Testing Prospective Effects in Longitudinal Research: Comparing Seven Competing Cross-Lagged Models

Ulrich Orth
University of Bern
D. Angus Clark
University of Michigan
M. Brent Donnellan
Michigan State University
Richard W. Robins
University of California, Davis

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2020). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000358

Author Note

Ulrich Orth, Department of Psychology, University of Bern; D. Angus Clark, Department of Psychiatry, University of Michigan; M. Brent Donnellan, Department of Psychology, Michigan State University; Richard W. Robins, Department of Psychology, University of California, Davis.

This research was supported by Swiss National Science Foundation Grant PP00P1-123370 to Ulrich Orth, National Institute on Drug Abuse Grant DA017902 to Richard W. Robins, and National Institute on Aging Grant R01AG060164 to Richard W. Robins. We thank Laurenz L. Meier for allowing us to use data from the "My Work and I" study for this article. The present research has been preregistered on December 3, 2017, on the Open Science Framework (OSF, https://osf.io/tzjmd). For five of the samples, data are available on the OSF (https://osf.io/5rjsm). The OSF folder also includes information on how the data of the other five samples can be accessed, Mplus scripts and output for all models, the R script used for the meta-analytic computations, and sample Mplus scripts and output for multiple-indicator versions of the CLPM and RI-CLPM.

Correspondence concerning this article should be addressed to Ulrich Orth, Department of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland. E-mail: ulrich.orth@psy.unibe.ch

Abstract

In virtually all areas of psychology, the question of whether a particular construct has a prospective effect on another is of fundamental importance. For decades, the cross-lagged panel model (CLPM) has been the model of choice for addressing this question. However, CLPMs have recently been critiqued, and numerous alternative models have been proposed. Using the association between low self-esteem and depression as a case study, we examined the behavior of seven competing longitudinal models in 10 samples, each with at least four waves of data and sample sizes ranging from 326 to 8,259. The models were compared in terms of convergence, fit statistics, and consistency of parameter estimates. The traditional CLPM and the random intercepts cross-lagged panel model (RI-CLPM) converged in every sample, whereas the other models frequently failed to converge or did not converge properly. The RI-CLPM exhibited better model fit than the CLPM, whereas the CLPM produced more consistent cross-lagged effects than the RI-CLPM. We discuss the models from a conceptual perspective, emphasizing that the models test conceptually distinct psychological and developmental processes, and address the implications of the empirical findings with regard to model selection. Moreover, we provide practical recommendations for researchers interested in testing prospective associations between constructs, and suggest using the CLPM when focused on between-person effects and the RI-CLPM when focused on within-person effects.

Keywords: cross-lagged panel models; longitudinal; prospective effects; statistical analyses; structural equation modeling

Cross-lagged regression models are by far the most commonly used method to test the prospective effect of one construct on another (Biesanz, 2012; McArdle, 2009; Wu, Selig, & Little, 2013). However, the traditional cross-lagged panel model (CLPM) has been critiqued because it does not distinguish between-person and within-person variance (e.g., Berry & Willoughby, 2017; Hamaker, Kuiper, & Grasman, 2015; Usami, Hayes, & McArdle, 2016; Usami, Todo, & Murayama, 2019). A number of newer models have been proposed with the goal of providing more valid insights about prospective effects between constructs (see Usami, Murayama, & Hamaker, 2019). These newer models differ in a number of ways, in particular with regard to how they account for between-person differences.

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 (in most models, between-person effects are modeled as correlations between trait factors). However, in many fields researchers are also interested in gaining information about the consequences of between-person differences (e.g., Do individuals with low self-esteem have a higher risk of developing depressive symptoms compared to individuals with high self-esteem?). Thus, although it seems useful to distinguish within-person and between-person differences, the suggested new models do not allow testing for prospective between-person effects.

Besides this conceptual issue, another concern with the available models is that there is the risk that researchers select one of the models in an arbitrary way, without considering conceptual differences between the models in the actual psychological or developmental process being tested. Moreover, the choice of model could reflect selective reporting (i.e., choosing the best model after seeing the results from several competing models), which would reduce the confirmatory nature of the research and increase the rate of false-positive findings (Simmons,

Nelson, & Simonsohn, 2011). Thus, it is important for the field to seek consensus about standard models that should be used for specific types of research questions (e.g., analyses of between-person or within-person prospective effects).

The goal of the present research was to compare seven competing longitudinal models that have all been proposed as ways to estimate cross-lagged effects between constructs. We assessed the frequency of convergence problems, the fit of the models, and the consistency of parameter estimates across ten longitudinal samples, each with at least four waves of data. We used real datasets—not simulated data—to test the models, because simulated data are necessarily based on assumptions about the data generating process, which may favor some models over others and thereby confound conclusions about the behavior, validity, and replicability of the models. We used the relation between low self-esteem and depression as a case study, given that there is a large body of research examining this substantive question (for a review, see Orth & Robins, 2013) and given that numerous longitudinal multi-wave datasets were available that include well-validated measures of both constructs. The existing research suggests that low self-esteem has a prospective effect on depression, but this finding is based almost exclusively on the traditional CLPM, and whether it emerges when other longitudinal models are used is largely unknown.

Description of the Models Tested in the Present Research

We examined the following seven models, all of which estimate cross-lagged effects between two constructs: the CLPM, that is, the traditional version of the cross-lagged model (e.g., Finkel, 1995); random intercepts cross-lagged panel model (RI-CLPM; Hamaker et al., 2015); autoregressive latent trajectory model (ALT; Bollen & Curran, 2004; for an earlier version of this model, see Curran & Bollen, 2001); latent curve model with structured residuals

(LCM-SR; Curran, Howard, Bainter, Lane, & McGinley, 2014); bivariate latent change score model (LCS; McArdle, 2001); bivariate latent change score model with changes-to-changes extension (LCS-CC; Grimm, An, McArdle, Zonderman, & Resnick, 2012); and bivariate cross-lagged trait-state-error model (Kenny & Zautra, 1995), also called the STARTS model (Kenny & Zautra, 2001). Figure 1 illustrates the models examined in this research, Table 1 provides an overview of the meaning of the cross-lagged effects in each model, and the Supplemental Material includes sample Mplus scripts of the models. In specifying the models, we followed standard specifications in the methodological literature.

Here, we briefly describe the models (see also Usami, Murayama, et al., 2019). The CLPM tests for the prospective effect of individual differences in a construct on change in individual differences in the other construct (the CLPM tests for change in individual differences, since the effects are controlled for the autoregressive effects in the constructs). Thus, the CLPM focuses on relative change in the constructs. For example, in the case study of low self-esteem and depression, the hypothesized causal effect is, “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” (see Table 1). The CLPM does not place constraints on the means of the variables, that is, the means are allowed to vary freely across waves.

The RI-CLPM is similar to the CLPM in some respects, but includes trait factors (called random intercept factors) that capture the stable between-person variances in the constructs. However, unlike intercept factors in a latent curve model, the observed mean structure is not explained by the random intercept factors, and indicator means are allowed to vary freely across waves. The autoregressive and cross-lagged effects are then modeled on the residualized scores,

not the observed scores. Consequently, in the RI-CLPM, a cross-lagged effect informs about 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 (the RI-CLPM tests for change in the within-person deviation from the trait level, since the effects are controlled for the autoregressive effects in the deviations). In the case study of low self-esteem and depression, the hypothesized causal effect is, “When individuals have lower self-esteem than usual, they will experience a subsequent increase in depression.”

The ALT is similar to the CLPM in some respects, but includes intercept and linear slope factors. The autoregressive and cross-lagged effects are estimated using the observed scores. The observed variable at Time 1 is excluded from the modeling of intercept and linear slope (though there are alternative approaches to addressing the Time 1 observation; see Jongerling & Hamaker, 2011; Ou, Chow, Ji, & Molenaar, 2017). Therefore, in the ALT, the mean structure of the variables is explained not only by the intercept and slope factors, but also by the means of the Time 1 variables. It is important to note that in the ALT, the intercept and slope factors are so-called accumulating factors (Usami, Murayama, et al., 2019), which means that the exact combination of intercept factors, slope factors, Time 1 means, autoregressive effects, and cross-lagged effects is needed to describe the average and individual trajectory for each of the constructs. For this reason, the intercept and slope factors of the ALT cannot be interpreted as simple growth factors as in growth curve models. In the case study of low self-esteem and depression, the hypothesized causal effect is, “When individuals have low self-esteem (relative to others), they will experience a subsequent increase in their rank-order value in depression, controlling for the accumulating intercept and slope factor of depression.”

The LCM-SR is similar to the RI-CLPM in many respects, but—in addition to the random intercept factors—includes slope factors that detrend the observed means. In the LCM-SR, the mean structure of the variables is explained by the intercept and slope factors, as common in latent growth models. In the LCM-SR, the autoregressive and cross-lagged effects are modeled on the residualized scores (as in the RI-CLPM), not the observed scores (as in the ALT). Applied to the case of low self-esteem and depression, the hypothesized causal effect is, "When individuals have lower self-esteem than would be expected from the self-esteem trajectory they follow, they will experience a subsequent increase in depression."

In the LCS, the observed scores are decomposed into latent wave-specific scores and error terms. The latent wave-specific scores are explained by several sources, including an intercept (also called level factor), a slope factor (accounting for constant change), an autoregressive effect (accounting for proportional change), and a cross-lagged effect (i.e. the effect of the predicted score in the other construct at the previous assessment). Thus, the LCS includes latent growth factors, similar to the ALT and LCM-SR, but an important difference is that change in the latent wave-specific scores is modeled explicitly in the LCS and that the autoregressive and cross-lagged effects predict latent change scores (or latent difference scores), not the observed scores (as in the ALT) or residualized scores (as in the LCM-SR). In other words, the LCS focuses on absolute change in the constructs. In the LCS, the mean structure of the variables is explained by the intercept and slope factors. Applied to the case of low self-esteem and depression, the hypothesized causal effect is, "When individuals have low self-esteem, they will experience a subsequent increase in depression."

The LCS-CC is very similar to the LCS. However, the LCS-CC also includes cross-lagged effects of the change scores on subsequent change scores, in addition to the cross-lagged

effects of the latent wave-specific scores on subsequent change scores. In the LCS-CC, the mean structure of the variables is explained by the intercept and slope factors. Applied to the case of low self-esteem and depression, the hypothesized causal effect (i.e., the change-to-change effect) is, “When individuals have decreased in self-esteem, they will experience a subsequent increase in depression.”

The STARTS is similar to the RI-CLPM in some respects, as the observed variables are explained by a trait factor and the autoregressive and cross-lagged effects are modeled on residualized scores. In the STARTS, the residualized scores are called state factors (Kenny & Zautra, 1995) or autoregressive trait factors (Kenny & Zautra, 2001). As in the RI-CLPM, the mean structure is not explained by the trait factors, and the means are allowed to vary freely across waves. However, important differences from the RI-CLPM are that the STARTS models error terms of the observed variables (i.e., the error terms are modeled in addition to the residualized scores) and includes complex constraints on the variances and covariances of the residualized variables to impose stationarity (such constraints might be unreasonable in developmental studies; see Donnellan, Kenny, Trzesniewski, Lucas, & Conger, 2012). In the case study of low self-esteem and depression, the hypothesized causal effect is, “When individuals have lower self-esteem than usual, they will experience a subsequent increase in depression.” Note that the description of the hypothesized causal effect is identical for the RI-CLPM and the STARTS; although the models differ in their exact specification as described above, the interpretation of the causal effect does not differ across the two models.

The models described above are often used to address the same general question, “What is the prospective effect of one variable on another?” Thus, researchers often have the same general causal process in mind when using one of these models. However, it is important to

emphasize that the models do not address the same psychological or developmental process, but rather test conceptually distinct processes (see Table 1). Therefore, the models should not be considered replications of each other, and the inferences that can be drawn from the cross-lagged effects depend directly on the model used to estimate the effects. Before selecting a model, researchers should carefully consider the psychological or developmental process they would like to examine in their research, and then select a model that best estimates that process.

There is an important caveat to this recommendation. Few, if any, psychological theories have the precision to perfectly match the hypothesized process to, for example, the RI-CLPM versus the ALT model or STARTS model. As noted above, we even suggest the same wording for the hypothesized causal effect in the RI-CLPM and STARTS model. Thus, we do not think that the differences in the specification of these two models lead to clear differences in the interpretation of the cross-lagged effects. Consequently, we consider it unlikely that some theories would provide sufficient nuance about the process to favor the RI-CLPM over the STARTS model, whereas other theories would provide compelling reasons for selecting the STARTS model. Even if the cross-lagged effect can be conceptually distinguished in other models, such as the LCM-SR, we think that theories rarely provide sufficient guidance to favor any particular within-person model over the others. In contrast, some psychological theories do distinguish between two types of change processes, that is, between-person and within-person effects. Given that multiple alternative models are available, especially for the analysis of within-person effects, we contend that the field would benefit from consensus about standard models that should be used for testing these two types of research questions, that is, questions related to prospective effects at the between-person versus within-person level.

The Case of Low Self-Esteem and Depression

For decades, the question of whether low self-esteem is a risk factor for depression has been the focus of theory and empirical research in the fields of social-personality and clinical psychology (e.g., Beck, 1967; Blatt, D'Afflitti, & Quinlan, 1976; Brown & Harris, 1978; Roberts & Monroe, 1994). A number of theoretical models of the relation between low self-esteem and depression have been proposed in the literature (for a review, see Orth & Robins, 2013). One of these, the *vulnerability model*, states that low self-esteem leads to depression. For example, Beck's cognitive theory of depression posits that negative beliefs about the self—as indicated by low self-esteem—are not just symptoms of depression, but are causally involved in the etiology of depressive symptoms and disorders (see also Abramson, Seligman, & Teasdale, 1978). In contrast, the *scar model* hypothesizes that low self-esteem is a consequence, rather than a cause, of depression, because depressive episodes might persistently alter a person's self-concept (Coyne, Gallo, Klinkman, & Calarco, 1998; Shahar & Davidson, 2003). In turn, these "scars" in the self-concept might account for decreased self-esteem even after the depressive episode has already remitted. It is important to note that both processes (i.e., low self-esteem leading to depression, and depression eroding self-esteem) are not mutually exclusive. Consequently, the *reciprocal relation model* is another possible model of the link between low self-esteem and depression (Orth & Robins, 2013). Moreover, the empirical relation between low self-esteem and depression could be spurious and caused by third factors, such as prior occurrence of stressful life events or underlying temperament factors (Hankin, Lakdawalla, Carter, Abela, & Adams, 2007; Watson, Suls, & Haig, 2002). Thus, the *spurious relation model* assumes that low self-esteem does not lead to depression and depression does not lead to low self-esteem.

In recent years, a growing number of longitudinal studies have supported the vulnerability model, that is, the hypothesis that low self-esteem leads to depression (e.g.,

Masselink, Van Roekel, & Oldehinkel, 2018; Orth, Robins, & Meier, 2009; Orth, Robins, & Roberts, 2008; Rieger, Göllner, Trautwein, & Roberts, 2016; Steiger, Allemand, Robins, & Fend, 2014; van Tuijl, de Jong, Sportel, de Hullu, & Nauta, 2014; Wouters et al., 2013). Moreover, a meta-analysis of 77 longitudinal studies provided strong support for the vulnerability model and weaker support for the scar model; the standardized prospective effect sizes for the vulnerability and scar effect were estimated as −.16 and −.08, respectively (Sowislo & Orth, 2013). The vulnerability model is highly robust, holding for men and women, for different age groups from childhood to old age, for different measures of self-esteem and depression, for cognitive-affective and somatic symptoms of depression, and for different time lags between assessments, ranging from one week to many years (Kuster, Orth, & Meier, 2012; Orth, Robins, Trzesniewski, Maes, & Schmitt, 2009; Sowislo & Orth, 2013; Steiger et al., 2014). Finally, the vulnerability effect holds after controlling for factors that might lead to both low self-esteem and depression, such as stressful life events (Orth, Robins, & Meier, 2009; Orth, Robins, Widaman, & Conger, 2014), low social support (Orth et al., 2014), and neuroticism (Sowislo, Orth, & Meier, 2014).

Given the consistent level of support in the literature, researchers have tested more detailed specifications of the vulnerability model. First, longitudinal data suggest that rumination partially mediates the effect; that is, low self-esteem prospectively predicted rumination, which in turn prospectively predicted depression (Kuster et al., 2012). Second, the vulnerability effect is driven predominantly by global self-esteem rather than domain-specific self-evaluations (Orth et al., 2014; Steiger et al., 2014). Third, the crucial vulnerability factor is a low level of self-esteem rather than instability or contingency of self-esteem (Sowislo et al., 2014). Fourth, the

vulnerability effect is accounted for by a lack of genuine self-esteem and not a lack of narcissistic self-enhancement (Orth, Robins, Meier, & Conger, 2016).

Nearly all of the studies reviewed above used the same statistical model, that is, the CLPM. Therefore, it is important to note that the current state of knowledge on the topic is based on the assumption that the CLPM provides a valid test of the vulnerability model. As discussed earlier in this article, alternative cross-lagged models differ in the specific meaning of the prospective effect from one construct on another (see Table 1, which also includes statements about the meaning of effects for the example of low self-esteem and depression). Since the present research uses the relation between low self-esteem and depression as a case study, the findings have the potential to deepen our understanding of the substantive issue under investigation, in addition to providing important insights into the statistical models being compared. We are aware of only one published article that has used a model other than the CLPM to test prospective associations between low self-esteem and depression. Specifically, Masselink, Van Roekel, Hankin, et al. (2018) used the CLPM and RI-CLPM, and found that the vulnerability effect of low self-esteem on depression was significant in both models (standardized effects ranged from −.11 to −.15 across waves and models), whereas the scar effect of depression on self-esteem was significant in the CLPM but nonsignificant in the RI-CLPM (standardized effects ranged from −.06 to −.10 across waves and models). However, the prospective association between low self-esteem and depression has not been tested yet with any of the other alternative models, that is, the ALT, LCM-SR, LCS, LCS-CC, and STARTS.

The Present Research

The goal of the present study was to gain insight into the behavior of seven longitudinal models that provide different ways of estimating cross-lagged effects between constructs. We

compared these models in terms of frequency of convergence problems, model fit, and consistency of parameter estimates. By replicating the findings across 10 longitudinal studies, we sought to better understand the behavior of the models when applied to datasets that are common in longitudinal research in psychology (i.e., data from samples with several hundred participants assessed at a moderate number of waves, such as 3-5 waves).

We used real datasets for the analyses because simulated data are necessarily based on assumptions about the data generating process. Given that the goal was to compare several competing models, it would be unclear how the data should be generated for simulation analyses. For example, if we would compare the competing models on the basis of data generated by one of the models (e.g., the LCM-SR), this model would likely show the best behavior in terms of model convergence, fit, and consistency of estimates. In short, we believe that analyses of actual data provides valuable information about how the models compare in those cases where researchers are uncertain about the data generating process.

Although the present analyses focus on empirical aspects of the models, we believe that conceptual considerations are important in selecting statistical models, and in many cases should have priority over empirical considerations. In the Discussion section, we will therefore explicitly distinguish between conceptual and empirical issues suggested by the present research. Nevertheless, even if models should be selected for *a priori* theoretical reasons, it is important to consider whether the data will yield model estimates with sufficient precision and whether there is significant risk of convergence issues when estimating the model. Thus, empirical data on convergence rates, model fit, and consistency of estimates specific to statistical models may provide important information for model selection and about how models behave in typical research applications.

Finally, the findings will contribute to a deeper understanding of the substantive question addressed by the present research, namely how are low self-esteem and depression associated over time. Given that nearly all prior studies on the topic were based on the CLPM, the present analyses with 10 samples may provide robust insights into the link between the constructs from both a between-person and within-person perspective.

Method

The present research has been preregistered on December 3, 2017, on the Open Science Framework (OSF, https://osf.io/tzjmd). For five of the samples, data are available in the OSF folder to this article (https://osf.io/5rjsm).[1] The OSF folder also includes Mplus scripts and output for all models, the R script used for the meta-analytic computations, and sample Mplus scripts and output for multiple-indicator versions of the CLPM and RI-CLPM.

Samples

The data come from six longitudinal studies including 10 samples. The Berkeley Longitudinal Study (BLS) has been approved by the Institutional Review Board of the University of California, Davis (529790-3, "Longitudinal Personality Study"). The California Families Project (CFP) has been approved by the Institutional Review Board of the University of California, Davis (217484-25, "Mexican Family Culture and Substance Use Risk and Resilience"). The data used from the Family Transitions Project (FTP) are archival data, available at the Inter-University Consortium for Political and Social Research (https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/26721). Likewise, the data from the National Longitudinal Survey of Youth 1979 (NLSY79) are archival data, available at https://www.nlsinfo.org. The studies My Work and I (MWI) and Your Personality (YP) were conducted in Switzerland and at the time of data collection (2009-2011) approval from an Ethics

Committee was not required in accordance with national law. Since we used anonymized data from the FTP, NLSY79, MWI, and YP, the present analyses were exempt from approval by the Ethics Committee of the first author's institution (Faculty of Human Sciences, University of Bern) in accordance with national law.

We used data from studies that fulfilled the following two criteria. First, the study was longitudinal, including at least four repeated assessments that were equally spaced across time (i.e., the time intervals between assessments were identical). Second, at each wave used, assessments of both self-esteem and depression were available (as noted above, in this research we used the link between low self-esteem and depression as a case study). The reasons for the first criterion were as follows: For most of the models tested (i.e., ALT, LCM-SR, LCS, LCS-CC, and STARTS), the minimum number of waves is four and it will be important that each model can be tested in each sample (it is a fairer comparison between models if all of them are tested based on the same set of samples).[2] Including studies with a moderate number of waves, such as four, is important because these are quite common in the literature. Moreover, using assessments that are equally spaced across time allows constraining structural coefficients (i.e., autoregressive and cross-lagged effects) to be equal across intervals, which increases the precision of estimates and keeps the models simple. Therefore, within studies, we used those waves that were equally spaced but not waves with differing time lags. A deviation from the preregistered research plan was that we included a fourth sample from the FTP, that is, the siblings sample, in addition to the children, mothers, and fathers samples. This sample inadvertently was not included in the research plan, but when we noticed this, we decided on its inclusion before conducting any analyses.

Table 2 provides an overview of descriptive information on the studies. The BLS is a study of a cohort of 508 individuals (57% female) who entered the University of California at Berkeley in 1992 (Robins, Hendin, & Trzesniewski, 2001). Four waves fulfilled the inclusion criteria for the present research: end of first, second, third, and fourth year of college. The CFP is an ongoing study of 674 Mexican-origin families from Northern California, who have been assessed annually since 2006 (Robins & Conger, 2017). For the children sample (50% female), Waves 1, 3, 5, and 7 fulfilled the inclusion criteria, and for the mothers sample, Waves 3, 5, 7, and 9. The FTP is a study of 451 families from Iowa, who have been assessed since 1989 (Conger et al., 2011). In the present research, we could use data from four samples: target children (52% female), siblings (52% female), mothers, and fathers. For all four samples, four waves of data could be used (1989 to 1992). The MWI is a German-language study of 663 individuals (51% female) reporting on work experiences and well-being (Meier & Spector, 2013). The study includes five waves of data, all of which could be used; assessments were conducted in 2009 and 2010. The NLSY79, Young Adults Section, is a nationally representative study of 11,521 adolescents and young adults (49% female), who have been assessed biannually since 1986 (https://www.nlsinfo.org). In the present research, data from 11 waves from 1994 to 2004 could be used. The YP is a German-language study of 344 young adults (49% female) living in Switzerland (Orth & Luciano, 2015). The study includes four waves of data, all of which could be used; assessments were conducted in 2010 and 2011.

Five of the samples were used in previous studies examining the relation between self-esteem and depression. However, in all of these studies, the CLPM was tested but none of the other models examined in the present research.[3]

Measures

Self-esteem was assessed in nine samples with the Rosenberg Self-Esteem Scale (RSE; Rosenberg, 1965), which is the most commonly used and well-validated measure of self-esteem (Donnellan, Trzesniewski, & Robins, 2015). The internal consistency of the RSE is typically in the .80s to .90s, which was also the case in the present samples (Table 2). In the CFP children sample, self-esteem was assessed with the Self-Description Questionnaire (Marsh, Ellis, Parada, Richards, & Heubeck, 2005), which is an established measure of self-esteem for children and adolescents (Donnellan et al., 2015).

Depression was assessed with a range of established multi-item measures, including the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977), a 7-item short form of the CES-D, the Mini Mood and Anxiety Symptom Questionnaire (MASQ; Casillas & Clark, 2000), the Symptom Checklist 90 (SCL-90; Derogatis & Savitz, 1999), and the Early Adolescent Temperament Questionnaire–Revised (EATQ; Capaldi & Rothbart, 1992; Ellis & Rothbart, 2001). As reported in Table 2, coefficient alpha estimates were generally high, with the exception of two samples in which brief depression measures were used (i.e., CFP children sample and NLSY79).[4]

Model Specifications

For all models, we used scale scores (i.e., mean scores across items included in a measure) to measure the constructs. Thus, the constructs were examined as observed variables, not latent variables, corresponding to how the new cross-lagged models have been described in the literature. For many of the models, versions with latent variables have not been introduced and, moreover, for most of the models nearly all applications use observed variables. Thus, it was important to examine the models in the way they are actually used in the field.

For some models (e.g., LDS), it is common practice to constrain residual variances and residual covariances to be equal across waves, although these constraints are often not described as a required part of the model and often not explicitly included in model descriptions. In contrast, for other models (e.g., CLPM), these constraints are typically not used by researchers. Thus, usage of constraints on residual variances and covariances varies across models (for a similar situation regarding latent growth models, see Grimm & Widaman, 2010). For this reason, we systematically tested four different versions of each model: (a) basic version, that is, without constraints on residual variances and covariances, (b) with constraints on residual variances, (c) with constraints on residual covariances, and (d) with constraints on both residual variances and residual covariances. We did not preregister testing these four versions, but expected that the constraints would reduce the frequency of convergence issues in more complex models and potentially reduce model fit. However, we did not have hypotheses about the influence of the constraints on the consistency of estimates across samples.

In the main analyses, we constrained structural coefficients (i.e., autoregressive and cross-lagged effects) to be equal across waves, because equality constraints across waves often facilitate proper convergence. Since we used these constraints in all models, any observed differences between models cannot be attributed to different usage of cross-wave equality constraints on structural coefficients, which is important for the validity of the conclusions. Moreover, we note that cross-wave equality constraints on structural coefficients are appropriate when the waves of data are equally spaced across time and when there is no theoretical reason to expect changes over time in the strength of the structural coefficients (Cole & Maxwell, 2003; Little, Preacher, Selig, & Card, 2007). Nevertheless, to gain information about the influence of

these constraints, in supplemental analyses we examined the convergence of models when omitting the constraints (see also Clark, Nuttall, & Bowles, 2018).

Moreover, to gain information about the influence of number of waves, in supplemental analyses we examined the convergence of models with 3-wave versions of all datasets by restricting the data to the first 3 waves. These analyses were conducted for those models that are identified with as few as three waves (i.e., CLPM and RI-CLPM).

Statistical Analyses

The analyses of structural equation models were conducted using Mplus version 8 (Muthén & Muthén, 2017). To deal with missing data, we used full information maximum likelihood estimation to fit models directly to the raw data, which produces less biased and more reliable results compared with conventional methods of dealing with missing data, such as listwise deletion (Schafer & Graham, 2002; Widaman, 2006). For all models, we used 20 random sets of starting values, which is useful when estimating complex models. Although the simpler models would not require this procedure, we treated all statistical models in the same way as our focus was on comparing the models. Fit was assessed by the comparative fit index (CFI) and the root-mean-square error of approximation (RMSEA), based on the recommendations of Hu and Bentler (1999) and MacCallum and Austin (2000). Hu and Bentler (1999) suggest that good fit is indicated by values greater than or equal to .95 for CFI, and less than or equal to .06 for RMSEA.

The meta-analytic computations were made with R 3.4.3 (R Core Team, 2017), using the metafor package (Viechtbauer, 2010). In the effect size analyses, we used random-effects models, following recommendations by Borenstein, Hedges, Higgins, and Rothstein (2009) and Raudenbush (2009). Between-study heterogeneity (i.e., τ^2) was estimated with the method of

moments, also called DerSimonian–Laird method (DerSimonian & Laird, 1986; Viechtbauer, 2010). Meta-analytic computations with effect sizes were made using Fisher's z_r transformations. Following Borenstein et al. (2009), the within-study variance of Fisher's z_r is given as $v = 1 / (n - 3)$.

Results

Convergence of Models

First, we examined the frequency with which the models showed convergence issues, such as nonconvergence or improper solutions (e.g., negative variances).[5] Table 3 shows the relative frequency of proper convergence across the 10 samples, for each of the models and model versions. The CLPM and RI-CLPM converged in every sample, whereas the other models frequently did not converge properly or did not converge at all. Convergence issues were most prevalent for the basic versions of the models (i.e., when no constraints on residual variances and covariances were used). But even when the full set of constraints were used, convergence issues were present in 40–70% of the samples (except for the LCS, with convergence issues in 10%). Although we expected that the newer models would converge less often than the CLPM, we were surprised by the frequency of convergence issues in the newer models, except for the RI-CLPM.

We then replicated the analyses without cross-wave equality constraints on autoregressive and cross-lagged effects. The LCS-CC includes these constraints for both the effects of levels on changes and the effects of changes on changes, so we omitted all of these constraints. In the STARTS, the constraints on the autoregressive effects are required for other constraints that are a fixed part of the model, so in this model we omitted only the constraints on the cross-lagged effects. The LCS-CC and STARTS never converged properly, the ALT, LCM-

SR, and LCS rarely converged properly, and the CLPM and RI-CLPM converged nearly always (Supplemental Table S1). Interestingly, for the CLPM and RI-CLPM the only convergence issues occurred in the NLSY79, which is the sample with the largest size and largest number of waves. The likely reason for this is the relatively large degree of missing data across the 11 waves of the NLSY79 (these missing data are partially by design), so cross-wave equality constraints on structural coefficients may help estimating the models in this particular situation.

For models that are theoretically identified with as few as three waves (i.e., CLPM and RI-CLPM), we replicated the analyses with 3-wave versions of all samples by restricting the datasets to the first three waves. The results showed that the CLPM and RI-CLPM converged properly in every sample (Supplemental Table S2).

Fit of Models

Next, we focused on model fit. As an example of the fit values in individual studies, Table 4 shows the fit for the basic versions of the seven competing models. The table also illustrates the pattern of convergence across models and samples. As reported above, the ALT, LCM-SR, LCS, LCS-CC, and STARTS often did not converge properly; however, when they converged, their fit was good. Overall, the fit of the RI-CLPM was roughly as good as the fit of these models. Fit values were lowest for the CLPM.[6]

Table 5 summarizes the results by showing mean fit values across the 10 samples. Given the relatively low frequency of proper convergence for all models except for the CLPM and RI-CLPM, it would be misleading to report average fit values for the other models. The reason is that for each of the other models the mean fit values would be based on a different subset of samples, and the fit values do not only depend on the model but also on the sample (thus, the means would not allow for a valid comparison of models). Therefore, we computed average fit

values only for the CLPM and RI-CLPM. The comparison clearly suggests that the RI-CLPM fits better than the CLPM, for all model versions. When including constraints on residual variances and covariances, fit was only slightly reduced for both the CLPM and RI-CLPM (in the case of the CLPM, the RMSEA actually indicated slightly better fit when including the constraints). Supplemental Tables S3-S5 show the fit values in individual studies for all competing models for the versions with constraints on residual variances and covariances. Similar to the pattern of results shown in Table 5, the results suggest that using these constraints leads to only minor reductions in model fit.

Consistency of Estimates

Finally, we examined the consistency of the structural estimates (i.e., cross-lagged and autoregressive effects) across the 10 samples. In these analyses, we focused again on the CLPM and RI-CLPM. Given the relatively low frequency of proper convergence for all other models, it would be misleading to examine the consistency of coefficients for these models. The coefficients from these models are drawn from a different subset of studies, which would not allow for a valid comparison of models, because differences between models could result from differences between samples in which the models converged. To illustrate the estimates in individual studies, Table 6 shows the results for the CLPM and RI-CLPM from the basic model. Moreover, Supplemental Tables S6-S12 report the estimates in individual studies for all competing models and all model versions. The findings shown in Table 6 suggest that the cross-lagged effects were more consistent for the CLPM compared to the RI-CLPM. A first indication is provided by the range of effect sizes. For the RI-CLPM, the range of cross-lagged effects (i.e., the difference between the largest and smallest effect size) was about twice as large compared to the CLPM, both for the self-esteem effect on depression (.19 for the RI-CLPM vs. .09 for the

CLPM) and the depression effect on self-esteem (.18 vs. .11). The range of the autoregressive effects was at about the same size for both models.

We used meta-analytic methods to examine the consistency of coefficients more systematically (Table 7). The consistency of the estimates across samples can be assessed on the basis of the heterogeneity indices τ and I^2, and the 95% prediction interval (Borenstein, Higgins, Hedges, & Rothstein, 2017). The findings showed that the cross-lagged effects were less consistent in the RI-CLPM than in the CLPM. For the CLPM, there was no evidence that the true effects differed across samples at all, as indicated by zero estimates for τ (i.e., the estimate of the standard deviation of true effects). Correspondingly, I^2 (i.e., the ratio of true heterogeneity by observed variability) was zero for the CLPM. In contrast, for the RI-CLPM, estimates of τ were at about .03 and of I^2 at about 40%. The difference between the CLPM and RI-CLPM was also reflected in the prediction interval (i.e., an estimate of where 95% of the true effects would fall). Whereas the prediction interval was very narrow for the CLPM (the difference between the lower and upper bound was .03 for both cross-lagged effects), it was about four times as large for the RI-CLPM (here, the lower and upper bounds differed by .14 and .12 for the two cross-lagged effects). For the autoregressive effects, the findings suggested that the estimates were more consistent in the RI-CLPM than in the CLPM, but there was significant heterogeneity for both models. We also meta-analyzed the structural coefficients of the CLPM and RI-CLPM for the other three versions of the models (i.e., when including constraints on residual variances and covariances). The findings on the consistency of estimates were very similar to the findings for the basic model, leading to the same conclusions (Supplemental Tables S13-S15).

It is possible that the inconsistency of coefficients across samples does not reflect pure imprecision, but rather that the more complex models (such as the RI-CLPM) are more sensitive

to true differences between samples than simpler models (such as the CLPM). If so, then a low level of inconsistency across samples (as for the CLPM) could indicate a methodological problem rather than a strength of a model. In exploratory analyses (not preregistered), we therefore compared inconsistency across samples with inconsistency within samples for the CLPM and RI-CLPM, using the vulnerability and scar effects from the basic model versions as an example. To gain information about inconsistency within samples, we computed bootstrapped bias-corrected 95% confidence intervals (CIs; based on 1,000 replications) and averaged the size of these CIs across samples. For the CLPM, the CIs had an average size of .121 (vulnerability effect) and .099 (scar effect); for the RI-CLPM, the CIs had an average size of .229 (vulnerability effect) and .197 (scar effect). Thus, as regards within-sample inconsistency, the CIs were 1.89 times (vulnerability effect) and 1.99 times (scar effect) larger for the RI-CLPM than for the CLPM. Corresponding information about inconsistency across samples is provided by the 95% CIs in the meta-analytic computations (Table 7). For the CLPM, the CIs had a size of .034 (vulnerability effect; rounded to 3 decimals) and .035 (scar effect); for the RI-CLPM, the CIs had a size of .069 (vulnerability effect) and .059 (scar effect). Thus, with regard to between-sample inconsistency, the CIs were 2.03 times (vulnerability effect) and 1.69 times (scar effect) larger for the RI-CLPM than for the CLPM. In other words, the degree to which between-sample inconsistency was larger for the RI-CLPM than for the CLPM was similar to the degree to which within-sample inconsistency was larger for the RI-CLPM than for the CLPM. Given that within-sample inconsistency is pure noise, the findings suggest that the larger inconsistency of estimates from the RI-CLPM across samples indicates lower precision of estimation compared to the CLPM. Thus, the findings do not support the hypothesis that the RI-CLPM reflects genuine differences between samples more accurately than the CLPM.

Discussion

We compared seven competing models that have been proposed for testing cross-lagged effects between constructs by examining model convergence, model fit, and consistency of estimates across 10 longitudinal samples. Specifically, we tested the CLPM (i.e., the traditional version of a cross-lagged model; e.g., Finkel, 1995), the RI-CLPM (Hamaker et al., 2015), the ALT (Bollen & Curran, 2004), the LCM-SR (Curran et al., 2014), the LCS (McArdle, 2001), the LCS-CC (Grimm et al., 2012), and the STARTS (Kenny & Zautra, 2001). For each model, we tested four versions with different cross-wave equality constraints on residual variances and covariances. The CLPM and the RI-CLPM converged in every sample, whereas the other five models frequently showed convergence issues. The RI-CLPM exhibited better model fit than the CLPM. For the CLPM, the cross-lagged effects were more consistent across samples than for the RI-CLPM.

Conceptual Discussion of the Models

Before we discuss the implications of the empirical findings, we first address conceptual considerations about the models. Given that in the present research the CLPM and RI-CLPM were the only models that converged consistently, we focus on these two models.

CLPM. An important limitation of the CLPM is that the model implies that the rank-order stability of constructs drops to zero in the long term (i.e., as the number of waves and study duration increases). However, we know from an increasing body of evidence that this assumption is wrong for most individual-difference constructs. For example, Fraley and Roberts (2005) provide convincing theoretical arguments and empirical findings that, although the longterm rank-order stability of personality constructs first declines rapidly as the time lag between assessments increases, it asymptotically approaches a nonzero value (e.g., standardized

values of .40 to .50 for the Big Five personality constructs). Similar findings have been published for the longterm rank-order stability of affective dispositions, self-esteem, and life satisfaction (Anusic & Schimmack, 2016; Kuster & Orth, 2013). Thus, the fact that the CLPM implies that the rank-order stability of the constructs eventually reaches zero is a limitation of this model. If the goal is to model the rank-order stability of constructs, then latent trait-state models are needed (Cole, 2012; Kenny & Zautra, 2001).

Moreover, the CLPM does not distinguish within-person and between-person variance, which is considered a limitation by some researchers (e.g., Berry & Willoughby, 2017; Hamaker et al., 2015). Clearly, when the goal is to examine associations of within-person variance, then models that separate the within-person and between-person component (such as the RI-CLPM) are the models of choice. However, we argue that precisely because the prospective effects tested in the CLPM are also based on between-person variance, it may answer questions that cannot be assessed with models that focus on within-person effects. For example, consider the possible effects of warm parenting on children's self-esteem (Krauss, Orth, & Robins, 2019): A cross-lagged effect in the CLPM would indicate that children raised by warm parents would be more likely to develop high self-esteem than children raised by less warm parents. A cross-lagged effect in the RI-CLPM would indicate that children who experience more parental warmth than usual at a particular time point will show a subsequent increase in self-esteem at the next time point, whereas children who experience less parental warmth than usual at a particular time point will show a subsequent drop in self-esteem at the next time point. Although some developmental theorizing is at the level of the individual (consequently, these effects should be reflected by models that focus on within-person effects), other developmental theorizing addresses the consequences of differences between persons. For example, we would expect children raised by

warm parents to develop higher self-esteem than children raised by less warm parents. From this perspective, the RI-CLPM can be problematic because if parental warmth shows very high stability over time, then the cross-lagged effect from the RI-CLPM cannot show an effect of parental warmth on self-esteem development, which does not make sense from a theoretical perspective.

Thus, some research questions focus on within-person effects, whereas others focus on between-person effects. For example, much of the research on risk and resilience factors theorizes that people high on a risk factor will show a worse developmental outcome than people low on the risk factor, and conversely for resilience factors; this is a between-person process that in our view is best tested by the CLPM. For these reasons, we believe that the fact that the CLPM does not distinguish within-person and between-person variance is not an unqualified limitation of the model.

As was evident also in the present analyses, the fit of the CLPM is typically not as good as the fit of the RI-CLPM (Hamaker et al., 2015; Masselink, Van Roekel, Hankin, et al., 2018). It is important to note that the CLPM is nested in the RI-CLPM (for further information about how the models examined in this research are nested, see Usami, Murayama, et al., 2019). That is, the CLPM is a special case of the RI-CLPM, where the variances of the two random intercept factors and the covariance between the random intercept factors are constrained to zero (thus, the CLPM has three additional degrees of freedom). Consequently, with increasing sample size, the RI-CLPM necessarily fits significantly better than the CLPM (MacCallum, Browne, & Cai, 2006). However, does this mean that the RI-CLPM should be preferred in model selection? Given that the two models differ in their conceptual meaning (see the discussion on between- and within-

person effects above), we believe that the decision between the CLPM and RI-CLPM should not be based on model fit, but rather on theoretical considerations.

From a conceptual perspective, it is important to note that the debate about the CLPM is also a debate about some of the most basic and frequently conducted analyses in psychology. For example, two-wave longitudinal data are often analyzed by multiple regression, using a predictor measured at Time 1 to explain the Time 2 outcome, after controlling for the outcome at Time 1; the exact same model can be conducted using structural equation modeling with latent variables. Alternatively, the Time 1 predictors are used to explain change scores in the outcome between Times 1 and 2 (or, the Time 2 outcome residualized for the Time 1 level in the outcome). MANOVA would still be another possibility. Essentially, these are all variations of the same analysis and they are conceptually similar to the CLPM. Thus, the recent spate of criticisms of CLPM apply to virtually all other methods to analyze repeated measures data. Thus, we believe that the debate about the CLPM has far-reaching implications for how longitudinal data should be analyzed.

RI-CLPM. Although the distinction of between-person and within-person variance in the RI-CLPM can be considered an advantage (Hamaker et al., 2015), 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. However, the random intercept factors provide information only about correlational associations between the constructs (similar to cross-sectional correlations, with the difference that the intercept factors are based on information from several waves), but not about time-lagged (i.e., longitudinal) between-person effects. Consequently, the RI-CLPM does not allow testing what many researchers—in fields such as personality, developmental, clinical, and

industrial–organizational psychology, as well as other fields in the social sciences—are interested in: the prospective between-person effect. Thus, the RI-CLPM gives up on the idea of trying to address questions about causal effects of between-person differences.

Another characteristic of the RI-CLPM is that the residualized scores are occasion-specific deviations from a person's trait level. Thus, if the model is correct and a large number of assessments are available, then the residualized scores will repeatedly return to the trait level by definition, even if individuals could experience similar deviations (e.g., in a positive or negative direction) at a few consecutive assessments. Consequently, in the RI-CLPM the cross-lagged effects capture temporary effects of one construct on the other, but the RI-CLPM cannot detect sustained prospective effects. In our opinion, this is an important limitation of the model. Although many effects between constructs do not persist forever and may be short-lived, the ideal cross-lagged model would be able to identify persistent effects, or at least long-term effects, if they exist. For the reason described above, we believe that the cross-lagged effects captured by the RI-CLPM are typically less persistent and more short-term compared to the cross-lagged effects captured by the CLPM. This also suggests that when using the RI-CLPM, shorter time lags between assessments (e.g., a few days, weeks, or months) are needed to be able to detect prospective effects between constructs. We note that the issue discussed in this paragraph is not resolved by the LCM-SR. Although the LCM-SR accounts for linear change (including individual differences in linear change) in constructs, again the residualized scores are temporary deviations from the trajectory an individual follows, so the model includes the assumption that individual scores repeatedly return to the general trajectory of the individual as the number of waves becomes large.

We noted above that the CLPM includes the unrealistic assumption that the rank-order stability of constructs approaches zero over the longterm. The RI-CLPM includes a similarly unrealistic assumption, specifically that that the between-person variance in constructs is perfectly stable. However, the rank-order stability of most individual difference constructs drops substantially below unity as the time interval increases (e.g., when examining change over longer than several years). Thus, by modeling between-person variance exclusively in the form of random intercept factors, the RI-CLPM includes an assumption that is unrealistic over longer periods. For this reason, the RI-CLPM does not allow for a perfect distinction between within-person and between-person variance. Consequently, some portion of the systematic between-person variance will be included in the residualized factors, which suggests that the cross-lagged effects in the RI-CLPM are not pure within-person effects but partially confounded with between-person variance.

It is important to note that when only two waves of data are available in a longitudinal study, the RI-CLPM is not identified (with two waves, the only cross-lagged model that can be used is the CLPM). This is a critical recognition given that two-wave longitudinal designs are currently much more common than designs with three or more waves (Usami, Todo, et al., 2019). Although this aspect cannot replace considerations about the theoretical meaning of cross-lagged effects, the question of how prospective effects should be tested on the basis of two waves is nevertheless an important one, given the ubiquity of two-wave longitudinal studies. Moreover, even if three or four waves of data are available, it seems difficult to disentangle people’s trait level from their level at a particular wave with high precision, since the average level is strongly dependent on each time point.

In sum, although the RI-CLPM resolves some of the issues of the CLPM (e.g., the CLPM shows typically worse fit than the RI-CLPM), it does so by producing other issues (i.e., the RI-CLPM cannot be used to test prospective between-person effects; prospective effects are by definition temporary).

Conclusions for the CLPM and RI-CLPM. Although we have discussed several conceptual issues of the CLPM and the RI-CLPM, we argue that—based on the current state of knowledge—both models should continue to be used in longitudinal research. If researchers are interested in understanding the prospective effects of between-person differences in a construct, then the CLPM should be selected. If researchers are interested in understanding the prospective effects of within-person deviations from the trait level in a construct, then the RI-CLPM should be selected. As suggested by the present findings, these two models are better suited than the other models to the typical design characteristics of longitudinal studies, which almost always have fewer than 4 or 5 waves of data. Moreover, most of the issues discussed for the RI-CLPM also apply to other models that focus on within-person effects, such as the LCM-SR. With regard to our recommendation that researchers use the RI-CLPM when focusing on within-person effects, we note the following qualification. In the rare research situations in which (a) such high theoretical precision exists that leads researchers to prefer a more complex model, such as the LCM-SR, over the RI-CLPM and (b) favorable design characteristics (e.g., a large number of waves) and power considerations suggest that the use of the more complex model is justified, researchers should use the more complex model in addition to the RI-CLPM.

Of course, it is important that researchers keep in mind what can, and cannot, be inferred from the results of each model. As described in Table 1, the cross-lagged effects generated by each model differ in their conceptual meaning, and should not be considered replications of each

other. Put differently, if the cross-lagged effects differ across models, this should not be considered a failed replication, but rather evidence that more nuanced processes are at work. Thus, Table 1 illustrates that the inferences that can be drawn from data depend directly on the model selected.

Finally, we would like to address a misconception when comparing CLPM and RI-CLPM cross-lagged effects. Researchers sometimes assume that it is easier to find large, or statistically significant, cross-lagged effects in the CLPM than in the RI-CLPM, given that the CLPM does not distinguish within-person and between-person variance. However, the cross-lagged effect in the CLPM should not be understood as an aggregate of the within- and between-person effects. Although the total variance in a construct across waves is the sum of between-person and within-person variance, the time-varying construct factors have different meaning in the CLPM and RI-CLPM. Thus, it is not an anomaly if the RI-CLPM cross-lagged effect is larger than the corresponding CLPM cross-lagged effect when the models are applied to the same data (for an example, see Study 3 in Masselink, Van Roekel, Hankin, et al., 2018).

Bivariate latent growth models. It may be worth briefly discussing another common approach for modeling longitudinal data. Researchers sometimes use bivariate latent growth models, also called parallel process models, with the goal of testing prospective effects between constructs (e.g., by correlating the intercept of one construct with the slope of another). Although bivariate latent growth models are certainly useful models for many research questions, they are not suitable for testing prospective effects between constructs for at least two reasons. First, there is no clear temporal order between intercept and slope, because the slope is defined by all waves of measurement (Zyphur et al., 2019). Therefore, the association between the intercept of one construct and the slope of the other construct cannot be meaningfully interpreted. Second, the

correlation between an intercept and slope depends on the location of the intercept. Often, Time 1 is used for locating the intercept, but Time 1 is typically an arbitrary starting point that cannot be interpreted as the starting point of a developmental process (Grimm, 2007). Importantly, however, the size, and even the sign, of the intercept–slope correlation may differ when locating the intercept at Time 1 versus another time point. Consequently, it is very difficult to interpret intercept–slope correlations in bivariate growth curve models, unless a strong argument can be made that the intercept is positioned at the beginning of a developmental process (e.g., transition into a new school, moving out of the parental home, etc.).

Implications of the Methodological Findings

As discussed above, ideally models should be selected for a priori theoretical reasons. However, practical considerations are also important, including the likelihood that problems will arise when using the model with actual data. Below, we discuss the implications of the present empirical findings on the choice of models.

Convergence of models. In the ten longitudinal samples examined in the present study, the only consistently converging models were the CLPM and RI-CLPM. In defense of the models that frequently showed problems converging, it is possible that some of these issues could be overcome by introducing constraints, such as fixing the variance, the mean, or both variance and mean of an intercept factor or linear slope to zero. For example, if there is not much mean-level change in a construct or little between-person variance in change, then this can lead to estimation issues in models including latent slope factors such as the ALT, LCM-SR, or LCS. Then, these issues could be resolved by adding constraints and simplifying the models. However, if this frequently happens, as in the present analyses, then this suggests that these models are too complex for the typical multi-wave panel design used in psychological research (see also Clark

et al., 2018; Voelkle, 2008). Thus, it might be a better strategy to select a simpler model from the beginning. In this research, it was precisely one of the research questions whether the models converge properly or frequently show convergence issues. Moreover, another important aspect to consider in this context is that when researchers use a model that frequently has convergence problems, then they are likely to tweak it in idiosyncratic ways to reach convergence, which makes the results of these models more difficult to compare across studies and risks capitalizing on chance patterns in the data. In other words, each model modification that is carried out in response to convergence issues reduces the confirmatory character of the analyses and raises concerns about the replicability of the findings.

Interestingly, in a recent simulation study testing the CLPM, RI-CLPM, and STARTS, the CLPM did not show any convergence issues (regardless of the hypothesized true model), whereas the RI-CLPM and STARTS were susceptible to improper solutions (Usami, Todo, et al., 2019). Although the RI-CLPM showed fewer convergence issues than the STARTS, improper solutions in the RI-CLPM emerged even when the RI-CLPM had been used to generate the data.

Model fit. The RI-CLPM exhibited better model fit than the CLPM. Moreover, if the other models (i.e., the ALT, LCM-SR, LCS, LCS-CC, and STARTS) converged properly, their fit was roughly as good as the fit of the RI-CLPM and consistently better than the fit of the CLPM. Thus, the results on fit values might suggest that the more complex models should be preferred in model selection. However, one important question in this context is whether “fit rules” or whether researchers should consider additional criteria such as theoretical meaning and replicability of the coefficients. Thus, does model fit have primacy over other criteria in model selection? It is important to emphasize that differences in model fit do not necessarily indicate which of two models is better from a substantive perspective. To provide an example, when the

goal is to model mean-level change in a construct, researchers should select a latent growth model, even if other types of models (such as latent trait-state models) would show better model fit. If other models would not help the researcher to gain empirical answers to the research question, then applying them is simply not useful, regardless of fit.

It is important to note that the simulation study by Usami, Todo, et al. (2019) did not yield clear-cut answers to the question of which model is better in terms of fit, using the Akaike information criterion and Bayesian information criterion. When the data were generated based on the CLPM, the CLPM showed better fit than the RI-CLPM and STARTS. When the data were generated based on the RI-CLPM, in some conditions the RI-CLPM fit best, whereas in other conditions the CLPM showed the best fit. And, when the STARTS was the hypothesized true model, the RI-CLPM typically showed the best fit.

Also, given that in the present samples some of the fit values of the CLPM did not meet the thresholds for good fit, it is important to note that the fit could be improved by using multiple indicators of the constructs and modeling the constructs as latent variables (for further information on latent-variable models, see below). In our experience, a large number of primary studies suggests that the fit of latent-variable versions of the CLPM is often acceptable according to current conventions. For example, a study that employed four of the present samples and measures of self-esteem and depression showed good fit for latent-variable versions of the CLPM, with CFI ranging from .98 to 1.00 and RMSEA ranging from .029 to .039 (Orth et al., 2016). In our opinion, the fact that the fit of the CLPM was below the commonly used thresholds in some of the present samples should not be taken as argument against the CLPM.

Consistency of estimates. For the CLPM, the structural coefficients were much more consistent compared to the RI-CLPM, both across and within samples, which is important with

regard to replicability of research findings (Asendorpf et al., 2013). Even if theoretical considerations have priority over replicability in model selection, replicability of estimates is—other things being equal—an important criterion of the quality of statistical models. The larger complexity of the RI-CLPM (i.e., resulting from inclusion of random intercept factors and residualizing the observed variables) likely is the reason for greater imprecision of estimates, resulting in a larger range of observed effect sizes across samples and a larger estimate of true heterogeneity, as found in the present research. Thus, compared to the CLPM, the present research suggests that when using the RI-CLPM, more data (in terms of number of waves and sample size) are needed for obtaining precise and replicable results.

The simulation study by Usami, Todo, et al. (2019) yielded similar conclusions: In the RI-CLPM, standard errors of estimates were 1.3 to 2.6 times larger than in the CLPM, depending on the specific conditions. In the STARTS, standard errors were 3.3 to 38.7 times larger than in the CLPM.

Constraints on residual variances and covariances. We also examined the consequences of using cross-wave equality constraints on residual variances and covariances for model convergence, model fit, and consistency of estimates. For each model, we compared a basic version (i.e., without constraints on residual variances and covariances) with versions that included constraints on residual variances, residual covariances, and both residual variances and covariances. Overall, using these constraints increased the rate of proper convergence. For the LCS and LCS-CC, the improvements were substantial (i.e., convergence improved from 30 to 90% and from 10 to 50%, respectively). However, for the ALT, LCM-SR, and STARTS, the improvements were only minor and negligible. For the CLPM and RI-CLPM, improvements in the rate of convergence was not an issue, because the rate of convergence was perfect for all

model versions. With regard to model fit, the findings suggested that there was a slight decrease in model fit by introducing the constraints, but overall the decrease was not substantial. Finally, the degree of consistency of estimates was very similar across model versions. Also, the weighted mean effect sizes replicated well across the four model versions.

These findings raise the question of whether constraints on residual variances and covariances should be used in cross-lagged models or not. We believe that using these constraints should not be a post-hoc decision, to maintain the confirmatory character of the analyses (Simmons et al., 2011). Thus, the field should find consensus about whether these constraints should be used. Such a consensus does not necessarily need to apply to all types of models in the same way. Using the constraints could be a standard for some models but not others. With regard to the CLPM and RI-CLPM, our recommendation is that the constraints should not be used, for the following reasons. First, the present results show that the constraints are not required for model convergence. Second, the results suggest that the constraints are not fully supported by the data, as indicated by reductions in model fit. Third, our experience and reading of the literature suggests that the constraints are not commonly used for the CLPM and RI-CLPM, so it might be simpler to continue with the tradition of not using constraints on residual variances and covariances.

Implications of the Substantive Findings

Besides providing insights into the statistical models, the present findings contribute to our understanding of the substantive issue that was used as a case study, that is, the relation between low self-esteem and depression. Although few question the empirical association between low self-esteem and depression, there has been a longstanding debate about the nature of this association (Beck, 1967; Blatt et al., 1976; Brown & Harris, 1978; Roberts & Monroe,

1994). A number of competing models have been proposed to explain why depression is associated with low self-esteem (see Orth & Robins, 2013). Importantly, these models differ with regard to the underlying causal process and the presumed causal direction of the link between the constructs. For example, whereas the vulnerability model states that low self-esteem is a causal risk factor for depression, the scar model assumes that causality runs in the opposite direction (i.e., experiences of depression lower the individual's self-esteem). Given that experimental designs that substantially alter people's levels of self-esteem and depression, and that are ecologically valid, are difficult for practical and ethical reasons, researchers have used observational longitudinal designs to study the relation between the constructs.

However, as reported in the Introduction, almost all prior longitudinal studies of self-esteem and depression were based on the CLPM (for an exception, see Masselink, Van Roekel, Hankin, et al., 2018). The present meta-analytic estimates based on the CLPM correspond closely to the results of the Sowislo and Orth (2013) meta-analysis, which were also based on the CLPM. Specifically, the vulnerability effect of low self-esteem on depression was −.13 in the present study and −.16 in Sowislo and Orth; the scar effect of depression on self-esteem was −.06 in the present study and −.08 in Sowislo and Orth. For the RI-CLPM, the present estimates differed from this pattern. The effect sizes were smaller (−.03 and −.04 for the vulnerability and scar effect, respectively), of about equal size, and the vulnerability effect was not significant (although marginally significant, as indicated by the upper bound of the confidence interval, which was .00). Nevertheless, both effects were in the same direction as in the CLPM.

It is important to remember that the cross-lagged effects have a different meaning in the CLPM versus RI-CLPM, so it is not surprising that the estimates differed across models (see Table 1). Whereas in the CLPM, the cross-lagged effect indicates a between-person effect, in the

RI-CLPM it is a within-person effect. In other words, in the CLPM the negative cross-lagged effect from self-esteem to depression means that individuals who have low self-esteem at Time 1 (i.e., relative to others) show a rank-order value in depression at Time 2 that is higher than would have been expected from their rank-order value in depression at Time 1 (i.e., the negative effect indicates that the rank order of individuals in depression increases as a function of the rank order in self-esteem). In contrast, in the RI-CLPM, the negative cross-lagged effect means that individuals who have lower self-esteem than they usually have (i.e., relative to their trait level) at Time 1 experience a subsequent increase in depression from Time 1 to Time 2. Thus, the CLPM findings suggest that individual differences in self-esteem predict changes in individual differences in depression, consistent with the vulnerability model. However, the RI-CLPM findings suggest that temporary fluctuations in self-esteem (around a person's trait level) have only small and nonsignificant prospective effects on fluctuations in depression (again, around a person's trait level). Nevertheless, we note that Masselink, Van Roekel, Hankin, et al. (2018) found stronger, and statistically significant, within-person vulnerability effects of low self-esteem on depression when using the RI-CLPM.

Limitations and Future Directions

An important limitation of the present research is that the results and, consequently, our conclusions may depend on the specific datasets used for the analyses. First, most of the studies used four-wave designs (only one study included five waves and one study 11 waves). Model convergence, fit, and consistency of estimates may be significantly altered as the number of waves increases. Nevertheless, we believe that the present set of longitudinal studies is useful because it is a good representation of the longitudinal studies typically designed by researchers and typically available. Possibly, the present set of studies is even positively biased (in terms of

number of waves), because the most common longitudinal study designs might still include only two or three waves.

Second, the present findings may depend on the specific measures used in the studies, as regards reliability or other characteristics of the measures. It should be noted, however, that the internal consistency estimates of the present measures were generally good, which improves convergence and fit of models. Moreover, prior research suggests that many of the measures included in the present studies typically show measurement invariance across time, age, and gender; some of these tests have even been conducted for the samples used in the present research (e.g., Orth et al., 2016; Orth et al., 2008). Thus, there is evidence that the present measures have favorable measurement properties, which reduces concerns that the convergence issues in some of the models could have resulted from measurement issues.

Third, the present findings could depend on the particular research question (Is low self-esteem prospectively associated with depression?) used as a case study. Some of the statistical models may be inappropriate to the substantive issue under investigation. For example, the fact that some models estimate a growth factor for the constructs would be problematic if no growth exists in self-esteem and depression in the present samples. Consequently, it would be useful to replicate the current analyses using another substantive issue, to gain broader and more robust insights into the behavior of the models in real datasets. Nevertheless, we believe that using real datasets instead of simulated data is an important feature of the present research, because it provides information about the behavior of models when using data that cannot be influenced by assumptions about the data generating process.

Future research on the behavior of cross-lagged models should also examine the issue of the timing of waves. Convergence, fit, and consistency of estimates might depend on whether the

time lag between waves is appropriate for the substantive research question (Dormann & Griffin, 2015; Gollob & Reichardt, 1987). Also, the inconsistency of estimates across studies could be partially explained by differences in the time lag in the present set of studies. In fact, statistical theory suggests that the size of cross-lagged effects depends on the time lag between assessments (Gollob & Reichardt, 1987). Unfortunately, although the present research used data from 10 samples, the number of studies does not provide sufficient power to systematically test time lag as a moderator in the meta-analytic computations. However, as reported in the Introduction, even a meta-analysis with 77 longitudinal samples did not find evidence for a significant moderator effect of time lag on the prospective effect of low self-esteem on depression (Sowislo & Orth, 2013). Moreover, whereas in the meta-analysis by Sowislo and Orth (2013) time lag varied strongly across studies (ranging from 1 week to 13 years), in the present research variability of time lag was much smaller (ranging from 2 months to 2 years, with most studies having a 1- or 2-year lag). For these reasons, we assume that in the present set of studies variability in time lag did not critically confound conclusions about the consistency of estimates. Nevertheless, future research should examine this issue in more detail. For example, continuous time modeling of panel data would allow examining this issue more systematically (Voelkle, Oud, Davidov, & Schmidt, 2012). Finally, the timing of waves may also vary between participants within studies. Although in longitudinal panel analyses, it is a common simplifying assumption that the time lags were identical for all participants, if there is significant variability in time lag between participants, this may contribute to imprecision in estimates of prospective effects.

Practical Recommendations

Based on the conceptual considerations and empirical findings of this research, we propose the following guidelines for testing prospective effects between constructs based on longitudinal data.

Recommendation 1: Researchers should select the model in an *a priori*, theory-driven way, depending on the specific research question(s) and the hypothesized psychological or developmental process. Ideally, researchers should preregister the selected model, providing as much detail as possible about how the model will be specified. For example, researchers should register which model(s) they will test, which measures will be used, whether constructs will be measured by multiple indicators (see Recommendation 6), and whether cross-wave equality constraints will be imposed on residual variances and residual covariances (see Recommendation 4) and/or structural coefficients (see Recommendation 5).

Recommendation 2: Of the seven models tested, we recommend using either the CLPM or the RI-CLPM, depending on whether the research question concerns between- or within-person effects. When the research question concerns prospective effects of between-person differences, the CLPM should be selected. When the research question concerns prospective effects of within-person deviations from trait levels, the RI-CLPM should be selected. In contexts where both research questions are of theoretical interest, we recommend that researchers fit both the CLPM and RI-CLPM to the same data, to examine between- and within-person effects, to compare the direction and magnitude of these effects, and to gain a richer understanding of the nature of the longitudinal association between the constructs (for examples, see Krauss et al., 2019; Masselink, Van Roekel, Hankin, et al., 2018). Given that in the present research most of the within-person models did not perform well enough (using datasets that resemble the longitudinal data available to most researchers), we believe that the

RI-CLPM (which did perform well) should be used when testing within-person effects. We recommend that the more complex within-person models only be used, in addition to the RI-CLPM, when (a) theory provides clear guidance in selecting a particular within-person model and when (b) the sample size and number of waves is substantially larger than in the present set of studies, given the risk of convergence issues. If the model selected does not converge, then researchers need to modify the model or select a different model, both of which reduce the confirmatory status of the research and the replicability of the findings. In sum, we recommend that researchers use the CLPM for testing prospective between-person effects and the RI-CLPM for testing prospective within-person effects. Using the same standard models would have the added benefit of facilitating comparison of results across studies.

Recommendation 3: Researchers should appropriately describe the conceptual meaning of the coefficients tested in whichever model is selected. Table 1 provides an overview of the conceptual meaning of the cross-lagged effects in different models. It is important to emphasize that the inferences that can be drawn from the cross-lagged effects depend directly on the model selected. Also, researchers should be explicit about the assumptions of the model used, emphasizing that the results are contingent upon the assumptions made by the model.

Recommendation 4: For the CLPM and RI-CLPM, researchers should not routinely use cross-wave equality constraints on residual variances and residual covariances. As discussed above, the present results show that residual variance/covariance constraints are not required for model convergence. Moreover, our reading of the literature suggests that these constraints are not commonly used for the CLPM and RI-CLPM, so we suggest continuing the practice of not using constraints on residual variances and covariances.

Recommendation 5: Researchers should use cross-wave equality constraints on structural coefficients (e.g., autoregressive and cross-lagged effects), unless there is a clear rationale against doing so. When the intervals between waves have the same (or approximately the same) length, typically there is no reason to expect systematic differences in the structural coefficients across intervals (Cole & Maxwell, 2003; Little et al., 2007). Exceptions are, for example, study designs that span different developmental periods (e.g., a study with yearly assessments from age 6 to 18 years) and study designs that include experimental treatments or significant transitions during some but not all intervals (e.g., a four-wave study in which all participants became retired during the first interval). In the absence of these reasons, using cross-wave equality constraints on structural coefficients has three important advantages. First, the information is aggregated across waves, which increases the precision of estimates (as indicated by smaller confidence intervals) and, consequently, the power of significance tests. Second, using these constraints generally improves model convergence and thereby helps in the estimation of more complex models (e.g., when controlling for covariates) and when missing data is a relevant issue. Third, using these constraints reduces the complexity of the results, because there is only one estimate per effect instead of one for each interval. This reduces the risk that researchers search for explanations of between-interval differences, when in truth these differences are simply caused by chance. Thus, having only one estimate per effect simplifies the reporting of the results and improves the clarity of interpretation.[7]

Nevertheless, often it is advisable to test whether cross-wave equality constraints are empirically supported, by using model comparison and testing whether the constraints significantly reduce model fit or not. If there is no theoretical reason why the effects should differ across waves but the constraints are not empirically supported, then the constraints should

not be used; however, in this situation the research report should note that the decision was not made a priori but driven by the data.

Recommendation 6: Researchers should use multiple indicators, if possible, to measure the constructs as latent variables. In the present research, all constructs were examined as observed variables (i.e., using scale scores). This decision was based on the fact that most of the models have been introduced as observed-variable models in the literature and consequently most researchers use these models with observed (not latent) variables. That is, the goal of the present research was to test the models as they are typically used in the field. Nevertheless, using latent-variable versions of the models would have important advantages (and, moreover, it would be interesting to replicate the present analyses with multiple-indicator models). Measuring the constructs as latent variables with multiple indicators allows controlling for measurement error, which increases the validity of the estimates (Cole & Preacher, 2014). Moreover, using multiple indicators of the constructs allows testing for measurement invariance, an assumption that cannot be tested on the basis of scale scores (Little et al., 2007; Schmitt & Kuljanin, 2008; Widaman, Ferrer, & Conger, 2010).

General information on how to use multiple indicators to measure the constructs as latent variables is available in the literature on structural equation modeling (e.g., Hoyle, 2012; Kline, 2016; Little, 2013). Hamaker (2018) provides a description of how to construct a multiple-indicator RI-CLPM. Information on multiple-indicator versions of the STARTS model can be found in Cole (2012) and Donnellan et al. (2012). In the OSF project folder to this article, we provide sample Mplus scripts and output for multiple-indicator versions of the CLPM and RI-CLPM (https://osf.io/5rjsm). The OSF folder also includes sample Mplus scripts for measurement models (for the case of two constructs), which can be used for testing configural,

weak, and strong measurement invariance across waves. Moreover, the folder also includes a model with strong measurement invariance that uses effects-coding, which is a non-arbitrary method of scaling the latent construct factors (Little, Slegers, & Card, 2006). This measurement model is then used as the basis for the CLPM and RI-CLPM. Specifically, when using effects-coding, the latent variables are measured with the same metric as the observed indicators (e.g., on a scale ranging from 1 to 4). Having a non-arbitrary metric for latent variables is helpful when researchers want to interpret means and intercepts of the latent variables, for example when using cross-lagged models that explicitly model growth, such as the LCM-SR or LCS.

Recommendation 7: When collecting longitudinal data, a large number of repeated assessments is desirable. Research suggests that model convergence is generally improved as the number of waves increases (e.g., Clark et al., 2018). Thus, researchers should not only seek to improve power by collecting large samples, but also by conducting a large number of assessments. Relatively few waves might be generally sufficient to test CLPMs (e.g., three or four) and RI-CLPMs (e.g., four or five), but precise recommendations will require future research – both simulations and analyses of real data – to determine the number of waves needed for each model to function effectively.

Conclusion

The present research suggests that two cross-lagged models—the CLPM and RI-CLPM—converge reliably when using data from longitudinal studies with designs that are common in the field of psychology, including a moderate number of repeated assessments (e.g., 3–5) and relatively large samples (e.g., 300–700). Depending on whether the research questions concern between-person or within-person effects, researchers should select in a priori, theory-based way the CLPM (when focusing on prospective effects of between-person differences), the RI-CLPM

(when focusing on prospective effects of within-person deviations from trait levels), or both. The present findings also suggest that it may be problematic to select other cross-lagged models when using datasets similar to those used in the present research, because there is a risk that they will not converge properly or will not converge at all.

References

Abramson, L. Y., Seligman, M. E. P., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology, 87*, 49-74. http://dx.doi.org/10.1037/0021-843X.87.1.49

Anusic, I., & Schimmack, U. (2016). Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology, 110*, 766-781. http://dx.doi.org/10.1037/pspp0000066

Asendorpf, J., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119. http://dx.doi.org/10.1002/per.1919

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. New York, NY: Harper and Row.

Berry, D., & Willoughby, M. T. (2017). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development, 88*, 1186-1206. http://dx.doi.org/10.1111/cdev.12660

Biesanz, J. C. (2012). Autoregressive longitudinal models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 459-471). New York, NY: Guilford.

Blatt, S. J., D'Afflitti, J. P., & Quinlan, D. M. (1976). Experiences of depression in normal young adults. *Journal of Abnormal Psychology, 85*, 383-389. http://dx.doi.org/10.1037/0021-843X.85.4.383

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models: A synthesis of two traditions. *Sociological Methods and Research, 32*, 336-383. http://dx.doi.org/10.1177/0049124103260222

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*, 5-18. http://dx.doi.org/10.1002/jrsm.1230

Brown, G. W., & Harris, T. (1978). *Social origins of depression: A study of psychiatric disorder*. New York, NY: Free Press.

Capaldi, D. M., & Rothbart, M. K. (1992). Development and validation of an early adolescent temperament measure. *Journal of Early Adolescence, 12*, 153-173. http://dx.doi.org/10.1177/0272431692012002002

Casillas, A., & Clark, L. A. (2000). *The Mini Mood and Anxiety Symptom Questionnaire (Mini-MASQ)*. Poster presented at the 72nd Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Clark, D. A., Nuttall, A. K., & Bowles, R. P. (2018). Misspecification in latent change score models: Consequences for parameter estimation, model evaluation, and predicting change. *Multivariate Behavioral Research, 53*, 172-189. http://dx.doi.org/10.1080/00273171.2017.1409612

Cole, D. A. (2012). Latent trait-state models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 585-600). New York, NY: Guilford.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558-577. http://dx.doi.org/10.1037/0021-843X.112.4.558

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*, 300-315. http://dx.doi.org/10.1037/a0033805

Conger, R. D., Lasley, P., Lorenz, F. O., Simons, R., Whitbeck, L. B., Elder, G. H., & Norem, R. (2011). *Iowa Youth and Families Project, 1989-1992 [Data file and codebook]*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research. http://dx.doi.org/10.3886/ICPSR26721.v2.

Coyne, J. C., Gallo, S. M., Klinkman, M. S., & Calarco, M. M. (1998). Effects of recent and past major depression and distress on self-concept and coping. *Journal of Abnormal Psychology, 107*, 86-96. http://dx.doi.org/10.1037/0021-843X.107.1.86

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 107-135). Washington, DC: American Psychological Association.

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*, 879-894. http://dx.doi.org/10.1037/a0035297

Derogatis, L. R., & Savitz, K. L. (1999). The SCL-90-R, Brief Symptom Inventory, and matching clinical rating scales. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (pp. 679-724). Mahwah, NJ: Erlbaum.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177-188. http://dx.doi.org/10.1016/0197-2456(86)90046-2

Donnellan, M. B., Kenny, D. A., Trzesniewski, K. H., Lucas, R. E., & Conger, R. D. (2012). Using trait-state models to evaluate the longitudinal consistency of global self-esteem from adolescence to adulthood. *Journal of Research in Personality, 46*, 634-645. http://dx.doi.org/10.1016/j.jrp.2012.07.005

Donnellan, M. B., Trzesniewski, K. H., & Robins, R. W. (2015). Measures of self-esteem. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 131-157). London, UK: Elsevier.

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods, 20*, 489-505. http://dx.doi.org/10.1037/met0000041

Ellis, L. K., & Rothbart, M. K. (2001). *Revision of the Early Adolescent Temperament Questionnaire*. Poster presented at the biennial meeting of the Society for Research in Child Development, Minneapolis, MN.

Finkel, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112*, 60-74. http://dx.doi.org/10.1037/0033-295X.112.1.60

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development, 58*, 80-92. http://dx.doi.org/10.2307/1130293

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

Grimm, K. J., An, Y., McArdle, J. J., Zonderman, A. B., & Resnick, S. M. (2012). Recent changes leading to subsequent changes: Extensions of multivariate latent difference score models. *Structural Equation Modeling, 19,* 268-292. http://dx.doi.org/10.1080/10705511.2012.659627

Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent growth curve modeling. *Structural Equation Modeling, 17*, 424-442. http://dx.doi.org/10.1080/10705511.2010.489006

Hamaker, E. L. (2018). *How to run a multiple indicator RI-CLPM with Mplus*. Retrieved from http://www.statmodel.com/download/RI-CLPM.pdf

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*, 102-116. http://dx.doi.org/10.1037/a0038889

Hankin, B. L., Lakdawalla, Z., Carter, I. L., Abela, J. R. Z., & Adams, P. (2007). Are neuroticism, cognitive vulnerabilities and self-esteem overlapping or distinct risks for depression? Evidence from exploratory and confirmatory factor analyses. *Journal of Social and Clinical Psychology, 26*, 29-63. http://dx.doi.org/10.1521/jscp.2007.26.1.29

Hoyle, R. H. (Ed.) (2012). *Handbook of structural equation modeling*. New York, NY: Guilford.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. http://dx.doi.org/10.1080/10705519909540118

Jongerling, J., & Hamaker, E. L. (2011). On the trajectories of the predetermined ALT model: What are we really modeling? *Structural Equation Modeling, 18,* 370-382. http://dx.doi.org/10.1080/10705511.2011.582004

Keller, A. C., Meier, L. L., Elfering, A., & Semmer, N. K. (2019). Please wait until I am done! Longitudinal effects of work interruptions on employee well-being. *Work and Stress*, Advance online publication. http://dx.doi.org/10.1080/02678373.2019.1579266

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

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: American Psychological Association.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Krauss, S., Orth, U., & Robins, R. W. (2019). Family environment and self-esteem development: A longitudinal study from age 10 to 16. *Journal of Personality and Social Psychology*, Advance online publication. http://dx.doi.org/10.1037/pspp0000263

Kuster, F., & Orth, U. (2013). The long-term stability of self-esteem: Its time-dependent decay and nonzero asymptote. *Personality and Social Psychology Bulletin, 39*, 677-690. http://dx.doi.org/10.1177/0146167213480189

Kuster, F., Orth, U., & Meier, L. L. (2012). Rumination mediates the prospective effect of low self-esteem on depression: A five-wave longitudinal study. *Personality and Social Psychology Bulletin*, *38*, 747-759. http://dx.doi.org/10.1177/0146167212437250

Kuster, F., Orth, U., & Meier, L. L. (2013). High self-esteem prospectively predicts better work conditions and outcomes. *Social Psychological and Personality Science, 4*, 668-675. http://dx.doi.org/10.1177/1948550613479806

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*, 357-365. http://dx.doi.org/10.1177/0165025407077757

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13*, 59-72. http://dx.doi.org/10.1207/s15328007sem1301_3

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201-226. http://dx.doi.org/10.1146/annurev.psych.51.1.201

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods, 11*, 19-35. http://dx.doi.org/10.1037/1082-989X.11.1.19

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analysis. *Psychological Assessment, 17*, 81-102. http://dx.doi.org/10.1037/1040-3590.17.1.181

Masselink, M., Van Roekel, E., Hankin, B. L., Keijsers, L., Lodder, G. M. A., Vanhalst, J., . . . Oldehinkel, A. J. (2018). The longitudinal association between self-esteem and depressive symptoms in adolescents: Separating between-person effects from within-

person effects. *European Journal of Personality, 32,* 653-671. http://dx.doi.org/10.1002/per.2179

Masselink, M., Van Roekel, E., & Oldehinkel, A. J. (2018). Self-esteem in early adolescence as predictor of depressive symptoms in late adolescence and early adulthood: The mediating role of motivational and social factors. *Journal of Youth and Adolescence, 47,* 932-946. http://dx.doi.org/10.1007/s10964-017-0727-z

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic analysis. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 341-380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60,* 577-605. http://dx.doi.org/10.1146/annurev.psych.60.110707.163612

Meier, L. L., & Cho, E. (2019). Work stressors and partner social undermining: Comparing negative affect and psychological detachment as mechanisms. *Journal of Occupational Health Psychology, 24,* 359-372. http://dx.doi.org/10.1037/ocp0000120

Meier, L. L., & Spector, P. E. (2013). Reciprocal effects of work stressors and counterproductive work behavior: A five-wave longitudinal study. *Journal of Applied Psychology, 98,* 529-539. http://dx.doi.org/10.1037/a0031732

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide: Eighth edition*. Los Angeles, CA: Muthén and Muthén.

Orth, U., & Luciano, E. C. (2015). Self-esteem, narcissism, and stressful life events: Testing for selection and socialization. *Journal of Personality and Social Psychology, 109,* 707-721. http://dx.doi.org/10.1037/pspp0000049

Orth, U., & Robins, R. W. (2013). Understanding the link between low self-esteem and depression. *Current Directions in Psychological Science, 22*, 455-460. http://dx.doi.org/10.1177/0963721413492763

Orth, U., Robins, R. W., & Meier, L. L. (2009). Disentangling the effects of low self-esteem and stressful events on depression: Findings from three longitudinal studies. *Journal of Personality and Social Psychology, 97*, 307-321. http://dx.doi.org/10.1037/a0015645

Orth, U., Robins, R. W., Meier, L. L., & Conger, R. D. (2016). Refining the vulnerability model of low self-esteem and depression: Disentangling the effects of genuine self-esteem and narcissism. *Journal of Personality and Social Psychology, 110*, 133-149. http://dx.doi.org/10.1037/pspp0000038

Orth, U., Robins, R. W., & Roberts, B. W. (2008). Low self-esteem prospectively predicts depression in adolescence and young adulthood. *Journal of Personality and Social Psychology, 95*, 695-708. http://dx.doi.org/10.1037/0022-3514.95.3.695

Orth, U., Robins, R. W., Trzesniewski, K. H., Maes, J., & Schmitt, M. (2009). Low self-esteem is a risk factor for depressive symptoms from young adulthood to old age. *Journal of Abnormal Psychology, 118*, 472-478. http://dx.doi.org/10.1037/a0015922

Orth, U., Robins, R. W., Widaman, K. F., & Conger, R. D. (2014). Is low self-esteem a risk factor for depression? Findings from a longitudinal study of Mexican-origin youth. *Developmental Psychology, 50*, 622-633. http://dx.doi.org/10.1037/a0033817

Ou, L., Chow, S. M., Ji, L., & Molenaar, P. C. M. (2017). (Re)evaluating the implications of the autoregressive latent trajectory model through likelihood ratio tests of its initial conditions. *Multivariate Behavioral Research, 52*, 178-199. http://dx.doi.org/10.1080/00273171.2016.1259980

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401. http://dx.doi.org/10.1177/014662167700100306

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 295-315). New York, NY: Russell Sage Foundation.

Rieger, S., Göllner, R., Trautwein, U., & Roberts, B. W. (2016). Low self-esteem prospectively predicts depression in the transition to young adulthood: A replication of Orth, Robins, and Roberts (2008). *Journal of Personality and Social Psychology, 110*, e16-e22. http://dx.doi.org/10.1037/pspp0000037

Rindskopf, D. (1984). Using phantom and imaginary latent variables to parameterize constraints in linear structural models. *Psychometrika, 49,* 37-47. http://dx.doi.org/10.1007/BF02294204

Roberts, J. E., & Monroe, S. M. (1994). A multidimensional model of self-esteem in depression. *Clinical Psychology Review, 14*, 161-181. http://dx.doi.org/10.1016/0272-7358(94)90006-X

Robins, R. W., & Conger, K. J. (2017). *California Families Project [Sacramento and Woodland, California]: Item-level (producer) codebook.* Ann Arbor, MI: Inter-University Consortium for Political and Social Research. http://dx.doi.org/10.3886/ICPSR35476.v1.

Robins, R. W., Donnellan, M. B., Widaman, K. F., & Conger, R. D. (2010). Evaluating the link between self-esteem and temperament in Mexican origin early adolescents. *Journal of Adolescence, 33,* 403-410. http://dx.doi.org/10.1016/j.adolescence.2009.07.009

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27,* 151-161. http://dx.doi.org/10.1177/0146167201272002

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177. http://dx.doi.org/10.1037/1082-989X.7.2.147

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18,* 210-222. http://dx.doi.org/10.1016/j.hrmr.2008.03.003

Shahar, G., & Davidson, L. (2003). Depressive symptoms erode self-esteem in severe mental illness: A three-wave, cross-lagged study. *Journal of Consulting and Clinical Psychology, 71,* 890-900. http://dx.doi.org/10.1037/0022-006X.71.5.890

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366. http://dx.doi.org/10.1177/0956797611417632

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin, 139,* 213-240. http://dx.doi.org/10.1037/a0028931

Sowislo, J. F., Orth, U., & Meier, L. L. (2014). What constitutes vulnerable self-esteem? Comparing the prospective effects of low, unstable, and contingent self-esteem on depressive symptoms. *Journal of Abnormal Psychology, 123*, 737-753. http://dx.doi.org/10.1037/a0037770

Steiger, A. E., Allemand, M., Robins, R. W., & Fend, H. A. (2014). Low and decreasing self-esteem during adolescence predict adult depression two decades later. *Journal of Personality and Social Psychology, 106*, 325-338. http://dx.doi.org/10.1037/a0035133

Usami, S., Hayes, T., & McArdle, J. J. (2016). Inferring longitudinal relationships between variables: Model selection between the latent change score and autoregressive cross-lagged factor models. *Structural Equation Modeling, 23*, 331-342. http://dx.doi.org/10.1080/10705511.2015.1066680

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods, 24*, 637-657. http://dx.doi.org/10.1037/met0000210

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

van Tuijl, L. A., de Jong, P. J., Sportel, B. E., de Hullu, E., & Nauta, M. H. (2014). Implicit and explicit self-esteem and their reciprocal relationship with symptoms of depression and social anxiety: A longitudinal study in adolescents. *Journal of Behavior Therapy and Experimental Psychiatry, 45*, 113-121. http://dx.doi.org/10.1016/j.jbtep.2013.09.007

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48. http://dx.doi.org/10.18637/jss.v036.i03

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

Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods, 17,* 176-192. http://dx.doi.org/10.1037/a0027543

Watson, D., Suls, J., & Haig, J. (2002). Global self-esteem in relation to structural models of personality and affectivity. *Journal of Personality and Social Psychology, 83,* 185-197. http://dx.doi.org/10.1037/0022-3514.83.1.185

Widaman, K. F. (2006). Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development, 71,* 42-64. http://dx.doi.org/10.1111/j.1540-5834.2006.00404.x

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4,* 10-18. http://dx.doi.org/10.1111/j.1750-8606.2009.00110.x

Wouters, S., Duriez, B., Luyckx, K., Klimstra, T., Colpin, H., Soenens, B., & Verschueren, K. (2013). Depressive symptoms in university freshmen: Longitudinal relations with contingent self-esteem and level of self-esteem. *Journal of Research in Personality, 47,* 356-363. http://dx.doi.org/10.1016/j.jrp.2013.03.001

Wu, W., Selig, J. P., & Little, T. D. (2013). Longitudinal data analysis. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Statistical analysis* (Vol. 2, pp. 387-410). New York, NY: Oxford University Press.

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., . . . Diener, E. (2019). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods*, Advance online publication. http://dx.doi.org/10.1177/1094428119847280

Footnotes

[1] We do not have the right to make the data from the other samples available on OSF. However, these samples are archival data and the OSF folder to this article provides information on how the data can be accessed (https://osf.io/5rjsm).

[2] The models differ with regard to the minimum number of waves required for estimation. The CLPM is the only model that can be estimated with two waves only. The RI-CLPM requires three waves of data, and the ALT, LCM-SR, LCS, LCS-CC, and STARTS require four waves.

[3] The samples examined in the present research were used in the following previous studies on the relation between self-esteem and depression. Data from the BLS were used in Orth et al. (2008), Orth, Robins, and Meier (2009), and Orth et al. (2016). Data from the CFP, children sample, were used in Robins, Donnellan, Widaman, and Conger (2010), Orth et al. (2014), and Orth et al. (2016). Data from the study MWI were used in Kuster et al. (2012) and Orth et al. (2016). Data from the NLSY79, Young Adults Section, were used in Orth et al. (2008) and Orth, Robins, and Meier (2009). Data from the study YP were used in Orth et al. (2016). The samples have been used in numerous studies addressing other research questions. However, previous studies using these data did not examine questions related to comparing longitudinal models for testing prospective effects (i.e., the central research questions of the present article). For the BLS, a publication list is available at https://osf.io/abq3r. For the CFP, a publication list is available at https://www.californiafamiliesproject.org/publications.html. For the FTP, a publication list is available at https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/26721. In addition to the studies cited above, the MWI has been used in Keller, Meier, Elfering, and Semmer (2019), Kuster, Orth, and Meier (2013), Meier and Cho (2019), and Meier and Spector

(2013). For the NLSY79, a publication list is available at https://nlsinfo.org/bibliography-start. In addition to the studies cited above, the YP has been used in Orth and Luciano (2015).

[4] For the BLS, MWI, and YP, the materials are available in the OSF folder to this article (https://osf.io/5rjsm). For the CFP, FTP, and NLSY79, the materials are available at the following URLs: https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/35476 (CFP), https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/26721 (FTP), and https://www.nlsinfo.org (NLSY79).

[5] For the statistical analyses, we note the following deviations from the preregistered research plan. Convergence of models: We do not only report the rate of improper solutions, but also the rates of proper convergence and nonconvergence. Model fit: We did not compute standard deviations of fit values and frequency of fit values that meet cutoffs. However, complete fit information (for models that converged properly) is available in tables. Consistency of estimates: Instead of computing means and standard deviations of coefficients, and frequencies of coefficients with theory-consistent sign and significant coefficients, we used meta-analytic methods to summarize the findings because these methods have important statistical advantages and provide additional information on heterogeneity of estimates.

[6] Fit was particularly low for the CLPM in the NLSY79, as indicated by CFI, which likely is a consequence of the large number of waves in this dataset. As discussed in the Introduction, the CLPM underestimates the long-term stability of constructs, and this issue leads to misfit as the number of waves increases. In contrast, the RMSEA showed a relatively good value (.040) for the CLPM in this sample, so it is possible that the RMSEA is not sensitive to the issue of underestimating the longterm stability. However, we note that, like the CFI, the RMSEA was considerably worse for the CLPM compared to the other models.

[7] When using cross-wave equality constraints on structural parameters, the constraints are typically imposed on unstandardized coefficients, which leads to slight variation in the resulting standardized coefficients and requires averaging of standardized coefficients. Alternatively, the construct variances can be reparameterized to be on a common, standardized metric by using phantom variables (Rindskopf, 1984). By doing so, the standardized coefficients become estimated parameters in the model, which allows putting equality constraints on the standardized coefficients (instead of on the unstandardized coefficients). Little (2013) provides a detailed example and rationale for why rescaling construct variances can be helpful when using equality constraints across waves.

Table 1

Meaning of the Cross-Lagged Coefficients in the Models

Model	Conceptual meaning of cross-lagged coefficient	Hypothesized causal effect for low self-esteem and depression
CLPM	Prospective effect of individual differences in Construct X on change in individual differences in Construct Y.	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.
RI-CLPM	Prospective effect of temporary deviation from the trait level in Construct X on change in the temporary deviation from the trait level in Construct Y.	When individuals have lower self-esteem than usual, they will experience a subsequent increase in depression.
ALT	Prospective effect of individual differences in Construct X on change in individual differences in Construct Y, controlling for the accumulating intercept and slope factor in Construct Y.[a]	When individuals have low self-esteem (relative to others), they will experience a subsequent increase in their rank-order value in depression, controlling for the accumulating intercept and slope factor of depression.
LCM-SR	Prospective effect of temporary deviation from the individual trajectory in Construct X on change in the temporary deviation from the developmental trajectory in Construct Y.	When individuals have lower self-esteem than would be expected from the self-esteem trajectory they follow, they will experience a subsequent increase in depression.
LCS	Prospective effect of latent score in Construct X on change in latent score in Construct Y.[b]	When individuals have low self-esteem, they will experience a subsequent increase in depression.
LCS-CC	Prospective effect of change in latent score in Construct X on change in latent score in Construct Y.[b]	When individuals have decreased in self-esteem, they will experience a subsequent increase in depression.
STARTS	Prospective effect of autoregressive trait factor in Construct X on change in the autoregressive trait factor in Construct Y.[c]	When individuals have lower self-esteem than usual, they will experience a subsequent increase in depression.[d]

Note. CLPM = cross-lagged panel model; RI-CLPM = random intercepts cross-lagged panel model; ALT = autoregressive latent trajectory model; LCM-SR = latent curve model with structured residuals; LCS = latent change score model; LCS-CC = latent change score model with changes-to-changes extension; STARTS = trait-state-error model.

[a] In the ALT, the intercept and slope factors are so-called accumulating factors (Usami, Murayama, et al., 2019), which means that the exact combination of intercept factors, slope factors, Time 1 means, autoregressive effects, and cross-lagged effects is needed to describe the average and individual trajectory for each of the Constructs X and Y. For this reason, the intercept and slope factors of the ALT cannot be interpreted as simple growth factors as in growth curve models. In contrast, in the LCM-SR the growth curve part and the cross-lagged part can be interpreted in isolation (i.e., without considering the other part) because the cross-lagged part is modeled on the residualized scores, not the observed scores as in the ALT.

[b] In the LCS and LCS-CC, the latent construct scores (e.g., lx1 to lx4 in Panels E and F of Figure 1) are predicted, model-implied individual scores. The unexplained variance in the observed construct variables (e.g., X1 to X4) is captured by the residuals (e.g., e1 to e4).

[c] In the STARTS, the autoregressive trait factors (e.g., xr1 to xr4) can be understood as state factors that are interlinked by first-order autoregressive effects. As described in the section "Description of the Models Tested in the Present Research," the STARTS includes complex constraints on the variances and covariances of the autoregressive trait factors to impose stationarity (for the exact specification of these constraints, see the Mplus sample scripts in the Supplemental Material).

[d] Note that the description of the hypothesized causal effect is identical for the RI-CLPM and the STARTS; although the models differ in their exact specification as described above, the interpretation of the causal effect does not differ across the two models.

Table 2

Descriptive Information on Samples

Study, sample	*N*[a]	Developmental period	Number of waves	Time interval	Measure of self-esteem (# items, alpha)	Measure of depression (# items, alpha)
BLS	404	Young adulthood	4	1 year	RSE (10 items, .90)	CES-D (20 items, .91)
CFP, children	674	Adolescence	4	2 years	SDQ (25 items, .88)	EATQ (6 items, .66)
CFP, mothers	636	Adulthood	4	2 years	RSE (10 items, .80)	MASQ (13 items, .91)
FTP, children	451	Adolescence	4	1 year	RSE (10 items, .87)	SCL-90 (12 items, .88)
FTP, siblings	451	Adolescence	4	1 year	RSE (10 items, .87)	SCL-90 (12 items, .90)
FTP, mothers	451	Adulthood	4	1 year	RSE (10 items, .89)	SCL-90 (12 items, .90)
FTP, fathers	451	Adulthood	4	1 year	RSE (10 items, .87)	SCL-90 (12 items, .90)
MWI	663	Adulthood	5	2 months	RSE (10 items, .89)	CES-D (20 items, .89)
NLSY79	8,259	Adolescence, young adulthood	11	2 years	RSE (10 items, .87)	CES-D (7 items, .70)
YP	326	Young adulthood	4	6 months	RSE (10 items, .91)	CES-D (20 items, .91)

Note. Number of waves is the number of equally spaced waves available for analysis. For coefficient alpha, the table reports average values across waves. BLS = Berkeley Longitudinal Study; CFP = California Families Project; FTP = Family Transitions Project; MWI = My Work and I; NLSY79 = National Longitudinal Survey of Youth 1979, Young Adults Section; YP = Your Personality; RSE = Rosenberg Self-Esteem Scale; SDQ = Self-Description Questionnaire; CES-D = Center for Epidemiologic Studies Depression Scale; EATQ = Early Adolescent Temperament Questionnaire; MASQ = Mood and Anxiety Symptom Questionnaire; SCL-90 = Symptom Checklist 90.

[a] Sample size used to compute the study weights for the meta-analytic computations, reflecting the number of participants who provided data on at least one of the study variables at one of the waves.

Table 3

Frequency (in Percent) of Proper Convergence, Improper Solution, and Nonconvergence of Models Across the 10 Samples

Model version, convergence	CLPM	RI-CLPM	ALT	LCM-SR	LCS	LCS-CC	STARTS
Basic model							
Proper convergence	100	100	10	30	30	10	10
Improper solution	0	0	80	70	30	40	70
Nonconvergence	0	0	10	0	40	50	20
Constraints on residuals							
Proper convergence	100	100	10	40	70	50	20
Improper solution	0	0	90	60	30	20	80
Nonconvergence	0	0	0	0	0	30	0
Constraints on covariances							
Proper convergence	100	100	30	40	40	10	10
Improper solution	0	0	20	60	30	50	70
Nonconvergence	0	0	50	0	30	40	20
Constraints on residuals and covariances							
Proper convergence	100	100	30	40	90	50	10
Improper solution	0	0	70	60	10	20	90
Nonconvergence	0	0	0	0	0	30	0

Note. The models with constraints included cross-wave equality constraints on residuals and/or covariances. CLPM = cross-lagged panel model; RI-CLPM = random intercepts cross-lagged panel model; ALT = autoregressive latent trajectory model; LCM-SR = latent curve model with structured residuals; LCS = latent change score model; LCS-CC = latent change score model with changes-to-changes extension; STARTS = trait-state-error model.

Table 4

Fit Values of the Models (Basic Version of Models)

Fit indicator, sample	CLPM	RI-CLPM	ALT	LCM-SR	LCS	LCS-CC	STARTS
CFI							
BLS	.950	.998					
CFP, children	.938	.985			.973		
CFP, mothers	.916	.994					
FTP, children	.917	.980		.979			
FTP, siblings	.937	.987		.988			
FTP, mothers	.915	.990			.994		
FTP, fathers	.901	1.000					
MWI	.932	.987					.996
NLSY79	.781	.980	.983	.983	.971	.972	
YP	.894	.976					
RMSEA							
BLS	.075	.015					
CFP, children	.076	.041			.061		
CFP, mothers	.104	.030					
FTP, children	.113	.063		.070			
FTP, siblings	.096	.050		.053			
FTP, mothers	.136	.053			.045		
FTP, fathers	.133	.006					
MWI	.105	.050					.027
NLSY79	.040	.012	.011	.011	.014	.014	
YP	.134	.075					

Note. Empty cells indicate that model did not converge properly or did not converge at all. CLPM = cross-lagged panel model; RI-CLPM = random intercepts cross-lagged panel model; ALT = autoregressive latent trajectory model; LCM-SR = latent curve model with structured residuals; LCS = latent change score model; LCS-CC = latent change score model with changes-to-changes extension; STARTS = trait-state-error model; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; BLS = Berkeley Longitudinal Study; CFP = California Families Project; FTP = Family Transitions Project; MWI = My Work and I; NLSY79 = National Longitudinal Survey of Youth 1979, Young Adults Section; YP = Your Personality.

Table 5

Mean Fit Values of the CLPM and RI-CLPM Across the 10 Samples

	CLPM		RI-CLPM	
Model version	CFI	RMSEA	CFI	RMSEA
Basic model	.908	.101	.988	.040
Constraints on residuals	.903	.095	.982	.045
Constraints on covariances	.907	.097	.986	.043
Constraints on residuals and covariances	.901	.093	.981	.046

Note. The models with constraints included cross-wave equality constraints on residuals and/or covariances. CLPM = cross-lagged panel model; RI-CLPM = random intercepts cross-lagged panel model; CFI = comparative fit index; RMSEA = root-mean-square error of approximation.

Table 6

Structural Coefficients of the CLPM and RI-CLPM (Basic Version of Models)

Sample	CLPM				RI-CLPM			
	SE→D	D→SE	SE→SE	D→D	SE→D	D→SE	SE→SE	D→D
BLS	−.18*	−.04	.75*	.37*	−.06	−.04	.47*	.04
CFP, children	−.10*	−.03	.53*	.42*	−.09*	−.07	.26*	.26*
CFP, mothers	−.13*	−.14*	.51*	.46*	−.07	−.06	.16*	.13*
FTP, children	−.11*	−.07*	.63*	.47*	−.02	−.03	.29*	.12*
FTP, siblings	−.14*	−.09*	.58*	.42*	−.08	−.13*	.22*	.19*
FTP, mothers	−.14*	−.04	.75*	.55*	.00	.05	.19*	.21*
FTP, fathers	−.10*	−.08*	.64*	.53*	.06	−.02	.13*	.13*
MWI	−.18*	−.08*	.79*	.60*	.00	−.11*	.19*	.26*
NLSY79	−.12*	−.06*	.61*	.38*	−.05*	−.02	.23*	.12*
YP	−.19*	−.08*	.75*	.40*	.10	−.02	.24*	.04

Note. The table shows standardized coefficients. CLPM = cross-lagged panel model; RI-CLPM = random intercepts cross-lagged panel model; SE = self-esteem; D = depression; BLS = Berkeley Longitudinal Study; CFP = California Families Project; FTP = Family Transitions Project; MWI = My Work and I; NLSY79 = National Longitudinal Survey of Youth 1979, Young Adults Section; YP = Your Personality.

* $p < .05$.

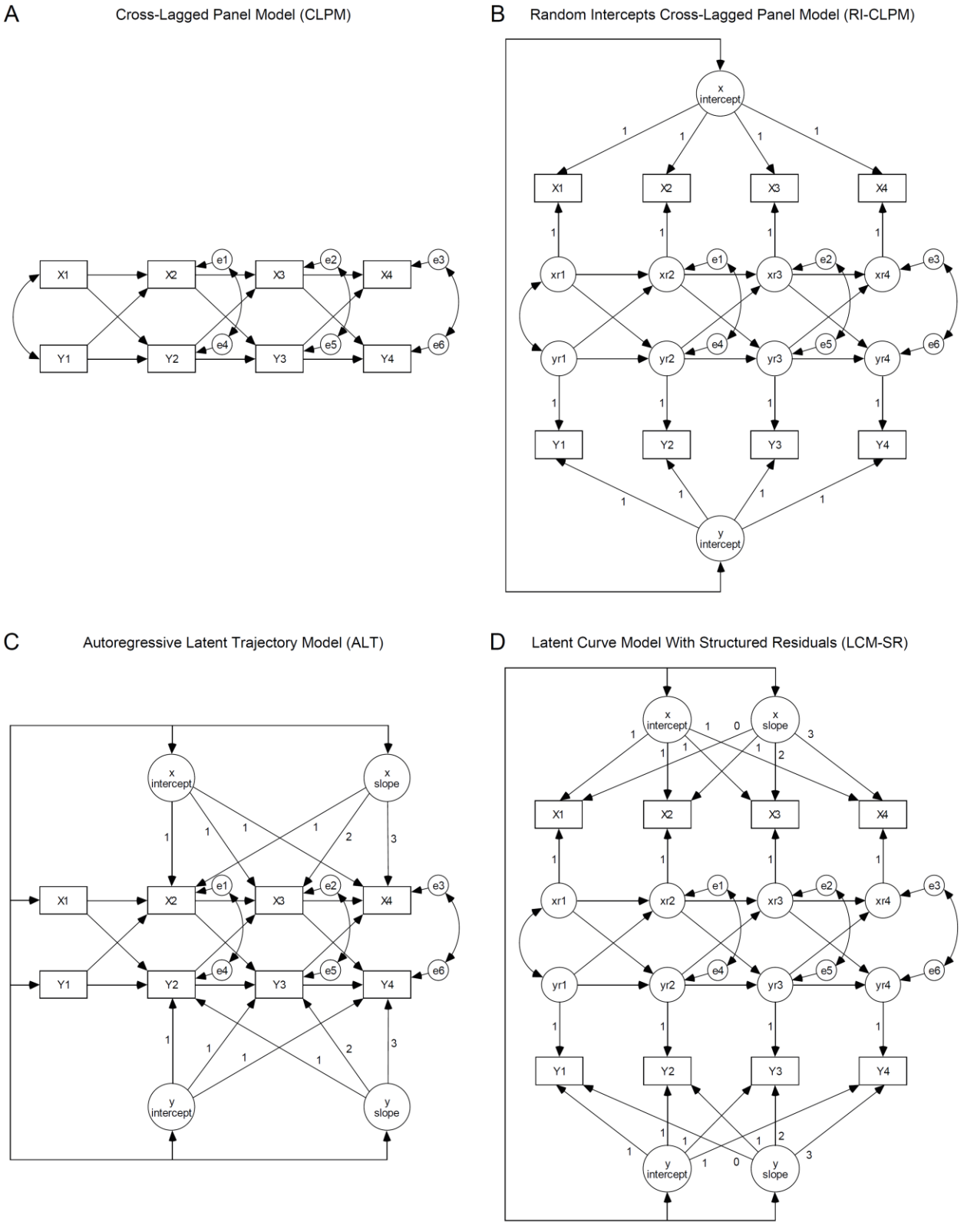
Table 7

Meta-Analytic Aggregation of Structural Coefficients for the CLPM and RI-CLPM Across the 10 Samples (Basic Version of Models)

Model, coefficient	Weighted mean effect size	95% CI	Heterogeneity			
			Q	τ	95% PI	I^2
CLPM						
SE→D	−.13*	[−.14, −.11]	6.0	.000	[−.14, −.11]	0.0
D→SE	−.06*	[−.08, −.05]	5.8	.000	[−.08, −.05]	0.0
SE→SE	.66*	[.60, .72]	181.9*	.155	[.45, .81]	95.1
D→D	.46*	[.41, .52]	85.2*	.103	[.28, .61]	89.4
RI-CLPM						
SE→D	−.03	[−.06, .00]	16.8	.033	[−.10, .04]	46.3
D→SE	−.04*	[−.07, −.01]	14.2	.027	[−.10, .02]	36.8
SE→SE	.24*	[.19, .29]	43.2*	.069	[.10, .37]	79.2
D→D	.15*	[.11, .20]	35.3*	.061	[.03, .27]	74.5

Note. Computations were made with random-effects models. For all computations, the number of studies was $k = 10$ and the total number of participants was $N = 12{,}766$. CLPM = cross-lagged panel model; RI-CLPM = random intercepts cross-lagged panel model; Effect size = weighted mean standardized coefficient; CI = confidence interval (indicates the accuracy of the weighted mean effect size); *Q* = statistic used in heterogeneity test; τ = estimate of the standard deviation of true effects; PI = prediction interval (estimates where 95% of the true effects would fall); I^2 = ratio of true heterogeneity by observed variability (given in percent); SE = self-esteem; D = depression.

* $p < .05$.



[Figure continued on next page.]

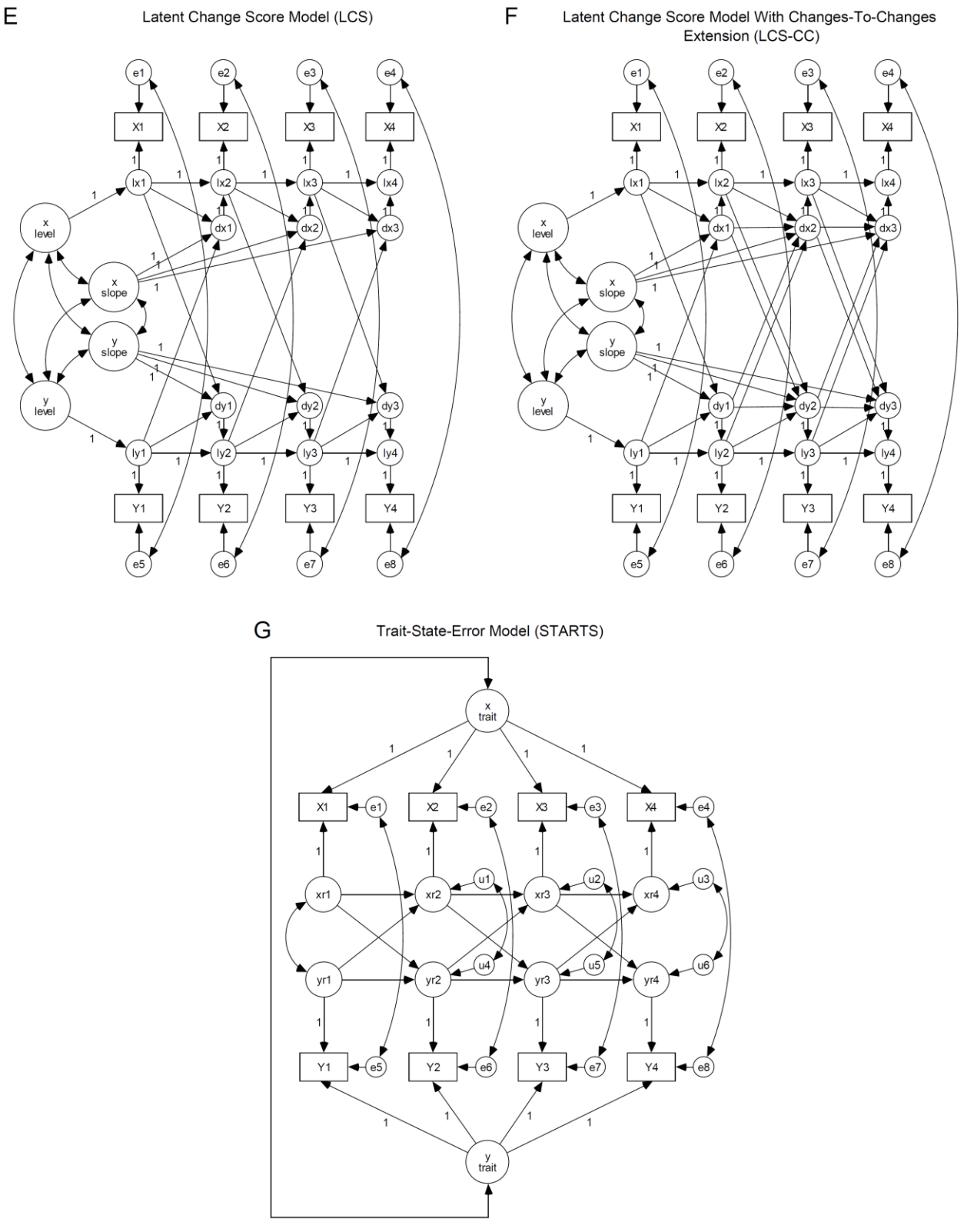


Figure 1. The figure illustrates the models tested in this research, for the case of four waves of data, including the cross-lagged panel model (CLPM, Panel A), random intercepts cross-lagged

panel model (RI-CLPM, Panel B), autoregressive latent trajectory model (ALT, Panel C), latent curve model with structured residuals (LCM-SR, Panel D), latent change score model (LCS, Panel E), latent change score model with changes-to-changes extension (LCS-CC, Panel F), and trait-state-error model (STARTS, Panel G). See the sample Mplus scripts in the Supplemental Material for additional details (e.g., constraints included in the models).