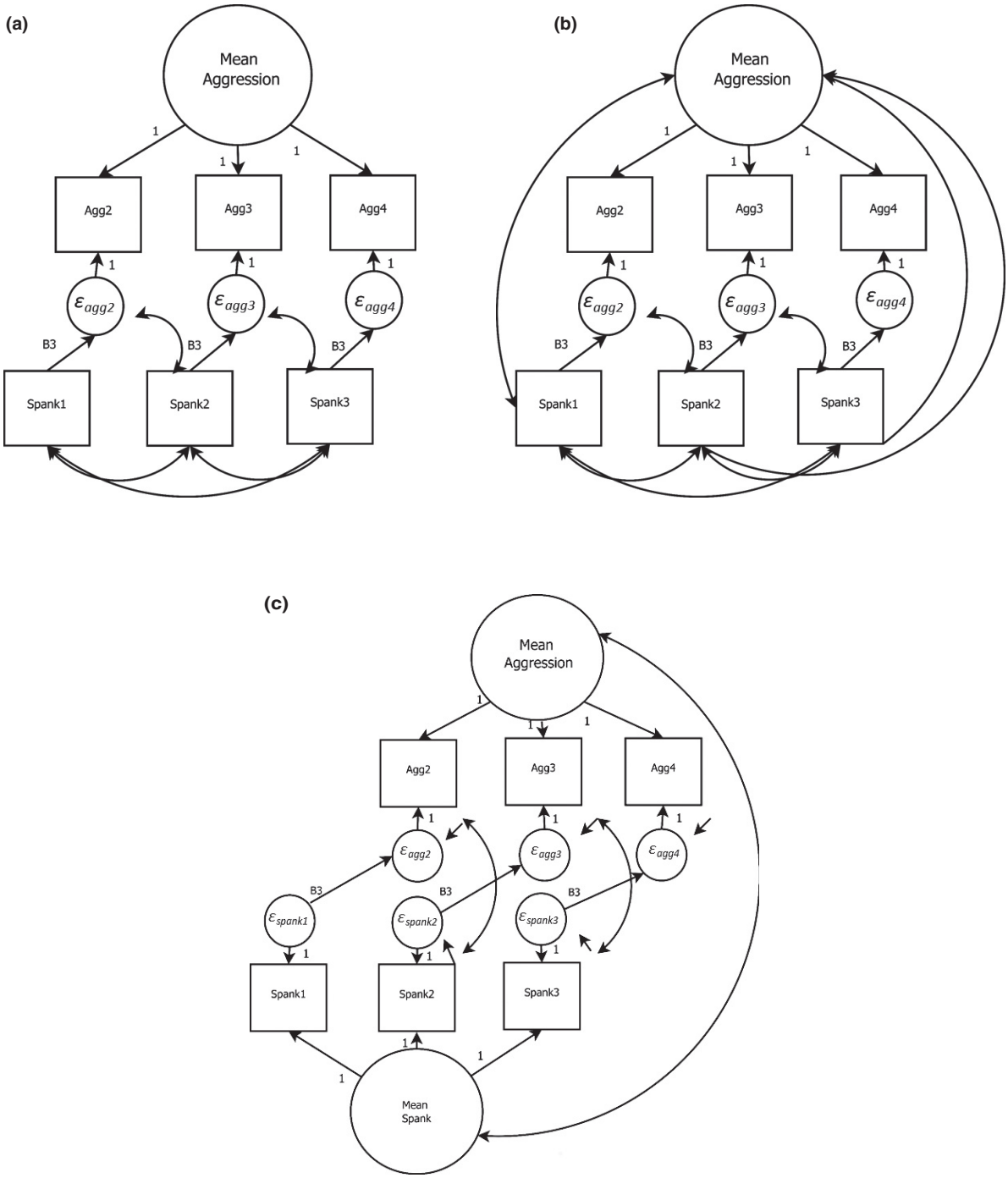

Figure 3. (a–c) Exemplar random intercept models displaying the lagged relation ($t - 1$) between time-varying spanking and children's subsequent aggression at time t, in which time-varying observed spanking is assumed to be (a) orthogonal to children's true mean levels of aggression (i.e., convergence) or (b) covary with children's true mean levels of aggression (i.e., disaggregating within-person effect from the between-person contextual effect), or in which (c) true mean spanking is assumed to covary with children's true mean levels of aggression (i.e., disaggregating the within-child effect from the total between-person effect).

Group-mean centering the time-varying predictor is another common alternative in the context of MLM. Here, the group mean is simply the person mean (i.e., person-mean centering [PMC]). PMC physically expunges between-person variability from the time-varying predictor, rendering the between- and within-person variation orthogonal. This parameterization also changes the meaning of

the between-person relation. Unlike the contextual effect, which reflects the between-person association, controlling for the absolute value of the time-varying predictor, here, the between-person effect represents the total between-person effect, *not* controlling for a given time-specific value of the time-varying predictor. For example, this would represent, between-child differences in spanking that are explained by one's mean levels of spanking, as well as the fact that a unit deviation from home own mean inherently carries greater value for those with higher mean levels of spanking (i.e., between-person effect + within-person effect).

Although some questions lend themselves more readily to one or the other interpretation, either is reasonable. In fact, providing that the time-varying effect is fixed (i.e., not random across children), model fit will be identical across parameterizations and one can easily recover one interpretation from the other ($B_{\text{contexual}} = B_{\text{TotalBP}} - B_{\text{WP}}$; $B_{\text{TotalBP}} = B_{\text{contexual}} + B_{\text{WP}}$; Bryk & Raudenbush, 1992; Hoffman, 2015). Critically, though, *one of these tests is required to justify the convergence assumption*. If the assumption is violated, then the effects of the time-varying predictors will be an uninterpretable weighted blend of between- and within-person effects.

The rub is that neither approach is estimable using SEM (see Curran et al., 2012 for a lucid discussion). Group (i.e., person)-mean centering poses difficulties for maximum likelihood estimation in SEM because it renders the sum of each individual's time-varying predictor values equal to zero (i.e., ipstatic). As such, this will always lead to a singular sample covariance matrix. Controlling for person-mean levels of the time-varying predictor as a between-person variable is also problematic because it builds in an implicit linear dependency between this person-specific mean and the time-varying representation of the same variable—thus, again, leading to a singular sample covariance matrix (Curran et al., 2012). In short, neither of the typical MLM approaches for disaggregating within- and between-person variation is possible in SEM.

Informed by the fixed-effect regression tradition most commonly found in econometrics (e.g., Allison, 2009; Bollen & Brand, 2010), one way to avoid a "convergence" effect in the context of SEM is to simply allow the time-varying *X*s to correlate with the random aggression intercept (Figure 3b). Indeed, although not typically discussed in terms of its disaggregating properties (however, Curran et al., 2012; Hoffman & Stawski, 2009), these covariances are often proposed as a default in latent growth models with time-varying covariates (i.e., Bollen & Curran, 2006; Curran et al., 2012). These pathways are critical because they allow a direct way for the time-varying *X*s to covary with systematic aspects of aggression—this is what expunges the between-person variability from the time-specific associations from spanking$_{t-1}$ to aggression$_t$. As above, without these pathways, between-person associations between systematic aspects of spanking and aggression—that is, aspects that cannot be time-specific, by definition (e.g., person-mean levels)—will "sneak in" and masquerade as time-specific, lagged associations. Thus, these pathways are essential to making any sense of the lagged effect, which can now be interpreted as the lagged within-person relation between spanking and children's subsequent levels of aggression. An ancillary benefit of these within-person inferences is that they control for all possible time-invariant confounds. Indeed, this is what has historically made these so-called "person-fixed effect" estimates so attractive to economists.

The downside of the above specification, though, is that—without some serious algebraic legwork (see Curran et al., 2012)—systematic components that fail to vary within people (e.g., *chronic* spanking, *chronic* aggression) are essentially absent from the model. Given that relations between these more trait-like aspects of individuals and contexts are often of great interest to developmentalists, this is a serious limitation.

Figure 3c represents one way to parsimoniously disaggregate the within- and between-person effects while explicitly modeling the latter. Specifically, rather than estimating each covariance between time-varying spanking and the random aggression intercept, we model the systematic between-person components for both spanking and aggression. Here, for simplicity, we model the systematic components as random intercepts—that is, between-child variability in children's mean levels of spanking and aggression (respectively) around their respective estimated population averages. Also, for simplicity, we constrain the effect of the time-varying predictors and the time-specific (within-person) residuals to be constant over time and function in one direction. As we will see, however, this model can be readily extended to capture other systematic person-level components (e.g., growth), as well as bidirectional lagged relations.

The between-person relation is represented by the connection between the latent factors representing "true" mean spanking and mean aggression over time. When there is reason to hypothesize a directional effect, one can regress one on the other.

When one has neither a substantive reason nor the methodological rigor to invoke a directional effect—as is the present case for the link between spanking and aggression—it is often most sensible to model the between-person relation as a covariance between the true mean spanking and true mean aggression. When standardized, the between-person covariance can be interpreted as the *total* between-person correlation between mean spanking and mean aggression.

The one-sided arrows connecting the structured residuals of aggression at time$_t$ with the structured residuals of spanking at time$_{t-1}$ represent the lagged within-person effect (see Figure 3c). This is where the structured residuals introduced above become useful. The structured residuals for spanking specified in Figure 3c explicitly disaggregate the within-person spanking variation (i.e., the structured residuals) from the between-person variation (see Curran et al., 2014). This is conceptually akin to PMC common in MLM—it renders between- and within-person spanking variability orthogonal. This is what allows us to interpret the between-person relation as the *total* between-person effect rather than a *contextual* effect. Critically, the inclusion of this covariance between the two latent factors is what expunges the between-person effect from the time-lagged, within-person effect of spanking on aggression. Without the pathway connecting the between-person variation in children's mean spanking and aggression levels, this covariation would be forced into the within-person, cross-lagged effect—thus, making it a "convergence" effect that blends the two levels of analysis.

How do we know if we *need* to disaggregate the two levels of inference? If the covariance linking the two latent between-person variables is statistically significant and the resulting between-person regression parameter (e.g., latent covariance divided by the latent spanking variance) differs statistically from the within-person regression parameter (i.e., this difference is the "contextual effect"), then the within- and between-person effects differ in magnitude and cannot be collapsed. Although our interpretations will become more complex as we add autoregressive and bidirectional cross-lagged relations, the take-home message remains the same—the interpretability of the parameter estimates depends on accurately disaggregating the between- and within-person sources of variation.

For instance, in Figure 4, we expand our model to include a fixed effect for linear growth and autoregressive and bidirectional cross-lagged relations between the structured residuals. Our example is based on the ALT model with structured residual (ALT-SR) proposed by Curran et al. (2014). Each child is specified to have his or her own spanking and aggression intercept, respectively. As time is centered on the first observation point (i.e., see growth factor loadings), this represents random between-child variation in spanking and aggression (respectively) at Time 1. Each is assumed to be independent and normally distributed. Our model for the means specifies estimates of the population average linear spanking and aggression growth rates, respectively. However, for pedagogical simplicity, we constrain the growth rates to be identical across children (i.e., linear growth fixed effect). That is, initial levels of spanking and aggression vary randomly between children, yet the rank order of individuals is preserved over time. Notably, this need not be the case; the growth rates could also be freed to vary randomly between children. Indeed, establishing a plausible population growth model—both in terms of the functional form (e.g., linear, quadratic, exponential) and the model for the (co)variances—is a prerequisite for unbiased estimates of the within-person autoregressive and cross-lagged effects (Voelkle, 2008).

The within-person component of the model allows each child to have his or her own time-specific (structured) residual—that is, the time-specific deviation from his or her own trajectory (i.e., ε_{spank_t}, ε_{agg_t}). Borrowing terminology from time-series analysis, systematic aspects of spanking and aggression over time are "detrended" from the longitudinal spanking and aggression variables. The respective growth models capture the systematic components of spanking and aggression over time, and the structured residuals capture the time-specific variation that remains. Given our aim of disaggregating the within-person effects from the between-person effects, we specify the autoregressive and cross-lagged components of the model to occur between the structured residuals. The within-person residual variances are assumed to be normally distributed and homogenous over time, with the exception of Time 1. This specification reflects that the Time 1 residuals carry different meanings than the residuals thereafter—the Time 1 residuals are residualized on only their respective growth processes; the post Time 1 residuals are residualized on both the growth processes and the autoregressive and cross-lagged relations.

Collectively, this specification explicitly disaggregates the two levels of variation: between-person and within-person. The covariance between

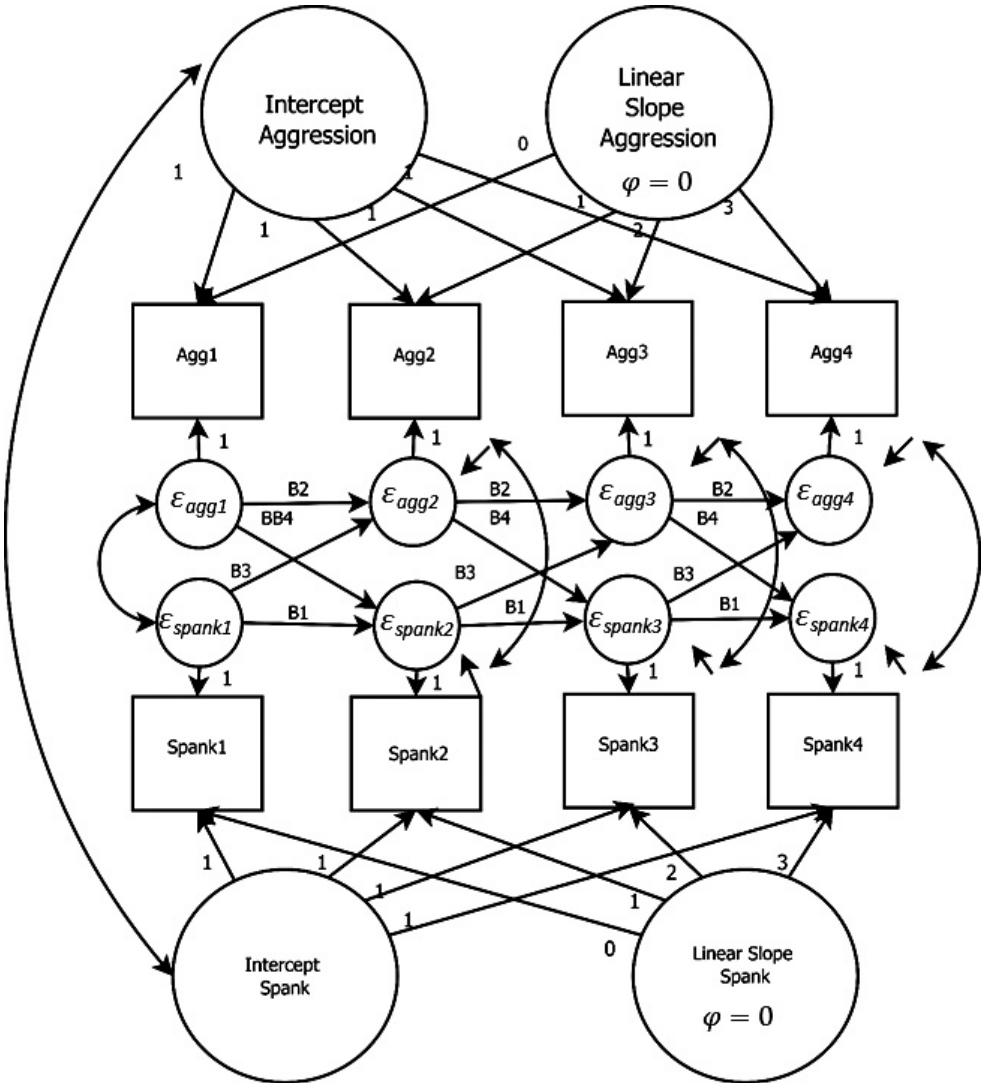

Figure 4. Autoregressive latent trajectory model with structured residuals (Curran et al., 2014), disaggregating the between-person association between intercepts (implied to be constant over time) from the respective within-person cross-lagged ($t - 1$) relations between spanking and child aggression over time.

the latent factors represents the *total* between-person association (i.e., *not* adjusting for within-person autoregressive or lagged effects), and the cross-lagged parameters represent (conditional) within-person associations. For instance, a positive cross-lagged spanking effect would be interpreted as the extent to which an increase from one's own typical spanking growth trajectory is predictive of a subsequent increase from her own aggression trajectory, on average, after adjusting for one's own preceding aggression level and all possible time-invariant covariates. Although this is undeniably a mouthful, the upside to the ALT-SR specification is that the relation is actually interpretable.

In contrast, the interpretations of the cross-lagged estimates produced by the typical ARCL specification are often far more murky. We illustrate why this is the case using simulated and actual data below. However, conceptually, we can consider the typical ARCL specification (Figure 1) as a special case of the ALT-SR (Figure 4) in which the systematic aspects about individuals (i.e., latent intercepts and slopes) as well as the connection between these systematic between-person aspects (i.e., covariance between the latent aggression and spanking intercepts) are constrained to zero. This is tantamount to ignoring the existence of and the interrelation between systematic between-person aspects of spanking and aggression over time. Of

course, simply ignoring them does not make them go away, if they do, in fact, exist. Rather, it forces these systematic between-person components to be captured by the autoregressive and cross-lagged components of the model. As such, these estimates become a weighted blend of between- and within-person effects that are only readily interpretable in the rare case that they are empirically indistinguishable—that is, cross-lagged "convergence" effects.

In the following section we use simulated data as well as actual data concerning the longitudinal relations between maternal spanking and child aggression to illustrate (a) how the typical ARCL specification represents a weighted amalgam of between- and within-person effects and (b) how accurately disaggregating these different levels of inference can have dramatic impacts on one's substantive conclusions about reciprocal relations. For readers interested in the more technical nuances, we describe how the "convergence" effects of the ARCL specification can be recovered from the disaggregated within- and between-person effects of the ALT-SR in the Supporting Information (with relevant Mplus syntax).

Monte Carlo Simulations

Given that population parameters are rarely known in the context of "real data," we conducted a series of Monte Carlo simulations to test the impact of fitting a typical ARCL specification to population data with distinct between- and within-person relations between spanking and child aggression (Mplus. 7.4; Muthén & Muthén, 2015). This allows us to address the question: What does the typical ARCL specification give us, when we *know* it is the wrong model (i.e., we defined the population model that generated the data). Specifically, we defined our true population parameters as those emerging from a population in which spanking and aggression were measured at the same points in time, longitudinally across 4 (balanced) consecutive years (see Figure 4). The population parameters represent a scenario in which the between- and within-person effects are disaggregated, and in which the within-person cross-lagged (i.e., $\beta = .02–.05$) and autoregressive (i.e., $\beta = .10–.15$) relations are quite modest. The respective (i.e., spanking and aggression) population within-person residual variances were specified to be identical over time (with the exception of t_1), and the respective systematic between-person differences were specified to be best represented by random intercepts (centered at t_1), mean linear growth rates constrained to be identical between persons (for simplicity), with moderate between-person correlation between t_1 intercept spanking and t_1 intercept aggression ($\psi_{stand} = .32$; Table 1). The sample size is 501. There is reason to suspect that these are plausible values, as the population parameters were generated from the covariance matrix published in a well-cited longitudinal study of parental physical discipline and externalizing problems (Lansford et al., 2011; Study 1). In M1 (Table 1), we display the average of the parameter estimates across 50,000 resamples of model when it is properly specified. In M2, we display the average of the resampled estimates from the same population model, yet *misspecified* as a "typical" ARCL model.

Simulation Results

As expected, when specified correctly, the ALT-SR model recovers the population parameters and standard errors quite nicely (Table 1, M1; i.e., Bias = (estimate − population)/population). When the model was misspecified, such that the between-person variances and relations between the random intercepts were constrained to zero, it expectedly fit comparatively less well than the correctly specified model. In absolute sense, though, the fit statistics were reasonable enough that many would go on to relax constraints to improve model fit rather than reject the model outright. Nonetheless, the biases were substantial. For instance, the cross-lagged population parameters were specified to be small and statistically indistinguishable from zero. Yet, the cross-lagged estimates from the misspecified ARCL model were positive and rather substantial—around double to four times larger than the true population values.

These biases make good sense. Whereas the population model includes within-person and between-person effects that are cleanly delineated, the M2 misspecification forces the covariance between systematic between-person aspects of spanking and aggression over time (i.e., random intercepts) to manifest through the cross-lagged effects.

What does this mean in practical terms? It means that (a) the autoregressive and cross-lagged estimates from the ARCL model—given this specification—comprise a blend of between-person and within-person effects, and (b) what would be substantively interpreted as a lagged reciprocal feedback process in the ARCL model is, in actuality, explained by the fact that children who are

Table 1
Monte Carlo Simulation Illustrating the Bias Introduced by Specifying An Autoregressive Cross-Lagged Panel (ARCL) Model to Data in Which the Population Data Were Generated by a Model Comprising Distinct Between- and Within-Person Processes (50,000 Iterations)

	M1						M2			
	True disaggregated as disaggregated (i.e., ALT-SR)						True disaggregated as ARCL			
	Pop. Parm.	Est. Parm	Bias Parm.	Pop. *SE*	Est. *SE*	Bias *SE*	Est. Parm	Bias Parm.	Est. *SE*	Bias *SE*
Fixed effects										
Agg_t on Agg_{t-1}	0.095	0.095	−0.004	.043	.043	−.007	0.610	5.417	.020	−.093
Agg_t on $Spank_{t-1}$	0.037	0.036	−0.024	.085	.085	−.004	0.177	3.789	.047	.054
$Spank_t$ on $Spank_{t-1}$	0.121	0.120	−0.006	.041	.040	−.005	0.600	3.962	.020	−.101
$Spank_t$ on Agg_{t-1}	0.022	0.022	0.000	.016	.016	−.006	0.041	0.873	.009	.024
(Co)variances										
Agg_{int} with $Spank_{int}$	0.261	0.260	−0.004	.048	.048	−.010				
Agg_{int}	1.921	1.914	−0.004	.151	.151	−.001				
$Spank_{int}$	0.350	0.349	−0.004	.029	.028	−.017				
Agg_{slope}										
$Spank_{slope}$										
Residual (co)variances										
$Spank_{\varepsilon it1}$	0.328	0.328	−0.001	.026	.026	.016	0.677	1.063	.043	−.023
$Agg_{\varepsilon it1}$	1.344	1.342	−0.002	.115	.114	−.010	3.257	1.423	.206	.000
$Spank_{\varepsilon it2-\varepsilon it4}$	0.233	0.233	−0.002	.012	.012	−.008	0.345	0.480	.013	−.023
$Agg_{\varepsilon it2-\varepsilon it4}$	1.359	1.356	−0.002	.069	.069	−.003	1.989	0.463	.073	−.032
M	0.261	0.260	−0.004	.048	.048	−.010	0.000			
Agg_{int}	1.901	1.901	0.000	.076	.077	.011	1.901	0.000	.077	.079
$Spank_{int}$	1.010	1.010	0.000	.034	.034	−.009	1.010	0.000	.035	.036
Agg_{slope}	−0.032	−0.032	0.009	.024	.024	.004	−0.032	0.009	.024	.033
$Spank_{slope}$	−0.124	−0.124	0.000	.011	.011	−.009	−0.124	−0.001	.011	.014
Fit statistics										
χ^2	25.398						280.300			
RMSEA	.01						.13			
SRMSEA	.03						.09			

Note The means model for the ARCL specific differs somewhat from the "typical" specification, which includes no systematic aspect of mean change—we include it presently to make cleaner comparisons across specifications; our conclusions are unaffected by eliminating the latent means model and freely estimating the measurement intercepts. Contemporaneous within-person residual covariance are included in the models (constrained to equality across time points), yet excluded from the table for ease of display. ALT-SR = autoregressive latent trajectory model with structured residuals; Pop. = population; Est. = estimated; Bias = (estimate − population)/population; Agg = child aggression; Spank = maternal spanking; Int = latent intercept (centered at time 1); Slope = latent linear growth rate; ε_{it} = structured residual for individual *i* at time *t*; RMSEA = root mean square error of approximation; CFI = comparative fit index.

typically spanked more over this period also tend to be children who typically show worse aggression problems. Thus, like most convergence effects, the regression parameters from the run-of-the-mill ARCL model would be virtually impossible to interpret. Furthermore, one is back to square one, with respect to the functional (causal) relation between spanking and aggression—namely, we are left with a zero-order between-person correlation between the respective true means of spanking and aggression over time. Does this mean that there is no causal relation between the two? Not necessarily. As the adage goes, absence of evidence is not evidence of absence. However, it does mean that a good deal more legwork is needed to understand these causal processes, if they do exist.

Empirical Examples With "Real" Data

To illustrate the generality of our results, we fit ARCL and ALT-SR models to two more empirical examples. The first is from the recent literature. We align our models as closely as possible to those from the original published article—though we expect slight differences to emerge as a function of fitting our models to summary (rather than raw) data, as well as (noted) differences in our

specifications. The second is based on novel data. For the sake of continuity' we use the terms "spanking" and "aggression" throughout; however, the actual measures often tap more comprehensive constructs (e.g., externalizing problems, physical discipline, etc.). Due to space restrictions, we provide a third example in the Supporting Information. It is based on the data from the *Fragile Families and Child Family Well-Being Study* (Reichman, Teitler, Garfinkel, & McLanahan, 2001), which has been used several times to test similar ARCL models.

Example 1: Lansford et al. (2011, Study 2)

In this example we use data reported by Lansford et al. (2011, Study 2). We selected these data because we found the study to be particularly thoughtful and well-articulated. In this work, the authors tested a series of ARCL models in a sample of 290 boys from low-income families who were followed longitudinally between the ages of 10 and 15 years. At 10, 11, 12, and 15 years of age, parents rated (1–4, Likert) the extent to which they used physical discipline (i.e., "spank" and "slap or hit with hand, fist, or object") and adolescents self-reported their levels of overt, covert, destructive, and nondestructive behaviors (e.g., "hit other students or gotten into a physical fight," "taken something from a store without paying for it"). We refer the reader to the original study for full coverage of the sample and methods.

Informed by a series of preliminary specifications to establish plausible growth models for each construct, we fitted a model essentially identical to the one we tested in our simulation (Figure 4). For both spanking and aggression, growth was modeled as a linear function of time. However, the slopes were constrained to be equal across children. The intercepts were centered on age 10 and assumed to vary randomly between children. The within-person residuals were modeled as structured residuals, with the cross-lagged and autoregressive parameters fitted to these residuals. Like the residual variances after Time 1, these respective regression parameters were constrained to be equal over time. We adopt these constraints for pedagogical simplicity. It is worth noting, however, that the varying nature of the temporal lags across the longitudinal observations makes these assumptions rather questionable. We provide annotated Mplus syntax for the a series of constraints that more accurately adjust for the observed lag structure in the Supporting Information.

Results

As shown in Table 2, neither the ARCL nor the ALT-SR specification fit the data terribly well, as is. However, this was driven largely by the numerous equality constraints imposed on the model. As evidenced by Lansford and colleagues' (2011) original findings, an unconstrained ARCL specification fit the data much better. Nonetheless, our more constrained ARCL specification yielded autoregressive and cross-lagged estimates that were quite similar to those published by the original authors. We found modest yet statistically significant autoregressive and cross-lagged relations. On average, higher levels of spanking at time_{t-1} were predictive of subsequently higher levels of spanking ($B = .56$, $p < .001$, $\beta = .57$) and aggression at time_t ($B = .038$, $p < .001$, $\beta = .11$), holding prior levels of aggression constant. With respect to the inverse relations—similar to the original findings—higher levels of time_{t-1} aggression were predictive of subsequently higher levels of aggression ($B = .49$, $p < .001$, $\beta = .49$), but not spanking at time_t. Collectively, one might quite reasonably interpret this as an indication that spanking exacerbates children's subsequent aggressive behavior over time.

However, consider the same data and a similar ARCL model, yet with the within- and between-child variation disaggregated, such that the systematic between-child aspects of children's respective spanking and aggression trajectories are expunged from the time-specific reciprocal feedback processes (Figure 4). As shown by the moderate, statistically significant covariance between the spanking and aggression random intercepts ($\psi_{\text{stand}} = .40$; Table 2), at any given point in time, on average, children who were spanked more tended to be more aggressive than children who were spanked less. However, once these systematic aspects about children's experiences and behavior over time are accounted for, there was no evidence of any time-specific reciprocal process, within-children ($B_{\text{spank}} = .021$, $p = .234$; $B_{\text{agg}} = .093$, $p = .49$). Or, taken another way, once the blend of between- and within-person effects from the ARCL specification were disentangled, what appeared to be time-specific lagged effects were largely manifestations of *systematic* aspects about individuals that are stable over time.

Do the between- and within-person spanking effects need to be disaggregated? Tantamount to a contextual effect from the MLM literature or a Hausman test from econometrics, we can address this by fitting a model constraint that tests the equality of the total between-person spanking effect

Table 2
Reciprocal Relations Between Spanking and Child Aggression, Specified as a Typical Autoregressive Cross-Lagged Panel (ARCL) Model and ALT-SR Model and Fitted to Summary Data From Lansford et al. (2011, Study 2; n = 209)

	ARCL (i.e., convergence)			ALT-SR (i.e., disaggregated)		
	B	*SE*	Stand.	*B*	*SE*	Stand.
Fixed effects						
Agg_t on Agg_{t-1}	0.484***	.031	.47	0.153**	.057	.15
Agg_t on $Spank_{t-1}$	0.038***	.010	.11	0.021	.017	.06
$Spank_t$ on $Spank_{t-1}$	0.555***	.028	.57	0.185**	.058	.13
$Spank_t$ on Agg_{t-1}	0.110	.086	.04	0.093	.134	.06
(Co)variances						
Agg_{int} with $Spank_{int}$				0.019***	.005	.40
Agg_{int}				0.014***	.002	
$Spank_{int}$				0.166***	.021	
Agg_{slope}						
$Spank_{slope}$						
Residual (co)variances						
$Spank_{\varepsilon it1}$	0.384***	.032		0.202***	.022	
$Agg_{\varepsilon it1}$	0.032***	.003		0.017***	.002	
$Spank_{\varepsilon it1-\varepsilon it4}$	0.220***	.011		0.173***	.012	
$Agg_{\varepsilon it1-\varepsilon it4}$	0.029***	.001		0.024***	.001	
M						
Agg_{int}	0.276	.010		0.276***	.010	
$Spank_{int}$	1.760***	.035		1.764***	.033	
Agg_{slope}	0.013***	.005		0.013***	.004	
$Spank_{slope}$	−0.120***	.015		−0.120***	.012	
Fit statistics						
χ^2	227.549			160.186		
df	30			27		
RMSEA	.15			.13		
CFI	.74			.828		

Note Contemporaneous within-person residual covariances are included in the models (constrained to equality across time points), yet excluded from the table for ease of display. ALT-SR = autoregressive latent trajectory model with structured residuals; Stand. = standardized coefficient; Agg = child aggression; Spank = maternal spanking; Int = latent intercept (centered at time 1); Slope = latent linear growth rate; ε_{it} = structured residual for individual *i* at time *t*; RMSEA = root mean square error of approximation; CFI = comparative fit index. **$p < .01$. ***$p < .001$.

and the total lagged within-person spanking effect (i.e., $B_{bp} - B_{wp} = B_{contextual} = 0$; see Supporting Information for Mplus syntax). Although the between-person effect of spanking on child aggression is not provided directly by the model, it can be obtained by scaling the covariance between the random intercepts on the variance of the latent spanking intercept (i.e., $B_{bp\_spank} = .019/.166 = .115$). Second, we can recover the total within-person lagged relation between spanking and aggression by summing the product of all possible direct and indirect paths linking Time 1 spanking with aggression over time (i.e., Figure 4; $B_3 + (B_3 \times B_1) + (B_3 \times B_2) + (B_3 \times B_2 \times B_2) + (B_3 \times B_1 \times B_1) + (B_3 \times B_1 \times B_2) + (B_1 \times B_3 \times B_2) = .029$). Fitting a constraint that total between-person and total within-person effects are equal indicated that we must reject this null hypothesis, $B_{contextual} = (.115 - .029 = .086;\ p = .04)$. As the two levels of inference differ from each other, they cannot be meaningfully collapsed. Or, in other words, the parameter estimates yielded by the ARCL specification are pretty much uninterpretable.

Example 2: The Family Life Project

In this second example, we use data from the *Family Life Project*, a prospective longitudinal study of children growing up in predominantly low-income families in rural regions of Pennsylvania and North Carolina (see Vernon-Feagans, Cox, & the FLP Key Investigator 2013). Maternal reports with regard to spanking and child behavior were collected when the child was 36, 48, and 58 months of age and in first grade. Spanking data were based

Table 3
Reciprocal Relations Between Spanking and Child Aggression, Specified as a Typical Autoregressive Cross-Lagged Panel (ARCL) and ALT-SR Model (Respectively) and Fitted to Raw Data From the Family Life Project *(n = 1,018)*

	ARCL (i.e., convergence)			ALT-SR (i.e., disaggregated)		
	B	*SE*	Stand.	*B*	*SE*	Stand.
Fixed effects						
Agg_t on Agg_{t-1}	0.512***	.015	.54	0.046~	.027	.05
Agg_t on $Spank_{t-1}$	0.020***	.004	.07	0.009	.008	.03
$Spank_t$ on $Spank_{t-1}$	0.551***	.014	.58	0.184***	.033	.19
$Spank_t$ on Agg_{t-1}	0.283***	.050	.09	0.150	.083	.05
(Co)variances						
Agg_{int} with $Spank_{int}$				0.082***	.082	.34
Agg_{int}				0.072 ***	.004	
$Spank_{int}$				0.787***	.054	
Agg_{slope}						
$Spank_{slope}$						
$Spank_{\varepsilon it1}$	1.874***	.032		0.972***	.060	
$Agg_{\varepsilon it1}$	0.174***	.003		0.100***	.006	
Residual (co)variances						
$Spank_{\varepsilon it2-\varepsilon it4}$	0.109***	.019		0.722 ***	.027	
$Agg_{\varepsilon it2-\varepsilon it4}$	0.039***	.006		0.060***	.002	
M						
Agg_{int}	0.573***	.013		0.549 ***	.012	
$Spank_{int}$	1.332***	.043		1.350 ***	.041	
Agg_{slope}	−0.078 ***	.005		−0.071***	.004	
$Spank_{slope}$	−0.205 ***	.017		−0.208***	.012	
Fit statistics						
χ^2	740.845			391.906		
df	30			27		
RMSEA	.15			.12		
CFI	.76			.88		

Note Contemporaneous within-person residual covariance are included in the models (constrained to equality across time points), yet excluded from the table for ease of display. ALT-SR = autoregressive latent trajectory model with structured residuals; Stand. = standardized coefficient; Agg = child aggression; Spank = maternal spanking; Int = latent intercept (centered at time 1); Slope = latent linear growth rate; ε_{it} = structured residual for individual *i* at time *t*; RMSEA = root mean square error of approximation; CFI = comparative fit index. ~ $p < .10$, ***$p < .001$.

on the mean of two items from a modified version of the Conflict Tactics Scale developed by the *Fast-Track* study (Strassberg et al., 1994). Mothers rated the extent to which they spanked the target child with a hand and with an object, on a 7-point scale ranging from 0 (*never*) to 7 (*almost every day*). Externalizing-type behavior was based on maternal ratings on the five-item conduct problems scale from the Strengths and Difficulties Questionnaire (Goodman, 2001). Higher scores reflect more frequent spanking and more problematic behavior, respectively. Though conduct problems can extend beyond aggression, as before, we use "aggression" for continuity. We fitted an identical series of models as those described in the first empirical example.

Results

As above, the ARCL specification does not fit the data terribly well (Table 3). Specification checks indicated that, in addition to the longitudinal constraints, ill fit was driven largely by the fact we constrained autoregressive and cross-lagged parameters at more distal lags to 0. Nonetheless, similar to the extant literature, we find modest, positive cross-lagged relations. On average, children who were spanked more at $time_{t-1}$ tended to show worse $time_t$ aggression ($B = .02$, $p < .001$; $\beta = .09$), and children who showed worse aggression at $time_{t-1}$ tended to be spanked more at $time_t$ ($B = .28$, $p < .001$; $\beta = .09$), adjusting for the respective autoregressive relations. One might readily

interpret this as a feedback loop, whereby spanking and aggression exacerbate each other over time (albeit, modestly).

However, once we disaggregate the within- and between-person relations, this conclusion seems less tenable—at least, with respect to within-person processes. Although there is evidence of modest to moderate between-child association between children's average levels of spanking and aggression over time ($\psi_{stand} = .34$), both of the within-person cross-lagged relations were statistically nonsignificant ($B_{t-1spank} = .009$, $p = .24$; $\beta = .03$; $B_{t-1agg} = .15$, $p = .07$; $\beta = .05$). As above, what appeared to be time-specific lagged effects were largely manifestations of *systematic* aspects about individuals over time. Of course, we cannot rule out the possibility that the contemporaneous between-person correlation between mean spanking and mean CP over time is, in actuality, driven by some unobserved feedback loop not accounted for by our model. However, there is no evidence of such process occurring within-person given the available data. Do the two levels of inference need to be disaggregated to be readily interpretable? As above, the answer seems to be, yes. Fitting a model constraint to recover the contextual effect indicated that the between- and within-person effects were, indeed, different ($B_{contextual} = .09$, $p < .001$).

Discussion

Autoregressive cross-lagged panel (ARCL) models have been a workhorse in developmental psychology for decades and are ubiquitous in contemporary developmental science. The popularity of the approach is presumably explained by the apparent alignment between the ARCL model and the reciprocal process that lay at the heart of many contemporary developmental theories. Informed by an increasingly vocal concern for the importance of disaggregating within- and between-person relations in the social sciences (e.g., Allison, 2009; Bollen & Brand, 2010; Curran & Bauer, 2011; Hoffman, 2015; Hoffman & Stawski, 2009; Molenaar & Newell, 2010; Voelkle et al., 2014), we have demonstrated that the typical ARCL specification (a) largely fails to align with the developmental theory it is often intended to test and (b) typically yields parameter estimates that are difficult to interpret meaningfully.

Specifically, we found repeated support for a positive *between-person* relation, such that children who were spanked more, on average, tended to be more aggressive than were children who were typically spanked less. However, no reciprocal relations were evident within-person. As such, what the typical ARCL specification suggested were reciprocal feedback effects were in actuality explained entirely by the between-person correlation between children's *overall* levels of spanking and aggression over time.

What does this mean? On substantive level it suggests that the reciprocal relation is not evident at the level of analysis that is arguably the most relevant to developmental theory (i.e., within-person). On a methodological level, it means that the cross-lagged associations disappeared after accounting for all possible time-invariant confounds—an inherent methodological strength of "fixing" the effect to reflect only within-child variation. To be clear, we are not suggesting that this provides definitive evidence that there is *no* causal relation between spanking and aggression (or the inverse). Causal links between spanking and aggression may be driven by chronicity or dose—aspects that would be missed by our within-person specification, yet captured by the between-person component of the model. Ultimately, though, we are left with a between-person correlation. Thus, we are back to square one with respect to the internal validity of our inference and the reciprocal processes we originally set out to test.

What role might statistical power play in these conclusions? When there are multiple levels of inference, such as the between- and within-person effects of the ALT-SR, each relation carries its own statistical power. Establishing power for such multilevel data is no small feat. Power is affected by a complex array of factors. For instance, in addition to effect size and sample size, the relative magnitudes of the respective between- and within-person variances have an especially profound effect on power, as they establish the extent to which observations are independent. In addition, statistical power can also be affected by the number, timing, and balance of the longitudinal observations over time, the reliability of one's variables and the shapes of their distributions, model size, missing data, and so forth (see Hox, 2010; Muthén & Curran, 1997; Snijders & Bosker, 2012). Given this complexity, valid "rules of thumb" regarding sample size and number of longitudinal observations are difficult to establish.

In terms of more general observations, with longitudinal data, one can often maximize predictor variability and/or reliability (thus, power) by adding individuals or longitudinal observations to the

data (Bolger & Laurenceau, 2014; Willett, 1989). Some simulation work suggests that the former typically affords greater benefits (Bolger & Laurenceau, 2014). All things being equal, within-person inferences tend to carry greater statistical power than do between-person inferences (Bellemare, Bissonnette, & Kröger, 2014; Cohen, 1977). Also, in series of "back of the envelope" simulations (see Supporting Information), we found some indication that, on average, across multiple sample and effect sizes, the statistical power of within-person inferences were also less impacted by reductions in predictor variance than were between-person inferences. Given that the cross-lagged estimates of the ALT-SR reflect within-person relations, these benefits are noteworthy.

Ultimately, though, an underpowered analysis is an underpowered analysis. A modest effect size of .10 would be woefully underpowered, for say, the sample of 290 boys (with four observations) from our first empirical example, irrespective of whether the effect is between-person or within-person—even if we doubled the number of longitudinal observations (see Supporting Information). As such, the broad take-home messages with respect to the ARCL/ALT-SR models and statistical power are: (a) establish that one's model yields interpretable parameter estimates (a weakness of the ARCL model); (b) if this requires disaggregating within- and between-person effects (a strength of the ALT-SR model), then establish plausible effect sizes for *each* level of analysis, along with plausible ICCs; and (c) conduct a multilevel power analysis, being realistic about the sample sizes and longitudinal observations required by the small effect sizes that are common in our field.

Is the run-of-the mill ARCL model inherently "wrong?" It is wrong only to the extent to which the ARCL specification is a bad representation of the population processes that gave rise to the data. Like others (e.g., Hoffman, 2015; Snijders & Bosker, 2012), we have proposed that the "convergence" assumptions underlying the model are typically implausible; however, it need not be the case. Indeed, although the true population processes are unknown, models fitted to "real" data can often be compared, as a presumed proxy for the plausibility of their respective representations of population processes. For instance, we presented Curran and colleagues' (2014) ALT-SR model as a more plausible specification for the population processes of interest because it disaggregates between- and within-person relations using statistical tools that are already familiar to many developmentalists. As the typical ARCL specification is nested within the ALT-SR, these models (and mixed variants therein) can be compared using likelihood ratio tests. In addition, although the ARCL specification may not always be directly recoverable, a variety of different models are available test complex dynamic relations between children and contexts over time, between- and within-persons (e.g., latent difference score, McArdle, 2001; state–trait, Kenny & Zautra, 2001; latent differential equations, Boker & Laurenceau, 2007; state–space, Molenaar, 2003; see Grimm, 2007 for overview). At the end of the day, it is difficult to know whether one has specified the "right" model. However, with some theory-driven comparisons, one hopes to feel increasing confident about specifying the *least wrong* model.

Of course, there are many ways to get the model wrong, despite disaggregating within- and between-person effects. Using the present ALT-SR specification as just one example, at a minimum, our model assumes reliable measures, factorial invariance, stationarity, discrete and balanced time, appropriate growth functions, a temporal lag structure that is causally appropriate (in both directions), simultaneity in the residuals but not the structural relations, no cross-lagged associations between the residuals (i.e., strict exogeneity), no retest effects, no unobserved confounds (including contemporaneous causal effects), no within-person interactions with time or cross-level interactions between within- and between-person spanking and/or aggression, and that the within-person autoregressive and cross-lagged estimates do not vary randomly across children (Curran & Bauer, 2011; Hertzog & Nesselroade, 2003; Wooldridge, 2010). Many of these assumptions are testable and can be relaxed. Like any other model, they should be tested to the extent to which they can be, balancing substantive theory with empirical specification checks.

Other assumptions are more difficult to test. For instance, practical and methodological realities often make the assumption that one's temporal lags are on their true causal time scales rather dubious and tricky to test. Furthermore, like the vast majority of quantitative work conducted by developmentalists, our models assume that the developmental processes we are attempting to model are generalizable beyond single individuals—so called, ergodicity. That is, although our specification disaggregates within- and between-person relations, the within-person estimates nonetheless reflect *pooled* estimates *across* individuals. With the exception of those working with time-series data (i.e., ~100+ longitudinal observations), most

developmentalists lack the variability to conduct single-subject analyses. However, as argued eloquently by others (e.g., Molenaar, 2004; Molenaar & Newell, 2010), the ergodicity assumption may well be inappropriate for many developmental phenomena. Simply stated, what is true in aggregate may rarely be true for the individual. As technology allows us to collect increasingly intensive developmental data, this will become an important and fascinating assumption to test more regularly.

Until then, though, those of us working with more typical longitudinal data will need to do the best we can. Increasing work highlights the fact that accurately disaggregating within- and between-person processes is a necessary (though not sufficient) part of this aim. As such, it may (again) be time to put the ARCL workhorse out to pasture. The good news is that the remarkable rate of innovation in longitudinal data science has made it far greater than a two-horse race.

References

Allison, P. D. (2009). *Fixed effects regression models* (Vol. 160). London, UK: Sage.

Bellemare, C., Bissonnette, L., & Kröger, S. (2014). *Statistical power of within and between-subjects designs in economic experiments*. Working Paper #8583. Bonn, Germany: IZA.

Boker, S. M., & Laurenceau, J. P. (2007). Coupled dynamics and mutually adaptive context. In T. D. Little, J. A. Boivard, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 299–324. Mahwah, NJ: Erlbaum.

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, *89*, 1–34. doi:10.1353/sof.2010.0072

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective* (Vol. 467). Hoboken, NJ: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. London, UK: Sage.

Campbell, D. T. (1963). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 212–242). Madison, WI: University of Wisconsin Press.

Chou, C. P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*, 247–266. doi:http://dx.doi.org/10.1080/10705519809540104

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, *115*, 74. doi:10.1037/0033-2909.115.1.74

Cronbach, L., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude X treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, *67*, 717–724. doi:10.1037/0022-0663.67.6.717

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583–619. doi:10.1146/annurev.psych.093008.100356

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*, 8–94. doi:10.1037/a0035297

Curran, P. J., Lee, T. H., Howard, A. L., Lane, S. T., & MacCallum, R. C. (2012). Disaggregating within-person and between-person effects in multilevel and structural equation growth models. In J. Harring & G. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 217–253). Charlotte, NC: Information Age.

Duncan, O. D. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, *72*, 177. doi:http://dx.doi.org/10.1037/h0027876

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, *43*, 557–572. Retrieved from http://www.jstor.org/stable/2094779

Gershoff, E. T. (2013). Spanking and child development: We know enough now to stop hitting our children. *Child Development Perspectives*, *7*, 133–137. doi:10.1111/cdep.12038

Gershoff, E. T., Lansford, J. E., Sexton, H. R., Davis-Kean, P., & Sameroff, A. J. (2012). Longitudinal links between spanking and children's externalizing behaviors in a national sample of White, Black, Hispanic, and Asian American families. *Child Development*, *83*, 838–843. doi:10.1111/j.1467-8624.2011.01732.x

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child Adolescent Psychiatry*, *40*, 1337–1345. doi:doi:10.1097/00004583-200111000-00015

Grimm, K. J. (2007). Multivariate longitudinal methods for studying developmental relationships between depression and academic achievement. *International Journal of Behavioral Development*, *31*, 328–339. doi:10.1177/0165025407077754

Gromoske, A. N., & Maguire-Jack, K. (2012). Transactional and cascading relations between early spanking and children's social-emotional development. *Journal of Marriage and Family*, *74*, 1054–1068. doi:10.1111/j.1741-3737.2012.01013.x

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102. doi:10.1037/a0038889

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251–1271. Retrieved from http://hdl.handle.net/1721.1/64309

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging*, *18*, 639. doi:10.1037/0882-7974.18.4.639

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*, 97–120. doi:10.1080/15427600902911189

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. London, UK: Routledge.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York, NY: Seminar Press.

Kenny, D. A. (1973). Cross-lagged and synchronous common factors in panel data. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 153–167). New York, NY: Seminar Press.

Kenny, D. A., & Harackiewicz, J. M. (1979). Cross-lagged panel correlation: Practice and promise. *Journal of Applied Psychology*, *64*, 372. doi:10.1037/0021-9010.64.4.372

Kenny, D. A., & Zautra, A. (2001). *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, US: American Psychological Association. doi: http://dx.doi.org/10.1037/10409-008

Lansford, J. E., Criss, M. M., Laird, R. D., Shaw, D. S., Pettit, G. S., Bates, J. E., & Dodge, K. A. (2011). Reciprocal relations between parents' physical discipline and children's externalizing behavior during middle childhood and adolescence. *Development and Psychopathology*, *23*, 225–238. doi:10.1017/S0954579410000751

Lee, S. J., Altschul, I., & Gershoff, E. T. (2013). Does warmth moderate longitudinal associations between maternal spanking and child aggression in early childhood? *Developmental Psychology*, *49*, 2017–2028. doi:10.1037/a0031630

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. H. C. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

Molenaar, P. C. (2003). *State space techniques in structural equation modeling: Transformation of latent variables in and out of latent variable models*. Retrieved from http://www.hhdev.psu.edu/hdfs/faculty/molenaar.html

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201–218. doi:10.1207/s15366359mea0204_1

Molenaar, P., & Newell, K. M. (2010). *Individual pathways of change: Statistical models for analyzing learning and development*. Washington, DC: American Psychological Association.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371. doi:10.1037/1082-989X.2.4.371

Muthén, L. K., & Muthén, B. O. (2015). *Mplus. Statistical Analysis with Latent Variables: User's Guide*. Los Angeles: Muthén & Muthén.

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, *23*, 303–326. doi:doi:10.1016/S0190-7409(01)00141-4

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, *38*, 337–341. doi:10.1093/ije/dyn357

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*, 245. doi:10.1037/0033-2909.88.2.245

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY: Oxford University Press.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Strassberg, Z., Dodge, K. A., Pettit, G. S., & Bates, J. E. (1994). Spanking in the home and children's subsequent aggression toward kindergarten peers. *Development and Psychopathology*, *6*, 445–461. doi:10.1017/S0954579400006040

Vernon-Feagans, L., Cox, M.; the FLP Key Investigators. (2013). The Family Life Project: An epidemiological and developmental study of young children living in poor rural communities. *Monographs of the Society for Research in Child Development*, *78*(Serial No. 5), 1–150. doi:10.1111/mono.12046

Voelkle, M. C. (2008). Reconsidering the use of autoregressive latent trajectory (ALT) models. *Multivariate Behavioral Research*, *43*, 564–591. doi:10.1080/00273170802490665

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, *49*, 193–213. doi:10.1080/00273171.2014.889593

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*, 587–602. doi:10.1177/001316448904900309

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of

individual change over time. *Psychological Bulletin*, *116*, 363. doi:10.1037/0033-2909.116.2.363

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Recovering the Autoregressive Cross-Lagged Panel (ARCL) "Convergence Effect" Estimates From the Autoregressive Latent Trajectory Model With Structured Residuals (ALT-SR) Disaggregated Estimates.

Appendix S2. Example 3: Fragile Families and Child Well-Being Study.

Appendix S3. Data and Exemplar Mplus Syntax for the Main Empirical Examples.

Appendix S4. Population Parameters and Mplus Syntax for Monte Carlo Simulations.

Appendix S5. "Back of the Envelope" Monte Carlo Power Curves for a Simple Random Intercepts Model With a Single Time-Varying Covariate (X).