Journal of Behavioral Data Science, 2025, *5* (1), 1–16.
DOI: https://doi.org/10.35566/jbds/larzelere

An Innovation to Test Treatment X Pretest Interactions within Difference-in-Differences

Robert E. Larzelere* and Hua Lin

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

Abstract. We introduce a way to test Treatment X Pretest interactions within difference-in-differences (DID). Mathematically adding a Treatment X Pretest interaction to DID transforms the treatment estimate to an ANCOVA-type estimate, which differs from DID's estimate and is often biased against at-risk cases. Dual-centered ANCOVA duplicates DID's treatment estimate and can test whether that estimate varies by pretest scores. To illustrate, we test a Treatment X Pretest interaction for the effects of therapy for depression using the Fragile Families and Child Wellbeing longitudinal dataset. After centering posttest and pretest outcome data on pretest group means, DID and ANCOVA estimates are both equivalent to the original DID treatment estimate. ANCOVA of these dual-centered data can then test a Treatment X Pretest interaction.

Keywords: Difference-in-differences · Treatment X Pretest interaction · Longitudinal analyses · Causal validity · ANCOVA

1 Introduction

Longitudinal analyses that control for pre-existing differences with ANCOVA-type controls are biased against corrective actions (Larzelere, Lin, Payton, & Washburn, 2018) unless the covariates predict treatment condition perfectly (as in regression discontinuity designs). By definition, corrective actions are selected to reduce the poor prognosis of a presenting problem. Subsequent outcomes therefore constitute an unknown combination of the original poor prognosis of the problem and the extent to which the corrective action modified that prognosis. Controlling statistically for pre-existing differences via regression or matching reduces that selection bias, but rarely eliminates it. For example, a recent

* Correspondence concerning this article should be addressed to Robert E. Larzelere, Professor, Department of Human Development & Family, Science, 233 NRD Bldg., Oklahoma State University, Stillwater, OK 74078, United States. Email: robert.larzelere@okstate.edu Phone: (405) 744-2053. Fax: (405) 744-6344.

meta-analysis of efforts to improve low-performing schools found that analyses that used matching or regression methods predicted significantly worse effects than randomized studies on high-stakes math exams and marginally worse on language arts exams (Schueler, Asher, Larned, Mehrotra, & Pollard, 2021). That may be why 47% of the studies qualifying for that meta-analysis used difference-in-differences instead of regression-type controls to adjust for pre-existing differences.

There are two basic ways to analyze change in two-occasion longitudinal analyses: ANCOVA predicting residualized change scores, $Y_1|Y_0$ (Y_1 controlling for baseline Y_0), and difference-in-differences predicting simple difference scores, $Y_1 - Y_0$. ANCOVA has more statistical power (van Breukelen, 2013) but produces biased treatment estimates in non-randomized studies from invariant between-person differences (Berry & Willoughby, 2016; Hamaker, Kuiper, & Grasman, 2015) and measurement error (Huitema, 2011). Difference-in-differences overcomes these two biases, but is biased by any variations from its parallel-trends assumption. Some have recommended running both types of change-score analyses, either to bracket the true causal effect given some assumptions (Angrist & Pischke, 2009; Ding & Li, 2019) or to test robustness across alternative analyses (Duncan, Engel, Claessens, & Dowsett, 2014). A limitation of difference-in-differences has been its inability to test Treatment X Pretest interactions. For example, the meta-analysis of efforts to improve low-performing schools tested many moderators, but not whether the success of these efforts varied by the schools' previous performance on the outcomes (e.g., high-stakes testing). This article introduces a method to test whether treatment effects vary by pretest levels using difference-in-differences without inadvertently changing the treatment estimate to ANCOVA's estimate.

This article focuses on two-occasion data for two reasons. Many longitudinal studies have only two occasions (Usami, Todo, & Murayama, 2019), and these two change-score analyses are basic building blocks for more complex longitudinal analyses (Lin & Larzelere, 2024).

Treatment estimates become identical for the two change-score analyses after pretest means are equalized across treatment groups, but these robust estimates are not necessarily less biased. Different methods of equating pretest group means yield different treatment estimates (Lin & Larzelere, 2020). Pretest matching produces robust results that are equivalent to the original ANCOVA (Reichardt, 2019), which is unbiased only if the assumptions of the original ANCOVA are met (e.g., no measurement error in the covariates, equality of true pretest group means with each other: van Breukelen, 2013). Centering both posttest and pretest scores on pretest group means preserves everyone's difference score, rendering the treatment estimates robust and equivalent to the original difference-in-differences, which is unbiased under the assumption of parallel slopes under the null hypothesis. The two pairs of robust results therefore differ from each other as much as the original discrepancy between the two change-score analyses. But the dual-centered data can be analyzed with ANCOVA to

test a Treatment X Pretest interaction in a model duplicating the treatment effect from difference-in-differences (Lin & Larzelere, 2020).

1.1 Basics

Assume $X_{ij} = 1$ for treatment ($j = 2$), and $X_{ij} = 0$ for control ($j = 1$). Occasions are $t = 0$ (pretest) and $t = 1$ (posttest), with outcome variable Y_{ijt} for individual i within group j at occasion t. The equation for ANCOVA is:

$$Y_{ij1} = b_0 + b_1 X_{ij} + b_2 Y_{ij0} + e_{ij}. \tag{1}$$

The equation for difference-in-differences is:

$$Y_{ij1} - Y_{ij0} = \gamma_0 + \gamma_1 X_{ij} + \varepsilon_{ij}. \tag{2}$$

By adding Y_{ij0} to both sides of Equation (2), it can be shown that its treatment effect γ_1 is identical to the treatment effect b_1 in Equation (1) when b_2=1 in Equation (1). This is possible only when all $e_{ij} = 0$ or the variance of Y_{ij} is increasing over time. For the purposes of this article, we assume that some $e_{ij} > 0$ and that the variance of Y_{ij} is stable over time. Then the two treatment effect sizes equal each other ($b_1 = \gamma_1$) only if the pretest group means are equal to each other.

Dual-centered ANCOVA centers pretest and posttest scores on the pretest group means:

$$Y_{ij1} - \widehat{\mu}_{j0} = \omega_0 + \omega_1 X_{ij} + \omega_2 (Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}, \tag{3}$$

where the group-mean-centered pretest scores are the residuals τ_{ij} in the following equation:

$$Y_{ij0} = \widehat{\mu}_{j0} + \tau_{ij}. \tag{4}$$

Lin and Larzelere (2020) showed that, under the assumption of no pretest group mean differences, the treatment estimate in Equation (3) is identical to the treatment effect in difference-in-differences Equation (2), i.e., $\omega_1 = \gamma_1$. The $(Y_{ij0} - \widehat{\mu}_{j0})$ term is a generated regressor, however, which biases the standard error for the treatment effect ω_1 downward (Brorsen, Lin, & Larzelere, 2025; Pagan, 1984). The correct standard error can be obtained from Equation (2) or by analyzing Equations (3) and (4) together via two-stage least squares (Brorsen et al., 2025). Next we consider adding Treatment X Pretest interactions to the above analyses.

1.2 Treatment X Pretest Interactions

Standard ANCOVA. When there is a significant Treatment X Pretest interaction, treatment effects vary in magnitude and significance at different pretest scores (Huitema, 2011). Because significant interactions apply to both component predictors, the auto-regressive slope b_2 will then also vary significantly

across groups. A significant Treatment X Pretest interaction violates the ANCOVA assumption of homogeneity of the regression slope across groups. We follow Huitema (2011) and Lin (2020) in interpreting a significant Treatment X Pretest interaction.

Consider standard ANCOVA with a significant Treatment X Pretest interaction:

$$Y_{ij1} = b_0 + b_1 X_{ij} + b_2 Y_{ij0} + b_3 X_{ij} Y_{ij0} + e_{ij}. \tag{5}$$

Equation (5) can be re-arranged to indicate how the effect of Treatment X_{ij} varies by the pretest score (Lin, 2020):

$$Y_{ij1} = (b_0 + b_2 Y_{ij0}) + (b_1 + b_3 Y_{ij0}) X_{ij} + e_{ij}. \tag{6}$$

Reciprocally, the effect of the pretest Y_{ij0} on the posttest Y_{ij1} also varies by treatment condition (heterogeneity of regression slopes):

$$Y_{ij1} = (b_0 + b_1 X_{ij}) + (b_2 + b_3 X_{ij}) Y_{ij0} + e_{ij}. \tag{7}$$

One way to interpret significant Treatment X Pretest interactions is the Johnson and Neyman (1936) technique, which calculates regions of significant treatment effects at all pretest values. Alternatively, the picked-points analysis (Huitema, 2011; Lin, 2020) shows the estimated treatment effects at picked pretest values.

Equation (6) indicates that the estimated conditional effect of treatment X_{ij} on the posttest at any pretest score is:

$$\widehat{b}^*_{Tx} = b_1 + b_3 Y_{ij0}. \tag{8}$$

Reciprocally the conditional effect of the pretest on the posttest for either treatment condition according to Equation (7) is:

$$\widehat{b}^*_{lag1} = b_2 + b_3 X_{ij}. \tag{9}$$

Difference-in-Differences. To our knowledge, there is no generally accepted method of testing a Treatment X Pretest interaction within difference-in-differences without changing the main effect of treatment to ANCOVA's estimate. The reason is that tests of Treatment X Pretest interactions require both main effects to be included in the regression equation:

$$Y_{ij1} - Y_{ij0} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 Y_{ij0} + \gamma_3 X_{ij} Y_{ij0} + \varepsilon_{ij}. \tag{10}$$

But adding the pretest to both sides of Equation (10) yields the following:

$$Y_{ij1} = \gamma_0 + \gamma_1 X_{ij} + (1 + \gamma_2) Y_{ij0} + \gamma_3 X_{ij} Y_{ij0} + \varepsilon_{ij}. \tag{11}$$

Equation (11) is the same as Equation (5) for standard ANCOVA with a Treatment X Pretest interaction, with $b_2 = 1 + \gamma_2$. Therefore the treatment

effect γ_1 in Equations (10) and (11) is equivalent to ANCOVA's treatment effect b_1 in Equation (5). Omitting the auto-regressive term $\gamma_2 Y_{ij0}$ from Equation (10) is equivalent to fixing γ_2 to 0, which is usually nonsensical, since the pretest Y_{ij0} is one of the two components of the difference score being predicted. Fixing the slope coefficient γ_2 to 0 in Equation (10) is also equivalent to fixing the coefficient $(1+\gamma_2)$ to 1 in Equation (11), which makes the equations for ANCOVA and difference-in-difference identical. This is possible, however, only when the variance of the outcome scores is increasing over time or unless pretest scores predict posttest scores perfectly.

To add a Treatment X Pretest interaction to dual-centered ANCOVA in Equation (3), we apply the same steps as in Equations (5) through (9) for standard ANCOVA. In both cases, a significant interaction changes the unconditional marginal effects in Equations (1) and (2) to conditional effects that vary with pretest scores.

Adding a Treatment X Pretest interaction to Equation (3) for dual-centered ANCOVA yields:

$$Y_{ij1} - \widehat{\mu}_{j0} = \omega_0 + \omega_1 X_{ij} + \omega_2 (Y_{ij0} - \widehat{\mu}_{j0}) + \omega_3 X_{ij} (Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}. \tag{12}$$

Because dual-centered ANCOVA predicts the same treatment effect as difference-in-differences, the Treatment X Centered Pretest interaction can be interpreted in the same way as a Treatment X Pretest interaction in standard ANCOVA. Analyzing Equation (12) by itself yields the correct standard error for ω_3, according to our simulation (Lin, 2023). Equation (12) can be re-arranged to indicate how the treatment effect varies by the group-mean-centered pretest score (Lin, 2020).

$$Y_{ij1} - \widehat{\mu}_{j0} = (\omega_0 + \omega_2 [Y_{ij0} - \widehat{\mu}_{j0}]) + (\omega_1 + \omega_3 [Y_{ij0} - \widehat{\mu}_{j0}]) X_{ij} + \nu_{ij}. \tag{13}$$

Reciprocally, the effect of the group-mean-centered pretest score also varies by treatment condition:

$$Y_{ij1} - \widehat{\mu}_{j0} = (\omega_0 + \omega_1 X_{ij}) + (\omega_2 + \omega_3 X_{ij}) (Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}. \tag{14}$$

Equation (13) indicates that the effect of treatment on the pretest-group-mean-centered posttest at any group-mean-centered pretest score is:

$$\widehat{\omega}^*_{Tx} = \omega_1 + \omega_3 (Y_{ij0} - \widehat{\mu}_{j0}). \tag{15}$$

Reciprocally the effect of the group-mean-centered pretest on the centered posttest at either level of treatment according to Equation (14) is

$$\widehat{\omega}^*_{lag1} = \omega_2 + \omega_3 X_{ij}. \tag{16}$$

1.3 Illustrative Example

The following example estimates the effect of psychotherapy to treat depression in mothers from the Fragile Family & Child Wellbeing (FFCW) dataset. We selected this corrective action because its effectiveness has been documented in meta-analyses of hundreds of randomized trials (Cuijpers et al., 2023). Although these effect sizes shrink over time (Miguel et al., 2021) and in typical field implementations (Ormel, Hollon, Kessler, Cuijpers, & Monroe, 2022), there is no reason to think that such therapies are harmful on average.

We expect therapy to look more effective with difference-in-differences than with ANCOVA, as is typical for longitudinal analyses of corrective actions (Larzelere et al., 2018). We then illustrate how to use dual-centered ANCOVA to test the Treatment X Pretest interaction corresponding to the treatment estimate from the difference-in-differences model.

2 Methods

2.1 Participants.

The FFCW dataset started with baseline data on at-risk couples whose children were 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 (Time 1), and later when the children were approximately 1, 3, 5, and 9 years old (Times 2 to 5). The current example investigated the apparent effects of psychotherapy for maternal depression when their children were five years old, using data on maternal depression symptoms when their children were 5 and 9 years old (Time 4 and Time 5). At baseline (when the focal child was born), the 4566 mothers were 25.2 years old and had some college on average, and consisted of 21.0% White, 47.6% Black, 27.4% Hispanic, and 4.0% others. The sample size for this study consisted of the 3285 mothers with complete data on therapy for depression at Time 4 and on depression symptoms at Times 4 and 5. The data are available on the Open Science Framework Home website (`https://osf.io/532xt/?view_only=5857097b48034e7786a8933b4af22e3a`).

2.2 Measures

Depression treatment was based on maternal responses to questions about whether they had received any counseling or therapy in the past twelve months. "Yes" answers led to the question "Was this counseling or therapy for depression?" Mothers who reported receiving therapy for depression were contrasted with mothers who responded "No" to either of these questions.

Depression symptoms were assessed by maternal self-reports of relevant symptoms from the Composite International Diagnostic Interview--Short Form (CIDI-SF), Section A (Kessler, Andrews, Mroczek, Ustun, & Wittchen, 1998), a standardized survey instrument for assessing mental disorders. It uses two stem ques-

tions and four follow-up questions to identify possible eligibility for a Major Depressive Episode. Eligibility then led to eight symptom questions to determine depression severity. Sub-eligibility symptoms resulted in possible scores from 1 to 4. Four points were added to the number of the eight symptoms associated with a possible Major Depressive Episode. This produced a 13-point scale (0 to 12) for depression severity, with the majority of the scores being 0 (73.8% at Time 4; 73.9% at Time 5).

3 Results

Table 1 provides the means, standard deviations, and other descriptive statistics for therapy at Time 4 of the FFCW dataset and for depression symptoms at Times 4 and 5.

Table 1: Descriptive Statistics

	Treatment T4	N	Mean	SD	Minimum	Maximum
Depress T4	0	3078	1.54	3.36	0	12
	1	207	7.33	4.48	0	12
	Total	3285	1.9	3.72	0	12
Depress T5	0	3078	1.61	3.45	0	12
	1	207	5.1	4.94	0	12
	Total	3285	1.83	3.66	0	12

Note. T4 = Time 4 of the FFCW dataset. T5 = Time 5. Depress = Depression symptoms.

Prior to adding an interaction term, standard ANCOVA and difference-in-differences produced contradictory estimates of treatment effects, as is typical of longitudinal analyses of corrective actions (Larzelere et al., 2018). According to ANCOVA, therapy for depression led to more depression symptoms at Time 5 than predicted by initial symptoms at Time 4, b_1= 1.74, t (3284) = 6.59, $p <$ .001. In contrast, difference-in-differences indicated that depression symptoms decreased more following therapy than otherwise, γ_1= -2.31, t (3284) = -7.70, $p <$ .001. Because psychotherapy for depression has been shown to be effective in many randomized trials (Cuijpers et al., 2023), difference-in-differences may be less biased against corrective actions than ANCOVA. Most researchers, however, would also want to know whether these treatment effects vary by the level of presenting depression symptoms. We will illustrate the use of dual-centered ANCOVA to test a Treatment X Pretest interaction in difference-in-differences after summarizing a Treatment X Pretest interaction in standard ANCOVA for comparative purposes.

3.1 Standard ANCOVA

Analyzing the data with standard ANCOVA led to the following result from Equation (5):

$$Y_{ij1} = 1.13 + 2.54X_{ij} + .314Y_{ij0} - .119X_{ij}Y_{ij0} + e_{ij}. \tag{17}$$

indicating that therapy predicted worsening depression symptoms than controls, b_1= 2.54, $p < .001$, a harmful-looking effect that was reduced for those with worse initial symptoms, b_3= -.119, $p < .05$. Plugging coefficients into Equation (8) gives the magnitude of the estimated treatment effect for each pretest score:

$$\widehat{b}^*_{Tx} = 2.54 + (-.119)Y_{ij0}. \tag{18}$$

This signifies that the harmful-looking effect of therapy varied from 2.54 for those with pretest depression scores of 0 to a reduced harmful-looking treatment effect of only 1.11 for those with maximum pretest scores of 12. These effect sizes varied around the average treatment effect of 1.74 from standard ANCOVA before adding the interaction term.

Figure 1 uses picked-points analysis to show the conditional treatment effects predicted at the mean pretest scores for the treatment and control groups and at the maximum depression score (Lin, 2020). Figure 4 in Appendix A shows the 95% confidence intervals of these coefficients and the significance of these treatment effects at each pretest score. Next we illustrate similarities and differences in testing the same Treatment X Pretest interaction within difference-in-differences.

3.2 Difference-in-Differences via Dual-Centered ANCOVA

Using Equation (12), the results from dual-centered ANCOVA from the same data after centering all depression scores around their pretest group means, $Y_{ijt} - \widehat{\mu}_{j0}$, are

$$Y_{ij1} - \widehat{\mu}_{j0} = .10 + (-2.30)X_{ij} + .314(Y_{ij0} - \widehat{\mu}_{j0}) + (-.119)X_{ij}(Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}, \tag{19}$$

indicating that, for those with initial depression symptoms at their group mean ($Y_{ij0} - \widehat{\mu}_{j0} = 0$), depression symptoms decreased more for women in therapy than controls, ω_1= -2.30, $p < .001$, a beneficial-looking effect that was enhanced further for those with worse initial symptoms, ω_3= -.119, $p < .05$.

Using Equation (15), the estimated effect of therapy on the pretest-group-mean-centered posttest for any group-mean-centered pretest score was

$$\widehat{\omega}^*_{Tx} = -2.30 + (-.119)(Y_{ij0} - \widehat{\mu}_{j0}). \tag{20}$$

This signifies that therapy led to steeper decreases in depression symptoms than in controls, with that beneficial-looking effect varying from -2.12 for the minimum possible centered pretest score for the comparison group (0 – 1.5 = -1.5,

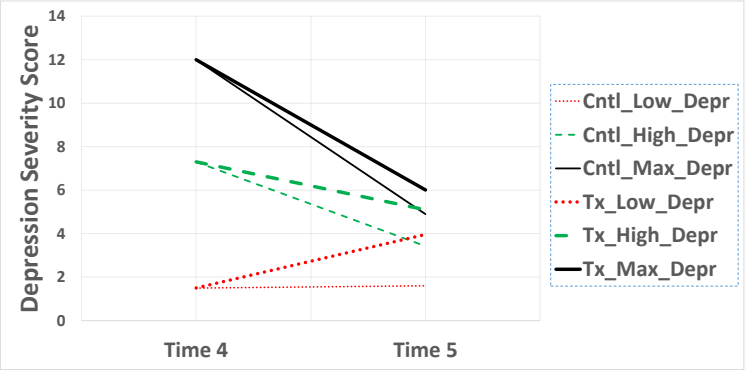

Figure 1: Predicted changes from Time 4 to Time 5 at different pretest scores according to standard ANCOVA (Low = mean pretest for controls, High = mean pretest for treatment, Max = maximum depression score) to illustrate the Treatment X Pretest interaction.

subtracting their mean pretest score) to a stronger beneficial-looking treatment effect of -2.86 for the maximum possible centered pretest score for the treatment group (12 – 7.3 = 4.7, subtracting their mean pretest score). These effect sizes varied around the average treatment effect of -2.31 from difference-in-differences before adding the interaction term.

This result and its confidence intervals are displayed in Figure 5 of Appendix A (Lin, 2020). Figure 2 uses picked-points analysis to illustrate how estimated treatment effects varied across the range of centered pretest scores that are possible in both treatment and comparison groups. Figure 3 illustrates the same treatment effects at the same picked pretest points after decentering all depression scores. This illustrates a potential problem with difference-in-differences in that its parallel-slopes assumption is less tenable at minimum and maximum scores. When centered pretest scores were at the minimum for the control group, they could not decrease further for that group, but could decrease further in the treatment group (a floor effect for the control group). In this case, however, this floor-effect bias is in the opposite direction of the Treatment X Pretest interaction and therefore does not invalidate it. (Therapy at a centered pretest of -1.5 [originally 5.8] decreased to a posttest mean of -2.49 [4.81 on original scale]. Controls at a centered pretest of -1.5 [originally 0.0] could not decrease, artificially increasing the extent to which therapy looked relatively effective at low depression levels. If controls could have decreased their centered depression pretest scores, the differential effectiveness would have been even smaller at low

depression levels, increasing the Treatment X Pretest interaction even more.) The ceiling effect bias was in the same direction as the Treatment X Pretest interaction, but was relatively minor as only 15 women in the therapy group had maximum posttest depression scores of 12 (7.2% of the therapy group, vs. 76.2% of controls with minimum posttest scores of 0).

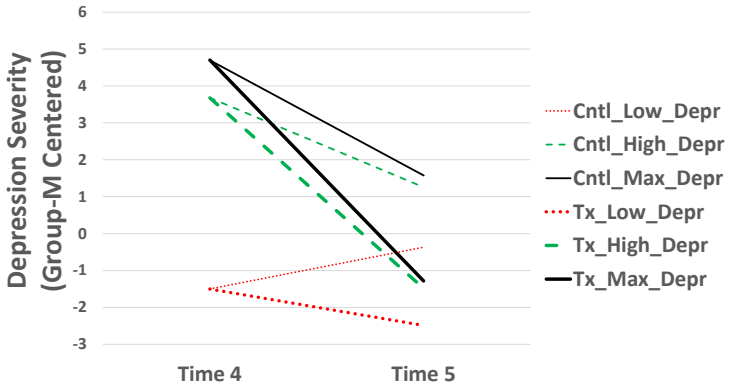

Figure 2: Predicted simple change scores from Time 4 to Time 5 for treatment vs. comparison groups at three levels of group-mean centered pretest scores, based on dual-centered ANCOVA (Low = minimum possible centered score for controls; High = one SD above the group mean pretest scores; Max = maximum possible centered pretest score for treatment).

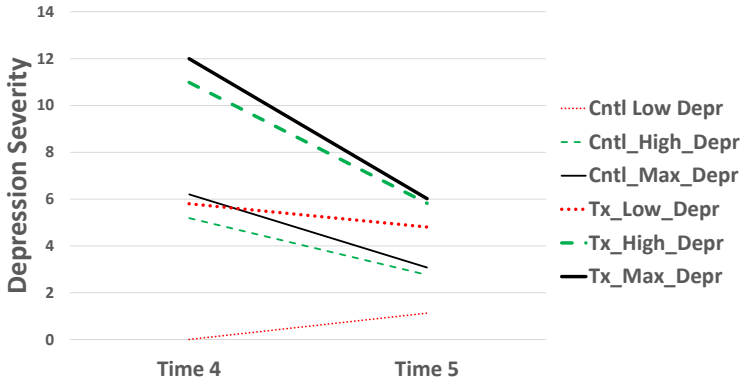

Figure 3: Predicted simple change scores from Time 4 to Time 5 for treatment vs. comparison groups at three levels of group-mean centered pretest scores according to dual-centered ANCOVA after decentering all scores (Low, High, & Max defined as in Figure 2).

4 Discussion

ANCOVA-type controls have been shown to be biased in longitudinal analyses (Berry & Willoughby, 2016; Hamaker et al., 2015; Hoffman, 2015), usually biased against corrective actions such as medical treatments and psychotherapy (Larzelere et al., 2018). This study demonstrates a novel way to overcome one disadvantage of the main alternative, difference-in-differences, which otherwise cannot test Treatment X Pretest interactions without changing the treatment effect to the estimate from ANCOVA. This innovation takes advantage of the fact that centering all longitudinal data around pretest group means makes the treatment effects of ANCOVA equal to estimates from difference-in-differences (Lin & Larzelere, 2020). This is called dual-centered ANCOVA in two-occasion analyses, which is used herein to test a Treatment X Pretest interaction corresponding to a difference-in-differences model.

We do not know of a better way to test Treatment X Pretest interactions in difference-in-differences. Without Treatment X Pretest interactions, difference-in-differences are limited to assuming that the estimated treatment effects are identical at every pretest score, an untenable assumption without sufficient evidence. When regression slopes are heterogeneous across treatment conditions, the effect of treatment also varies with the pretest score. For this situation, (Huitema, 2011, Chapter 11) showed how to calculate the conditional treatment effect at each level of pretest scores in standard ANCOVA. The lack of a parallel way to test Treatment X Pretest interactions in difference-in-differences appears

to be a limitation in such analyses, one that can be overcome after centering all data on the pretest group means.

Unless ANCOVA clearly produces a less-biased treatment estimate in longitudinal analyses, difference-in-differences should be used to test the robustness of the estimated treatment effect (Duncan et al., 2014), if not a less-biased estimate. The least-biased estimate is generally the one whose assumptions are best satisfied. From our experience, it is helpful to compare the plausibility of the no-treatment effect implied by their respective null hypotheses. A no-treatment effect in difference-in-differences assumes that the groups' average trends from pretest to posttest will be parallel to each other, with no shrinkage of the difference between group means. In contrast, the null hypothesis in ANCOVA assumes that any group difference on the pretest will spontaneously shrink from pretest to posttest according to regression toward the grand mean. This shrinkage is plausible in randomized trials when initial differences on the pretest group means are due only to random fluctuations (i.e., no true difference between the pretest group means). ANCOVA is also unbiased if the covariates fully determine treatment group assignment (van Breukelen, 2013). In many other applications, however, pretest group means reflect true differences as well as random fluctuations, and the covariates do not fully explain treatment assignment. The remaining bias is recognized as *residual confounding* by epidemiologists (Rothman, Greenland, & Lash, 2008), which often makes corrective actions such as therapy for depression look more harmful than they are (Larzelere et al., 2018). In contrast, difference-in-differences' treatment estimates are not biased by true differences that do not change from pretest to posttest nor by measurement error in the pretest, but it has its own biases in non-randomized studies (e.g., any variations from the parallel-slopes assumption not due to the treatment effect). Unless the original ANCOVA is less biased, difference-in-differences provides either a less biased treatment estimate or a test of that estimate's robustness (Duncan et al., 2014). Dual-centered ANCOVA can then be used to test a Treatment X Pretest interaction within difference-in-differences.

Acknowledgments

We gratefully acknowledge funding from research grant #R03 HD107307 from the National Institute of Child Health and Human Development and from the Oklahoma State University Foundation.

References

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

Berry, D., & Willoughby, M. T. (2016). 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

Brorsen, B. W., Lin, H., & Larzelere, R. E. (2025). Critique of enhanced power claimed for Quasi-ANCOVA and Dual-Centered ANCOVA. *PLOS ONE*, *20*(1), e0317860. doi: https://doi.org/10.1371/journal.pone.0317860

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Papola, D., ... Karyotaki, E. (2023). Psychological treatment of depression: A systematic overview of a 'meta-analytic research domain'. *Journal of Affective Disorders*, *335*, 141–151. doi: https://doi.org/10.1016/j.jad.2023.05.011

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, *27*(4), 605–615. doi: https://doi.org/10.1017/pan.2019.25

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. P. 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.

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Wiley. doi: https://doi.org/10.1002/9781118067475

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., & Wittchen, H. (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

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. (2020). Probing two-way moderation effects: A review of software to easily plot Johnson-Neyman figures. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(3), 494–502. doi: https://doi.org/10.1080/10705511.2020.1732826

Lin, H. (2023). *Simulation test of standard errors of regression coefficients in DC-ANCOVA with Treatment X Pretest interaction* (Tech. Rep.). Department of Human Development and Family Science, Oklahoma State University.

Lin, H., & Larzelere, R. (2024). Lord's paradox illustrated in three-wave longitudinal analyses: Cross lagged panel models versus linear latent growth models. *Journal of Behavioral Data Science*, *4*(2), 51–63. doi: https://doi.org/10.35566/jbds/lin

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*(1), 135–147. doi: https://doi.org/10.1016/j.adolescence.2020.11.001

Miguel, C., Karyotaki, E., Ciharova, M., Cristea, I. A., Penninx, B. W., & Cuijpers, P. (2021). Psychotherapy for comorbid depression and somatic disorders: a systematic review and meta-analysis. *Psychological Medicine*, *53*(6), 2503–2513. doi: https://doi.org/10.1017/s0033291721004414

Ormel, J., Hollon, S. D., Kessler, R. C., Cuijpers, P., & Monroe, S. M. (2022). More treatment but no less depression: The treatment-prevalence paradox. *Clinical Psychology Review*, *91*, 102111. doi: https://doi.org/10.1016/j.cpr.2021.102111

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, *25*(1), 221–247. doi: https://doi.org/10.2307/2648877

Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford.

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

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Wolters Kluwer.

Schueler, B. E., Asher, C. A., Larned, K. E., Mehrotra, S., & Pollard, C. (2021). Improving low-performing schools: A meta-analysis of impact evaluation studies. *American Educational Research Journal*, *59*(5), 975–1010. doi: https://doi.org/10.3102/00028312211060855

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. doi: https://doi.org/10.1371/journal.pone.0209133

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

Appendix A Supporting Information

Whereas Figures 1, 2, and 3 in the main article use picked-points analysis to illustrate the significant Treatment X Pretest interaction at selected pretest scores, the following Supporting Figures illustrate the magnitude and significance of the estimated treatment effect for each possible pretest score. These plots are based on the Johnson-Newman technique, including 95% confidence intervals for the estimated treatment effect at each pretest score (Lin, 2020). The two figures illustrate the contradictory results from the two basic change-score analyses that are typical of longitudinal analyses of corrective actions (Larzelere et al., 2018). According to ANCOVA, therapy for maternal depression at Time 4 in the Fragile Families dataset appears to result in increased depression symptoms four years

later at Time 5, controlling for depression symptoms at Time 4 (Supporting Figure 4). In contrast, dual-centered ANCOVA duplicates difference-in-differences by indicating that therapy for maternal depression reduces depression scores more than for the comparison group (Supporting Figure 5). In both cases, therapy appears to be significantly more effective at high levels of initial depression than at low levels of initial depression (reducing the harmful-looking effect in standard ANCOVA, but increasing the beneficial-looking effect in dual-centered ANCOVA).

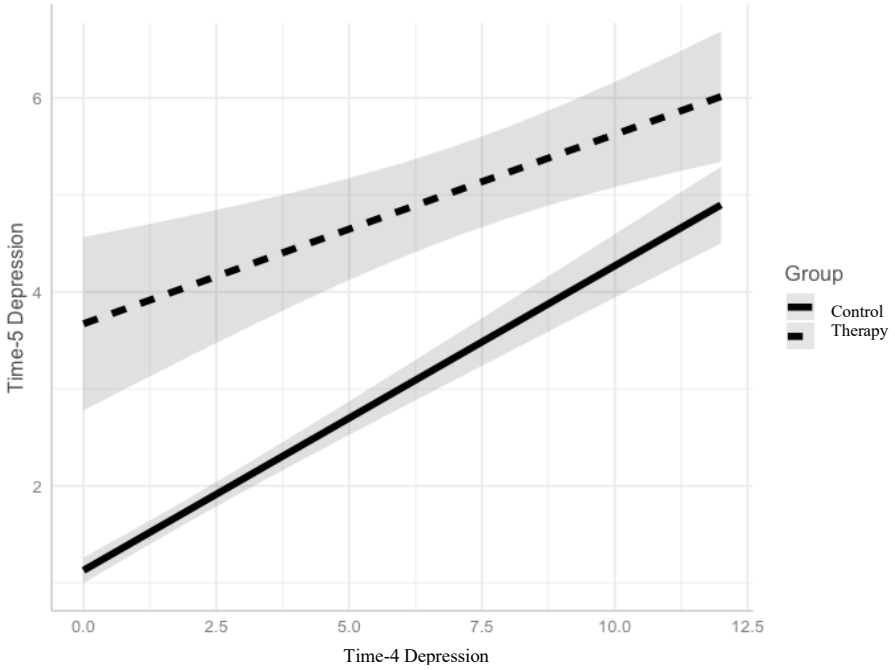

Figure 4: Predicted posttest depression scores for each pretest depression score for Therapy (dashed upper line) or Control (solid lower line) according to standard ANCOVA.

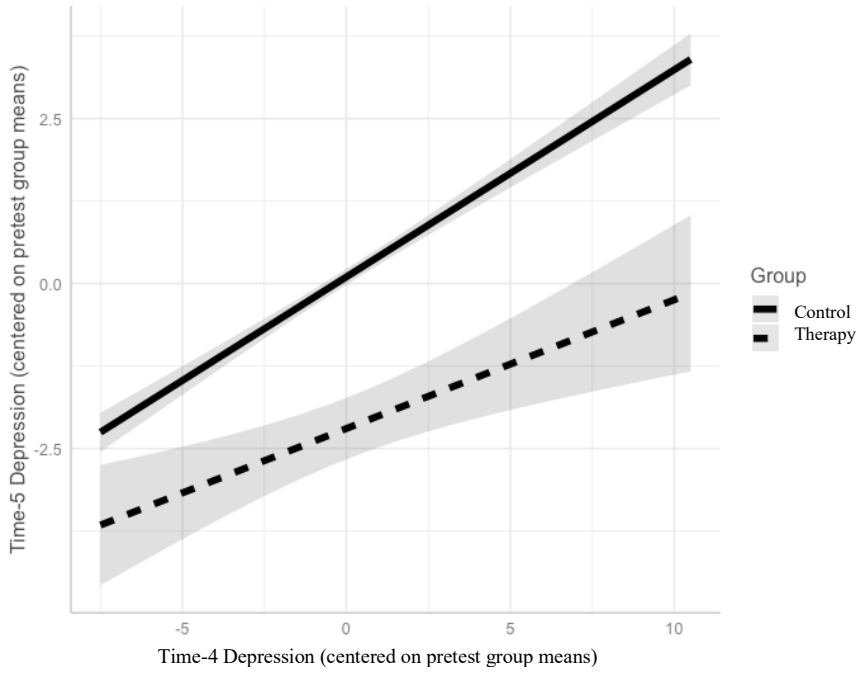

Figure 5: Predicted posttest depression score at Time-5 (centered on pretest group means at Time-4) for each pretest depression score (centered on pretest group means) according to dual-centered ANCOVA for Therapy (dashed lower line) and Control (solid upper line).