Difference-in-Slopes: Causal Effects on Descriptive Associations*

Alexander Coppock† Gustavo Diaz‡

May 13, 2026

Abstract

Randomized treatments may change not only the levels of outcomes, but also the associations between them. This article examines the difference-in-slopes (DiS) estimand: the within-arm slope of the best linear predictor of one post-treatment outcome (Y_2) from another (Y_1), taken in the treatment group minus the control group. The corresponding regression estimator $\hat{Y}_2 = \beta_0 + \beta_1 Y_1 + \beta_2 Z + \beta_3 Z \cdot Y_1 + \varepsilon$ looks forbidden to researchers trained to avoid post-treatment bias. However, the DiS is identified under standard experimental assumptions (random assignment and SUTVA), and we prove that its plug-in estimator is numerically equal to the OLS interaction coefficient $\hat{\beta}_3$. We illustrate the method with reanalyses of three experiments: a persuasion experiment in which treatment tightened ideological constraint ($\widehat{\text{DiS}} = 0.43$, $p < 0.001$), an affective polarization megastudy in which an identity intervention loosened the Democrat–Republican thermometer coupling ($\widehat{\text{DiS}} = 0.204$, $p < 0.001$), and a social media experiment in which Facebook deactivation sharpened the attitude-behavior association ($\widehat{\text{DiS}} = 0.032$, $p = 0.003$).

*We thank Hakeem Jefferson for the spirited discussion that sparked this research. Replication materials are available at [REPOSITORY].

†Associate Professor, Department of Political Science, Northwestern University. a.coppock@northwestern.edu.

‡Assistant Professor of Instruction, Department of Political Science, Northwestern University. gustavo.diaz@northwestern.edu.

1 Introduction

Consider a researcher who collects two post-treatment outcomes, Y_1 and Y_2, in an experiment where treatment Z is randomly assigned. The researcher wants to know whether Z changed not just the level of each outcome but the association between them: does treatment make Y_2 more or less correlated with Y_1? Following the literature on regression adjustment for experiments (Bowers, 2011; Lin, 2013), one natural strategy is to estimate

$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 Z + \beta_3 Z \cdot Y_1 + \varepsilon \tag{1}$$

via OLS regression and inspect the coefficient β_3 on the interaction term.

This regression looks dangerous: Y_1 is post-treatment. Previous work has correctly warned that conditioning on a post-treatment variable biases estimates of treatment effects (Montgomery, Nyhan, & Torres, 2018; Rosenbaum, 1984). The warning is right, and it applies to the coefficient on Z in this regression.

The interaction coefficient, however, estimates a coherent estimand that we call the difference-in-slopes (DiS). The DiS is a function only of the joint distribution of $(Y_1(z), Y_2(z))$ within each arm z. The estimand does not involve any cross-counterfactual quantities like $\text{Cov}(Y_1(1), Y_2(0))$, $\text{Cov}(Y_1(1), Y_1(0))$, or any moment that mixes potential outcomes across arms. Random assignment identifies each arm's joint distribution separately, so the DiS is identified without any extra assumptions beyond those typically invoked for randomized experiments.

What's possibly unfamiliar is how the same regression produces one contaminated coefficient and one clean one. Researchers who want to estimate the ATE of Z on Y_2 should use the difference-in-means, the marginal regression $Y_2 = \beta_0 + \beta_1 Z + \varepsilon$, or an adjustment approach that does not condition on any post-treatment variables. Researchers who want to estimate the DiS should use $Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 Z + \beta_3 Z \cdot Y_1 + \varepsilon$ to estimate the interaction coefficient and ignore the other terms. The post-treatment bias literature has made

researchers reluctant to run such regressions; the point of this paper is that the reluctance should be estimand-specific.

The DiS is an interesting scientific estimand in its own right. Researchers studying ideological constraint want to know whether a treatment tightens the relationship between attitudes on related issues (Converse, 1964). Researchers studying persuasion want to know whether a treatment aligns attitudes more closely with behavior (Glasman & Albarracín, 2006). Researchers studying affective polarization want to know whether a treatment changes how strongly positive affect toward one party predicts negative affect toward the other (Iyengar, Sood, & Lelkes, 2012). In each setting, the scientific question concerns how the structure of the joint outcome distribution changes, not just how the mean of each outcome changes.

This paper makes three contributions. First, we give the DiS a precise potential outcomes definition and show why random assignment identifies the $Z \cdot Y_1$ coefficient while leaving the Z coefficient biased. Second, we prove that the $Z \cdot Y_1$ coefficient from OLS regression is numerically equivalent to the plug-in estimator. Third, we apply the procedure in reanalyses of three experimental datasets covering ideological constraint within a persuasion experiment (Coppock, Ekins, & Kirby, 2018; Coppock & Green, 2022), structural decoupling in an affective polarization megastudy (Voelkel et al., 2024), and attitude-behavior consistency in a Facebook deactivation experiment (Allcott, Braghieri, Eichmeyer, & Gentzkow, 2020). Section 2 defines the estimand and situates it among related treatment effects. Section 3 proves identification, derives the OLS equivalence, and clarifies the interpretation of the remaining regression coefficients. Section 4 presents the applications. Section 5 concludes.

2 The Estimand

Let $Z_i \in \{0, 1\}$ denote random treatment assignment for unit i, and let $Y_{1i}(z)$ and $Y_{2i}(z)$ denote potential outcomes under assignment $z \in \{0, 1\}$. We assume SUTVA throughout.

Both outcomes may be affected by Z in any direction and magnitude; we impose no restriction on treatment effects on Y_1 or Y_2 individually. The sequencing between Y_1 and Y_2 is trivial and their subscripts could be flipped.

For each arm z, the slope of the best linear predictor of $Y_2(z)$ from $Y_1(z)$ is

$$\beta(z) = \frac{\text{Cov}(Y_1(z),\, Y_2(z))}{\text{Var}(Y_1(z))}, \tag{2}$$

provided $\text{Var}(Y_1(z)) > 0$. This coefficient predicts $Y_2(z)$ from $Y_1(z)$ within arm z. It is a feature of the arm-specific joint distribution of potential outcomes, a descriptive summary of how $Y_1(z)$ and $Y_2(z)$ co-vary, and not a causal claim about the effect of Y_1 on Y_2.

The difference-in-slopes estimand is

$$\text{DiS} = \beta(1) - \beta(0) = \frac{\text{Cov}(Y_1(1),\, Y_2(1))}{\text{Var}(Y_1(1))} - \frac{\text{Cov}(Y_1(0),\, Y_2(0))}{\text{Var}(Y_1(0))}. \tag{3}$$

A positive DiS means the conditional relationship between Y_1 and Y_2 is steeper in the treatment arm. $\text{DiS} = 0$ means treatment did not change the linear conditional relationship, even if it shifted both means. Just as $E[Y_i(1)]$ and $E[Y_i(0)]$ are each descriptive summaries but their difference is causal, the within-arm slopes are descriptive and the difference-in-slopes across arms is causal. In this sense, the DiS belongs to the family of treatment effects on distributional features beyond means: effects on variances (Heckman, Smith, & Clements, 1997), quantiles, and correlations (Bloome & Schrage, 2021) all involve identified arm-specific comparisons. Unlike mediation analysis, the DiS does not decompose the effect of Z on Y_2 into a path through Y_1 and a direct path; such a decomposition would require sequential ignorability (Imai, Keele, & Tingley, 2010), which no experimental design guarantees.

A closely related estimand is the difference in within-arm Pearson correlations, $\rho_1 - \rho_0$. Because $\beta(z) = \rho_z \cdot \text{SD}(Y_2(z))/\text{SD}(Y_1(z))$, the slope and correlation diverge when treatment shifts marginal variances. This paper focuses on the slope; the correlation alternative and its inference are discussed in Appendix A2.

3 Identification and Estimation

3.1 The plug-in estimator

Quantities that require mixing potential outcomes within a unit, like individual treatment effects $Y_i(1)-Y_i(0)$, are not point-identified from observed data. The DiS is not such a quantity: it combines two arm-specific joint distributions, each delivered by random assignment without confounding. Identifying DiS therefore reduces to identifying the arm-specific moments in equation (3).

Under random assignment,[1] for each $z \in \{0, 1\}$:

$$E[Y_1(z) \cdot Y_2(z)] = E[Y_1 \cdot Y_2 \mid Z = z], \tag{4}$$

$$E[Y_1(z)] = E[Y_1 \mid Z = z], \tag{5}$$

$$E[Y_2(z)] = E[Y_2 \mid Z = z], \tag{6}$$

$$E[Y_1(z)^2] = E[Y_1^2 \mid Z = z]. \tag{7}$$

All four quantities are identified from observed data, and the covariance and variance follow immediately:

$$\mathrm{Cov}(Y_1(z), Y_2(z)) = E[Y_1 Y_2 \mid Z = z] - E[Y_1 \mid Z = z] \cdot E[Y_2 \mid Z = z],$$
$$\mathrm{Var}(Y_1(z)) = E[Y_1^2 \mid Z = z] - \big(E[Y_1 \mid Z = z]\big)^2.$$

The plug-in estimator replaces each population moment with its sample analog computed over units in arm z:

$$\widehat{\mathrm{DiS}} = \hat{\beta}(1) - \hat{\beta}(0), \qquad \hat{\beta}(z) = \frac{\widehat{\mathrm{Cov}}(Y_1, Y_2 \mid Z = z)}{\widehat{\mathrm{Var}}(Y_1 \mid Z = z)}. \tag{8}$$

[1] We state identification for randomized assignment mechanisms in which Z is marginally independent of the potential outcomes (Bernoulli, complete, or block-randomized with constant per-block assignment probability). For unequal-probability designs (e.g., block-randomized with varying rates), the same logic identifies the moments arm-by-arm using inverse-probability weighting or block-stratified analysis.

Under random assignment, arm-specific sample moments converge to their population analogs, so $\widehat{\text{DiS}}$ is consistent for DiS.

3.2 Numerical equivalence with OLS

Computing $\widehat{\text{DiS}}$ from equation (8) requires extracting arm-specific sample covariances and variances separately. We show next an easier way: the $Z \cdot Y_1$ coefficient from a single OLS regression is numerically identical to $\widehat{\text{DiS}}$.

Proposition 1. *Let $\hat{\beta}_3$ denote the OLS coefficient on $Z \cdot Y_1$ from the regression $Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 Z + \beta_3 Z \cdot Y_1 + \varepsilon$. Then $\hat{\beta}_3 = \widehat{DiS}$.*

The proof is in Appendix A1. The key step is that the column space of the pooled design matrix is equivalent to the column space of two arm-specific design matrices stacked, so the pooled OLS criterion decomposes into two independent arm-specific OLS problems whose solutions are exactly $\hat{\beta}(0)$ and $\hat{\beta}(1)$. This decomposition parallels the argument in Ding, Feller, and Miratrix (2019) showing that randomization-based and regression-based estimators of treatment effect variation are numerically equivalent.

In practice, regressing Y_2 on Y_1, Z, and $Z \cdot Y_1$ and reading the coefficient on the interaction term gives $\widehat{\text{DiS}}$; in R, `lm_robust(Y2 ~ Y1 + Z + Y1:Z, data = df)` from the `estimatr` package. HC2 standard errors provide asymptotically valid inference and handle the structural heteroskedasticity that arises with binary or discrete outcomes. For the related difference-in-correlations estimand, the nonparametric bootstrap achieves correct coverage while the Fisher z formula is severely miscalibrated for non-normal outcomes (Appendix A2).

The Z coefficient in the same regression does not estimate the ATE. Its probability limit is $E[Y_2(1) \mid Y_1(1) = 0] - E[Y_2(0) \mid Y_1(0) = 0]$: the gap in Y_2 between arms at the reference level $Y_1 = 0$. The ATE is $E[Y_2(1)] - E[Y_2(0)]$. The two quantities differ because the conditioning events $\{Y_1(1) = 0\}$ and $\{Y_1(0) = 0\}$ select different subpopulations: units for whom treatment produces a low Y_1 are not the same as units for whom control produces a low Y_1, and these

groups differ in the unobserved characteristics that also affect Y_2. The resulting bias is generally non-zero and can have the opposite sign from the ATE. Of course, the marginal regression $Y_2 = \beta_0 + \beta_1 Z + \varepsilon$, which conditions on no post-treatment variables, identifies the ATE of Z on Y_2.

We also caution against possible misinterpretations of the coefficient on Y_1 in the same regression, which estimates the association between Y_1 and Y_2 in the control group. As in most any observational setting, that association might reflect the causal effects of Y_1 on Y_2, Y_2 on Y_1, or the influence of third variables both observed and unobserved.

The difference-in-slopes regression has the same algebraic structure as the standard heterogeneous treatment effects regression $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 Z \cdot X + \varepsilon$, which uses a pre-treatment moderator X (Brambor, Clark, & Golder, 2006; Kam & Franzese, 2007). The key distinction is that X is pre-treatment and Y_1 is post-treatment. For the interaction coefficient β_3, the algebra is the same in both cases. The interpretation differs: with pre-treatment X, β_3 measures how the conditional average treatment effect of Z varies with X; with post-treatment Y_1, it is the DiS, measuring how the within-arm predictive slope changes across arms. For the Z coefficient, the settings differ sharply, as shown above.

4 Applications

We apply the difference-in-slopes regression estimator in reanalyses of the datasets produced by three experimental designs.

4.1 Ideological constraint: the op-ed experiment

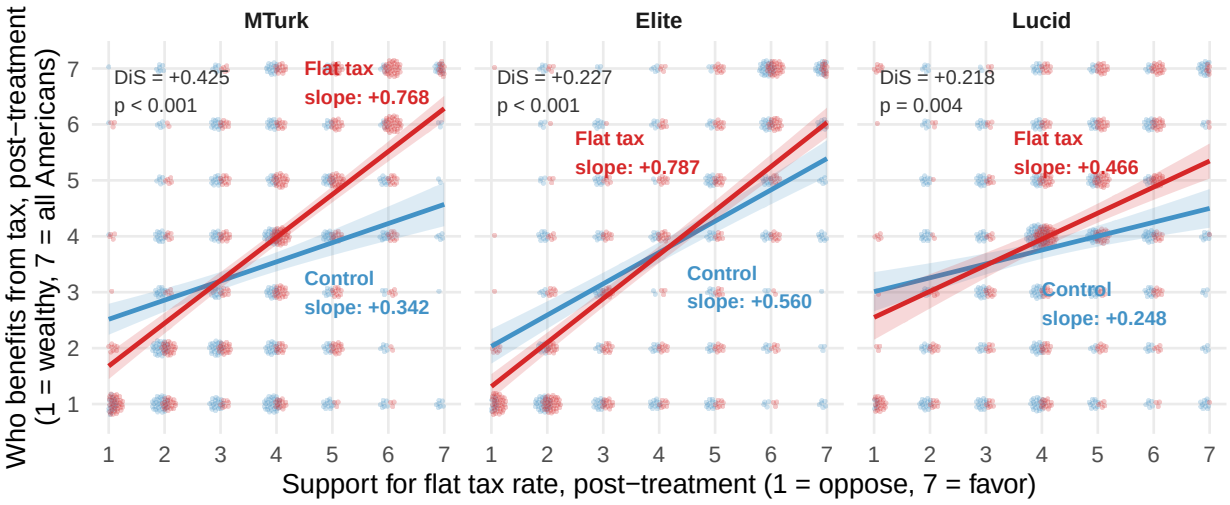
Design. Coppock et al. (2018) originally reported a newspaper op-ed experiment in which subjects were randomly assigned to read one of four real op-eds or a control condition in an MTurk sample ($N_{\text{treatment}} = 587$, $N_{\text{control}} = 622$) and an elite sample ($N_{\text{treatment}} = 463$, $N_{\text{control}} = 448$); Coppock and Green (2022) replicated the study on a Lucid sample

($N_{\text{treatment}} = 496$, $N_{\text{control}} = 581$). We estimate the difference-in-slopes to consider whether the flat-tax op-ed (penned by Senator Rand Paul) tightened the within-domain attitude structure: $\text{DiS} > 0$ means the op-ed made one flat-tax attitude more predictive of another. Y_1 = support for a 14.5% flat corporate tax rate (1–7, post-treatment) and Y_2 = belief about who benefits from the tax (1–7, post-treatment).

Results. Figure 1 plots post-treatment Y_2 against post-treatment Y_1 separately by arm in each sample. In the control arm, the slope of the benefits belief on the rate-support question is modest (MTurk: 0.34, Elite: 0.56, Lucid: 0.25). In the treatment arm, after reading the Rand Paul op-ed, the slope is substantially steeper in all three samples (MTurk: 0.77, Elite: 0.79, Lucid: 0.47). The difference-in-slopes is $\widehat{\text{DiS}} = 0.43$ (SE = 0.06, $p < 0.001$) in MTurk, 0.23 (SE = 0.06, $p < 0.001$) in the elite sample, and 0.22 (SE = 0.08, $p = 0.004$) in Lucid. The op-ed tightened the within-domain attitude structure in all three independent samples, at or below the 0.004 level.

The vertical gap between the two fitted lines at any given value of Y_1 does not estimate the conditional average treatment effect of the op-ed for units with that Y_1 value. Y_1 is a post-treatment outcome: it conflates units from different principal strata defined by the joint potential outcomes $(Y_1(0), Y_1(1))$ (Frangakis & Rubin, 2002). $\widehat{\text{DiS}}$ is the only causally-identified quantity in the figure.

Figure 1: Flat tax op-ed tightens within-domain ideological constraint.



Note: Data from Coppock et al. (2018) (MTurk $N = 1{,}209$; Elite $N = 911$) and Coppock and Green (2022) (Lucid $N = 1{,}069$). Each panel plots responses to the 14.5% flat corporate tax question (Y_1, x-axis) against responses to who benefits from the flat tax (Y_2, y-axis) for the flat-tax-treatment arm (red) and the control arm (blue). Lines are OLS fits within each arm with 95% confidence bands. DiS denotes the difference-in-slopes estimate; we estimate HC2 standard errors.

4.2 Affective polarization structure: the Strengthening Democracy megastudy

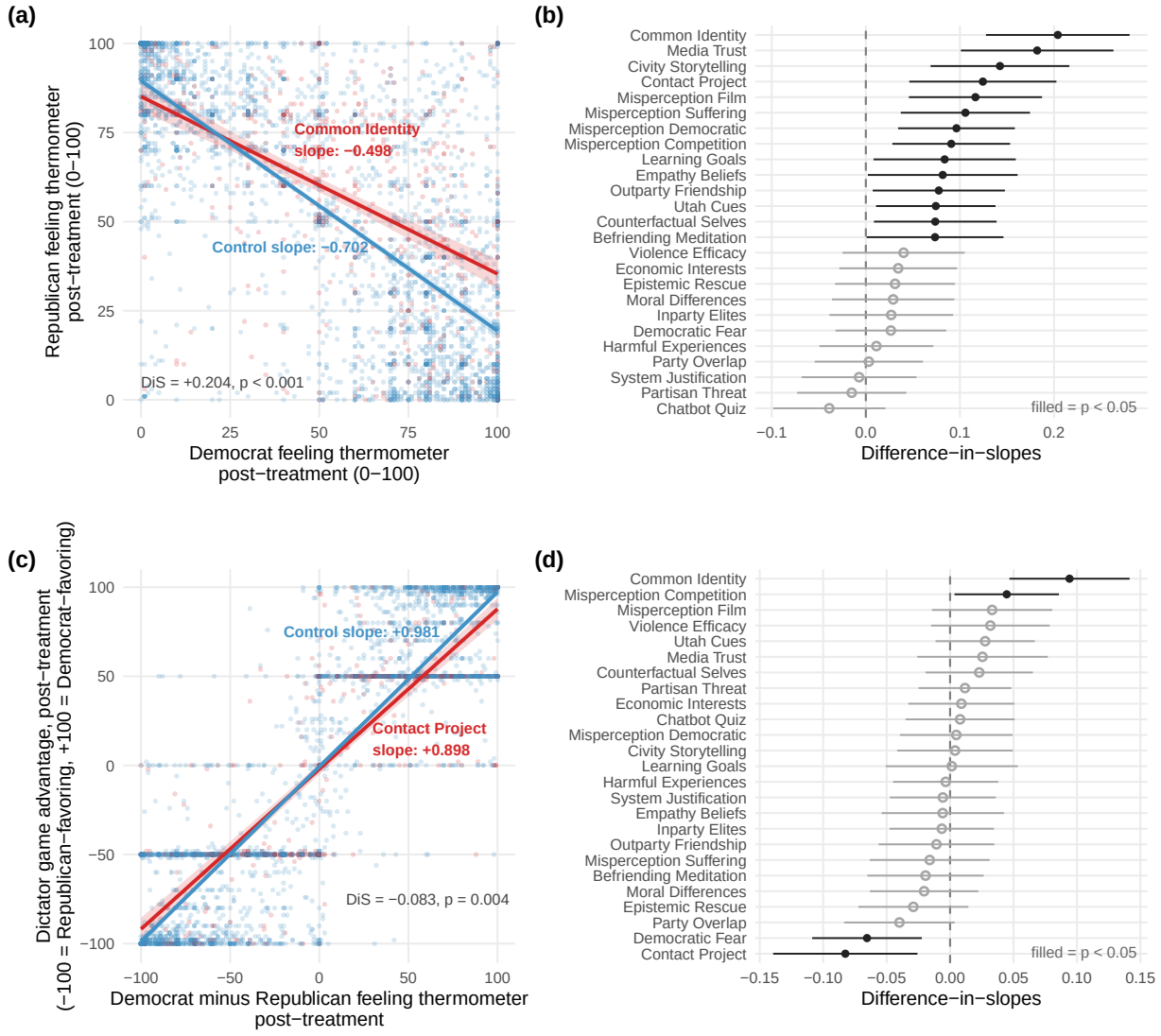
Design. Standard affective polarization experiments measure whether a treatment shifts the difference between in-party and out-party feeling thermometers (Iyengar et al., 2012); here we consider the different question of whether the treatment loosens the strong negative association between the two thermometers. Voelkel et al. (2024) ran a pre-registered megastudy testing 25 brief online interventions (readings, videos, and exercises) designed to reduce antidemocratic attitudes and partisan animosity. Participants were recruited via Lucid and quota-sampled for approximate national representativeness on age, gender, race, and region, yielding $N = 32{,}059$ US Democrats and Republicans. Each participant was randomly assigned to one of the 25 treatment arms or to a no-intervention control; the control group ($N = 5{,}691$) is used as the comparison for every arm. Each treatment arm contains approximately 1,100–1,150 respondents.

We consider two Y_1, Y_2 pairs. In the first, Y_1 = Democratic feeling thermometer (0–100, higher = warmer toward Democrats) and Y_2 = Republican feeling thermometer (0–100, higher = warmer toward Republicans), each derived from the respondent's in- and out-party coldness scores by party membership. In the second, Y_1 = Democratic minus Republican thermometer difference (-100 to $+100$) and Y_2 = dictator-game advantage: for Democrats, the amount kept from the Republican allocatee (0–100); for Republicans, the negative of the amount kept from the Democratic allocatee, so that positive values indicate Democratic-favoring behavior throughout.

Results. Figure 2 presents results for two specifications, plotting post-treatment Y_2 against post-treatment Y_1; as with Figure 1, the vertical gap between fitted lines at any given Y_1 value is not a conditional average treatment effect. The first specification (panels a–b) examines the Democratic–Republican thermometer association. In the control arm, the slope is -0.702 (SE = 0.012): unsurprisingly, respondents who feel warmer toward Democrats tend to feel colder toward Republicans. Panel (a) shows this alongside the common identity intervention, which emphasizes shared American identity. Under that treatment the slope rises to -0.498 ($\widehat{\text{DiS}} = 0.204$, SE = 0.039, $p < 0.001$). Panel (b) plots all 25 arms: 14 of 25 produce significant loosening of the thermometer coupling.

The second specification (panels c–d) considers the attitude-behavior association. The control slope is $+0.981$ (SE = 0.008): respondents who prefer Democrats over Republicans on the thermometer also show more Democratic-favoring dictator game behavior. Four arms significantly alter this correlation. Contact project ($\widehat{\text{DiS}} = -0.083$, $p = 0.004$) and democratic fear ($\widehat{\text{DiS}} = -0.065$, $p = 0.003$) reduce it. The common identity intervention, which most strongly loosens the thermometer coupling in the first specification, significantly strengthens the attitude-to-behavior correlation ($\widehat{\text{DiS}} = +0.094$, $p < 0.001$) in the second.

Figure 2: Affective polarization megastudy: two difference-in-slopes specifications.



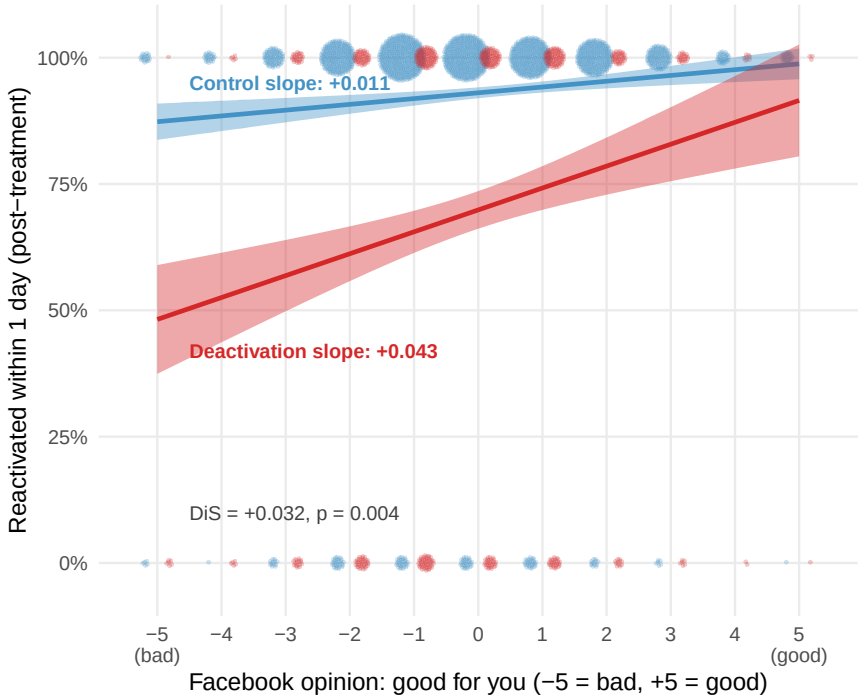
Note: Data from Voelkel et al. (2024); control $N \approx 5{,}550$, each treatment arm $N \approx 1{,}130$. Panels (a)–(b): Y_1 = Democratic feeling thermometer, Y_2 = Republican feeling thermometer (both 0–100). Control slope $= -0.702$; the winning arm (common identity) raises it to -0.498; 14 of 25 arms significant, all positive. Panels (c)–(d): Y_1 = Democratic minus Republican thermometer difference, Y_2 = dictator-game advantage signed DEM-positive. Control slope $= +0.981$; the winning arm (contact project) reduces it to $+0.898$; 4 of 25 arms significant. In the coefplots, filled points indicate $p < 0.05$; error bars are 95% confidence intervals. Democratic and Republican thermometers are derived from in- and out-party warmth scores by respondent party. Dictator-game advantage = amount kept (0–100) for Democrats and negative of amount kept for Republicans, so that positive values indicate Democratic-favoring economic behavior.

4.3 Attitude-behavior consistency: the Facebook deactivation experiment

Design. Allcott et al. (2020) randomized US Facebook users to deactivate their accounts for four weeks (treatment) or continue as normal (control), using a willingness-to-accept (WTA) mechanism: participants who valued deactivation at less than a randomly assigned price offer ($0 or $102) were required to deactivate. The analysis sample consists of participants with WTA $<$ $102, within whom the price offer is a pure randomized instrument; the treatment group was deactivated on 90% of automated compliance checks ($N_{\text{treatment}} = 578$, $N_{\text{control}} = 2{,}133$). At endline, both arms completed a Facebook opinion survey and then deactivated for a mandatory 24 hours, after which they were free to reactivate. Y_1: "To what extent do you think Facebook is good or bad for you?" (0–10 scale, recentered to -5 to $+5$, higher = more positive). Y_2: whether the participant reactivated within one day of the post-endline 24-hour window (1 = yes; 0 = no), observed for all participants with no missing data.

Results. Figure 3 shows that in the control arm, platform opinion and reactivation within one day are nearly unrelated (slope = 0.011): 93% of controls reactivate across the full range of platform opinions. In the treatment arm the slope is four times larger (slope = 0.043): platform opinion and rapid return co-occur more strongly after four weeks of required deactivation. The difference-in-slopes is $\widehat{\text{DiS}} = 0.032$ (SE = 0.011, $p = 0.003$), consistent with the attitude-behavior link strengthening through direct experience (Ajzen & Fishbein, 1977).

Figure 3: Facebook deactivation sharpens the attitude-behavior link.



Note: Data from Allcott et al. (2020); $N_{\text{treatment}} = 578$, $N_{\text{control}} = 2{,}133$. The plot shows the probability of reactivating within one day (Y_2) against opinion about Facebook (Y_1, -5 = bad, $+5$ = good) for the deactivation arm (red points) and the control arm (blue points). Lines are OLS fits with 95% confidence bands. Y_1 recentered so that 0 = neutral. $Y_2 = 1$ if account reactivated within one day of the post-endline 24-hour deactivation window.

5 Conclusion

As experimenters, we're most familiar with estimating the effects of treatments on the *levels* of multiple post-treatment outcomes; in this paper we have explored how to estimate the effects of treatments on the *associations* between multiple post-treatment outcomes. We showed that under standard experimental assumptions, the difference-in-slopes estimand is identified. Estimation is straightforward: the $Z \cdot Y_1$ coefficient from an OLS interaction regression is numerically equivalent to the plug-in estimator and can be read from standard output. While our method does not require downloading any new software, it does require interpretive care: the same regression that estimates the DiS also includes coefficients that suffer from post-treatment bias or other inferential obstacles.

We think this mode of analysis is applicable to a broad range of experimental designs. In

American politics, we might study the effects of racial cues on the correlation between racial resentment and policy support. In comparative politics, we might study how performance news shifts the association between vote intentions and retrospective evaluations. In international relations, we might study how threat information changes how tightly threat perceptions predict policy preferences. Since as a discipline, we are unused to posing theoretical questions about the causal effects of treatments on descriptive associations, we think it is difficult to anticipate at this moment many or even most of the possible future uses of this method.

We consider several extensions in the appendix. Appendix A4 characterizes the family of conditional difference-in-slopes estimands that arise when the population is stratified by a pre-treatment covariate X. In a slope analog of Simpson's paradox, sign reversals between the marginal and conditional DiS are possible. Whereas covariate adjustment typically increases the precision of estimates of the ATE, for the DiS, covariate adjustment actually changes the estimand (to a conditional DiS) and not just the precision with which it is estimated. Second, researchers who prefer the scale-free difference-in-correlations $\rho_1 - \rho_0$ can estimate it in each arm; the nonparametric bootstrap rather than the Fisher z formula is required for inference when outcomes depart from bivariate normality, as the Fisher z standard error can be off by a factor of nearly four for binary outcomes (Appendix A2). Third, the DiS as defined uses a linear projection; researchers who fit more flexible conditional expectation functions (polynomials, LOWESS, splines) obtain a pointwise difference in curves rather than a single number, which is displayable but not reducible to $\widehat{\text{DiS}}$ (Appendix A5). Fourth, on statistical power: when both outcomes are normalized to unit variance, the DiS standard error equals $\sqrt{1-\rho^2}$ times the ATE standard error, where ρ is the control-arm correlation between Y_1 and Y_2. DiS is therefore always at least as powerful as an ATE test of the same standardized magnitude, with the advantage growing as $|\rho|$ increases. Power simulations calibrated to the empirical range of ρ across the three applications appear in Appendix A3.

References

Ajzen, I., & Fishbein, M. (1977). Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, *84*(5), 888–918.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–676. doi: 10.1257/aer.20190658

Bloome, D., & Schrage, D. (2021). Covariance regression models for studying treatment effect heterogeneity across one or more outcomes: Understanding how treatments shape inequality. *Sociological Methods & Research*, *50*(3), 1034–1072. doi: 10.1177/0049124119882449

Bowers, J. (2011). Making effects manifest in randomized experiments. In *Cambridge handbook of experimental political science* (pp. 459–480). Cambridge University Press. doi: 10.1017/cbo9780511921452.032

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14*(1), 63–82.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.

Coppock, A., Ekins, E., & Kirby, D. (2018). The long-lasting effects of newspaper op-eds on public opinion. *Quarterly Journal of Political Science*, *13*(1), 59–87.

Coppock, A., & Green, D. P. (2022). Do belief systems exhibit dynamic constraint? *Journal of Politics*, *84*(2), 725–738. doi: 10.1086/716294

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, *114*(525), 304–317.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29.

Glasman, L. R., & Albarracín, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude–behavior relation. *Psychological Bulletin*, *132*(5), 778–822.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies*, *64*(4), 487–535. doi: 10.2307/2971729

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431.

Kam, C. D., & Franzese, R. J. (2007). *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor: University of Michigan Press.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, *7*(1), 295–318.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A*, *147*(5), 656–666.

Voelkel, J. G., Stagnaro, M. N., Chu, J. Y., Pink, S. L., Mernyk, J. S., Rand, D. G., ... others (2024). Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science*, *386*(6719), eadh4764. doi: 10.1126/science.adh4764