Journal of Behavioral Data Science, 2024, *4* (2), 51–63.
DOI: https://doi.org/10.35566/jbds/lin

Lord's Paradox Illustrated in Three-Wave Longitudinal Analyses: Cross Lagged Panel Models Versus Linear Latent Growth Models

Hua Lin and Robert E. Larzelere

Department of Human Development and Family Science
Oklahoma State University, USA
hua.lin@okstate.edu and robert.larzelere@okstate.edu

Abstract. Lord's (1967) paradox showed that two basic ways to analyze change longitudinally can produce contradictory results in 2-occasion nonrandomized studies. This study extends that paradox to difference-score and ANCOVA-type residualized change score analyses across three waves of data for four corrective actions thought to be effective: corrective disciplinary actions by parents (timeout and reasoning) and corrective actions by professionals (psychotherapy and hospitalization). All significant findings indicated that these corrective actions were harmful according to cross-lagged panel models but beneficial according to linear latent growth models. One type of analysis may not generalize to the other type of analysis. These results are consistent with recent recognition that ANCOVA-type analyses are biased by invariant between-person differences, but difference-score analyses can have their own biases. Recognition of these biases is needed to discriminate between stronger and weaker causal evidence in longitudinal analyses.

Keywords: Causal inference · Cross-lagged panel analysis · Latent growth modeling

1 Introduction

Most longitudinal studies have found that corrective actions by parents and by professionals appear to be harmful in analyses that control for initial differences with ANCOVA-type analyses of residualized change scores $Y_2|Y_1$ (i.e., Y_2 conditional on Y_1; Larzelere, Lin, Payton, & Washburn, 2018). Examples include parent-youth discussions about the risks of unprotected sex (Lin & Larzelere, 2020), psychotherapy for children, and methylphenidate (i.e., Ritalin: Larzelere, Ferrer, Kuhn, & Danelia, 2010). Although all of these corrective actions have looked harmful according to analyses of residualized change score analyses (i.e., predicting Wave-2 outcomes Y_2 while controlling for Wave-1 outcome scores Y_1),

difference-score analyses have often made them look beneficial from the same data (predicting $Y_2 - Y_1$; Larzelere et al., 2018).

This inconsistency is an example of Lord's (1967) paradox. In Lord's original hypothetical study, females' and males' weight gains were compared with each other using the two types of change-score analyses. Initial average weights differed significantly for females and males, and their average weights stayed the same from pretest to posttest for both genders. Difference-score analyses indicated no gender difference in weight gained, as expected. However, ANCOVA indicated that males gained more weight than females who started at the same weight. Although both results are correct for their corresponding *predictive* research questions, both cannot provide correct causal inferences about the effect of manipulating a causal variable (e.g., for a corrective action of interest). Consistent with Lord's original paradox, causally relevant coefficients from residualized change score analyses are generally biased in the direction of the pretest group means, *relative to* the difference-score coefficients, regardless of which analysis is least biased (Angrist & Pischke, 2009; Larzelere et al., 2018; Lin & Larzelere, 2020). This corresponds to recent documentations that longitudinal analyses of residualized change scores are biased by between-person differences that do not change during the study (Berry & Willoughby, 2017; Hamaker, Kuiper, & Grasman, 2015; Hoffman, 2015).

Despite being discussed for over 50 years, the implications of Lord's paradox have been insufficiently recognized in developmental psychology. Longitudinal analyses have preferred analyzing residualized change scores since Cronbach and Furby's 1970 recommendation. Two-wave residualized change score and difference-score analyses are building blocks for more complex models such as cross-lagged panel analyses and linear growth models. Therefore, this problem of contradictory, potentially biased estimates likely generalizes to advanced statistical models. However, little is known about how Lord's paradox applies to more complex statistical models (e.g., cross-lagged panel models and latent growth models) or how to minimize these biases to approximate valid causal estimates more closely. Like ANCOVA, cross-lagged panel models predict residualized change scores (e.g., predicting y_t controlling for y_{t-1}) between adjacent occasions across three or more occasions. Therefore, cross-lagged panel models could be considered a series of $T - 1$ ANCOVAs. In contrast, the most basic latent growth model typically predicts a simple difference score from Wave 1 to Wave T based on the best-fitting linear slope of the outcome scores across the T waves. In this article, we modify the latent growth model to predict simple difference scores between adjacent waves. This modified latent growth model is more similar to cross-lagged panel models by modeling change in the outcome scores from Wave $t - 1$ to Wave t across T waves.

1.1 Cross-Lagged Panel Model

Cross-lagged panel models estimate the bidirectional effects between the treatment condition and the outcome score over time (Selig & Little, 2012). The

cross-lagged panel model provides information about how variations in one variable (typically treatment vs. control) predict changes in another variable (the outcome) over time. The multi-wave cross-lagged panel model can be described as follows:

$$X_{i,t} = \alpha_0 + \alpha_1 X_{i,t-1} + \alpha_2 Y_{i,t-1} + \varepsilon_{i,xt}$$

$$Y_{i,t} = \theta_0 + \theta_1 X_{i,t-1} + \theta_2 Y_{i,t-1} + \varepsilon_{i,yt}$$

where X and Y represent the treatment and outcome variables at a given time t, predicted from these variables at the immediately preceding time $t-1$. These are adjacent-wave ANCOVA functions for both variables - as predictor and outcome at adjacent time points. The primary interest is the treatment effect θ_1 of X_{t-1} on the outcome at the next time point, controlling for the preceding outcome score $Y_t|Y_{t-1}$.

1.2 Latent Growth Model

Whereas cross-lagged panel models predict residualized change scores $Y_t|Y_{t-1}$, the linear growth model uses difference scores as its basic building block for analyzing change. Linear latent growth models analyze how individuals' scores change over time and how treatment conditions influence such changes using the difference-score approach:

$$\text{Level 1} : Y_{ti} = \beta_{0i} + \beta_{1i} T_{ti} + r_{ti}$$

$$\text{Level 2} : \beta_{0i} = \gamma_{00} + \gamma_{01} X_j + \epsilon_{0i}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} X_j + \epsilon_{1i}$$

where Level 1 represents how individual scores change linearly over time, and Level 2 predicts initial scores and within-individual changes from between-person differences in the causal variable of interest X_j. At Level 1, Y_{ti} represents individual i's outcome at time t; β_{0i} represents the starting point (when $T_{ti} = 0$) on individual i's best-fitting straight line across time; β_{1i} represents the individual's linear slope across time T_{ti}, and r_{ti} represents the unexplained error in the individual's outcome Y_{ti}. At level 2, γ_{00} represents the mean of the individual starting points on the outcome when $X_j = 0$; γ_{01} is the effect of the predictor X_j on the starting point (or intercept) β_{0i}; ϵ_{0i} represents the deviation of the individual's starting point from what is predicted by the rest of that equation (the fixed-effects part); X_j is the treatment condition, e.g., with $j = 2$ for the treatment group and $j = 1$ for the comparison group; γ_{10} is the mean linear slope across the waves when $X_j = 0$; γ_{11} is the effect of the predictor X_j on the average individual slope β_{1i}; and ϵ_{1i} is the deviation of the individual's slope from the slope predicted from the fixed-effects part of that equation. With

person-mean centering, the latent growth model estimates pure within-person changes at Level 1. Level 2 then estimates between-person differences in those changes. In two-wave analyses, the slope is the difference score from Wave 1 to Wave 2 ($Y_{2i} - Y_{1i}$). In three-wave analyses, each individual's slope is the estimated linear change per unit of time in that person's best-fitting straight line across their scores at all three waves. The primary interest of the latent growth model is the effect of the treatment on change in the slope γ_{11}.

1.3 The Current Study

The current study used four examples of corrective actions thought to be effective to illustrate Lord's paradox in three-wave longitudinal analyses. The four examples involve the apparent effect of (1) disciplinary time-out on subsequent child aggression, (2) disciplinary reasoning on subsequent child aggression, 3) psychotherapy on subsequent maternal depression, and 4) hospitalization on subsequent physical health. Each example was analyzed with a cross-lagged panel model and a latent growth model across three waves of data: Although standard latent growth models typically predict one linear slope from the first to the last wave, the two-slope latent growth model in this study was designed to be more similar to a cross-lagged panel by predicting simple difference scores between adjacent waves. The intercept was modeled as usual (all loadings set to 1), but Slope 1 specified the simple change from Wave 1 to Wave 2 (with loadings set at -1 and 0), whereas Slope 2 specified the simple change from Wave 2 to Wave 3 (loadings set at 0 and 1). The model then estimated the effect of each correction action at one wave (Wave 1 or 2) on simple change in the outcome from that wave to the next wave.

It was hypothesized that cross-lagged panel models would make corrective actions appear to be harmful, whether implemented by parents (time-out, reasoning) or by professionals (psychotherapy, hospitalizations). In contrast, latent growth models would indicate that all these corrective actions would lead to improvements in the same outcomes.

2 Method

2.1 Participants

This study used the Fragile Families and Child Wellbeing (FFCW) dataset which started with baseline data for mostly unmarried couples with children born from 1998 to 2000 in 20 large cities of the United States (Reichman, Teitler, Garfinkel, & McLanahan, 2001). It includes a wide range of data on household characteristics, physical and mental health, and parenting, first when the children were born, and later when the children were approximately 1, 3, 5, 9, 15, and 22 years old. The current study uses corrective action data when the children were 3 and 5 years old and outcome data when they were 3, 5, and 9 years old. At baseline (when the child was born), the 4588 mothers in these 3-wave analyses averaged

25.2 years old and had some college on average, and consisted of 21.2% White, 48.0% Black, 27.0% Hispanic, and 3.8% others. Missingness ranged from 8% to 28%. Full information maximum likelihood was used to adjust for missing data in the 3-wave analyses, which assumes that those data were missing at random. The FFCW data set (https://ffcws.princeton.edu/documentation) is available from Princeton University's Office of Population Research (OPR) data archive.

2.2 Measures

Time-out Disciplinary time-out was assessed by mothers' self-report on one item from the Parent-Child Conflict Tactics Scale (Straus, Hamby, Finkelhor, Moore, & Runyan, 1998), which asks how often in the past year mothers put their child in time-out or sent them to their room. The frequency was reported on a 8-point scale, ranging from never (0) to 11-20 times (6) to more than 20 times (7). We created a dummy variable indicating whether the time-out frequency was above the median frequency or not: 11 or more times (1), or less than 11 times (0).

Reasoning Disciplinary reasoning was also assessed by mothers' self-report from one item of the Parent-Child Conflict Tactics Scale (Straus et al., 1998), using the same response options. The item asks how often in the past year mothers explained why something was wrong. We created a dummy variable indicating whether reasoning occurred more frequently than the median or not: 11 or more times (1), or less than 11 times (0).

Child Aggression The FFCW measure of child aggression was a modified version of the aggression subscale of the Child Behavior Checklist (CBCL; Achenbach, 1991; Achenbach & Rescorla, 2000) with 19 items at age 3, 13 items at age 5, and 17 items at age 9. Mothers reported whether various behaviors were not true, somewhat/sometimes true, or often/very true of the child. Sample questions include destroying things, being disobedient, hitting others, getting in many fights, screaming a lot, and threatening people. The scale demonstrated excellent reliability with coefficient alphas of 0.88 (age 3), 0.82 (age 5), and 0.88 (age 9).

Psychotherapy for Depression Psychotherapy for depression was measured by two questions. Mothers reported whether they had received counseling/therapy for personal problems in the past year. If "yes," they were asked whether the counseling/therapy was for depression or for a range of other problems. Reported counseling/therapy for depression was coded 1, and other answers were coded 0. Two dummy codes indicated whether mothers received psychotherapy for depression when the child was 3 and 5 years old.

Depression Severity Depression severity was based on maternal self-reports about symptoms of a Major Depressive Episode, derived from the Composite International Diagnostic Interview—Short Version (Kessler, Andrews, Mroczek, Ustun, & Wittchen, 1998). The CIDI is a standardized survey for assessing mental disorders such as depression. The depression items included two stem questions and seven additional questions for those exceeding the threshold on the stem questions. We constructed a 13-point scale from none (0) through sub-threshold symptoms (1 to 4) to the number of symptoms above the threshold, including the stem questions (5 to 12).

Hospitalization Hospitalization was measured by a single dummy-coded item indicating whether mothers visited an emergency room or had an overnight hospital stay during the past year.

Physical Health Mothers' physical health was based on mothers' self-reports on their health condition on a five-point scale (0 = poor to 4 = great).

3 Results

The results showed contradictory results from cross-lagged panel models compared to linear latent growth models. The four examples estimate the apparent effects of corrective actions by parents (time-out and reasoning) and by professionals (psychotherapy and inpatient hospitalized treatments).

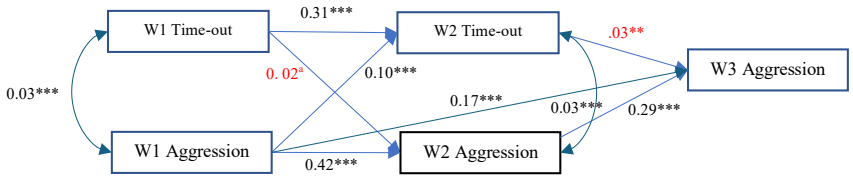
3.1 Time-out and Subsequent Child Aggression

Cross-lagged panel models made time-out at Wave 2 look significantly harmful by increasing child aggression at Wave 3, after controlling for the preceding aggression scores: $b = 0.03, p = 0.002$, Figure 1, Plot A. (Time-out at Wave 1 also predicted higher aggression at Wave 2 controlling for Time-1 aggression, but only marginally, $b = 0.02, p < 0.01$.) In contrast, two-slope latent growth models made time-out look helpful in reducing child aggression from each wave to the next wave: $b = -0.06$ (Wave 1 time-out predicting change in aggression from Wave 1 to Wave 2), and $b = -0.03$ (Wave 2 time-out predicting change in aggression from Wave 2 to Wave 3), ps < 0.01, Figure 1, Plot B.

3.2 Reasoning and Subsequent Child Aggression

Cross-lagged panel models also made disciplinary reasoning at Wave 2 look harmful by predicting more child aggression at Wave 3, after controlling for Wave 1 and Wave 2 aggression scores: $b = 0.04, p = 0.001$, Figure 2, Plot A. In contrast, the 2-slope latent growth model made reasoning at Wave 1 look helpful in reducing child aggression from Wave 1 to Wave 2: $b = -0.07, p < .001$, Figure 2, Plot B.

Plot A: CLPM



Plot B: LGM

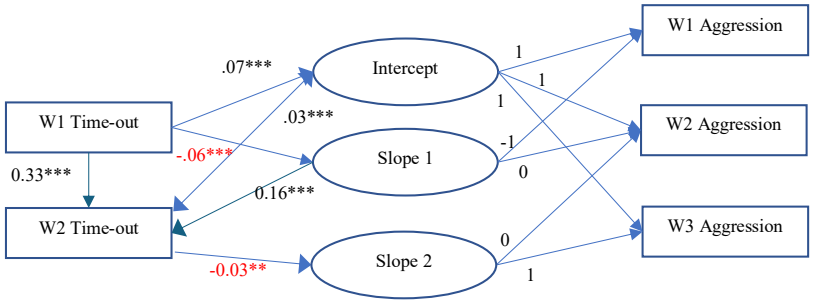


Figure 1. Cross-Lagged Panel Model (CLPM) and Latent Growth Model (LGM) of Time-out and Child Aggression across three waves of data. $^{a}p < .10;^{*}p < .05;^{**}p < .01;^{***}p < .001. N = 4153$

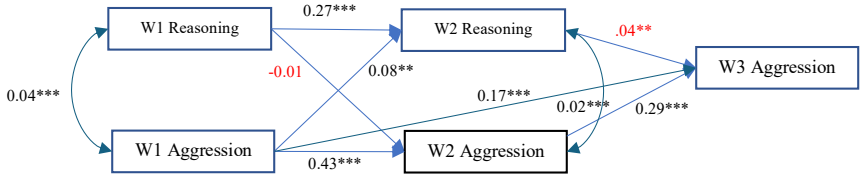
3.3 Psychotherapy and Subsequent Depression

The results followed a similar pattern for professional treatments. A cross-lagged panel model made therapy for depression look significantly harmful by predicting higher depression severity at the next wave, even after controlling for the preceding depression severity score: $b = 0.70$ (Wave 1 psychotherapy predicting Wave 2 depression severity) and $b = 1.60$ (Wave 2 psychotherapy predicting Wave 3 depression severity), all $ps < 0.05$, Figure 3, Plot A. In contrast, 2-slope latent growth models made therapy look helpful in reducing depression severity from each wave to the next wave: $b = -3.05$ (Wave 1 psychotherapy predicting a decrease in depression from Wave 1 to Wave 2) and $b = -0.73$ (Wave 2 psychotherapy predicting a decrease in depression from Wave 2 to Wave 3), $ps < 0.05$, Figure 3, Plot B.

3.4 Hospitalization and Subsequent Physical Health

A cross-lagged panel model made hospitalization look harmful by predicting worse physical health in mothers at the next wave, after controlling for mothers' preceding physical health score: $b = -0.15$ (Wave 1 hospitalization predicting worse health at Wave 2) and $b = -0.09$ (Wave 2 hospitalization predicting worse health at Wave 3), all $ps < 0.05$, Figure 4, Plot A. In contrast, a 2-slope latent

Plot A: CLPM



Plot B: LGM

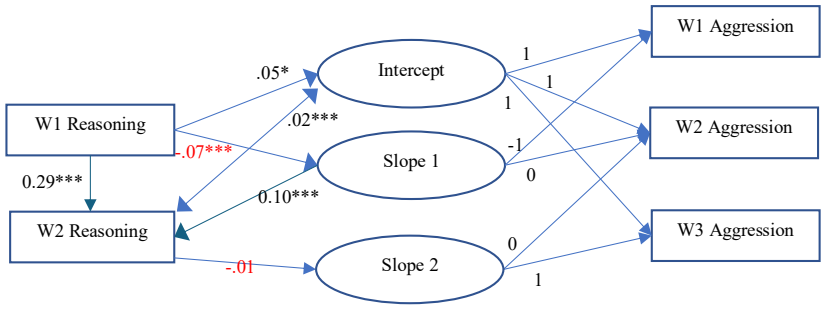


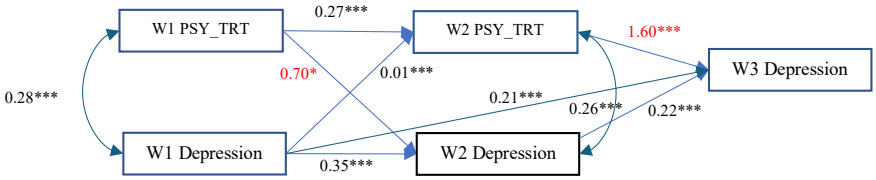
Figure 2. Cross-Lagged Panel Model (CLPM) and Latent Growth Model (LGM) of Reasoning and Child Aggression across three waves of data. $^{*}p < .05;^{**}p < .01;^{***}p < .001. N = 4153$

growth model made hospitalization look helpful in improving mothers' health from each wave to the next wave: $b = 0.20$ (Wave 1 hospitalization predicting improving health from Wave 1 to Wave 2) and $b = 0.12$ (Wave 2 hospitalization predicting improving health from Wave 2 to Wave 3), ps < 0.01, Figure 4, Plot B.

4 Discussion

Despite being well-known for over 50 years, the implications of Lord's (1967) paradox for multi-wave longitudinal analyses have not been well understood. The current study used four examples of corrective actions to illustrate Lord's paradox in three-wave longitudinal analyses. As expected, results from the difference-score approach (e.g., latent growth models) contradicted results from the residualized change score approach (cross-lagged panel models), just as in two-wave analyses. All four corrective actions looked effective according to latent growth models but harmful according to cross-lagged panel models. This may help explain why longitudinal analyses of residualized change scores have been unable to find effective parental responses to perceived child problems, such as persistent defiance, smoking, and precocious sex (Larzelere et al., 2018). The bias

Plot A: CLPM



Plot B: LGM

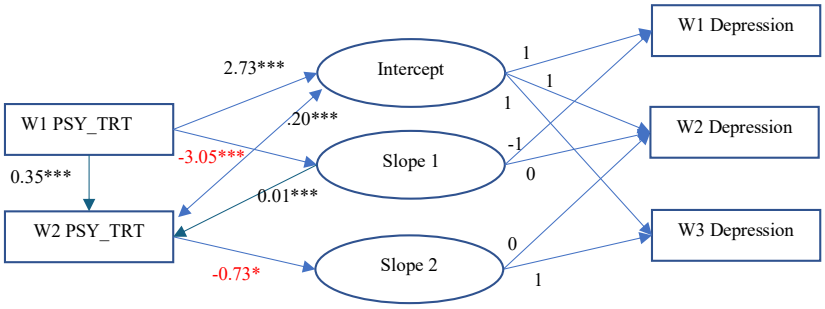
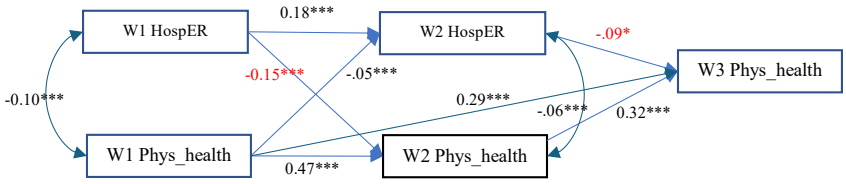


Figure 3. Cross-Lagged Panel Model (CLPM) and Latent Growth Model (LGM) of Mothers' Depression across three waves of data, predicted by Psychotherapy (PSY_TRT). $^{*}p < .05;^{**}p < .01;^{***}p < .001. N = 4588$

in residualized change score analyses helps parenting researchers confirm what they oppose (e.g., spanking), but hinders their efforts to document more effective corrective actions to replace it. The failure to find more effective corrective disciplinary responses in basic parental discipline research may help explain why clinical treatments for conduct problems in children (mostly implemented by parents) have not improved in effectiveness over the past 50 years (Weisz et al., 2019). In any case, it is worrisome that the kinds of analyses considered to be sufficient causal evidence to oppose harsh discipline practices such as spanking make most corrective actions by professionals look harmful also (Larzelere et al., 2018). These results can be explained by systematic biases recently elucidated in ANCOVA-type longitudinal analyses, because they confound within-person changes with invariant between-person differences, which are already reflected in the initial outcome scores (Berry & Willoughby, 2017; Hamaker et al., 2015; Hoffman, 2015).

Note that the four corrective actions in this study are all considered to be effective on average. Their effectiveness has been demonstrated in meta-analyses of randomized trials for psychotherapy for depression (Cuijpers et al., 2023) and time-out for oppositional defiance (Larzelere, Gunnoe, Roberts, Lin, & Ferguson, 2020) , whereas disciplinary reasoning and hospital-based treatments are widely considered to be effective in most cases. These results add to a wide range of

Plot A: CLPM



Plot B: LGM

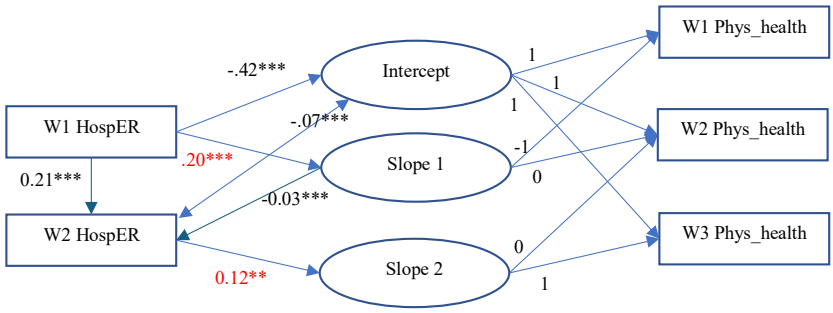


Figure 4. Cross-Lagged Panel Model (CLPM) and Latent Growth Model (LGM) of Mothers' Health across three waves of data. HospER = Overnight hospitalization or Emergency room visit. $^{*}p < .05;^{**}p < .01;^{***}p < .001. N = 4588$

corrective actions shown (incorrectly) to be significantly harmful in longitudinal analyses of residualized change scores whether implemented by parents or professionals (Larzelere et al., 2018).

Does this mean that difference-score analyses are always less biased than residualized change score analyses? Not necessarily. If the covariates account perfectly for selection into treatment conditions, ANCOVA is unbiased (Van Breukelen, 2013). The problem is that covariates fall short of this ideal in comprehensiveness, validity, and reliability in most longitudinal analyses. Steiner, Cook, Shadish, and Clark (2010) showed that ANCOVA can approximate unbiased causal effects when the covariates include baseline scores on the outcome and variables that account for self-selection into treatment conditions. Their study compared self-selection by college students into exercises to improve either math or vocabulary. It is unclear how well their results generalize to other situations in which self-selection is less well understood and poorly represented in the covariates. Note that the current study adjusted only for baseline scores on the outcome, with no additional covariates to account for why people with the same problem severity selected the corrective action of interest or not.

One factor is that the bias in ANCOVA-type analyses of residualized change scores is larger when the treatment conditions differ greatly in baseline scores on the outcome. When pretest group mean scores differ, the assumption of indepen-

dence between covariates and treatment of ANCOVA is violated. Violations of this assumption usually imply invariant between-group differences, which have been shown recently to bias analyses of residualized change scores (Berry & Willoughby, 2017; Hamaker et al., 2015). The bias occurs because within-person change following the corrective action is confounded or "smushed" with between-person differences that are unchanging (Hoffman, 2015).

In contrast, independence of treatment condition and baseline scores is not an assumption of the difference-score approach, making it free from that particular bias. Nor are difference-score analyses biased by measurement error in its baseline scores, whereas residualized change score analyses are known to be biased by that measurement error. In contrast to residualized change score approaches, the difference-score approach ignores between-person differences except for differences due to within-person changes in the time period studied. In randomized studies, between-person differences that precede the treatment are removed, so that any between-person differences at post-test are due *only* to within-person changes due to the treatment conditions. In non-randomized studies, difference-score analyses can have their own unique biases, such as regression toward the mean, but the results of this and other studies of corrective actions suggest that difference-score models are often less biased than are residualized change score models, such as cross-lagged panel models.

What can be done to improve the causal validity of longitudinal analyses? The first step is to recognize the problem. One improvement would be to follow the example of econometricians in checking the robustness of results across multiple types of analyses (Duncan, Engel, Claessens, & Dowsett, 2014). Angrist and Pischke (2009) showed that these two types of change-score analyses will bracket the true causal effect under some assumptions, but it can be difficult to tell whether those assumptions are satisfied. For example, the true effect of job training programs was outside this bracket for both men and women in Lalonde's (1986) classic study. Robustness across both types of change-score analyses is therefore consistent with an unbiased causal effect, but does not guarantee it (Lin & Larzelere, 2020).

Statisticians continue to expand the options for improving the capability of longitudinal analyses to approximate less biased causal inferences (e.g., Zyphur et al., 2020). Whereas simulations of statistical innovations are generally based on conditions that may not apply to real data (e.g., the possibility of avoiding all specification errors), illustrations with actual data have rarely shown which types of analyses can correctly recover the same direction of effectiveness for corrective actions that have been documented in randomized trials. Confidence in causal inferences from longitudinal analyses can be strengthened by showing that statistical innovations can make longitudinal analyses agree with unbiased causal evidence from randomized trials.

Acknowledgments

This research was supported by a National Institutes of Health (Grant number R03 HD107307) NICHD R03 grant.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. University of Vermont Department of Psychiatry.

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA Preschool Forms & Profiles*. University of Vermont Department of Psychiatry.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's approach*. Princeton University Press. doi: https://doi.org/10.1515/9781400829828

Berry, D., & Willoughby, M. T. (2017). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, *88*(4), 1186–1206. doi: https://doi.org/10.1111/cdev.12660

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68–80. doi: https://doi.org/10.1037/h0029382

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Ebert, D., & Karyotaki, E. (2023). Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: A comprehensive meta-analysis including 409 trials with 52,702 patients. *World Psychiatry*, *22*(1), 105–115. doi: https://doi.org/10.1002/wps.21069

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*(11), 2417–2425. doi: https://doi.org/10.1037/a0037996

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

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

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., & Wittchen, H. U. (1998). The World Health Organization Composite International Diagnostic Interview – Short Form (CIDI-SF). *International Journal of Methods in Psychiatric Research*, *7*(4), 171–185. doi: https://doi.org/10.1002/mpr.47

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, *76*(4), 604–620.

Larzelere, R. E., Ferrer, E., Kuhn, B. R., & Danelia, K. (2010). Differences in causal estimates from longitudinal analyses of residualized versus simple gain scores: Contrasting controls for selection and regression artifacts. *International Journal of Behavioral Development*, *34*(2), 180–189. doi: https://doi.org/10.1177/0165025409351386

Larzelere, R. E., Gunnoe, M. L., Roberts, M. W., Lin, H., & Ferguson, C. J. (2020). Causal evidence for exclusively positive parenting and for timeout: Rejoinder to Holden, Grogan-Kaylor, Durrant, and Gershoff (2017). *Marriage & Family Review*, *56*(4), 287–319. doi: https://doi.org/10.1080/01494929.2020.1712304

Larzelere, R. E., Lin, H., Payton, M. E., & Washburn, I. J. (2018). Longitudinal biases against corrective actions. *Archives of Scientific Psychology*, *6*(1), 243–250. doi: https://doi.org/10.1037/arc0000052

Lin, H., & Larzelere, R. E. (2020). Dual-centered ANCOVA: Resolving contradictory results from Lord's paradox with implications for reducing bias in longitudinal analyses. *Journal of Adolescence*, *85*, 135–147. doi: https://doi.org/10.1016/j.adolescence.2020.11.001

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305. doi: https://doi.org/10.1037/h0025105

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, *23*(4-5), 303–326. doi: https://doi.org/10.1016/S0190-7409(01)00141-4

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 265–278). New York, NY, US: Guilford Press.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250–267. doi: https://doi.org/10.1037/a0018719

Straus, M. A., Hamby, S. L., Finkelhor, D., Moore, D. W., & Runyan, D. (1998). Identification of child maltreatment with the Parent-Child Conflict Tactics Scales: Development and psychometric data for a national sample of American parents. *Child Abuse and Neglect*, *22*(4), 249–270. doi: https://doi.org/10.1016/s0145-2134(97)00174-9

Van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922. doi: https://doi.org/10.1080/00273171.2013.831743

Weisz, J. R., Kuppens, S., Ng, M. Y., Vaughn-Coaxum, R. A., Ugueto, A. M., Eckshtain, D., & Corteselli, K. A. (2019). Are psychotherapies for young people growing stronger? Tracking trends over time for youth anxiety, depression, attention-deficit/hyperactivity disorder, and conduct problems. *Perspectives on Psychological Science*, *14*(2), 216–237. doi: https://doi.org/10.1177/1745691618805436

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., ... Diener, E. (2020). From data to causes I: Building a general cross-lagged panel model (GCLM). *Organizational Research Methods*, *23*(4), 651–687. doi: https://doi.org/10.1177/1094428119847278