Title

Effect Size Guidelines for Cross-Lagged Effects

Journal

Psychological Methods, 29(2)

ISSN

1082-989X

Authors

Orth, Ulrich

Meier, Laurenz L

Bühler, Janina Larissa

et al.

Publication Date

2024-04-01

DOI

10.1037/met0000499

Peer reviewed

Psychological Methods

Effect Size Guidelines for Cross-Lagged Effects

Ulrich Orth, Laurenz L. Meier, Janina Larissa Bühler, Laura C. Dapp, Samantha Krauss, Denise Messerli, and Richard W. Robins

Online First Publication, June 23, 2022. http://dx.doi.org/10.1037/met0000499

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000499

Effect Size Guidelines for Cross-Lagged Effects

Ulrich Orth[1], Laurenz L. Meier[2], Janina Larissa Bühler[3], Laura C. Dapp[1], Samantha Krauss[1], Denise Messerli[2], and Richard W. Robins[4]

[1] Department of Psychology, University of Bern
[2] Department of Psychology, University of Neuchâtel
[3] Department of Psychology, Heidelberg University
[4] Department of Psychology, University of California, Davis

Abstract

Cross-lagged models are by far the most commonly used method to test the prospective effect of one construct on another, yet there are no guidelines for interpreting the size of cross-lagged effects. This research aims to establish empirical benchmarks for cross-lagged effects, focusing on the cross-lagged panel model (CLPM) and the random intercept cross-lagged panel model (RI-CLPM). We drew a quasirepresentative sample of studies published in four subfields of psychology (i.e., developmental, social–personality, clinical, and industrial–organizational). The dataset included 1,028 effect sizes for the CLPM and 302 effect sizes for the RI-CLPM, based on data from 174 samples. For the CLPM, the 25th, 50th, and 75th percentiles of the distribution corresponded to cross-lagged effect sizes of .03, .07, and .12, respectively. For the RI-CLPM, the corresponding values were .02, .05, and .11. Effect sizes did not differ significantly between the CLPM and RI-CLPM. Moreover, effect sizes did not differ significantly across subfields and were not moderated by design characteristics. However, effect sizes were moderated by the concurrent correlation between the constructs and the stability of the predictor. Based on the findings, we propose to use .03 (small effect), .07 (medium effect), and .12 (large effect) as benchmark values when interpreting the size of cross-lagged effects, for both the CLPM and RI-CLPM. In addition to aiding in the interpretation of results, the present findings will help researchers plan studies by providing information needed to conduct power analyses and estimate minimally required sample sizes.

Translational Abstract

Researchers in psychology and related disciplines often use longitudinal data to examine the effect of a construct measured at one point in time on another construct measured at a later time point. This article provides guidelines for interpreting the size of these prospective effects. We focused on two frequently used models: the cross-lagged panel model (CLPM) and the random intercept cross-lagged panel model (RI-CLPM). We examined the range of effect sizes reported for these models in a quasirepresentative sample of published articles drawn from four subfields of psychology (developmental, social–personality, clinical, and industrial–organizational). Average effect sizes were similar for the CLPM and RI-CLPM and did not differ significantly across subfields. Based on the findings, we recommend that researchers use .03 (small effect), .07 (medium effect), and .12 (large effect) as benchmark values when interpreting the size of cross-lagged effects for both the CLPM and RI-CLPM.

Keywords: cross-lagged panel model, random intercept cross-lagged panel model, effect size, empirical benchmarks, longitudinal research

Supplemental materials: https://doi.org/10.1037/met0000499.supp

Ulrich Orth https://orcid.org/0000-0002-4795-515X
Laurenz L. Meier https://orcid.org/0000-0002-5675-1562
Janina Larissa Bühler https://orcid.org/0000-0003-3684-9682
Laura C. Dapp https://orcid.org/0000-0003-3985-6976
Samantha Krauss https://orcid.org/0000-0003-0124-347X
Denise Messerli https://orcid.org/0000-0001-6042-5738
Richard W. Robins https://orcid.org/0000-0002-5088-3484
Janina Larissa Bühler is now at the Department of Psychology, Johannes Gutenberg University Mainz.

Data, materials, and code are available on the Open Science Framework (OSF; https://osf.io/4cwbr/). Richard W. Robins was supported by a grant from the National Institute on Aging (R01AG060164).

The data are available at https://osf.io/4cwbr/

The materials are available at https://osf.io/4cwbr/

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

In many fields of psychology and related disciplines, researchers are interested in testing the effect of one construct on another. When experimental designs are not feasible for practical or ethical reasons, researchers often analyze longitudinal data using cross-lagged models to gain insights into prospective effects between the constructs (Biesanz, 2012; McArdle, 2009; Wu et al., 2013). The cross-lagged panel model (CLPM; e.g., Finkel, 1995) and the random intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015) are arguably the two most frequently used cross-lagged models. The key coefficients from these models are the cross-lagged effects, which capture the degree to which one construct predicts the other construct over time, controlling for prior levels of the outcome and concurrent associations between the constructs. Typically, a cross-lagged effect is reported in the form of a standardized regression coefficient (also called beta coefficient or beta weight), which provides information about how many standard deviations the outcome will change when the predictor changes by one standard deviation (Cohen et al., 2003).

The Need for Effect Size Guidelines

An important issue in research with cross-lagged models is that no effect size conventions are available for interpreting cross-lagged effects. Sometimes, the effects are evaluated using effect size conventions suggested for correlation coefficients, such as .10, .30, and .50 indicating small, medium, and large effects (Cohen, 1988, 1992). However, there are important reasons why conventions established for correlations should not be applied to cross-lagged effects (Adachi & Willoughby, 2015).

First, cross-lagged effects are based on prospective associations using longitudinal data, whereas effect size conventions for correlations typically refer to concurrent associations estimated using cross-sectional data. In most cases, cross-lagged effects are estimated over long periods (e.g., months or years). Given that virtually all psychological constructs change over time to some degree, theory predicts that longitudinal associations are systematically smaller than concurrent correlations.

Second, not only are cross-lagged effects based on longitudinal data, they are also controlled for the prior level (i.e., the autoregressive effect) of the predicted variable. The autoregressive effect already accounts for a large portion of variance in the outcome (i.e., the variance that is stable over the time interval), which limits the theoretically-possible range of cross-lagged effects from other constructs. Thus, cross-lagged effects can explain only variance of the outcome that has changed between the assessments.

Third, cross-lagged effects are also controlled for the concurrent correlation between predictor and outcome at Time 1. More precisely, cross-lagged models take the bivariate correlation between the predictor at Time 1 and the outcome at Time 2, and partition it into two components (Kenny, 1979): (a) the path consisting of the concurrent correlation between the constructs at Time 1 and the autoregressive effect of the outcome between Times 1 and 2, and (b) the direct path from the predictor at Time 1 on the outcome at Time 2. Given that the bivariate correlation equals the sum of the two paths, there is no reason to expect that one path (i.e., the cross-lagged effect) would have the same magnitude as the bivariate correlation (in fact, the other path often accounts for a large portion of the bivariate correlation). To state this more conceptually, the correlation between two variables separated in time might be due to a prospective effect of one variable on the other or an alternative pathway—that is, that the two variables are concurrently associated *and* stable over time so what looks like a prospective effect is really a concurrent correlation in disguise. Cross-lagged models provide a way to examine the prospective effect while controlling for the alternative pathway.

Taken together, statistical theory predicts that cross-lagged effects should be substantially smaller than bivariate correlations between the constructs. However, many researchers and readers of research articles might not be aware that effect size conventions for correlations should not be used when evaluating cross-lagged effects. If cross-lagged effects are typically much smaller than correlations, then a cross-lagged effect of, for example, .10 might seem small, but it could indicate a much more meaningful effect than suggested by effect size conventions for correlations.

Therefore, the present research aims to establish empirical benchmarks for small, medium, and large cross-lagged effects that can guide the interpretation of findings from cross-lagged models. As we will explain in more detail below, we drew a quasirepresentative sample of studies in several broad subfields of psychology to examine the distribution of cross-lagged effects in the literature. Empirical knowledge about the distribution of cross-lagged effect sizes and the factors that moderate the size of cross-lagged effects will help researchers plan studies by providing information needed to conduct a priori power analyses and estimate minimally required sample sizes. Empirical approaches have been used to derive benchmarks for other effect size measures, including correlation coefficients (Bosco et al., 2015; Brydges, 2019; Gignac & Szodorai, 2016; Lovakov & Agadullina, 2021; Paterson et al., 2016) and Cohen's *d* (Kinney et al., 2020; Lovakov & Agadullina, 2021). For example, research on the empirical distribution of correlation coefficients suggests that .10, .20, and .30 (Brydges, 2019; Gignac & Szodorai, 2016; Paterson et al., 2016) or .10, .25, and .40 (Lovakov & Agadullina, 2021) are appropriate benchmarks for small, medium, and large correlations, indicating that the values proposed by Cohen (1988, 1992) are too large.

It may be useful to briefly review what can be anticipated about the typical size of cross-lagged effects. Although no research on the empirical distribution of cross-lagged effects is available, meta-analytic reviews have estimated mean values for specific effects using the CLPM (we are not aware of meta-analytic reviews using the RI-CLPM). For example, the effect of self-beliefs on academic achievement has been estimated as .08 (Valentine et al., 2004), the effect of low self-esteem on depression as .16 and on anxiety as .10, and the reversed effects both as .08 (Sowislo & Orth, 2013), the effect of low positive emotionality on depression as .08 and on anxiety as .06, and the reversed effects as .07 and .09 (Khazanov & Ruscio, 2016), the effect of work–family conflict on strain as .03–.08 and the reversed effect as .05–.08 (Nohe et al., 2015), the effect of low attachment security on substance use as .05 (Fairbairn et al., 2018), the effect of social relationships on self-esteem as .08 and the reversed effect also as .08 (Harris & Orth, 2020), and the effect of work experiences on self-esteem as .02–.05 and the reversed effect as .05–.10 (Krauss & Orth, 2021). These meta-analytic reviews are probably not representative of research conducted in psychology, because meta-analyses are commonly done when there is already a large literature documenting a particular effect, which likely introduces selectivity of stronger effects. For this reason, we believe that the approach taken in the present research (i.e., drawing a quasirepresentative sample of individual

studies) is a better strategy for establishing effect size benchmarks compared with sampling meta-analytic reviews.

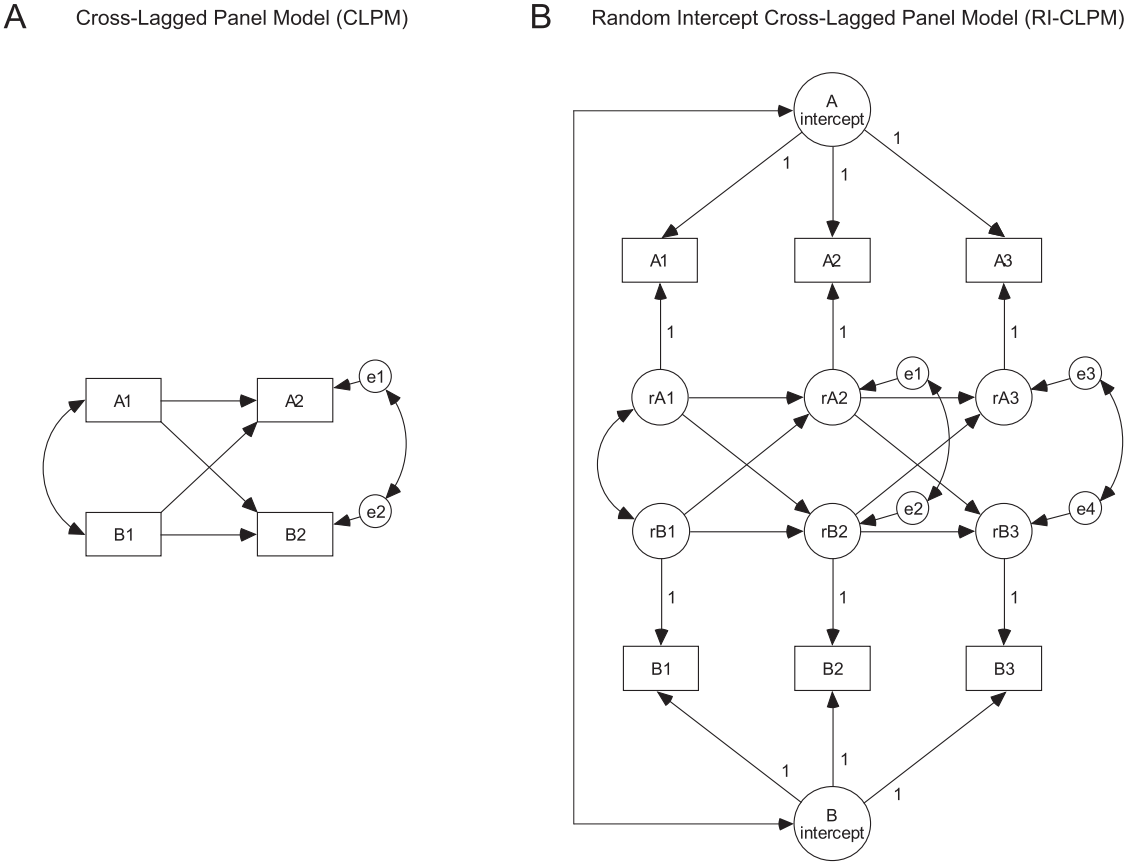
Description of the CLPM and RI-CLPM

The need for effect size conventions applies to all types of cross-lagged models used in the field (for overviews, see Mund & Nestler, 2019; Orth et al., 2021; Usami et al., 2019). However, the present study focused on the CLPM and the RI-CLPM for three reasons. First, both models are frequently used, allowing us to generate reliable estimates of the distribution of cross-lagged effects. The CLPM is by far the most frequently used cross-lagged model in psychology (Orth et al., 2021; Usami et al., 2019), and the RI-CLPM, although recently introduced in the literature (Hamaker et al., 2015), has seen a surge in use over the past few years (Mulder & Hamaker, 2021; Usami, 2021). Second, in typical applications in the field, estimating the CLPM or RI-CLPM rarely leads to convergence problems (e.g., improper solutions or nonconvergence), whereas other models often show convergence problems (Orth et al., 2021; Usami et al., 2019). Third, the models can be considered representative for testing two types of cross-lagged effects: between-person (CLPM) and within-person (RI-CLPM; Lüdtke & Robitzsch, 2021; Orth et al., 2021).

Figure 1A shows a generic illustration of the CLPM, which requires at least two waves of data. The CLPM tests for the cross-lagged effect of individual differences in one construct (e.g., Construct A at Time 1) on individual differences in the other construct (e.g., Construct B at Time 2), controlling for prior individual differences in the outcome (e.g., Construct B at Time 1). Thus, given that the cross-lagged effects are controlled for autoregressive effects of the constructs, the CLPM tests for *change* in individual differences. Consider the example of warm parenting and children's self-esteem given in Orth et al. (2021): In the CLPM, a cross-lagged effect of warm parenting on self-esteem would indicate that children raised by warm parents are more likely to develop high self-esteem than children raised by less warm parents.

Figure 1B shows a generic illustration of the RI-CLPM, which requires at least three waves of data. The RI-CLPM is similar to the CLPM, but includes random intercept factors (conceptually corresponding to trait factors) that capture the stable between-person variance in the constructs across assessments. The cross-lagged effects are then modeled between the residualized scores

Figure 1
Generic Illustrations of the CLPM and the RI-CLPM



Note. The models differ with regard to the minimum number of waves required for estimation. The CLPM (Panel A) requires two waves of data and the RI-CLPM (Panel B) requires three waves. In the RI-CLPM, the wave-specific factors (i.e., residualized scores) of Constructs A and B are denoted as rA1–rA3 and rB1–rB3. CLPM = cross-lagged panel model; RI-CLPM = random intercept cross-lagged panel model.

(in Figure 1B denoted as "rA1," "rA2," etc.), after removing the stable between-person variance associated with each construct. Consequently, the RI-CLPM tests for the cross-lagged effect of the within-person deviation from the trait level of one construct (e.g., "rA1") on the within-person deviation from the trait level of the other construct (e.g., "rB2"), controlling for the prior within-person deviation from the trait level of the predicted construct (e.g., "rB1"). Again, given that the effects are controlled for autoregressive effects in the deviations, the RI-CLPM tests for *change* in the within-person deviation from the trait level. Using the same example as above, an RI-CLPM cross-lagged effect of warm parenting on children's self-esteem 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 (Orth et al., 2021). Thus, although the CLPM and RI-CLPM have some similarities, they test conceptually distinct effects, which could result in divergent effects across models, as well as divergent average effect sizes.

Moderators of Cross-Lagged Effects

In this project, we also examined moderators of cross-lagged effects. First, we examined whether the size of cross-lagged effects differed between subfields of psychology. Then, we tested whether cross-lagged effects varied as a function of: (a) design characteristics (i.e., control of covariates, latent vs. observed measurement of constructs, presence of shared method variance, and time lag between assessments); and (b) other coefficients estimated in the model (i.e., Time 1 concurrent correlation between the constructs, stability over time of predictor, stability over time of outcome, and, in the RI-CLPM, the correlation between random intercepts of the constructs).

Statistical theory suggests that design characteristics could influence the size of cross-lagged effects. For example, controlling for covariates in the model could decrease the cross-lagged effect of a predictor, if the effects of the covariates are confounded with the effect of the predictor (Rohrer, 2018). Latent measurement can increase the validity of the construct factors by separating measurement error from construct variance (Cole & Preacher, 2014). Consequently, if manifest variables (rather than latent variables) are used as construct factors, then the lack of control for measurement error and other biases could lead to artificially attenuated cross-lagged effects. In contrast, the presence of shared method variance could lead to artificially inflated cross-lagged effects, even if shared method variance is already accounted for to a great extent by controlling for the concurrent correlation between prior levels of the outcome and the predictor (Podsakoff et al., 2012). Also, the time lag between assessments may influence the size of cross-lagged effects. Research on continuous time modeling suggests that with increasing time lag, the cross-lagged effect first increases, reaches a maximum, and then decreases over long periods (Dormann & Griffin, 2015; Voelkle et al., 2012). In the analyses, we therefore tested for linear and curvilinear effects of time lag. However, the time lag at which a cross-lagged effect reaches its maximum may vary substantially across constructs, ranging from very short (minutes, hours, or days) to very long periods (years or even decades), depending on the presumed causal process through which one variable influences another (for an empirical example of a cross-lagged effect that increased over long periods, see de Moor et al., 2021).

The size of cross-lagged effects could also be moderated by other coefficients in the model. For example, if the outcome shows high stability over time (i.e., a large autoregressive effect), then the maximum size of cross-lagged effects from other constructs is limited. In contrast, if the predictor shows high stability, then this could facilitate stronger cross-lagged effects. Also, if the predictor and the outcome are strongly correlated (as indicated by, e.g., the concurrent correlation at Time 1 or, in the RI-CLPM, the correlation between the random intercepts), then this could predict stronger cross-lagged effects between the constructs. However, evidence that other coefficients in the model moderate the size of the cross-lagged effects does not imply any kind of causal effect, for two reasons. First, theory allows one to derive opposite causal hypotheses. For example, a large concurrent correlation between the constructs could be a consequence, rather than a cause, of a large cross-lagged effect of the predictor on the outcome. Similarly, although high stability of an outcome might prevent other constructs from showing large cross-lagged effects on the outcome (i.e., stability of the outcome influences the cross-lagged effect), one could also argue that the stability of an outcome is reduced by large cross-lagged effects of another construct on the outcome (i.e., the cross-lagged effect influences the stability of the outcome). Second, given that the other coefficients are estimated in the same model as the cross-lagged effect, there is no temporal ordering of the coefficients and the moderator results can only indicate if other coefficients from the model covary with the cross-lagged effect. Nevertheless, we believe that testing for the moderator effects of other model coefficients can provide useful information for understanding heterogeneity in the size of cross-lagged effects across studies. In the moderator analyses, we therefore used a hierarchical strategy. In the first step, we tested only for the effects of design characteristics, and in the second step, we added the effects of the other model coefficients.

The Present Research

The goal of this research was to establish empirical benchmarks for cross-lagged effects in the CLPM and RI-CLPM. First, we examined the distribution percentiles. Then, we estimated mean effect sizes for the CLPM and RI-CLPM, using meta-analytic methods. We also examined whether effect sizes differed between the CLPM and RI-CLPM when both effects were estimated with the same data and whether effect sizes from the two models were correlated across studies. Next, we tested whether the size of cross-lagged effects was moderated by subfield of psychology, characteristics of the study design, and other coefficients in the models, again using meta-analytic methods. Finally, we compared the distribution of cross-lagged effects with the distribution of correlation coefficients estimated in the models.

We operationalized small, medium, and large effects as the 25th, 50th, and 75th percentile of the distribution of effect sizes, following the procedures used in prior studies on empirical benchmarks for effect sizes (e.g., Brydges, 2019; Gignac & Szodorai, 2016; Kinney et al., 2020; Lovakov & Agadullina, 2021). Moreover, we conceptualized these benchmarks as surrounding anchors (Bosco et al., 2015), not as cutoff values (i.e., minimum values). For example, if a correlation of .10 indicates a small effect and .25

a medium effect, then a correlation of .17 is interpreted as a small to medium effect.

Method

The present research used anonymized data and therefore was 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.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow the Journal Article Reporting Standards (Appelbaum et al., 2018). Data, materials, and code are available on the Open Science Framework (OSF; https://osf.io/4cwbr/). Computations were made with R (R Core Team, 2021), using the psych package Version 2.1.6 (Revelle, 2021) and the metafor package Version 3.0–2 (Viechtbauer, 2010). The present research was not preregistered.

Selection of Studies

Because the goal of the present research was to establish empirical benchmarks for cross-lagged effects in psychology, we collected the data by drawing a quasirepresentative sample of articles published in the following four subfields: developmental, social–personality, clinical, and industrial–organizational. We selected these subfields because they frequently use cross-lagged models and are among the largest areas of psychology. Collectively, the four subfields provide a relatively comprehensive coverage of psychological research that uses cross-lagged models. For each subfield, we selected three journals that could be considered representative of the field (developmental: *Developmental Psychology*, *Child Development*, *European Journal of Developmental Psychology*; social–personality: *Journal of Personality and Social Psychology*, *Personality and Social Psychology Bulletin*, *Personality and Individual Differences*; clinical: *Journal of Consulting and Clinical Psychology*, *Clinical Psychological Science*, *Journal of Affective Disorders*; industrial–organizational: *Journal of Applied Psychology*, *Journal of Organizational Behavior*, *European Journal of Work and Organizational Psychology*). We restricted the sample to articles published in 6 years (i.e., 2015 to 2020), which provided for a sufficiently large set of potentially eligible articles. We focused on the years 2015 to 2020 to maximize the number of articles that provided information on the RI-CLPM, which was introduced in the literature in 2015 (Hamaker et al., 2015).

For each of the 12 journals, we searched the full text of all articles published between 2015 and 2020. We selected all articles that included the term "cross-lagged" somewhere in the full text, which resulted in 1,048 potentially relevant articles. To decide whether the study should be included in the present analyses, the full text of these articles was evaluated by the third, fourth, fifth, or sixth author of the present research. In addition, 80 articles were rated by two of the authors to obtain estimates of interrater agreement. The interrater agreement on inclusion or exclusion in the dataset was high ($\kappa = .97$) and all disagreements were discussed until consensus was reached.

Studies were included if the following criteria were met: (a) the study was longitudinal; (b) the study used a relevant model (i.e., the CLPM and/or the RI-CLPM); (c) standardized estimates were reported for cross-lagged effects in both directions between constructs (i.e., effect of Construct A on Construct B, and vice versa); (d) the sample was not a treatment group of an intervention study (however, we used information from control groups if the control group did not undergo any treatment); (e) the article did not report inconsistent information on the effect sizes; and (f) the study was not a meta-analysis of cross-lagged effects (we used this exclusion criterion so that data from individual articles would not have a disproportionate influence on the benchmarks established in the present research). These procedures led to the inclusion of 144 articles, which provided data on 174 samples. A document with the references of these articles is available on OSF (https://osf.io/4cwbr/). Supplemental Figure S1 shows the PRISMA flow diagram of the selection procedure (adapted from Moher et al., 2009).

Coding of Studies

We coded the following sample characteristics: sample type (i.e., nationally representative, college/university students, clinical, or community), sample size, mean age at Time 1, and country in which the sample was collected. Moreover, we coded the following information regarding study design and effect sizes (this information was organized by pairs of constructs, for which cross-lagged effects were reported): Construct A, Construct B, presence of shared method variance (a dichotomous variable coded yes if measures of Constructs A and B were based on reports/ratings from the same person), latent measurement (a dichotomous variable coded yes if analyses were based on latent variables for at least one of the Constructs A and B), time lag between Times 1 and 2, correlation between Constructs A and B at Time 1, correlation between random intercepts (for the RI-CLPM only), cross-lagged effect of Construct A on B, cross-lagged effect of Construct B on A, stability effect of Construct A, stability effect of Construct B, and control of covariates (a dichotomous variable coded yes if at least one covariate was controlled in the model, over and above control for prior levels of the outcomes). For correlations and regression coefficients, standardized coefficients were coded. The effect size information was coded for both the CLPM and RI-CLPM, if available for the pair of constructs. Design and effect size information was coded for each pair of constructs available for a sample. Even if some studies included more than two waves of data (e.g., as noted above, the RI-CLPM requires at least three waves of data), the cross-lagged and stability effects were coded only with regard to the interval between Times 1 and 2, to avoid an overly complex structure of the dataset and analyses. This procedure is justified given that the cross-lagged and stability effects for Times 1 and 2 can be considered as unbiased sample from all intervals that would be available in multiwave studies.

The articles were coded by the third, fourth, fifth, or sixth author of the present research. In addition, 75 studies were coded by two of the authors to obtain estimates of interrater agreement. The interrater agreement was good (mean $r = .99$ for continuous variables and mean $\kappa = .89$ for categorical variables). All disagreements were discussed until consensus was reached.

Statistical Analyses

For each sample, the dataset included more than one cross-lagged effect, so there was a multilevel structure in the data (i.e., effect sizes nested in samples). Specifically, for each pair of constructs, there were two cross-lagged effects (i.e., Construct A predicting Construct B, and vice versa) and, moreover, for many of the samples information was available for multiple pairs of constructs. In all meta-analytic computations, we accounted for the multilevel structure by estimating multilevel meta-analytic models (using the "rma.mv" function in metafor; Viechtbauer, 2010). Specifically, we computed multilevel random-effects models (for estimating weighted mean effect sizes) and multilevel mixed-effects models (for testing moderators), following recommendations by Borenstein et al. (2009). Between-study heterogeneity was estimated with restricted maximum likelihood estimation, as recommended by Viechtbauer (2010). The meta-analytic computations with cross-lagged effects 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 $v = 1/(n - 3)$.

In the analyses, we examined the absolute values of cross-lagged effects, unless otherwise noted. The sign of a cross-lagged effect can be positive or negative, depending on the specific constructs examined in the research. For establishing effect size benchmarks, the sign of the cross-lagged effect is irrelevant and including it in the analyses would have led to meaningless results. For the same reason, we examined absolute values of correlation coefficients when tested as moderators of cross-lagged effects (i.e., the Time 1 correlations of the constructs in the CLPM and RI-CLPM, and the correlation between the random intercepts in the RI-CLPM).

Results

Description of Dataset

The dataset included information from 174 samples. Seventy-five percent were community samples, 12% were samples of college/university students, 8% were nationally representative, and 5% were clinical samples. Sample sizes ranged from 54 to 14,004 (M = 1,296.4, SD = 2,432.0, Mdn = 493.5; total number of participants = 225,577). Thirty-one percent of the samples were from the United States, 11% from China, 9% from Germany, 9% from the Netherlands, 7% from Switzerland, 5% from Canada, 5% from Norway, 3% from Finland, 2% from Australia, 2% from Poland, 2% from Sweden, 2% from the United Kingdom, 8% from other countries, and 4% were collected in multiple countries or the country was unknown. Mean age at Time 1 ranged from 2.5 to 74.4 years (M = 23.6, SD = 15.0). These data suggest that the samples included in the dataset were very heterogeneous.

Effect size information was coded for 592 pairs of constructs. As noted above, for each pair of constructs, there were two cross-lagged effects. For the analyses, we therefore restructured the dataset so that both directions of effects could be examined in the same analysis.[1] Thus, the dataset used in the analyses consisted of 1,184 cases (as noted above, the meta-analytic computations accounted for the multilevel structure of the data). In 48% of the models, constructs were measured as latent variables. Shared method variance was present in 74% of the models. Time lag between Times 1 and 2 ranged from .02 to 10.00 years (M = 1.17, SD = 1.08, Mdn = 1.00). Covariates were controlled for in 65% of the analyses with the CLPM and in 53% of the analyses with the RI-CLPM. The dataset included 1,028 effect sizes for the CLPM and 302 effect sizes for the RI-CLPM.

Distribution of Effect Sizes

First, we examined the distribution percentiles of the effect sizes, separately for the CLPM and RI-CLPM (Table 1; for histograms, see Figure 2). For the CLPM, the 25th, 50th, and 75th percentiles corresponded to values of .03, .07, and .12, respectively. For the RI-CLPM, the corresponding values were .02, .05, and .11. Then, we estimated weighted mean effect sizes by using multilevel random-effects models (see Table 2). For the CLPM and RI-CLPM, the mean cross-lagged effects were .095 and .083, respectively.

We also tested whether effect sizes differed between the CLPM and RI-CLPM if the cross-lagged effect had been estimated with both models using the same data (k = 146 effect sizes for each model). Importantly, in this analysis we did not examine absolute values but the originally observed values of cross-lagged effects (i.e., it was essential to account for the fact that the sign of the cross-lagged effect could differ between the CLPM and RI-CLPM). The mean difference between CLPM and RI-CLPM effects was −.001 [−.018; .016], which was nonsignificant, $t(145) = -.123$, $p = .903$. Moreover, the effect sizes in the CLPM and RI-CLPM were correlated at .37 ($p < .001$), suggesting some correspondence between the results obtained from these two models. It should be noted that this set of effect sizes was only a subset of the overall dataset and relatively small. Nevertheless, the fact that the mean difference was virtually zero and nonsignificant is in line with the mean effect sizes in the overall dataset reported above. Thus, the findings suggest that effect sizes from the CLPM and RI-CLPM did not differ systematically from each other, at least on average, and that the effect sizes were substantially correlated between the CLPM and RI-CLPM.

To test for publication bias, we used Egger's regression test (Egger et al., 1997). In these analyses, we again did not use absolute values but the originally observed values, which was required for a meaningful test of an asymmetric distribution of effect sizes. Moreover, because Egger's test is not available for the "rma.mv" function in metafor, we used the "rma" function, ignoring the multilevel structure of the data in these analyses. The tests were nonsignificant for both the CLPM ($z = -1.148$, $p = .251$) and RI-CLPM ($z = -.634$, $p = .526$), suggesting that the cross-lagged effects were not influenced by publication bias.

Moderator Analyses

Next, we tested whether the cross-lagged effects differed across subfields. These analyses were conducted only for the CLPM, due to the lower number of RI-CLPM effect sizes. To

[1] When describing the moderator analyses, we use the terms stability of predictor and stability of outcome, rather than stability of Construct A and stability of Construct B, because in half of the cases Construct A was the predictor and in the other half the outcome (and vice versa for Construct B).

Table 1
Distribution Percentiles for Cross-Lagged Effects in the CLPM and RI-CLPM

Percentile	CLPM (k = 1,028)	RI-CLPM (k = 302)
5	.01	.01
10	.01	.01
15	.02	.01
20	.03	.02
25	**.03**	**.02**
30	.04	.02
35	.05	.03
40	.05	.04
45	.06	.04
50	**.07**	**.05**
55	.08	.06
60	.09	.07
65	.10	.08
70	.11	.09
75	**.12**	**.11**
80	.14	.13
85	.17	.14
90	.20	.17
95	.27	.20

Note. The 25th, 50th, and 75th percentiles are shown in bold. CLPM = cross-lagged panel model; RI-CLPM = random intercept cross-lagged panel model; k = number of effect sizes.

examine differences across subfields, we followed the procedure described by Viechtbauer (2010). In the first model, the intercept was omitted, which allowed us to estimate weighted mean effect sizes for each of the four subfields. Table 3 shows the estimates, ranging from .092 to .101 (all ps $< .001$). The second model included an intercept but used only three dummy variables for the four subfields, which allowed us to test for *differences* across subfields. The differences were nonsignificant, $Q_{\text{Model}} = .646$, $df = 3$, $p = .886$. Thus, the effect sizes did not differ significantly by subfield and the point estimates were very similar.

We used a hierarchical strategy to test whether the cross-lagged effects were moderated by design characteristics and other coefficients in the models. In the first step, we tested for the effects of design characteristics (i.e., control of covariates, latent measurement, shared method variance, and time lag between assessments). In the second step, we added the effects of the other model coefficients (i.e., correlation at Time 1, stability of predictor, stability of outcome, and, for the RI-CLPM, the correlation between random intercepts). For time lag, we tested linear and quadratic effects to assess whether there was a curvilinear relation between time lag and the cross-lagged effects. To avoid collinearity between the linear and quadratic term, the quadratic effect was tested with an orthogonalized power polynomial (i.e., squared time lag residualized for linear time lag; Little et al., 2006). With regard to the CLPM, the quadratic effect was nonsignificant in both steps of the hierarchical analyses (ps = .216 and .409). With regard to the RI-CLPM, the quadratic effect was nonsignificant in the first step (p = .236), but significant in the second step ($p < .001$). However, in the second step the estimate for the quadratic time lag was clearly untrustworthy (that is, the regression coefficient was extremely large and the standard error was more than 10 times larger compared with the first step; moreover, the quadratic effect had a positive sign, implying a function opposite to what statistical theory would predict). We therefore did not interpret the findings from this second step and concluded that there was no reliable evidence for a curvilinear effect of time lag, for both the CLPM and RI-CLPM. In the remainder of the moderator analyses, we included only linear time lag but not quadratic time lag (for the results of the analyses with quadratic time lag, see Supplemental Tables S1 and S2).

Table 4 shows the results for the CLPM. None of the design characteristics had a significant moderator effect. Thus, although theory might predict that cross-lagged effects are larger, for example, when shared method variance is present or when studies do not control for covariates, the results did not support these hypotheses. However, when adding other coefficients to the model, all of these variables significantly explained heterogeneity in the cross-lagged effect. Specifically, the cross-lagged effect was larger when the constructs showed a larger concurrent correlation and when the predictor showed higher stability across time. In contrast, the cross-lagged effect was smaller when the outcome showed higher stability. As discussed in the Introduction, however, the evidence that other coefficients from the CLPM moderate the size of the cross-lagged effect does not imply any causal effects, but simply indicates that these other coefficients covary with the cross-lagged effect.

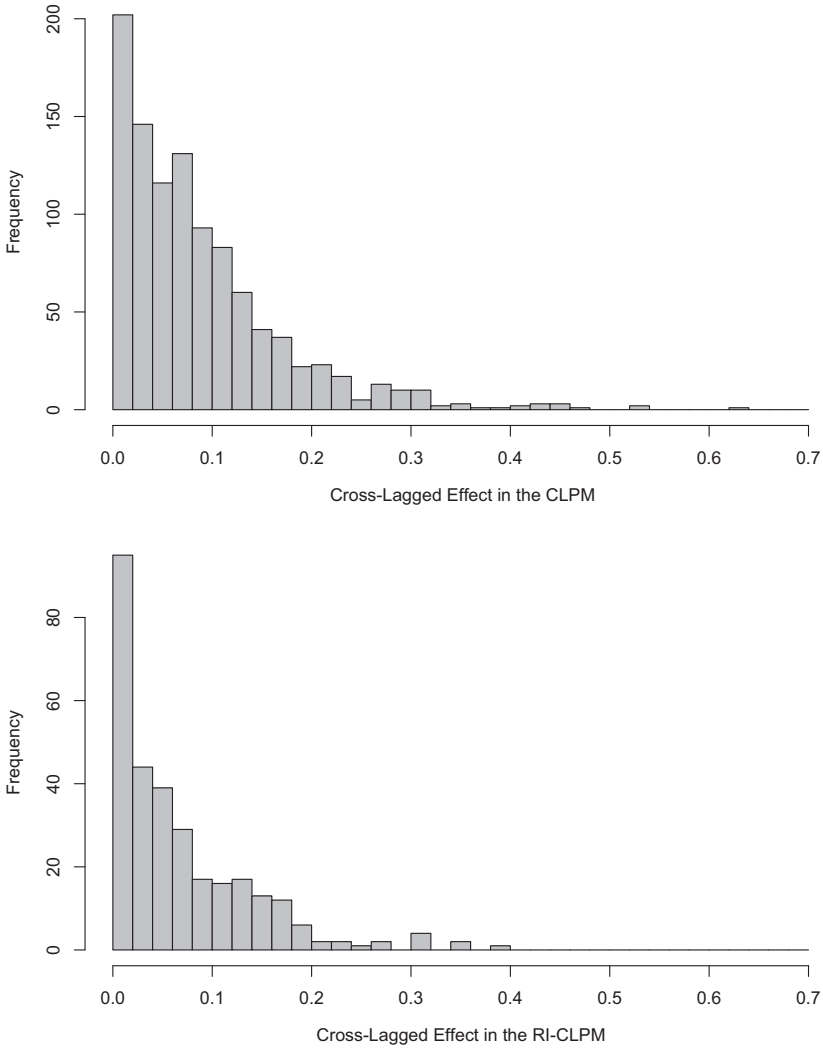
Table 5 shows the results for the RI-CLPM. Again, none of the design characteristics had a significant effect. When adding other coefficients of the model, two of these variables showed significant effects. Specifically, as for the CLPM, the cross-lagged effect was larger when the concurrent correlation at Time 1 and the stability of the predictor were stronger (again, we note that the evidence that these coefficients moderate the size of the cross-lagged effect does not imply any causal effects). In contrast to the CLPM, the stability of the outcome was not significantly related to the size of the cross-lagged effect, even if the point estimate of this moderator effect was in the same direction as for the CLPM. Moreover, the correlation between the random intercepts did not show a significant moderator effect.

Distributions of Correlations in the CLPM and RI-CLPM

Finally, we examined the distribution of correlations in the CLPM and RI-CLPM, to put the size of cross-lagged effects in context. As reviewed in the Introduction, empirical research suggests that correlations of about .10 are at the 25th percentile, correlations of .20–.25 at the 50th percentile, and correlations of .30–.40 at the 75th percentile, indicating small, medium, and large effects (Gignac & Szodorai, 2016; Lovakov & Agadullina, 2021; Paterson et al., 2016).

For the CLPM, we examined the concurrent correlation between the constructs at Time 1. The 25th, 50th, and 75th percentile corresponded to correlations of .12, .26, and .48; these values are slightly larger than empirical benchmarks reported in the literature. However, it should be noted that in about half of the studies the constructs were measured as latent variables, which typically increases the size of correlations between constructs. In fact, for studies in which the constructs were assessed as manifest variables, the 25th, 50th, and 75th percentile corresponded to correlations of .09, .20, and .40. Thus, we concluded that the concurrent

Figure 2
Distributions of Cross-Lagged Effects in the CLPM and RI-CLPM



Note. CLPM = cross-lagged panel model; RI-CLPM = random intercept cross-lagged panel model.

correlations were overall as expected, which strengthens confidence in the representativeness of the present set of studies.

For the RI-CLPM, we examined the Time 1 correlation and the correlation between the random intercepts. For the Time 1 correlation, the 25th, 50th, and 75th percentile corresponded to values of .07, .16, and .38, and for the random intercept correlation to .10, .23, and .51. It should be noted that in the RI-CLPM the Time 1 correlation is the correlation between the residualized construct factors,

Table 2
Meta-Analytic Estimates of Weighted Mean Cross-Lagged Effects in the CLPM and RI-CLPM

Model	*k*	*N*	Weighted mean Effect size	95% CI	*Q*	Variances σ_1^2	Variances σ_2^2
CLPM	1,028	198,446	.095*	[.087, .102]	7,112.5*	.0013	.0033
RI-CLPM	302	57,176	.083*	[.066, .101]	2,871.7*	.0018	.0018

Note . Computations were made with multilevel random-effects models. CLPM = cross-lagged panel model; RI-CLPM = random intercept cross-lagged panel model; *k* = number of effect sizes; *N* = number of participants, on which effect sizes are based; CI = confidence interval; *Q* = statistic used in heterogeneity test; σ_1^2 = variance component corresponding to the level of the grouping variable (i.e., between samples); σ_2^2 = variance component corresponding to the level nested within the grouping variable (i.e., within samples).
* $p < .001$.

Table 3
Meta-Analytic Estimates of Weighted Mean Cross-Lagged Effects in the CLPM by Subfield of Psychology

Subfield	*k*	*N*	Weighted mean Effect size	95% CI
Developmental	464	96,719	.097*	[.085, .110]
Social–personality	330	42,634	.092*	[.077, .106]
Clinical	170	47,999	.092*	[.075, .110]
Industrial–organizational	64	11,094	.101*	[.075, .126]

Note. Cross-lagged effects by subfield were examined with two multilevel mixed-effects models, following the procedure described by Viechtbauer (2010). In the first model (shown in the table), the intercept is omitted, which allows to estimate weighted mean effect sizes for each of the four categories. The second model includes an intercept but uses only three dummy variables for the four categories, which allows to test the differences between subfields. The differences between subfields were nonsignificant, $Q_{Model} = .646$, $df = 3$, $p = .886$. CLPM = cross-lagged panel model; *k* = number of effect sizes; *N* = number of participants, on which effect sizes are based; CI = confidence interval.
* $p < .001$.

and consequently is quite different conceptually from the typical cross-sectional correlation between constructs. In contrast, the correlation between the random intercepts is more comparable to the typical cross-sectional correlation between constructs, except that the random intercepts are latent variables based on multiple assessments of the constructs, which likely increases the size of correlations compared with the typical cross-sectional correlation. Overall, these considerations suggest that the percentiles of correlations in the RI-CLPM do not raise concerns that the present benchmarks for cross-lagged effects in the RI-CLPM underestimate the true distribution of cross-lagged effects.

Discussion

The goal of this research was to establish empirical benchmarks for the size of cross-lagged effects. We focused on the CLPM and RI-CLPM, given that these two cross-lagged models are the most frequently used in the field of psychology. To examine the distribution of effect sizes, we drew a quasirepresentative sample of studies published in developmental, social–personality, clinical, and industrial–organizational psychology. The dataset included 1,028 effect sizes for the CLPM and 302 effect sizes for the RI-CLPM.

Effect Size Conventions Suggested by the Present Research

We operationalized small, medium, and large effects as the 25th, 50th, and 75th percentile of the distribution of effect sizes, respectively, following the procedures used in prior studies on empirical benchmarks for effect sizes (e.g., Brydges, 2019; Gignac & Szodorai, 2016; Kinney et al., 2020; Lovakov & Agadullina, 2021). The 25th, 50th, and 75th percentiles corresponded to values of .03, .07, and .12, respectively, for the CLPM, and .02, .05, and .11, respectively, for the RI-CLPM. Given that the effect sizes did not differ significantly between the CLPM and RI-CLPM, we suggest that the same set of values should be used as benchmarks for the CLPM and RI-CLPM. Moreover, given that a much larger number of effect sizes was available for the CLPM compared to the RI-CLPM, we suggest that researchers use the values obtained for the CLPM, that is, .03 (small effect), .07 (medium effect), and .12 (large effect), regardless of whether the cross-lagged effects were estimated using the CLPM or the RI-CLPM.

The empirical benchmarks determined in the present research may be smaller than many researchers would expect. However, as reviewed in the Introduction, the findings from meta-analytic reviews of specific cross-lagged effects (e.g., the effect of self-beliefs on academic achievement) indicate that mean values of cross-lagged effects are typically in the range of .05 to .10. Thus, mean effect sizes from meta-analytic reviews of specific effects are comparable to the mean effect sizes found in the present research.

The size of the cross-lagged effects did not differ significantly across developmental, social–personality, clinical, and industrial–organizational psychology. Consequently, the same effect size conventions can be used across these subfields. Moreover, the

Table 4
Effects of Moderators on the Cross-Lagged Effect in the CLPM

	Model 1 (*k* = 1,028)			Model 2 (*k* = 416)		
Moderator	*B*	*SE*	*p*	*B*	*SE*	*p*
Design characteristics						
Control of covariates	−.002	.009	.830	−.006	.009	.533
Latent measurement	−.004	.008	.624	−.003	.009	.769
Shared method variance	.004	.009	.632	−.015	.010	.132
Time lag	−.005	.003	.101	−.004	.003	.103
Other CLPM coefficients						
Correlation at Time 1	—	—	—	.133	.019	<.001
Stability of predictor	—	—	—	.086	.022	<.001
Stability of outcome	—	—	—	−.189	.022	<.001

Note. Computations were made with multilevel mixed-effects models. Regression coefficients are unstandardized. Dash indicates that moderator was not included in the model. CLPM = cross-lagged panel model; *k* = number of effect sizes.

Table 5
Effects of Moderators on the Cross-Lagged Effect in the RI-CLPM

Moderator	Model 1 (k = 278)			Model 2 (k = 120)		
	B	*SE*	*p*	*B*	*SE*	*p*
Design characteristics						
Control of covariates	.018	.021	.399	−.085	.069	.222
Latent measurement	−.000	.020	.996	−.070	.081	.386
Shared method variance	−.005	.015	.724	−.015	.022	.513
Time lag	−.025	.019	.181	−.115	.158	.466
Other RI-CLPM coefficients						
Correlation at Time 1	—	—	—	.147	.064	.021
Correlation between random intercepts	—	—	—	−.059	.053	.260
Stability of predictor	—	—	—	.151	.042	<.001
Stability of outcome	—	—	—	−.053	.042	.202

Note. Computations were made with multilevel mixed-effects models. Regression coefficients are unstandardized. Dash indicates that moderator was not included in the model. RI-CLPM = random intercept cross-lagged panel model; k = number of effect sizes.

benchmarks identified in the present research may be useful for other disciplines, such as sociology, economics, education, medicine, and health sciences. As long as specific effect size benchmarks are missing for these disciplines, the conventional values suggested by the present research may provide a basis for interpreting the size of cross-lagged effects.

It is important to note that cross-lagged effects correspond conceptually to regression coefficients from more basic statistical models. Specifically, when two-wave longitudinal data are analyzed by multiple regression, using a predictor assessed at Time 1 to explain an outcome at Time 2, controlling for the outcome at Time 1, then the effect of the predictor is interpreted in the same way as a cross-lagged effect from the CLPM. Similarly, when the Time 1 predictor is used to explain change scores in the outcome between Times 1 and 2, then the effect also corresponds conceptually to the cross-lagged effect from the CLPM. Thus, in these situations, the benchmarks established in the present research can be used to interpret the size of the effects.

The mean effect sizes (i.e., .095 for the CLPM and .083 for the RI-CLPM, see Table 2) were larger than the median effect sizes (i.e., the 50th percentiles). Specifically, for both the CLPM and RI-CLPM, the means were located approximately at the 65th percentile, corresponding to the positively skewed distributions (as illustrated in Figure 2). Readers might wonder whether it would be more appropriate to use these means as indicators of a medium effect size. However, we believe that it is preferable to follow the procedures used in prior studies on empirical benchmarks for effect sizes, which were typically based on percentiles (including the median) rather than on means (e.g., Brydges, 2019; Gignac & Szodorai, 2016; Kinney et al., 2020; Lovakov & Agadullina, 2021).

Additional Implications of the Findings

The results of the moderator analyses suggest that studies that control for covariates do not find smaller cross-lagged effects, and this finding holds for both the CLPM and RI-CLPM (as noted above, covariates were controlled for in 65% of the analyses with the CLPM and in 53% of the analyses with the RI-CLPM). Theoretically, if an observed effect of a predictor is confounded by the predictor's association with a covariate (i.e., if the true effect is smaller than the observed effect), then controlling for the covariate should reduce the cross-lagged effect (Rohrer, 2018). Thus, an important question is whether studies that controlled for covariates actually controlled for theoretically-relevant confounding factors. It is beyond the scope of this article to determine whether this was the case, because it would require in-depth theoretical expertise to evaluate whether, in each study, the covariates were relevant confounding factors. This issue merits attention in future research.

Cross-lagged effects did not differ between studies that used latent versus manifest construct factors in the analyses. Although latent (compared with manifest) measurement often leads to larger associations between variables, this is not necessarily the case for all path coefficients included in a model (Cole & Preacher, 2014). Thus, it is possible that cross-lagged effects are not generally underestimated when models use manifest variables. In future research, it would be interesting to examine this issue more systematically based on simulation studies. Moreover, we emphasize that it is generally recommended to measure constructs as latent factors, to control for measurement error and systematic biases. However, the present findings suggest that it is not necessary to take this design characteristic into account when interpreting the size of cross-lagged effects.

Similarly, the size of cross-lagged effects was not influenced by shared method variance between the constructs. Although shared method variance often causes artificially inflated associations between variables when examining zero-order correlations (Podsakoff et al., 2012), the present findings suggest that this is not the case for cross-lagged effects. A possible explanation is that, when estimating cross-lagged effects, shared method variance is already controlled for to a great extent by inclusion of the concurrent correlation between the predictor and the outcome at Time 1 and the stability of the outcome between Times 1 and 2. Again, even if it is generally advised to control for shared method variance (e.g., by using multimethod assessment), the present findings suggest that it is not necessary to take this design characteristic into account when interpreting the size of cross-lagged effects.

Limitations and Strengths

The effect sizes examined in this research were sampled exclusively from published studies. Thus, an important question is

whether publication bias may have influenced the empirical benchmarks for cross-lagged effects. Generally, the literature on publication bias suggests that unpublished effect sizes tend to be smaller compared with published effect sizes (Sutton, 2009). Thus, if there is publication bias in research using cross-lagged models, then the empirical benchmarks determined in the present research would likely *overestimate*, not underestimate, the typical size of cross-lagged effects. Given that many researchers might feel that the conventional values proposed in this research are too small rather than too large, we believe that the possibility of publication bias is less problematic in the present context compared with other research contexts. Nevertheless, as reported above, we tested for publication bias but did not find significant evidence for either the CLPM or the RI-CLPM. Moreover, a recent meta-analysis tested but failed to find evidence that published and unpublished evidence on cross-lagged effects differed significantly from each other, in research on the effect of self-esteem on work outcomes (Krauss & Orth, 2021). In sum, these empirical findings do not support the hypothesis that the published evidence on cross-lagged effects is systematically inflated by publication bias.

The present research used a quasirepresentative sampling approach, by examining articles published in 12 different journals. Clearly, we do not know whether examining other journals would have yielded different findings. Nevertheless, the sampling approach covered four broad subfields of psychology and we assessed more than 1,000 articles for eligibility in the dataset, which resulted in a relatively heterogeneous set of samples and a relatively large number of effect sizes. Thus, we believe that the present dataset provided an appropriate basis for examining the distribution of effect sizes.

We did not code whether longitudinal measurement invariance was evaluated in the articles from which the effect sizes were obtained. However, cross-lagged models such as the CLPM and RI-CLPM require specific levels of longitudinal measurement invariance (Little et al., 2007; Schmitt & Kuljanin, 2008). If measurement invariance does not hold across waves, then this could reduce the magnitude of the stability coefficients, which could affect the cross-lagged effects, depending on whether lack of measurement invariance was found for the predictor or the outcome. In future research, it would be useful to assess this issue in detail.

When using a multilevel longitudinal model such as the RI-CLPM, standardized coefficients can be computed based on group or individual variances (Schuurman et al., 2016). In the present research, we used the standardized coefficients as reported in the published articles. However, in future research on the RI-CLPM, it would be interesting to test whether empirical benchmarks for effect sizes depend on the method of standardization used.

Recommendations

Based on the present findings, we propose the following guidelines for interpreting the size of cross-lagged effects. We suggest that .03 indicates a small effect, .07 a medium effect, and .12 a large effect. It is important to emphasize that these conventional values apply to standardized regression coefficients (i.e., beta coefficients or beta weights), not to unstandardized regression coefficients. We recommend that the conventional values are conceptualized as surrounding anchors rather than as cutoff values (Bosco et al., 2015). For example, if a cross-lagged effect is estimated as .10, it should be interpreted as a medium to large effect. Given that the effect sizes did not differ significantly between the CLPM and RI-CLPM, we recommend using the same set of conventional values for these two models. Moreover, it is possible that the conventional values are also appropriate for cross-lagged effects estimated with other types of models, such as the latent curve model with structured residuals (Curran et al., 2014), the STARTS model (Kenny & Zautra, 2001), and the latent change score model (McArdle, 2001). As long as empirical benchmarks are missing for other types of cross-lagged models, we tentatively suggest that the values above be used to gauge the magnitude of an effect. The values identified in the present research also apply to coefficients from multiple regression analyses of longitudinal data where prior levels of the outcome are controlled for. Finally, the present findings suggest that cross-lagged effects do not differ in size across four large subfields of psychology (developmental, social–personality, clinical, and industrial–organizational). We therefore suggest that researchers from all areas of psychology and related disciplines use the same benchmark values, which has the advantage that researchers can directly compare effect sizes across a wide range of research contexts.

References

Adachi, P., & Willoughby, T. (2015). Interpreting effect sizes when controlling for stability effects in longitudinal autoregressive models: Implications for psychological science. *European Journal of Developmental Psychology*, *12*(1), 116–128. https://doi.org/10.1080/17405629.2014.963549

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Biesanz, J. C. (2012). Autoregressive longitudinal models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 459–471). Guilford Press.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Brydges, C. R. (2019). Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging*, *3*(4), igz036. https://doi.org/10.1093/geroni/igz036

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Erlbaum.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

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

de Moor, E. L., Denissen, J. J. A., Emons, W. H. M., Bleidorn, W., Luhmann, M., Orth, U., & Chung, J. M. (2021). Self-esteem and satisfaction with social relationships across time. *Journal of Personality and Social Psychology*, *120*(1), 173–191. https://doi.org/10.1037/pspp0000379

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods*, *20*(4), 489–505. https://doi.org/10.1037/met0000041

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fairbairn, C. E., Briley, D. A., Kang, D., Fraley, R. C., Hankin, B. L., & Ariss, T. (2018). A meta-analysis of longitudinal associations between substance use and interpersonal attachment security. *Psychological Bulletin*, *144*(5), 532–555. https://doi.org/10.1037/bul0000141

Finkel, S. E. (1995). *Causal analysis with panel data*. Sage.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Harris, M. A., & Orth, U. (2020). The link between self-esteem and social relationships: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology*, *119*(6), 1459–1477. https://doi.org/10.1037/pspp0000265

Kenny, D. A. (1979). *Correlation and causality*. Wiley.

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). American Psychological Association. https://doi.org/10.1037/10409-008

Khazanov, G. K., & Ruscio, A. M. (2016). Is low positive emotionality a specific risk factor for depression? A meta-analysis of longitudinal studies. *Psychological Bulletin*, *142*(9), 991–1015. https://doi.org/10.1037/bul0000059

Kinney, A. R., Eakman, A. M., & Graham, J. E. (2020). Novel effect size interpretation guidelines and an evaluation of statistical power in rehabilitation research. *Archives of Physical Medicine and Rehabilitation*, *101*(12), 2219–2226. https://doi.org/10.1016/j.apmr.2020.02.017

Krauss, S., & Orth, U. (2021). Work experiences and self-esteem development: A meta-analysis of longitudinal studies. *European Journal of Personality*. Advance online publication. https://doi.org/10.1177/08902070211027142

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*, *13*(4), 497–519. https://doi.org/10.1207/s15328007sem1304_1

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*(4), 357–365. https://doi.org/10.1177/0165025407077757

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Lüdtke, O., & Robitzsch, A. (2021). *A critique of the random intercept cross-lagged panel model*. PsyArxiv. https://doi.org/10.31234/osf.io/6f85c

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). Scientific Software International.

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

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling*, *28*(4), 638–648. https://doi.org/10.1080/10705511.2020.1784738

Mund, M., & Nestler, S. (2019). Beyond the cross-lagged panel model: Next-generation tools for analyzing interdependencies across the life course. *Advances in Life Course Research*, *41*, 100249. https://doi.org/10.1016/j.alcr.2018.10.002

Nohe, C., Meier, L. L., Sonntag, K., & Michel, A. (2015). The chicken or the egg? A meta-analysis of panel studies of the relationship between work-family conflict and strain. *Journal of Applied Psychology*, *100*(2), 522–536. https://doi.org/10.1037/a0038012

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034. https://doi.org/10.1037/pspp0000358

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569. https://doi.org/10.1146/annurev-psych-120710-100452

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

Revelle, W. (2021). *psych: Procedures for personality and psychological research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods*, *21*(2), 206–221. https://doi.org/10.1037/met0000062

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin*, *139*(1), 213–240. https://doi.org/10.1037/a0028931

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 435–452). Russell Sage Foundation.

Usami, S. (2021). On the differences between general cross-lagged panel model and random-intercept cross-lagged panel model: Interpretation of cross-lagged parameters and model choice. *Structural Equation Modeling*, *28*(3), 331–344. https://doi.org/10.1080/10705511.2020.1821690

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods*, *24*(5), 637–657. https://doi.org/10.1037/met0000210

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

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

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*(2), 176–192. https://doi.org/10.1037/a0027543

Wu, W., Selig, J. P., & Little, T. D. (2013). Longitudinal data analysis. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Vol. 2. Statistical analysis* (pp. 387–410). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934898.013.0018

Received November 17, 2021
Revision received February 24, 2022
Accepted March 1, 2022 ■