Robust and Pseudo-Robust Solutions to Lord's Paradox

Robert E. Larzelere and Hua Lin
Oklahoma State University

ANCOVA-type analyses of residualized-change scores and difference-score analyses of simple change scores often produce discrepant and even contradictory results in non-randomized pre-post studies, a problem known as Lord's Paradox (Lord, 1967). Duncan et al. (2014) called for developmental science to emulate econometrics in prioritizing robust results across different types of analyses and datasets. Robustness is especially impressive when the two+ analyses have contrasting biases, as is often the case for the two types of change-score analyses. Unfortunately, consistency across analyses of residualized and simple change scores can often occur without reducing the bias when Lord's Paradox applies, a consistency of results we call *pseudo-robustness*. This paper identifies four pseudo-robust solutions to Lord's Paradox. We limit ourselves to 2-occasion 2-group cases, consistent with our building-block approach. We need to understand simple building blocks more thoroughly to interpret more complex longitudinal analysis appropriately.

Pseudo-Robustness #1. ANCOVA and difference-score analysis controlling for the pretest

Adding the pretest to a difference-score analysis makes the treatment effect identical to standard ANCOVA (Allison, 1990).

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

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

The equation for difference-score analysis is:

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

Adding the pretest as a covariate to Equation (2) yields:

$$Y_{ij1} - Y_{ijo} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 Y_{ij0} + \varepsilon_{ij} \tag{3}$$

Adding the pretest score Y_{j0} to both sides of Equation (3) yields:

$$Y_{ij1} = \gamma_0 + \gamma_1 X_{ij} + (1+\gamma_2) Y_{ij0} + \varepsilon_{ij} \tag{4}$$

But Equation (4) is identical to Equation (1) for ANCOVA, with $b_2 = 1 + \gamma_2$. Treatment effects will therefore equal each other, i.e., $\gamma_1 = b_1$.

Contrast Between Treatment Effects

Paper prepared for the Modern Modeling Methods conference, U. of Conn., Storrs, CT, 6/28/23. We are grateful for funding from NICHD grant #5 R03 HD107307.

Other examples of pseudo-robustness follow from an equation for the difference between the estimated treatment effects for ANCOVA (b_1) and for difference-score analysis (γ_1) as follows.

ANCOVA (Equation 1)

The expected outcome in the treatment group ($X_j = 1$) for ANCOVA in Equation (1) is

$$E(Y_{i21}) = E(b_0 + b_1 + b_2 Y_{i20}) \tag{5}$$

and for the control group ($X_j = 0$), it is

$$E(Y_{i11}) = E(b_0 + b_2 Y_{i10}) \tag{6}$$

The treatment effect b_1 is the difference between the two expected outcomes:

$$E(Y_{i21}) - E(Y_{i11}) = E(b_0 + b_1 + b_2 Y_{i20}) - E(b_0 + b_2 Y_{i10}) \tag{7}$$

$$E(Y_{i21}) - E(Y_{i11}) = b_1 + b_2[E(Y_{i20}) - E(Y_{i10})] \tag{8}$$

Solving for the treatment effect b_1

$$b_1 = [E(Y_{i21}) - E(Y_{i11})] - b_2[E(Y_{i20}) - E(Y_{i10})] \tag{9}$$

$$b_1 = (\bar{Y}_{i21} - \bar{Y}_{i11}) - b_2(\bar{Y}_{i20} - \bar{Y}_{i10}) \tag{10}$$

Difference-score analysis (Equation 2)

The expected outcomes in the treatment group ($X_j = 1$) for the difference-score Equation (2) is

$$E(Y_{i21} - Y_{i20}) = E(\gamma_0 + \gamma_1) \tag{11}$$

and for the control group ($X_j = 0$), it is

$$E(Y_{i11} - Y_{i10}) = E(\gamma_0) \tag{12}$$

The treatment effect γ_1 is the difference between the treatment and control groups

$$E(Y_{i21} - Y_{i20}) - E(Y_{i11} - Y_{i10}) = \gamma_1 \tag{13}$$

Thus the expected treatment effect is

$$\gamma_1 = (\bar{Y}_{i21} - \bar{Y}_{i11}) - (\bar{Y}_{i20} - \bar{Y}_{i10}) \tag{14}$$

The difference between the expected value of the treatment effect for the two types of change-score analyses is

$$\gamma_1 - b_1 = [(\bar{Y}_{i21} - \bar{Y}_{i11}) - (\bar{Y}_{i20} - \bar{Y}_{i10})] - [(\bar{Y}_{i21} - \bar{Y}_{i11}) - b_2(\bar{Y}_{i20} - \bar{Y}_{i10})] \tag{15}$$

$$\gamma_1 - b_1 = -(\bar{Y}_{i20} - \bar{Y}_{i10}) + b_2(\bar{Y}_{i20} - \bar{Y}_{i10}) \tag{16}$$

$$= (b_2 - 1)(\bar{Y}_{i20} - \bar{Y}_{i10}) \tag{17}$$

The two treatment estimates (γ_1 & b_1) are therefore equal whenever the pretest group means equal each other or when $b_2 = 1$, which can lead to pseudo-robustness.

Pseudo-robustness #2: Matching

The most common way to make the pretest group means equal is by matching. According to Equation (17), that can lead to pseudo-robustness, which can create an illusion of stronger causal validity from equivalent results from both change-score analyses, whether that consistent effect is unbiased or not. Lin (Lin, 2018; Lin & Larzelere, 2020) tested pretest matching with simulated data. To our knowledge, Lin (2018) was the first to investigate Lord's paradox with two datasets simulated to fit the null hypothesis for *both* analyses of the paradox (ANCOVA and difference-score analysis). The robust results produced by pretest matching replicated the treatment effect for ANCOVA (see Table 1). This is expected because matching and ANCOVA-type statistical controls are equivalent under some assumptions (Reichardt, 2019). The resulting robust treatment effect across the two change-score analyses were unbiased if the null hypothesis for ANCOVA was unbiased, but it remained as biased as ANCOVA if the difference-score null hypothesis was unbiased (i.e., parallel slopes from pretest to posttest). Consistent with other critiques of analyses of residualized change (Berry & Willoughby, 2017; Hamaker, Kuiper, & Grasman, 2015; Hoffman, 2015), the bias from matching generally makes corrective actions appear to be harmful, ranging from medication and therapy treatments for depression in mothers (Table 1, from Lin, 2018) to corrective actions by parents (Larzelere, Lin, Payton, & Washburn, 2018; Lin & Larzelere, 2020). Nonetheless, several studies have claimed to produce more causally valid results from matching, especially after entropy-score matching (Kang, 2022a, 2022b) or propensity-score matching (Haviland, Nagin, & Rosenbaum, 2007).

For example, Kang (2022a, 2022b) commendably implemented several innovations to enhance the causal validity of the effect of spanking on subsequent child outcomes. She dropped cases with overly frequent spanking (> 2 times per week), used entropy balancing for matching (an improvement on propensity-score matching), and demonstrated exact robustness across lagged-dependent-variable regression and difference-score regression analyses. Because entropy matching makes pretest group means equal, the difference-score analyses (Time-2 outcome score minus Time-1 outcome score) duplicated the analyses of residualized scores, as shown in Equation (4).

Indeed, Kang (2022b) reported identical causally relevant coefficients in difference-score analyses and ANCOVA after entropy-score matching (her Table 3 vs. Appendix for "spanked last week vs. not" in the full sample in Kang, 2022b). When the two types of change-score analyses were used on a subsample of the matched data ("spanking <2+ times" in Kang, 2022b), the agreement between the two change-score analyses was only approximate, probably because the pretest group means were no longer exactly identical. Kang (2022a, Table 3, p. 52) obtained identical treatment effects across residualized and difference-score analyses for the same reasons.

Like Kang (2022a, 2022b), Haviland et al. (2007) combined several statistical strategies to enhance the validity of causal inferences. They matched on propensity scores within two of the three developmental trajectories that had substantial overlapping cases on propensity scores. Because the propensity scores balanced those who joined a gang at age 14 on preceding self-reported violence at ages 10, 11, 12, and 13, propensity-score matching nearly equated the joiners vs. non-joiners on those four pre-treatment measures. After propensity-score matching, they found that the joiners were significantly more violent at ages 14 and 15 than the matched non-joiners. They also used a regression model of the propensity-matched data, controlling for the 12 covariates used to calculate propensity scores, including the four annual self-reported

violence scores from ages 10 to 13. Table 1 shows that the confidence intervals were almost exactly the same for the two change-score methods. (They actually reported group differences on the post-test, which is nearly equivalent to a difference-score estimator because the mean pretest scores were balanced across ages 10 to 13.) In Haviland et al. (2007), the *p* values for the two types of change were very close to each other at each post-treatment age from 14 to 17. The *p* values were slightly smaller for a nonparametric equivalent of predicting residualized scores ("Level of violence" in their Table 5) than for "change in violence," presumably reflecting the lower statistical power of difference-score analyses.

The robust effects in Haviland et al. (2007) seem plausible, because joining a gang seems likely to increase the violent actions of new gang members. However, Lin (2018) attempted to apply their methods to medication and psychotherapy treatments for depression in the mothers of the Fragile Families dataset. She also obtained robust results with propensity-score matched groups (see Table 1), but those robust results all indicated that medication and therapy made depression symptoms worse, nearly consistent with the results of applying ANCOVA to the original data. In both examples, propensity-score matching has the same bias as ANCOVA, a bias in the direction of the two group's initial differences in the outcome variable. (i.e., gang joiners were more violent at the pretest than non-joiners and mothers receiving treatment were more depressed initially than those not receiving treatment for depression, prior to propensity-score adjustments.)

The tendency for propensity-score and entropy-score matching to create consistent results that duplicate the bias in residualized-change scores may explain why three studies have found that interventions thought to reduce child abuse instead predict increased rates of child abuse and injuries requiring a hospital visit. After matching on propensity scores, Matone et al. (2012) found that home visiting was associated with a higher rate of hospital visits for physical injuries than the matched comparison group (415 per 1000 vs. 364 per 1000, $p < .0001$), a difference due mostly to superficial injuries. Later Matone et al. (2018) used entropy matching to compare hospital visits with abuse-related injuries from families in three preventive interventions compared to matched controls. Like their earlier study, the Nurse Family Practitioner home visiting program predicted significantly higher rates of severe injuries: OR = 1.32 (95% CI {1.08, 1.62}). Parents as Teachers and Early Head Start had larger odds ratios compared to matched controls: (OR = 4.11 [1.60, 10.55] and OR = 3.15 [1.412, 7.06], respectively). More recently, Holland et al. (2022) found evidence that surveillance bias could partially account for these results, but that that home visiting still predicted higher rates of child abuse reports that were considered worth investigating after home visiting ended (4.0% vs. 2.9% of families, $p < .001$). Home visiting was also linked to more children being removed from the home overall (2.8% vs. 2.1%, $p < .001$). This study also claimed to be the first to use difference-in-differences to estimate surveillance bias and child abuse allegations, but we have shown here that the results of difference-score analyses duplicate the results of residualized-score analyses after matching on pretest scores. These replicated studies illustrate the importance of recognizing the possibility of biases against corrective actions in studies that balance the treatment and comparison groups based on propensity-score or entropy-score matching. Otherwise these replicated studies could be used to discourage home visiting, Early Head Start, and Parents as Teachers due to a bias in residualized-score analyses that has not been eliminated by matching on propensity scores or entropy scores.

Pseudo-robustness #3: Dual-centered data

Lin (2018; Lin & Larzelere, 2020) extended Huitema's (2011) quasi-ANCOVA to dual-centered ANCOVA to adjust for pretest group differences. It retains everyone's difference score by centering the post-test scores as well as the pre-test scores around their pretest group means. Then the robust results for difference-score analyses and ANCOVAs of the dual-centered data replicate the estimated treatment effect from the original difference-score analyses. Therefore the robust results after this dual-centering is unbiased only when the original difference-score analysis is unbiased (Lin, 2018; Lin & Larzelere, 2020) .

Pseudo-robustness #4: Fan-shaped increase in variance over time

Equation (17) also implies that the two estimates of a treatment effect will agree with each other when the within-group slope coefficient $b_2 = 1$. When Y_0 and Y_1 are both standardized around their own means and standard deviations, b_2 is the within-group correlation between the pretest and the posttest, at least under the null hypothesis. In that case, $b_2 = 1.00$ only in the limiting case of perfect stability of each individual case from pretest to posttest z scores. Change-score analyses could be used, however, on raw outcome scores. The within-group unstandardized auto-regressive coefficient could equal 1.00 given increasing variance from Time 1 to Time 2. In that case, treatment estimates from the two change-score analyses will necessarily equal each other. When this occurs, the standardized regression coefficient predicting post-test scores Y_{ij1} from Y_{ij0} will equal the ratio of the standard deviations of the pretest and posttest scores, $\beta_2 = \frac{s_{Y_{ij0}}}{s_{Y_{ij1}}}$. An example in which the two change-score score analyses yield very similar treatment effects for this reason is the second example (Electoral Returns to Beneficial Policy) in Ding & Li (2019). The unstandardized slope coefficient was .997, which resulted in similar treatment effects: $d_1 = 7.14$ vs. $b_1 = 7.12$ in difference-score vs. residualized-score analyses, respectively. The equivalence of these results is not surprising because the treatment effects from ANCOVA and difference-score analyses are identical when the slope coefficient $b_2 = 1$. In addition, we can use the ratio above to test whether increasing variance over time explains robust results from the two change-score analyses, such as a study of the effects of parental disciplinary tactics on toddler outcomes over two months (Larzelere, Knowles, Henry, & Ritchie, 2018). In that study $s_{Y_{ij0}} < s_{Y_{ij1}}$, but the ratios were larger than the unadjusted correlations between the pretest and the posttest: externalizing: 7.39/8.00 = .92 vs. $r = .79$; internalizing: 5.95/6.52 = .91 vs. $r = .69$; total problems: 16.87/17.69 = .94 vs. $r = .75$. It would have been preferable to compare the ratios with the standardized coefficient used in the final models after all covariates were included. Nonetheless, this illustrates a case where the increasing variance of the outcome variable over time could have at least partially explained the unusual robustness in results across the two change-score analyses.

Table 2
Equivalence of Apparent Effects of Treatment Effects in Two Change-Score Analyses After Matching

	Pretest difference	*r(pre,post)*	ANCOVA	Difference-score analysis	*N*
Lord's paradox data (to fit null hypothesis for difference-score analysis)					
Original data	-29.99***	.48	-15.60***	-.02	1000
Matched data	-.09	.48	-15.58***	-15.53***	334
Dual-centered	0	.48	-.01	-.01	1000
"Reversed" Lord's paradox data (to fit null hypothesis for ANCOVA)					
Original data	-30.02	.48	.02	15.61***	1000
Matched data	-.09	.48	.06	.11	302
Dual-centered	0	.48	15.61***	15.61***	1000
Medication treatment for depression (for mothers in FFCW dataset)					
Original data	5.53		1.79***	–1.87***	3515
Matched data	.17		1.49**	1.38*	970
Dual-centered	0		–1.95***	–1.95***	3515
Propensity-score matching	.26		1.24*	1.12	388
Psychotherapy for depression (for mothers in FFCW dataset)					
Original data	5.53		1.74***	– 2.31***	3515
Matched data	.36		1.43**	1.18*	1049
Dual-centered	0		– 2.30***	– 2.31***	3515
Propensity-score matching	.28		1.24*	1.02*	802
Effect of joining a gang at age 14 on subsequent violence (Haviland et al., 2007)					
Propensity-score matching	$\bar{d}_0$ at ages 10-13: .24; $\bar{d}_0$ PrS at ages 10-13: .03		(CI: 14 | 12 covariates) (.25, 1.02)	(CI: post-test at 14) (.25, 1.00)	59 of 551 joiners
Propensity-score matching			(CI: 15 | 12 covariates) (.12, 1.14)	(CI: post-test at 15) (.14, 1.16)	53 joiners
Propensity-score matching			(CI: 14-17 | 12 covariates) (.02, .76)	(CI: 14-17 | 12 covariates) (.08, .79)	59 joiners

References

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. In C. Clogg (Ed.), *Sociological methodology 1990* (pp. 93-114). Oxford, UK: Blackwell.

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: 10.1111/cdev.12660

Ding, P., & Li, F. (2019). A Bracketing Relationship between Difference-in-Differences and Lagged-Dependent-Variable Adjustment. *Political Analysis, 27*(4), 605-615. doi: 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: 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: 10.1037/a0038889

Haviland, A. M., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*(3), 247-267. doi: 10.1037/1082-989X.12.3.247

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

Holland, M. L., Esserman, D., Taylor, R. M., Flaherty, S., & Leventhal, J. M. (2022). Estimating Surveillance Bias in Child Maltreatment Reporting During Home Visiting Program Involvement. *Child Maltreat*, 10775595221118606. doi: 10.1177/10775595221118606

Huitema, B. E. (2011). *The analysis of covariance and alternatives* (2nd ed.). Hoboken, NJ: Wiley.

Kang, J. (2022a). Spanking and children's early academic skills: Strengthening causal estimates. *Early Childhood Research Quarterly, 61*, 47-57. doi: 10.1016/j.ecresq.2022.05.005

Kang, J. (2022b). Spanking and children's social competence: Evidence from a US kindergarten cohort study. *Child Abuse Negl, 132*, 105817. doi: 10.1016/j.chiabu.2022.105817

Larzelere, R. E., Knowles, S. J., Henry, C. S., & Ritchie, K. L. (2018). Immediate and long-term effectiveness of disciplinary tactics by type of toddler noncompliance. *Parenting: Science & Practice, 18*(3), 141-171. doi: 10.1080/15295192.2018.1465304

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: 10.1037/arc0000052

Lin, H. (2018). *Revealing and resolving contradictory ways to reduce selection bias to enhance the validity of causal inferences from non-randomized longitudinal data.* (Doctoral Dissertation), Oklahoma State University, Stillwater, OK.

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: 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: 10.1037/h0025105

Matone, M., Kellom, K., Griffis, H., Quarshie, W., Faerber, J., Gierlach, P., . . . Cronholm, P. F. (2018). A mixed methods evaluation of early childhood abuse prevention within evidence-based home visiting programs. *Maternal and Child Health Journal, 22*(Suppl 1), 79-91. doi: https://doi.org/10.1007/s10995-018-2530-1

Matone, M., O'Reilly, A. L. R., Luan, X. Q., Localio, R., & Rubin, D. M. (2012). Home visitation program effectiveness and the influence of community behavioral norms: a propensity score matched analysis of prenatal smoking cessation. *BMC Public Health, 12*. doi: 10.1186/1471-2458-12-1016

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