Beyond Pretrends: A Discordance-Based Sensitivity Analysis for Difference-in-Differences*

Thomas Leavitt **thomas.leavitt@baruch.cuny.edu**
Marxe School of Public and International Affairs
Baruch College, City University of New York (CUNY)
New York, NY 10010, USA

Abstract

In the canonical Difference-in-Differences design, the control group's post-treatment change serves as an imputation of the treated group's counterfactual change in the same period, an imputation justified by parallel trends. However, differences in group composition can produce between-group differences in how outcomes would evolve over time, rendering this imputation vulnerable to confounding. An alternative imputation — such as one based on the treated group's pre-treatment change — avoids such between-group confounding but introduces the risk of confounding from within-group temporal shifts. Ideally, both imputations, each vulnerable to different sources of confounding, would have concordant values, thereby yielding the same causal conclusions. When the imputations are discordant, conclusions under parallel trends hinge more critically on that assumption since alternative imputations would point to different results. Yet in these scenarios, existing pretrends-based sensitivity analyses can show low sensitivity because they ignore post-treatment deviations from pretrends in the control group. This paper therefore proposes a discordance-based sensitivity model in which parallel pretrends are necessary but not sufficient for low sensitivity. I formally justify this model in terms of the expected distance between the ATT under parallel trends and under alternative assumptions, weighted by the joint plausibility of those assumptions. I then provide a decision-theoretic rationale for benchmarking violations of parallel trends using the worst-case discordance between the parallel trends imputation and alternative imputations. Finally, I apply both pretrends- and discordance-based sensitivity models to assess how a labor supply shock influenced electoral support for apartheid-era policies in South Africa, showing how the two approaches yield different results.

Keywords: Difference-in-Differences, Sensitivity Analysis, Parallel Trends, Partial Identification

1. Introduction

Difference-in-Differences (DID) is a canonical quasi-experiment, providing a powerful tool for inferring causal effects in observational studies. In a typical DID setup, one of two groups observed over time receives a treatment. To infer the average treatment effect on the treated group (the ATT), one compares the treated group's after-treatment minus before-treatment outcomes with the corresponding after-minus-before outcomes for a com-

[*] An earlier version of this paper circulated under the title "Beyond Parallel Trends: Improvements on Estimation and Inference in the Difference-in-Differences Design."

parison group. Because the treated group's counterfactual change in outcomes under no treatment is unobservable, the comparison group's observed change serves as an imputation of this counterfactual change. The parallel trends assumption states that this imputation equals the treated group's counterfactual change in outcomes under no treatment. In most applications, researchers treat the plausibility of this assumption, which they assess via pre-treatment trends, as the key criterion for the credibility of the DID design.

Although parallel trends is virtually synonymous with the DID framework itself, much of the foundational literature points to a different basis for credibility. This literature on quasi-experiments (Shadish et al., 2002; Cook and Campbell, 1979) and, more broadly, observational studies (Cochran, 1965, 1972) emphasizes that credibility often rests not on a single piece of evidence tied to a particular assumption. Rather, credibility depends on the alignment between multiple strands of evidence, each justified by different assumptions. As Rosenbaum (2015, p. 53) explains in "Cochran's Causal Crossword," the strength of quasi-experiments arises from their capacity to

> *cultivate varied strands of evidence within a single study, each strand being weak on its own, each strand vulnerable in a different way, but with the several strands gaining in strength if they agree in appropriate ways.*

In this view, quasi-experiments like DID function as a crossword puzzle (Haack, 1995, 1993): An answer is credible when it aligns with multiple clues — both across and down. An answer that fits one clue but contradicts another is less convincing than an answer that satisfies both.

This metaphor of the crossword puzzle is compelling in a DID setting, where the aim is to impute the treated group's after-minus-before average in the post-period. Consider two pieces of evidence, each vulnerable to a different source of confounding:

1. The first is the control group's after-minus-before average in the post-period. This serves as an imputation for the treated group's counterfactual change in outcomes in the same post-period. This imputation has the advantage of sharing the *same time period* as the counterfactual. However, because this imputation comes from a group with a different composition, the imputation may misrepresent the counterfactual if the groups would evolve differently over time or are exposed to different post-period shocks — i.e., if parallel trends fails.

2. The second is the treated group's after-minus-before average in the pre-period. This imputation avoids concerns about differences between groups because it uses data from the *same group* as the counterfactual. But since this imputation reflects a different time period, the imputation relies on the assumption that the treated group would have evolved similarly across time. If this assumption fails, then the imputation may misrepresent the counterfactual.

The first piece of evidence is vulnerable to *between-group differences*, and the second to *temporal shifts*. Ideally, both pieces of evidence would have similar values and thus support the same conclusion about the ATT. In that case, even if parallel trends were false and an alternative trend assumption were true, the resulting error from relying on parallel trends would be small. More generally, if an alternative trend assumption is at least somewhat plausible, concordance between the two pieces of evidence means one would expect inference based on parallel trends to have less error than it would if the two pieces of evidence were discordant. In short, concordance implies that causal conclusions are less sensitive to violations of parallel trends.

Despite this straightforward intuition, it is not reflected in existing sensitivity analyses to violations of parallel trends. As I go on to show, concordance implies parallel pretrends, but the converse is not true. Thus, pretrends-based sensitivity analyses can show high robustness to the failure of parallel trends, even if an alternative trend assumption, justified by a different strand of evidence, would yield a substantially different conclusion.

In this paper, I therefore propose a discordance-based sensitivity analysis. Since concordance implies parallel pretrends, this paper's sensitivity model will often give similar results to existing pretrends-based models. However, when conclusions diverge under alternative trend assumptions — such as when pretrends are parallel but the control group's average post-period outcome deviates from its pretrend — this paper's model will indicate greater sensitivity to violations of parallel trends.

In what follows, I develop a discordance-based sensitivity analysis that builds on existing frameworks. This development begins with a formal setup in Section 2, review of the pretrends-based model in Section 3, and the motivation for alternatives in Section 4, followed by the introduction of the discordance-based approach in Section 5. This approach bounds violations of parallel trends based on the discordance between the parallel trends imputation and alternative possible imputations for the treated group's counterfactual. As I show in Section 5.1, the model's use of discordance can be formally justified in terms of the expected distance between the ATT under parallel trends and the ATT under alternative identification assumptions, evaluated with respect to the joint plausibility of those assumptions (including that of non-identification). I also show that concordance implies parallel pretrends, though the converse does not hold. Section 5.2 offers a decision-theoretic justification for the model's use of the worst-case discordance between the parallel trends imputation and any alternative imputation. Section 6 then generalizes the argument — developed up to that point for a single post-treatment period — to settings with multiple post-periods. Section 7 then applies the discordance-based sensitivity analysis to a case study examining how a labor supply shock in apartheid-era South Africa affected electoral support for apartheid policies. The final section concludes.

2. Setup

Suppose two groups, a treated group ($G = 1$) and a comparison group ($G = 0$), where the index set of groups is denoted by $\mathcal{G} \coloneqq \{0, 1\}$. Also suppose $t = 1, \dots, T$ periods of which T is the only post-treatment period. That is, between periods $T - 1$ and T, the treated group is exposed to treatment and the comparison group is not. Denote the index set of time periods by $\mathcal{T} \coloneqq \{1, \dots, T\}$. For all $(g, t) \in \mathcal{G} \times \mathcal{T}$, let the treatment indicator for group g in period t be $Z_{g,t} \coloneqq \mathbb{1}\{G = g\}\mathbb{1}\{t = T\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function that equals 1 if its argument is true and 0 if not. For the treated group, $Z_{g,t} = 0$ for all $t < T$ and $Z_{g,T} = 1$. For the control group, $Z_{g,t} = 0$ for all periods.

To define the causal target in terms of potential outcomes, let $Y_{g,t}(0)$ denote the untreated potential outcome for any group-period $(g, t) \in \mathcal{G} \times \mathcal{T}$. Also let $Y_{1,T}(1)$ denote the treated potential outcome for group $G = 1$ in the post-treatment period, T. The causal target is the ATT, denoted by τ,

$$\tau \coloneqq \bar{Y}_{1,T}(1) - \bar{Y}_{1,T}(0), \tag{1}$$

where, for any arbitrary random quantity W, I write $\bar{W}$ to denote the expectation (mean) of W with respect to the population-level joint cumulative distribution function (CDF). For example, $\bar{Y}_{1,T}(1)$ is the expected treated potential outcome for group $G = 1$ in period T.

This basic setup features two groups and a single post-treatment period, but it generalizes easily to more complex settings. In staggered-adoption settings with variation in treatment timing, Callaway and Sant'Anna (2021) show that the analysis can be decomposed into a set of simple DID comparisons between each timing-defined group and a never-treated or not-yet-treated group. The same basic setup also extends naturally to settings with multiple post-treatment periods, as detailed in Section 6. In such cases, there are multiple post-treatment ATTs, and the target parameter can be any linear combination of them, with a common choice being the average ATT across all post-treatment periods (Rambachan and Roth, 2023, p. 2561).

Returning to the basic setting with two groups and one post-period, I first state the standard "consistency" assumption of Hernán and Robins (2020), which enables expressing the ATT in terms of observable quantities:

Assumption 1. *For all $(g, t) \in \mathcal{G} \times \mathcal{T}$,*

$$Y_{g,t} = Z_{g,t} Y_{g,t}(1) + (1 - Z_{g,t})\, Y_{g,t}(0). \tag{2}$$

Assumption 1 implies that the observed outcome in each period equals the potential outcome under the treatment status actually received. This assumption rules out treatment anticipation and interference, ensuring that each unit's observable outcome does not reflect any unit's future treatment status or the treatment status of other units at the same time. Moreover, by defining a single treated and a single untreated potential outcome, this setup assumes a single, well-defined version of the treatment.

Under Assumption 1, the ATT is

$$\tau = \bar{Y}_{1,T} - \bar{Y}_{1,T}(0). \tag{3}$$

Assumption 1 justifies substituting the treated group's observable outcome in the post-period for the treated group's potential outcome in the post-period. Hence, the expression for the ATT in (3) contains $\bar{Y}_{1,T}$ instead of $\bar{Y}_{1,T}(1)$.

The ATT in (3) is now composed of an observable quantity, $\bar{Y}_{1,T}$, and an unobservable one, $\bar{Y}_{1,T}(0)$. To fully express the ATT in only observable quantities, it is helpful to first express the ATT in (3) as

$$\begin{aligned} \tau = {} & \bar{Y}_{1,T} - \bar{Y}_{1,T-1}(0) - \left(\bar{Y}_{0,T}(0) - \bar{Y}_{0,T-1}(0)\right) \\ & - \left[\bar{Y}_{1,T}(0) - \bar{Y}_{1,T-1}(0) - \left(\bar{Y}_{0,T}(0) - \bar{Y}_{0,T-1}(0)\right)\right]. \end{aligned} \tag{4}$$

This expression for the ATT in (4) boils down to the canonical DID,

$$\text{DID} \coloneqq \bar{Y}_{1,T}(1) - \bar{Y}_{1,T-1}(0) - \left(\bar{Y}_{0,T}(0) - \bar{Y}_{0,T-1}(0)\right), \tag{5}$$

minus period T's Difference-in-Trends, where the Difference-in-Trends for any $t \geq 2$ is

$$\delta_t \coloneqq \bar{Y}_{1,t}(0) - \bar{Y}_{1,t-1}(0) - \left(\bar{Y}_{0,t}(0) - \bar{Y}_{0,t-1}(0)\right). \tag{6}$$

Therefore, the ATT in (4) is

$$\tau = \text{DID} - \delta_T. \tag{7}$$

The canonical point identification assumption about δ_T is parallel trends, which, for any $t \geq 2$, is

$$\delta_t = 0. \tag{8}$$

Under the assumption that $\delta_T = 0$, the ATT in (7) equals the DID in (5), making the unobservable ATT point identified in terms of observable quantities. However, any point assumption about the value of the counterfactual δ_T, not just $\delta_T = 0$, allows one to recover the ATT by subtracting off the assumed value of δ_T from the DID.

To simplify this framework going forward, define a detrended outcome for group g in time $t \geq 2$ as

$$\epsilon_{g,t} \coloneqq Y_{g,t} - \bar{Y}_{g,t-1}, \tag{9}$$

which is simply the observable outcome for group g in period t minus the mean (expectation) of group g's observable outcome in the previous period $(t-1)$. The detrended outcome $\epsilon_{g,t}$ therefore represents the period-to-period change in group g's average outcome. If the underlying outcome series for group g exhibits an upward or downward trend, the sequence of detrended outcomes in the pre-treatment periods, $\{\epsilon_{g,t}\}_{t=2}^{T-1}$, will generally reflect those changes. This transformation adjusts the level (i.e., the average outcome at each point in time) of the series but does not impose any shape restrictions — such as a constant slope or flat trend — on the detrended outcomes.

Using the detrended outcomes, the setup thus far admits an equivalent representation. By combining the definition in (9) with Assumption 1, the ATT in (1) can be rewritten exactly in terms of the detrended potential outcomes. The algebraic steps below make this equivalence clear.

First, recall that $Z_{g,t} \coloneqq \mathbb{1}\{G = g\}\mathbb{1}\{t = T\}$, so that $Z_{g,t} = 0$ for all $t < T$. In those periods, Assumption 1 implies that the detrended outcome in (9) can be expressed as $Y_{g,t} - \bar{Y}_{g,t-1}(0)$. Substituting the expression for $Y_{g,t}$ from (2) into (9) then yields $\epsilon_{g,t} = [Z_{g,t}Y_{g,t}(1) + (1 - Z_{g,t})\,Y_{g,t}(0)] - \bar{Y}_{g,t-1}(0)$. This expression is algebraically equivalent to

$$\epsilon_{g,t} = Z_{g,t}\left(Y_{g,t}(1) - \bar{Y}_{g,t-1}(0)\right) + (1 - Z_{g,t})\left(Y_{g,t}(0) - \bar{Y}_{g,t-1}(0)\right), \tag{10}$$

which equals $Y_{g,t}(1) - \bar{Y}_{g,t-1}(0)$ when $Z_{g,t} = 1$ and $Y_{g,t}(0) - \bar{Y}_{g,t-1}(0)$ when $Z_{g,t} = 0$. I refer to these as the detrended potential outcomes, defined as

$$\epsilon_{g,t}(z) \coloneqq Y_{g,t}(z) - \bar{Y}_{g,t-1}(0), \tag{11}$$

for $z \in \{0, 1\}$, corresponding to the detrended control $(z = 0)$ and treated $(z = 1)$ potential outcomes, respectively. Using this definition in (11), the detrended outcome in (10) can be written compactly as

$$\epsilon_{g,t} = Z_{g,t}\epsilon_{g,t}(1) + (1 - Z_{g,t})\epsilon_{g,t}(0), \tag{12}$$

analogous to the original expression of the "consistency" assumption in (2).

It now follows directly that the ATT in (1), defined in terms of the raw potential outcomes, is algebraically equivalent to its detrended form, $\bar{\epsilon}_{1,T}(1) - \bar{\epsilon}_{1,T}(0)$:

$$\begin{aligned} \tau &:= \bar{Y}_{1,T}(1) - \bar{Y}_{1,T}(0) = \bar{Y}_{1,T}(1) - \bar{Y}_{1,T-1} - \big(\bar{Y}_{1,T}(0) - \bar{Y}_{1,T-1}\big) \\ &= \bar{\epsilon}_{1,T}(1) - \bar{\epsilon}_{1,T}(0). \end{aligned} \tag{13}$$

Similarly, the DID in (5) and the Difference-in-Trends in (6) can both be expressed in terms of detrended potential outcomes as

$$\text{DID} := \bar{Y}_{1,T}(1) - \bar{Y}_{1,T-1}(0) - \big(\bar{Y}_{0,T}(0) - \bar{Y}_{0,T-1}(0)\big) = \bar{\epsilon}_{1,T}(1) - \bar{\epsilon}_{0,T}(0) \tag{14}$$

and

$$\delta_t := \bar{Y}_{1,t}(0) - \bar{Y}_{1,t-1}(0) - \big(\bar{Y}_{0,t}(0) - \bar{Y}_{0,t-1}(0)\big) = \bar{\epsilon}_{1,t}(0) - \bar{\epsilon}_{0,t}(0). \tag{15}$$

As in the earlier formulation based on the raw potential outcomes, it remains the case under the detrended-outcome notation that the ATT is the DID minus period T's Difference-in-Trends:

$$\text{DID} - \delta_T = \bar{\epsilon}_{1,T}(1) - \bar{\epsilon}_{0,T}(0) - [\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)] = \bar{\epsilon}_{1,T}(1) - \bar{\epsilon}_{1,T}(0) = \tau. \tag{16}$$

The parallel trends assumption states that, $\delta_T = 0$, i.e., that the unobservable $\bar{\epsilon}_{1,T}(0)$ equals the observable $\bar{\epsilon}_{0,T}(0)$. To distinguish the assumption itself from the specific counterfactual imputation the assumption justifies (regardless of whether it actually holds), I refer to $\bar{\epsilon}_{0,T}$ as the parallel trends imputation for $\bar{\epsilon}_{1,T}(0)$. I denote the corresponding imputed ATT by $\tau_{0,T}$, where the subscript $(0,T)$ indicates that the counterfactual $\bar{\epsilon}_{1,T}(0)$ is imputed using the comparison group's detrended outcome in period T, $\bar{\epsilon}_{0,T}$.

Leavitt and Hatfield (2025) extend the detrended-outcome approach to a much broader class of prediction functions, introducing a general point identification requirement they term equal expected prediction errors. They show that DID can be viewed as a special case of a more general two-stage procedure: The first stage uses pre-treatment data to *predict* untreated post-treatment outcomes, and the second stage uses the comparison group's observable post-treatment prediction errors to *correct* the treated group's predictions (see Leavitt and Hatfield, 2025, p. 1827). In standard DID, the parallel trends assumption implies equal expected prediction errors when the prediction function is $\bar{Y}_{g,t-1}$, which is precisely the detrended-outcome formulation in (9). This broader framework encompasses many other prediction models, including sequential DID, two-way fixed effects, and models with unit- or group-specific time trends, among others. The key implication is that the reasoning developed here for a simple prediction model extends naturally to a much wider class of models for predicting how outcomes evolve over time.

Finally, to complete the formal setup, let $\mathcal{V} \subseteq \{2, \ldots, T-1\}$ denote the set of pre-treatment validation periods selected for analysis, with each $v \in \mathcal{V}$ indexing a single period. The set $\mathcal{V}$ can be any nonempty subset of $\{2, \ldots, T-1\}$, such as $\{2\}$, $\{T-1\}$ or the full set $\{2, \ldots, T-1\}$. If one imputes the counterfactual using the detrended outcome from a particular period $v \in \mathcal{V}$ and group $g \in \mathcal{G}$, $\bar{\epsilon}_{g,v}$, the resulting imputed ATT is denoted $\tau_{g,v}$.

3. Pretrends-based Sensitivity Analysis

Researchers cannot directly evaluate whether parallel trends (i.e., $\delta_T = 0$) holds since untreated potential outcomes in period T are unobservable. However, researchers can directly evaluate parallel trends in periods before treatment, i.e., whether $\delta_t = 0$ for $2 \leq t < T$. Hence, researchers typically assess the plausibility of parallel trends by assessing whether pretrends are parallel. Presumably in response to this widespread practice, Rambachan and Roth (2023) derive a general sensitivity analysis framework that imposes restrictions on the extent to which parallel trends is violated based on violations in the pre-period. The framework is general and enables a wide class of set restrictions on the unobservable value of δ_T.

The leading set restriction on δ_T that Rambachan and Roth (2023) propose, which builds on the sensitivity analysis in Manski and Pepper (2018), stipulates that violations of parallel trends are no greater than a magnitude, $M \geq 0$, of the greatest absolute value of the Difference-in-Trends in the pre-period. Formally, one can write this sensitivity analysis as one in which the counterfactual Difference-in-Trends (δ_T) lies in a compact set given by

$$\left\{\delta_T : |\delta_T| \leq M \max_{v \in \mathcal{V}} |\delta_v| \right\}. \tag{17}$$

The sensitivity parameter $M \geq 0$ controls how tightly one constrains the identification assumption in terms of the greatest absolute violation of parallel trends in the pre-period. Point identification of the ATT holds under $M = 0$, which implies parallel trends. Otherwise, so long as $\max_{v \in \mathcal{V}} |\delta_v| > 0$, the ATT is set-identified rather than point-identified. In particular, under the set restriction in (17), the ATT belongs to the following set:

$$\left[\text{DID} - M \max_{v \in \mathcal{V}} |\delta_v|, \, \text{DID} + M \max_{v \in \mathcal{V}} |\delta_v| \right]. \tag{18}$$

To bound the post-period violation of parallel trends, one might want to use the average pre-period violation — or another function of the pre-period violations — instead of the maximum, as in (17). To accommodate these possibilities, I embed the leading sensitivity analysis of Rambachan and Roth (2023) as a special case of a sensitivity analysis in which

$$\left\{\delta_T : |\delta_T| \leq M \sum_{v \in \mathcal{V}} w_v |\delta_v| \right\}, \tag{19}$$

where $\{w_v\}_{v \in \mathcal{V}}$ are weights over the set $\mathcal{V}$ defined as

$$\{w_v\}_{v \in \mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|} \text{ with } w_v \geq 0 \text{ for all } v \in \mathcal{V} \text{ and } \sum_{v \in \mathcal{V}} w_v = 1, \tag{20}$$

where, in general, $\mathbb{R}^d$ denotes the set of d-dimensional vectors with real-valued entries and $|\cdot|$ of a set denotes its cardinality (i.e., the number of elements in that set).

These weights determine how much relative emphasis the analysis places on each validation period's Difference-in-Trends. The set restriction in (17) is a special case of the weighted model described in (19) and (20). In this special case, all weight is assigned to the pre-periods that attain the maximum absolute Difference-in-Trends, while every other validation period receives weight 0.

4. Motivating an Alternative to Pretrends-based Approaches

To motivate an alternative to a pretrends-based sensitivity analysis, Figure 1 below presents two stylized examples based on Keele et al. (2019) and Rosenbaum (2017, p. 164-165).

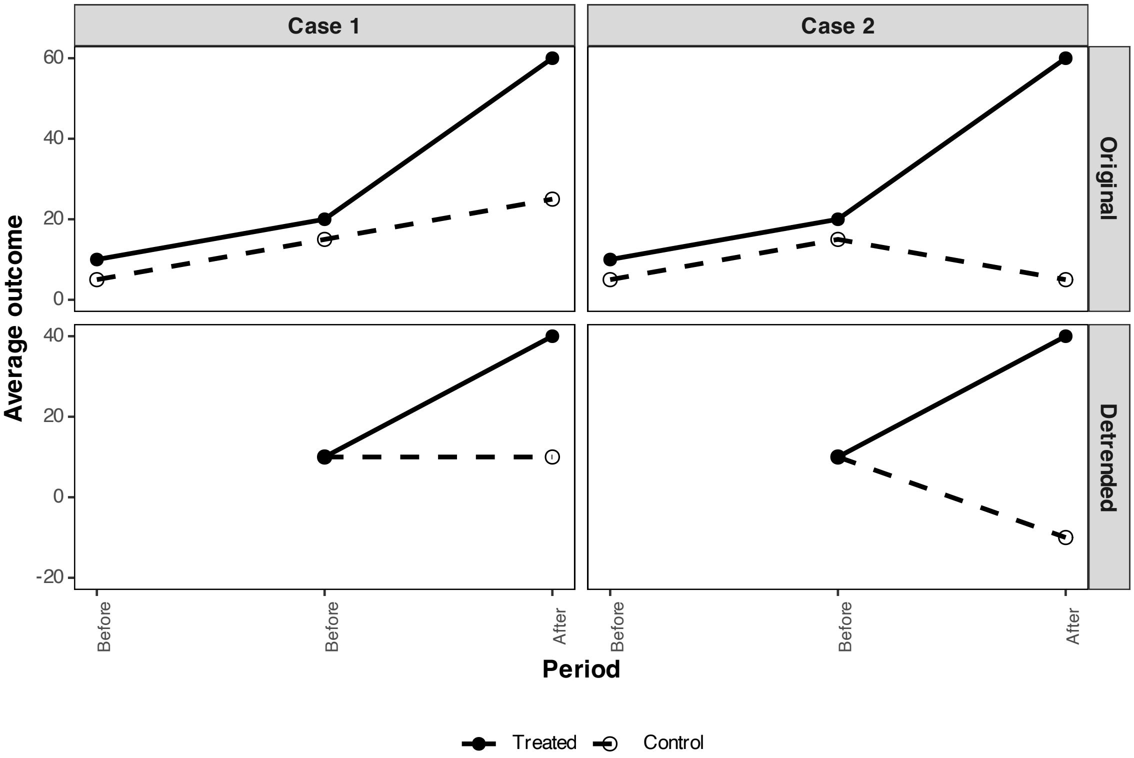

Figure 1: Two stylized scenarios (columns), each showing trends for treated and control groups across three time periods: two pre-treatment periods and one post-treatment period. The x-axis is common to all panels and reflects the time dimension, while the y-axis represents the average of the outcome variable, with the scale of the outcome varying by row. In the top row, outcomes are plotted on their original scale; in the bottom row, outcomes are transformed into changes relative to the group's mean outcome in the immediately preceding period.

These two cases both have perfectly parallel pretrends. However, in Case 2, the post-period outcome in the control group exhibits a shock in which the trend breaks from its pre-period pattern. In Case 1, by contrast, no such shock exists, as the control group's pre-trend continues uninterrupted into the post-period.

In Case 2, the control group's post-period deviation from the control group's pre-period trend is consistent with two possible scenarios:

- **Scenario 1 (common shock):** Both treated and control groups experience a shared shock in the post-period. In this scenario, the treated group's pre-period trend in average outcomes is a poor imputation for the treated group's counterfactual post-period average outcome under no treatment. Instead, the control group's change in average outcomes better captures how the treated group's average outcome would have evolved given that both groups experienced the same shock.

- **Scenario 2 (control-specific shock):** Only the control group experiences a post-period shock. In this scenario, the control group's change in average outcomes is a poor imputation for the treated group's counterfactual average outcome since the control group experienced a shock that the treated group did not. Instead, the treated group's pre-period trend in average outcomes better captures how the treated group's average outcome would have evolved given that the treated group does not experience a shock.

In Case 2, the ATT based on the parallel trends imputation ($\tau_{0,T}$) — appropriate under Scenario 1 (common shock) — differs from that based on the treated group's detrended pre-period mean ($\tau_{1,T-1}$), which is more appropriate under Scenario 2 (control-specific shock). In other words, in Case 2, the ATT implied by the parallel trends imputation can be substantially off when the parallel trends assumption is false and an alternative imputation is justified. This is not so in Case 1, where all such imputations yield the same ATT. This distinction captures the intuition that causal conclusions about the ATT depend more heavily on the parallel trends assumption in Case 2 than in Case 1.

This intuition is not reflected in a pretrends-based sensitivity analysis. By discarding information from the control group in the post-period, such sensitivity analyses fail to distinguish between these two cases. Because Case 1 and Case 2 have identical pretrends, both are perfectly robust — under the sensitivity model in (17) — to violations of parallel trends that are some magnitude, $M \geq 0$, of the greatest pre-period violation of parallel trends.

These two cases also suggest an important distinction between the plausibility of parallel trends and robustness to violations of parallel trends (due to concordance across multiple imputations). It is standard practice to interpret parallel pretrends as indicative of the plausibility of parallel trends. However, as a growing body of literature (Kahn-Lang and Lang, 2020; Roth and Sant'Anna, 2023; Egami and Yamauchi, 2023) points out, pretrends are informative about post-period parallel trends only under additional assumptions. In the absence of such assumptions, the distinction between pretrends and concordance highlights a potential misinterpretation: Parallel pretrends, often taken as evidence for the plausibility of parallel trends, may instead reflect concordance — that is, robustness of conclusions to certain violations of parallel trends.

5. Discordance-based Sensitivity Analysis

I now propose an alternative, discordance-based sensitivity analysis that reflects the intuition for why Case 2 is more robust than Case 1. The central insight is that the assumptions underlying different imputations are not necessarily mutually exclusive. When these imputations are concordant, the validity of using the treated group's pre-period trend to impute the treated group's counterfactual does *not* imply that the control group's outcome change is an invalid imputation. The assumptions supporting both imputations are mutually exclusive only when the imputations themselves are discordant. Therefore, to the extent that the conclusion about the ATT under the parallel trends imputation holds under multiple assumptions rather than just one, this conclusion depends less on the truth of the parallel trends assumption.

This general logic is closely related to that of bracketing relationships for DID (Ding and Li, 2019; Hasegawa and Small, 2017; Ye et al., 2024; Angrist and Pischke, 2008; Guryan, 2001; Demetrescu et al., 2025). Setting aside technical details, the general logic of a bracketing relationship is based on what the ATT would be under one assumption and what the ATT would be under another. In some cases, the ATT under one assumption is guaranteed to be larger than the ATT under another, thereby implying that, if one mistakenly assumes one assumption when the other is true, the ATT must be bounded between two values. The closer the ATTs under these two assumptions, the better.

This general logic remains important even without the bounding relationship. Whether or not the ATT under one assumption is guaranteed to be larger or smaller than under another, what matters is the degree of discordance. The closer the two ATTs are, the more robust the causal conclusion becomes under the parallel trends assumption.

To set up this framework, first define the discordance between the parallel trends imputation and any alternative imputation based on the average of detrended potential outcomes under no treatment as

$$\lambda_{g,v} := |\bar{\epsilon}_{g,v}(0) - \bar{\epsilon}_{0,T}(0)| \tag{21}$$

for any $(g, v) \in \mathcal{G} \times \mathcal{V}$. Under this notion of discordance, defined in terms of average detrended outcomes, concordance in detrended outcomes does not imply concordance in levels (i.e., on the original outcome scale). Conversely, concordance in levels does not imply concordance on the detrended scale.

The discordance in (21) provides a benchmark for bounding the error of the parallel trends imputation. This error of the parallel trends imputation, $|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)|$, can be bounded as follows:

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M\lambda_{g,v}. \tag{22}$$

The intuition behind (22) is straightforward. The true counterfactual is unobservable and, hence, so is the error of the parallel trends imputation, represented by the left-hand side of the inequality in (22). However, the observable quantity, $\bar{\epsilon}_{g,v}(0)$, can serve as a proxy for the unobservable counterfactual. With this proxy, one can calibrate the error in the parallel trends imputation based on the observable discordance between $\bar{\epsilon}_{g,v}(0)$ and $\bar{\epsilon}_{0,T}(0)$. Specifically, the model assumes that the error of the parallel trends imputation is bounded by some magnitude, $M \geq 0$, of the discordance $\lambda_{g,v}$. While this bound treats $\bar{\epsilon}_{g,v}(0)$ as the proxy for the counterfactual, the logic could be reversed: In principle, one could instead treat $\bar{\epsilon}_{0,T}(0)$ as a proxy and use it to bound the error in using $\bar{\epsilon}_{g,v}(0)$ as a counterfactual imputation.

The pointwise bound in (22) relies on a single proxy. However, a more flexible approach aggregates information across all possible proxies. Specifically, one can form a convex combination of the pointwise bounds to obtain the following model:

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w_{g,v}\lambda_{g,v}, \tag{23}$$

where the weights $\boldsymbol{w} \coloneqq \{w_{g,v}\}_{(g,v)\in\mathcal{G}\times\mathcal{V}}$, which lie in the unit simplex over $\mathcal{G}\times\mathcal{V}$ denoted by

$$\Delta_{\mathcal{G}\times\mathcal{V}} \coloneqq \mathbb{R}^{|\mathcal{G}\times\mathcal{V}|} \text{ with } w_{g,v} \geq 0 \text{ for all } (g,v) \text{ and } \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v} = 1, \tag{24}$$

represent the informativeness of each possible proxy for the counterfactual.

Proposition 4 in the appendix shows that this upper bound in (23) is sharp for every choice of weights $\boldsymbol{w}$. That is, for every $\boldsymbol{w} \in \Delta_{\mathcal{G}\times\mathcal{V}}$, there exists at least one configuration of $\bar{\epsilon}_{1,T}(0)$, $\bar{\epsilon}_{0,T}(0)$, and $\bar{\epsilon}_{g,v}(0)$ for all $(g,v)\in\mathcal{G}\times\mathcal{V}$ consistent with (23) in which the inequality in (23) holds with equality. Thus, no strictly smaller upper bound is compatible with the model.

This weighted formulation in (23) allows researchers to incorporate multiple proxies and to encode judgments about the relative informativeness of each one. For instance, if all proxies are equally informative, one could assign uniform weights:

$$w_{g,v} = \frac{1}{|\mathcal{G}\times\mathcal{V}|} \text{ for all } (g,v)\in\mathcal{G}\times\mathcal{V},$$

which leads to the average-bound model of

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M\frac{1}{|\mathcal{G}\times\mathcal{V}|}\sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} \lambda_{g,v}. \tag{25}$$

Other choices reflect different judgements: time-decaying weights can give more importance to proxies from periods closer to treatment, while group-specific weights can emphasize proxies from certain groups (e.g., the treated group).

In general, however, the choice of weights is difficult to justify, as any specification ultimately encodes prior judgments that may vary across analysts. To guard against undue reliance on subjective weighting choices, I adopt a weighting scheme for the discordance-based model in (23), analogous to the approach in (17) from Rambachan and Roth (2023). This weighting scheme places all weight on whichever group-periods attain the largest discordance from the parallel trends imputation.

Formally, this weighting scheme, $\boldsymbol{w}^{\max}$, is

$$w_{g,v}^{\max} \coloneqq \begin{cases} 1/|\mathcal{A}| & \text{if } (g,v)\in\mathcal{A} \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

The set $\mathcal{A}$ in (26), formally defined as

$$\mathcal{A} \coloneqq \{(g,v)\in\mathcal{G}\times\mathcal{V} : \lambda_{g,v} = \max_{(g',v')\in\mathcal{G}\times\mathcal{V}} \lambda_{g',v'}\},$$

consists of the group–period pairs that attain the maximum discordance. Although the weights in (26) are chosen to be uniform on $\mathcal{A}$, any probability distribution supported on this set would yield the same bound.

These weights in (26) produce the sensitivity model of

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M \max_{(g,v)\in\mathcal{G}\times\mathcal{V}} \lambda_{g,v}. \tag{27}$$

Under this model, the error of the parallel trends imputation is bounded by a magnitude, $M \geq 0$, of the largest discordance across all validation periods. This bound in (27) is the central discordance-based sensitivity model proposed in this paper.

Returning now to the two stylized examples in Figure 1, the value of the sensitivity model in (27) is clear. Figure 2 below shows that, while both models yield identical sensitivity in Case 1, they diverge substantially in Case 2.

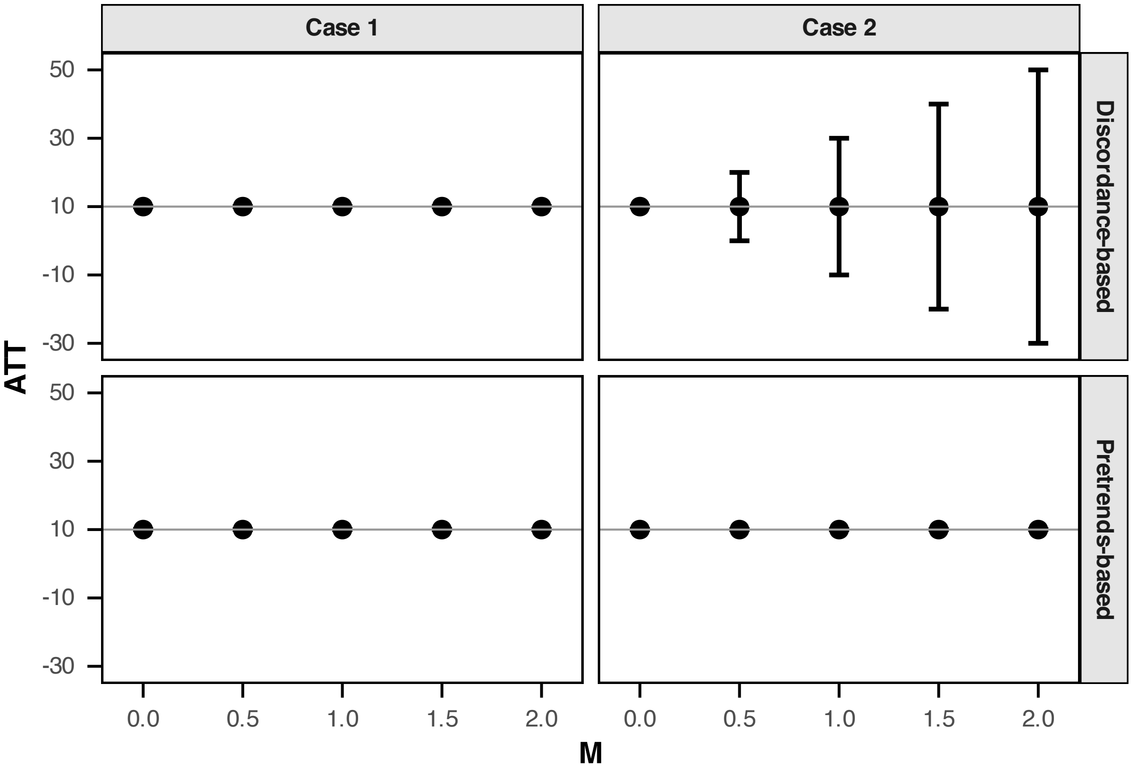

Figure 2: Sensitivity bounds for the two stylized scenarios in Figure 1, with columns representing the cases and rows indicating the sensitivity model. The x-axis shows values of the sensitivity parameter $M \in \{0, 0.5, 1, 1.5, 2\}$, and the y-axis shows the corresponding values of the ATT. Solid points represent the ATT under the parallel trends assumption. Vertical bars indicate the identified set for the ATT under violations of the parallel trends assumption bounded by M times the largest pre-period deviation.

Unlike the pretrends-based model in (17), the discordance-based model in (27) implies that the conclusion about the ATT depends more on parallel trends in Case 2 compared to Case 1.

The discordance-based model in (27) consists of two crucial components: a conception of discordance in (21) and a maximum-discordance weighting scheme in (26). I first justify the use of discordance as a principled foundation for the credibility of causal conclusions drawn under the parallel trends assumption, showing that agreement across multiple counterfactual imputations strengthens such conclusions. In doing so, this justification helps clarify two concepts often conflated in sensitivity analysis — (1) the plausibility of an identifying assumption and (2) the extent to which causal conclusions depend on it — and shows that concordance implies parallel pretrends, though not vice versa. I then provide a decision-theoretic justification demonstrating that the weighting scheme concentrating all mass on the maximal discordance is the minimax solution to a regret-based optimization problem. Together, these justifications provide a stronger foundation for the overall bound in (27).

5.1 Justification for Discordance

As noted above, one can use $\bar{\epsilon}_{g,v}(0)$ as a proxy for the counterfactual $\bar{\epsilon}_{1,T}(0)$, or vice versa, to bound the error of the counterfactual imputation. Following logic closely related to bracketing relationships, consider two scenarios:

1. The parallel trends imputation, $\bar{\epsilon}_{0,T}(0)$, is correct, but the proxy, $\bar{\epsilon}_{g,v}(0)$, is not.
2. The proxy, $\bar{\epsilon}_{g,v}(0)$, is correct, but the parallel trends imputation, $\bar{\epsilon}_{0,T}(0)$, is not.

In the first scenario, the parallel trends assumption holds exactly, and there is no error in the imputation. However, if one mistakenly assumes the first scenario when the second is actually true, then the size of the unobservable error will depend on the observable discordance between $\bar{\epsilon}_{0,T}(0)$ and $\bar{\epsilon}_{g,v}(0)$. Insofar as the second scenario could hold, then, all else equal, greater concordance between $\bar{\epsilon}_{0,T}(0)$ and $\bar{\epsilon}_{g,v}(0)$ implies a smaller overall error of the parallel trends imputation.

To formalize this reasoning within a sensitivity analysis framework based on set (rather than point) identification, consider the bound in (22), $|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M\lambda_{g,v}$. This bound is the set identification analogue of the point identification condition in the first scenario, whereby the parallel trends imputation is correct ($\bar{\epsilon}_{1,T}(0) = \bar{\epsilon}_{0,T}(0)$) and the proxy is not ($\bar{\epsilon}_{1,T}(0) \neq \bar{\epsilon}_{g,v}(0)$). Specifically, when $M = 0$ and $\lambda_{g,v} > 0$, the bound in (22) collapses to the point-identifying condition of Scenario 1.

The second scenario corresponds to an alternative bound,

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{g,v}(0)| \leq M\lambda_{g,v}. \tag{28}$$

This bound in (28) is the set identification analogue of the point identification condition in which the proxy is correct ($\bar{\epsilon}_{1,T}(0) = \bar{\epsilon}_{g,v}(0)$) and the parallel trends imputation is not ($\bar{\epsilon}_{1,T}(0) \neq \bar{\epsilon}_{0,T}(0)$). Specifically, when $M = 0$ and $\lambda_{g,v} > 0$, the bound in (28) collapses to the point-identifying condition of Scenario 2.

In (28), $\bar{\epsilon}_{g,v}(0)$ serves as the imputation and $\bar{\epsilon}_{0,T}(0)$ as the proxy. One could therefore redefine discordance as the distance between the imputation $\bar{\epsilon}_{g,v}(0)$ and the detrended mean of any other group-period index chosen to play the role of the proxy, not only $(0, T)$. Introducing this additional notation, however, would add complexity without changing the analysis: Because the proxy's group–period index in (28) is specifically $(0, T)$, and discordance in (21) is defined using an absolute value, the discordance in (28) is identical to that in (22).

Building on this alternative bound in (28), recall that $\tau_{0,T}$ and $\tau_{g,v}$ refer to the imputed values of the ATT when one imputes the counterfactual using $\bar{\epsilon}_{0,T}(0)$ or $\bar{\epsilon}_{g,v}(0)$, respectively.

When using the set restriction in (22), analogously denote the resulting lower and upper bounds of the imputed ATT by $\underline{\tau}_{0,T}$ and $\overline{\tau}_{0,T}$. By contrast, if one instead uses the set restriction in (28), I denote the corresponding bounds by $\underline{\tau}_{g,v}$ and $\overline{\tau}_{g,v}$.

Now suppose a prior distribution, π, over whether these two restrictions in (22) and (28) hold. The corresponding probabilities are defined as follows:

- $p_{(0,T),(g,v)}$ denotes the probability that both restrictions hold;
- $p_{(0,T),\neg(g,v)}$ and $p_{\neg(0,T),(g,v)}$ denote the respective probabilities that only one holds; and
- $p_{\neg(0,T),\neg(g,v)}$ denotes the probability that neither holds.

Proposition 1 below shows that, so long as $p_{\neg(0,T),(g,v)} > 0$, the expected errors of the ATT's bounds under the set restriction in (22) are increasing in discordance.

Proposition 1. *If $p_{\neg(0,T),(g,v)} > 0$, then, for any fixed $M \geq 0$, both $\mathrm{E}_\pi\left[|\overline{\tau} - \overline{\tau}_{0,T}|\right]$ and $\mathrm{E}_\pi\left[|\underline{\tau} - \underline{\tau}_{0,T}|\right]$ are strictly increasing in $\lambda_{g,v}$, where $\mathrm{E}_\pi\left[\cdot\right]$ denotes the expectation with respect to the joint prior distribution, π.*

Proposition 1 directly ties discordance to the robustness of causal conclusions.

Proposition 1 also clarifies the distinction between the plausibility of the parallel trends identification condition — formally, its marginal probability $p_{(0,T),(g,v)} + p_{(0,T),\neg(g,v)}$ — and the extent to which causal conclusions depend on that condition. To explicate this distinction between dependence on (rather than plausibility of) the parallel trends assumption, recall the two different set restrictions for the parallel trends imputation, $\bar{\epsilon}_{0,T}(0)$, and that of $\bar{\epsilon}_{g,v}(0)$:

- Equation (22): $|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M\lambda_{g,v}$,
- Equation (28): $|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{g,v}(0)| \leq M\lambda_{g,v}$.

If both restrictions are false, then the error of the bounds on the ATT under (22) is some fixed but unknown value. However, if (28) is true and (22) is false, then the error of the ATT's bounds under the incorrect restriction is simply the difference between the two bounding sets. This difference is driven entirely by the discordance between the parallel trends imputation, $\bar{\epsilon}_{0,T}(0)$, and the alternative imputation, $\bar{\epsilon}_{g,v}(0)$. Thus, discordance is not directly about plausibility. Rather, discordance captures the degree of dependence of inferences on the parallel trends identification condition; that is, if the condition were violated, how different would the conclusions be?

That said, concordance is informative only insofar as the identifying condition underlying the imputed quantity $\bar{\epsilon}_{g,v}(0)$ is itself at least somewhat plausible. A conclusion that would change little under a different identifying assumption is not meaningful if that alternative assumption is wholly implausible. This is why Proposition 1 includes in its if–then statement the antecedent condition $p\neg(0,T),(g,v) > 0$, reflecting a positive probability that the parallel trends assumption is false and that the alternative assumption holds.

Proposition 1 offers a formal justification for why one ought to quantify robustness to violations of parallel trends in terms of discordance between the parallel trends imputation and alternative imputations. Importantly, a discordance-based model accords with much of the intuition underlying parallel pretrends. I now show the connection between these two modes in which concordance implies parallel pretrends, but the converse is not true.

To formalize this argument, consider the detrended means of each group ($G = 1$ and $G = 0$) for any validation period in $\mathcal{V}$. The joint discordance is defined as a weighted sum of the absolute discordances of these detrended means from the benchmark $\bar{\epsilon}_{0,T}(0)$. Proposition 2 shows that, for any $v \in \mathcal{V}$, a small upper bound on joint discordance is sufficient to ensure a small violation of parallel pretrends.

Proposition 2. *For any pre-treatment validation period, $v \in \mathcal{V}$, the violation of parallel trends, $|\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,v}(0)|$, is bounded above by*

$$|\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,v}(0)| \leq 2 \max \left\{ \lambda_{1,v}, \, \lambda_{0,v} \right\}, \tag{29}$$

where the maximum on the right-hand side is an upper bound on the joint discordance between $\bar{\epsilon}_{0,T}(0)$ and the pair $(\bar{\epsilon}_{1,v}(0), \, \bar{\epsilon}_{0,v}(0))$.

The proposition establishes that when the joint discordance is constrained to be small — i.e., its upper bound is small — then the violation of parallel pretrends must also be small. In the limiting case in which the joint discordance is constrained to be zero, parallel pretrends cannot be violated. However, the converse does not hold: a small violation of parallel pretrends does not imply that the upper bound on discordance is small.

5.2 Justification for Maximum-Discordance Weighting

As noted above, the discordance-based sensitivity model in (27) has two key components: a conception of discordance in (21) and a maximum-discordance weighting scheme in (26). Thus far, I have provided a foundation justifying why discordance offers a reasonable basis for restricting violations of the parallel trends assumption. I now turn to justifying the specific weights in (27), which concentrate all mass on the $(g, v) \in \mathcal{G} \times \mathcal{V}$ that attain the maximum discordance.

The maximum-discordance weighting scheme has an important decision-theoretic justification as the solution to an adversarial game. In this game, both the worst-case violation of parallel trends and the bound implied by any proposed weighting scheme are determined under the same sensitivity model for a given $M \geq 0$. One then selects weights to minimize worst-case regret, defined as the largest possible gap between the worst-case violation of parallel trends and the bound implied by the selected weighting scheme. The weights in (26) are those that solve this minimax regret problem.

As others have noted in different sensitivity analysis frameworks (e.g., Cohen et al., 2020), it is intuitive to think of an adversarial game as a way to justify using the worst-case (widest) set restriction for any given $M \geq 0$. If the adversary represents future counterclaims against a study's causal conclusion, then adopting the widest bounds ensures that no counterclaim can overturn the researcher's conclusion for that value of $M \geq 0$. If a researcher can rule out certain values of the ATT under the worst-case (widest) sensitivity interval, then those values must also be ruled out under any narrower intervals.

To formally establish that the weights in (26) minimize worst-case regret, let $\psi := |\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)|$ denote the absolute error of the parallel trends imputation. Under the sensitivity model in (23), ψ is only partially identified: For all possible weights $\boldsymbol{w} \in \Delta_{\mathcal{G}\times\mathcal{V}}$, where $\Delta_{\mathcal{G}\times\mathcal{V}}$ denotes the unit simplex over $\mathcal{G} \times \mathcal{V}$ defined in (24), the model in (23) implies

$$0 \leq \psi \leq M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\lambda_{g,v}.$$

Thus, the set of all possible values for ψ under the sensitivity model is

$$\Psi := \left\{ \psi \in \mathbb{R}_{\geq 0} : \psi \leq M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\lambda_{g,v} \text{ for all } \boldsymbol{w} \in \Delta_{\mathcal{G}\times\mathcal{V}} \right\}, \tag{30}$$

where $\mathbb{R}_{\geq 0}$ is the set of nonnegative reals.

For any $\boldsymbol{w} \in \Delta_{\mathcal{G}\times\mathcal{V}}$, the corresponding worst-case regret is how much larger the absolute error of the parallel-trends imputation could be — given that ψ must lie in Ψ — relative to the bound implied by that weighting scheme. Formally, this worst-case regret is

$$R(\boldsymbol{w}) := \sup_{\psi\in\Psi} \left\{ \psi - M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\lambda_{g,v} \right\}. \tag{31}$$

The $\sup_{\psi\in\Psi}\{\cdot\}$ in (31) denotes the least upper bound of the set of all values that the expression in braces, $\{\cdot\}$, attains as ψ ranges over Ψ.

Proposition 3 now establishes that the model in (27), which places all weight on the group-periods that attain the maximum discordance, is equivalent to the model in (23) with weights that minimize this worst-case regret in (31).

Proposition 3. *The sensitivity model in* (27)*, which uses the weighting scheme* $\boldsymbol{w}^{\max}$ *in* (26)*, is equivalent to the general model in* (23) *with weights that solve the following minimax regret optimization problem:*

$$\boldsymbol{w} = \operatorname*{arg\,min}_{\boldsymbol{w}\in\Delta_{\mathcal{G}\times\mathcal{V}}} R(\boldsymbol{w}). \tag{32}$$

The optimization problem in (32), which is expressed in terms of the error of the parallel-trends imputation for a given weighting scheme, implicitly targets the width of the ATT's bounds implied by that scheme, effectively selecting the weighting scheme that yields the widest identified set for the ATT permitted by the model.

Crucially, for any $M \geq 0$, the identified intervals for all $(g, v) \in \mathcal{G} \times \mathcal{V}$ share the same center — the ATT implied by the parallel trends imputation. Different choices of weights, $\boldsymbol{w}$, affect only the width of the bounds, not their location. The minimax weights correspond to the maximum possible width across all feasible weighting schemes. Thus, any interval constructed using weights other than the worst-case weights must be bracketed within the worst-case interval. Consequently, no alternative weighting scheme can imply a value of the ATT that falls outside the researcher's bounds, and any such alternative necessarily implies less sensitivity than the minimax choice.

6. Extension to Multiple Post-Periods

Up to this point, the discordance-based sensitivity model has focused on the simple setting of a single post-treatment period. To accommodate settings with multiple post-treatment periods, it is helpful to begin by clarifying the causal estimands of interest. With several post-periods, the treated group $G = 1$ now has an ATT for each post-treatment period. I first consider the case in which the target is an aggregated estimand — namely, a linear combination of these period-specific ATTs — which requires only a minor modification of the notation introduced above. I then consider the case in which each post-period ATT is itself an estimand of interest, a setting that necessitates additional notation.

Let $T_0 \geq 1$ denote the number of pre-treatment periods and let $T_1 \geq 1$ denote the number of post-treatment periods, so that $T_0 + T_1 = T$. Define the sets of pre- and post-treatment periods as $\mathcal{T}_{\text{pre}} \coloneqq \{1, \dots, T_0\}$ and $\mathcal{T}_{\text{post}} \coloneqq \{T_0 + 1, \dots, T\}$, respectively. In the earlier setup with a single post-period, $T_1 = 1$ and $\mathcal{T}_{\text{post}} = \{T\}$.

The treated group $G = 1$ first enters treatment in period $T_0 + 1$. For each post-treatment period $s \in \mathcal{T}_{\text{post}}$, define the ATT for each post-period as

$$\tau^s \coloneqq \bar{Y}_{1,s}(1) - \bar{Y}_{1,s}(0). \tag{33}$$

With multiple post-treatment periods, the treated group $G = 1$ therefore has a distinct ATT for each $s \in \mathcal{T}_{\text{post}}$.

Denote the vector of post-treatment ATTs for group $G = 1$, with one component for each post-treatment period, by

$$\boldsymbol{\tau} \coloneqq \left(\tau^{T_0+1}, \dots, \tau^{T}\right)^{\top}, \tag{34}$$

where the superscript $^{\top}$ denotes transposition. Following Rambachan and Roth (2023, p. 2561), define an aggregated estimand as $\theta \coloneqq \boldsymbol{\ell}^{\top}\boldsymbol{\tau}$, where $\boldsymbol{\ell}$ is a known vector of the same dimension as $\boldsymbol{\tau}$. This formulation accommodates simple averages, weighted averages, and other linear combinations of the post-treatment ATTs.

When the estimand of interest is this aggregated quantity θ, we can replace the post-period quantity $Y_{g,T}(z)$ from the earlier framework with $\boldsymbol{\ell}^{\top}\boldsymbol{Y}_{g,\text{post}}(z)$, where $\boldsymbol{Y}_{g,\text{post}}(z)$ collects the treated ($z = 1$) or control ($z = 0$) potential outcomes for group g across all post-treatment periods $s \in \mathcal{T}_{\text{post}}$. With this replacement, all algebraic steps and arguments from the single-period setting apply directly. The sensitivity analysis then evaluates robustness for the aggregated estimand θ.

By contrast, when the analysis targets the ATTs for each post-treatment period individually, additional notation is in order. This need for additional notation is because outcomes should be detrended via a group's mean in pre-treatment periods so that we have an uncontaminated prediction of the treated group's mean counterfactual outcome in the post-period. Using a period already affected by treatment would presumably yield a distorted prediction of the average outcome that would have arisen under no treatment. For the control group, which is not treated in the post-period, the same reference period is used to ensure symmetry, so that differences between the treated group's counterfactual detrended mean and the control group's observable detrended mean reflect deviations from parallel trends rather than artifacts of differing baseline periods.

Therefore, redefine the detrended outcome for group g in period t using the most recent pre-treatment period, which, for any period $t \in \{1, \dots, T\}$, is $\min\{t-1, T_0\}$. Formally, this detrended outcome is

$$\epsilon_{g,t} \coloneqq Y_{g,t} - \bar{Y}_{g,\min\{t-1,T_0\}}. \tag{35}$$

When t is itself a pre-treatment period, (35) reduces to detrending the outcome using the group's mean in the immediately preceding period, just as in the single post-period setup. However, when t is a post-treatment period, the outcome is detrended using the most recent pre-treatment period, ensuring that the detrending baseline is not contaminated by treatment.

As before, detrending preserves the ATT. The ATT expressed in terms of these detrended outcomes in (35) is algebraically equivalent to the ATT defined using the raw potential outcomes. The difference now is that the treated group $G = 1$ has a distinct ATT for each post-treatment period, $s \in \mathcal{T}_{\text{post}}$.

For any such s, I analogously define a sensitivity model in which the violation of parallel trends is governed by some magnitude $M \geq 0$ of the maximum discordance between the parallel trends imputation and any other imputation based on pre-treatment validation periods. Let $\mathcal{V} \subseteq \mathcal{T}_{\text{pre}}$ denote the set of pre-treatment validation periods selected for analysis, with each $v \in \mathcal{V}$ indexing a single pre-treatment period. Denote the discordance between the parallel trends imputation and the average detrended outcome based on group g and validation period v as $\lambda^s_{g,v} \coloneqq |\bar{\epsilon}_{g,v} - \bar{\epsilon}_{0,s}(0)|$. Then an immediate extension of the discordance-based sensitivity model in (27) is

$$|\bar{\epsilon}_{1,s}(0) - \bar{\epsilon}_{0,s}(0)| \leq M \max_{(g,v)\in\mathcal{G}\times\mathcal{V}} \lambda^s_{g,v} \tag{36}$$

for all $s \in \mathcal{T}_{\text{post}}$.

Under this model in (36), the sensitivity analysis applies period-by-period in exactly the same manner as in the single post-period case. Various "shock" scenarios — such as a control-only shock, common shock, or more complex temporal patterns — remain fully encompassed by this framework. These scenarios merely alter the observed values of the discordance terms $\lambda^s_{g,v}$, which in turn determine the width of the sensitivity bounds on ATT^s. Thus, even with multiple post-treatment periods, the core principle persists: Sensitivity is governed by the extent to which the parallel trends imputation and alternative imputations based on pre-treatment periods disagree.

The distinction between targeting the aggregated estimand θ and the period-specific ATTs is important. Because the maximization in (36) is nonlinear, a sensitivity analysis applied separately to each post-period and then aggregated using $\boldsymbol{\ell}$ is generally not equivalent to one applied directly to θ. Therefore, when the parameter of interest is the linear combination θ, the sensitivity analysis should be applied directly to θ, as the worst-case discordance framework is designed to provide a single robustness assessment for a single estimand — whether aggregated or not — rather than an aggregation of robustness assessments computed separately for multiple estimands.

7. Application

I now present an applied example that illustrates how the discordance-based sensitivity analysis can yield different conclusions than existing approaches. The example is drawn from Wilse-Samson (2013), which investigates whether the sudden withdrawal of foreign mine labor from Malawi and Mozambique in the mid-1970s affected electoral support for apartheid policies in South African districts that depended on that labor. The aim of the analysis is to estimate the average effect of this labor supply shock on the voting behavior of White South Africans in mining districts. This application fits squarely within the canonical DID framework, featuring a well-defined treatment shock, variation in exposure across districts, and observations from both before and after treatment. Related studies have also leveraged this labor supply shock in their own DID designs (e.g., Dinkelman and Mariotti, 2016; Dinkelman et al., 2024).

As Wilse-Samson (2013) explains — and as comprehensive historical accounts detail (De Kiewiet, 1941; Thompson, 1990; Welsh, 2009; Dubow, 2014; O'Meara, 1996; Feinstein, 2005) — South Africa's apartheid regime faced a structural tension: On the one hand, the regime required large, concentrated labor forces to power urban industries like mining. Yet, on the other hand, the apartheid regime sought to limit Black urbanization out of fear of Black majority rule, instead relegating Black South Africans to fragmented "homelands." Presumably in an effort to manage this tension, the apartheid regime relied on foreign migrant labor to work in the mines, a strategy that many historians interpret as an attempt to reduce the political risks posed by a large, urbanized Black domestic workforce.

This balance was disrupted by a dramatic labor shock between South Africa's 1974 and 1977 national elections (Crush, 1993). In 1974, a plane carrying Malawian migrant laborers en route to South Africa crashed, killing 77 passengers. In response, Malawian President Hastings Banda suspended all migrant labor to South Africa. Around the same time, Portugal's *Revolução dos Cravos* in 1974 and Mozambique's subsequent independence in 1975 sharply curtailed the flow of Mozambican labor. These events — external to South Africa's mining sector — reduced the number of foreign mine workers from 336,000 in 1974 to 208,000 in 1977 (Wilse-Samson, 2013).

Likely in response to pressure from the mining industry, the apartheid regime allowed companies to expand recruitment of domestic Black South Africans. Their numbers rose between 1974 and 1977 from 86,000 to 214,000, often under longer contracts and with higher wages, contributing to a more stable urban workforce. As a result, the share of foreign miners fell from 78% in 1974 to under 50% in 1977 (Wilse-Samson, 2013). Given the apartheid regime's aforementioned structural tension, this shift may have provoked fears of Black majority rule and, in turn, heightened support for far-right parties seeking to entrench apartheid. The DID design in Wilse-Samson (2013) offers a way to investigate this possibility.

The data from Wilse-Samson (2013) include 153 electoral districts with right-wing vote share measured in elections both before and after the labor shock. Of these 153 districts, 14 contained at least one mine (and thus were directly exposed to the labor shock), while the remaining 139 did not. The pre-period includes the 1961, 1966, 1970, and 1974 national parliamentary elections. The post-period is collapsed into a single time point, denoted as 1977+. The outcome — right-wing vote share — is the proportion of voters in a district who supported parties to the right of the ruling National Party (NP), such as the Herstigte Nasionale Party (HNP) and the Conservative Party (CP).

Figure 3 below shows the trends in right-wing vote shares between districts with and without mines before and after the labor shock.

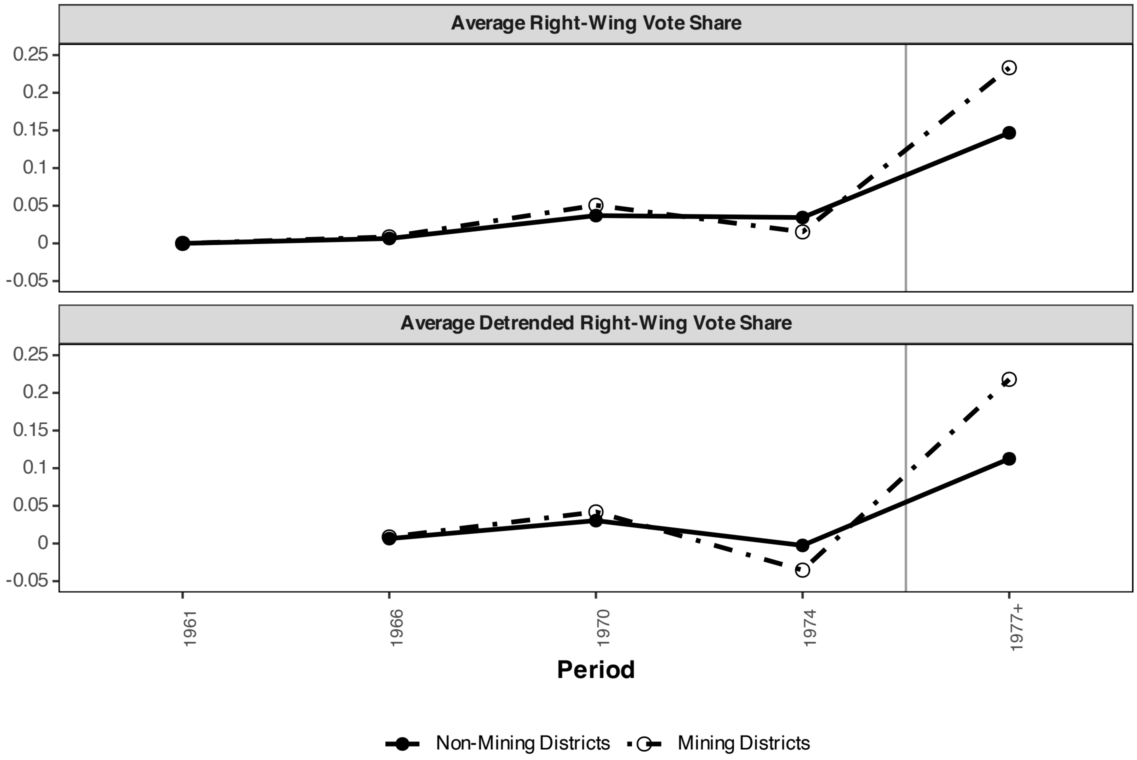

Figure 3: The average right-wing vote share in electoral districts with at least one gold mine (dashed line) and those without any gold mines (solid line). The top panel plots the average right-wing vote share on its original scale, while the bottom panel plots the detrended series, defined as each group's change in average right-wing vote share relative to its mean in the preceding election period. The x-axis shows national election years. The election years 1961, 1966, 1970, and 1974 are shown as distinct time points, while "1977+' represents the post-treatment election years (1977, 1981, and 1987). In the top panel, each point reflects the average right-wing vote share across all mining or non-mining districts for that election year. For "1977+", the value is computed by first averaging within each district across all post-treatment election years and then averaging these district-level means within each group. In the bottom panel, each point reflects the average detrended outcome, with detrending for "1977+" performed after first computing each district's average outcome across all post-treatment periods. The vertical grey line indicates the start of the treatment period for mining districts, beginning with the first election after 1974.

As Figure 3 shows, trends in right-wing vote share are relatively stable across mining and non-mining districts before 1974. After 1974, support increases in both groups, but the rise is substantially larger in mining districts.

Under the point identification condition that $M = 0$ in (27), which implies parallel trends, the estimated ATT is 0.11. Using a cluster bootstrap at the electoral district level, the estimated standard error (SE) is 0.02. Based on these ATT and SE estimates, the resulting 95% confidence interval (CI) is $[0.07, 0.15]$. Because inference is clustered at the individual level, each bootstrap draw resamples whole units and preserves the full pre- and post-treatment outcome paths for each unit. Detrending is therefore just a deterministic, unit-level transformation applied after resampling, which does not alter the sampling distribution produced by the clustered bootstrap.

The control group's large post-period deviation from pre-period trend suggests that the estimated ATT is highly dependent on the parallel trends assumption. In other words, suppose the average change from 1974 to 1977+ in the non-mining districts did *not* represent what the average change would have been in the mining districts without the labor shock. If instead the pre-period change in mining districts from 1970 to 1974 was the correct imputation for the counterfactual, then the ATT would be dramatically different (0.25 compared to 0.11).

In other words, the discordance is large. To see this directly from Figure 3, note first that the parallel trends imputation is 0.112, the average detrended post-treatment outcome for the control group (non-mining districts). When combined with the treated group's average detrended post-treatment outcome of 0.218, the resulting ATT point estimate is approximately 0.11. As can be seen directly from Figure 3, the remaining possible counterfactual imputations — ordered by group (0, then 1) and chronologically from 1966 through 1974 — are 0.007, 0.009, 0.030, 0.042, −0.003, and −0.035. Thus, the maximum discordance between the parallel trends imputation (0.112) and any other feasible imputation is approximately $|0.112 - (-0.035)| = 0.147$. Up to rounding, this 0.147 is exactly the difference between the two imputed ATTs (0.25 and 0.11) described in the paragraph above: the former obtained using the treated group's most recent pre-period average detrended outcome and the latter using the control group's post-treatment average detrended outcome.

This maximum discordance of approximately 0.15 implies that the ATT's estimated bounds for, say, $M = 1$ are roughly $[0.11 - 0.15,\ 0.11 + 0.15] = [-0.04,\ 0.26]$. Unlike the case when $M = 0$, accounting for sampling uncertainty in this setting with $M > 0$ is more challenging. Because the sample analogue of (27) includes a maximum operator, the usual bootstrap — appropriate in the $M = 0$ case above — may be inconsistent, such as when the population-level joint CDF assigns the maximal discordance to multiple group–periods.

I therefore draw on the "intersection–union' approach of Berger and Hsu (1996), which is known to produce valid, though often conservative, confidence intervals (see also Ye et al., 2024). In practice, I use the aforementioned district-level bootstrap procedure in which, for each bootstrap replication, I resample electoral districts with replacement and recompute both the DID estimator and the discordance for every group–period pair in $\mathcal{G} \times \mathcal{V}$ on the resulting resample. This yields a bootstrap distribution of the DID estimator and, for each group–period pair, a bootstrap distribution of discordance. For each pair, I then combine the bootstrap draws of the DID estimator and that pair's discordance to obtain bootstrap distributions for the corresponding lower and upper ATT bounds under the sensitivity model, from which I take the $\alpha/2$ and $1-\alpha/2$ quantiles. Finally, following the intersection–union logic of Berger and Hsu (1996), I combine these candidate intervals across all group-period pairs by taking the smallest lower endpoint and the largest upper endpoint. The resulting interval is valid, though conservative, because it covers the ATT bounds regardless of which group–period pair generates the maximal discordance in the population.

Figure 4 shows that the ATT's estimated bounds under the discordance-based model are highly sensitive to deviations from $M = 0$. Confidence intervals are computed using a cluster bootstrap together with the aforementioned "intersection-union" method from Berger and Hsu (1996).

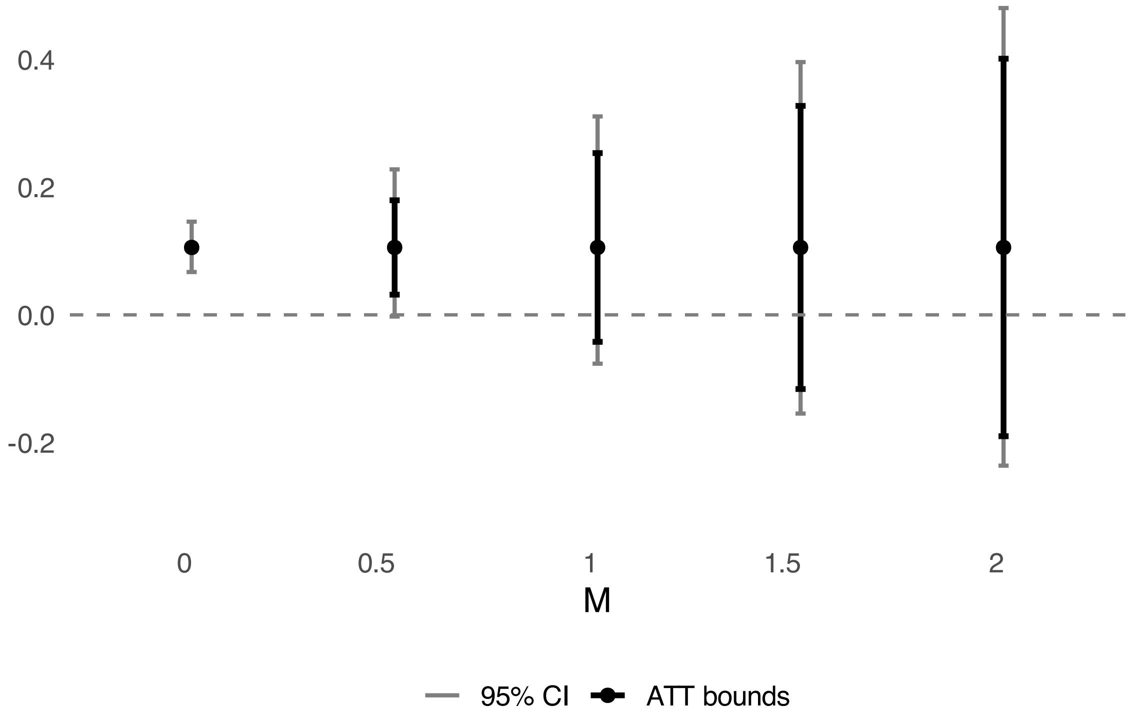

Figure 4: Estimates of the ATT under increasing values of $M \geq 0$. The x-axis is the parameter M, a user-specified magnitude of the maximum discordance, which implied a bound on the parallel trends violation. For each value of M, the black point marks the DID point estimate of the ATT. The black lines show the sample-based identified set for the ATT under that value of M, while the gray lines show 95% CIs constructed around the upper and lower bounds of the ATT.

This high sensitivity is expected given that, as shown in Figure 3, the maximum discordance reaches 0.147. As a result, even modest values of $M > 0$ permit violations of parallel trends large enough that the ATT's bounds include zero — meaning one can no longer rule out the possibility of no average effect on the treated. Ignoring sampling uncertainty, the changepoint value of M at which the bounds first include zero is $M = 0.71$. When one incorporates sampling uncertainty, this changepoint drops to $M = 0.48$.

Note that, as Figure 3 also shows, even though discordance is high, violations of parallel trends in the pre-period are small, with a maximum absolute difference of 0.03. This combination of mild pre-period violations and substantial discordance indicates that, although conclusions may be sensitive under the discordance-based model, they are robust under the pretrends-based sensitivity model. Figure 5 below shows the dramatically different conclusions one would reach depending on the sensitivity analysis model one uses.

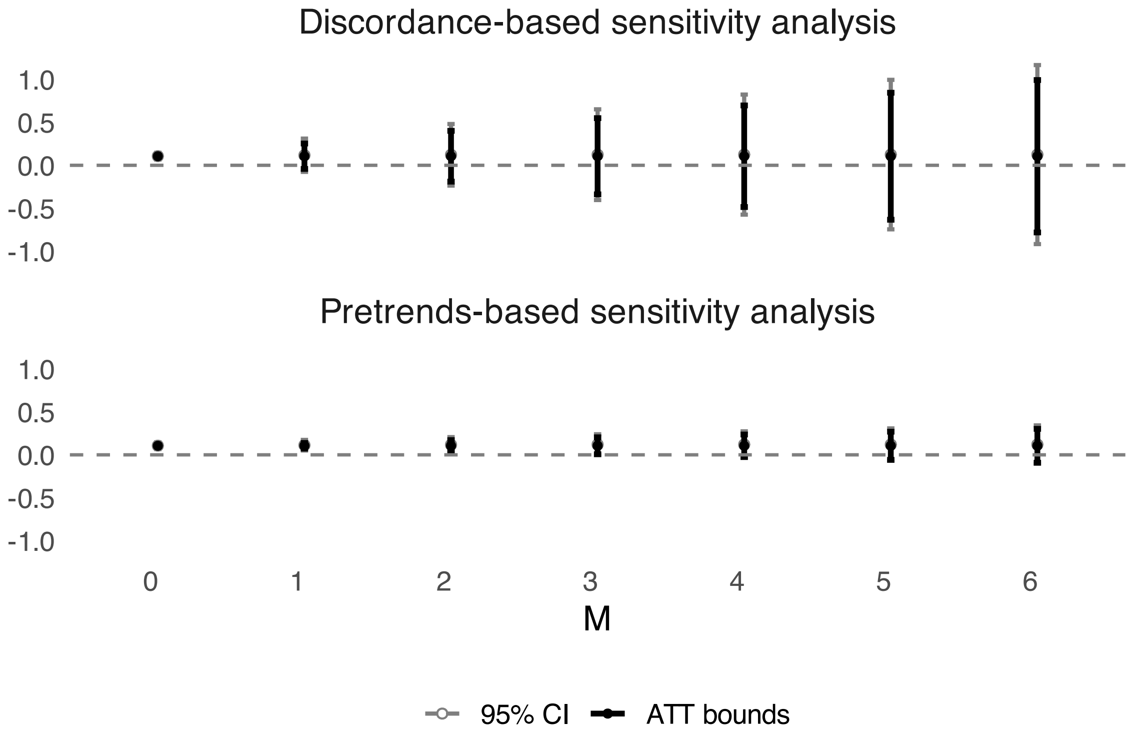

Figure 5: Estimates of the ATT under increasing values of $M \geq 0$. Each panel corresponds to a different approach for bounding violations of the parallel trends assumption: one based on discordance across group-time comparisons in the pre-period, and the other based on differences in trends across pre-treatment periods. The x-axis is the parameter M, which controls the maximum allowed deviation from parallel trends under each approach. For each value of M, the black point marks the DID point estimate of the ATT. The black lines show the identified set for the ATT given that value of M, while the gray lines show 95% CIs constructed around the upper and lower bounds of the ATT.

In both cases, the point estimate when $M = 0$ is 0.11. However, for the pretrends-based sensitivity analysis, the estimates of the ATT's lower and upper bounds bracket 0 at $M = 3.22$. The same changepoint value incorporating sampling uncertainty is $M = 2.98$. Both of these values are much greater than 0.71 and 0.48 in the discordance-based sensitivity analysis.

Although the discordance-based model suggests limited robustness to violations of parallel trends, the parallel trends assumption may still be plausible in the Wilse-Samson (2013) application. Between 1974 and 1977, South Africa experienced not only the mining labor shock, but also major nationwide events. The most notable of which was the Soweto uprisings of June 1976, the largest mass protest by Black South Africans against apartheid rule. To the extent that these uprisings and their media coverage heightened fears of majority rule among White South Africans, both mining and non-mining districts may have exhibited similar increases in right-wing vote shares. Such parallel movement across districts would support the plausibility of common shocks, suggesting that any post-treatment difference between groups arose from mining districts' exposure to the labor shock. Thus, although Figure 4 illustrates how inferences about the ATT would change under violations of parallel trends, the assumption of parallel trends itself could still be plausible in this setting.

8. Conclusion

It is well known that DID relies on the parallel trends assumption, under which the control group's post-treatment change is used to impute the treated group's counterfactual change. Foundational work on quasi-experiments emphasizes that the credibility of conclusions under this assumption hinges on whether they would persist under other plausible assumptions. Parallel trends provides only one imputation, and alternative trend assumptions — each vulnerable to different forms of confounding — can yield different implied ATTs. Existing pretrends-based sensitivity analyses focus exclusively on pre-treatment deviations from parallel trends. As a result, such analyses miss cases in which other plausible assumptions lead to discordant conclusions, such as when the control group's post-treatment trend departs from its pre-treatment pattern even with parallel pretrends.

This paper develops a discordance-based sensitivity analysis that formalizes the idea that credible DID inferences depend on concordance across multiple strands of evidence. The framework bounds violations of parallel trends using the worst-case discordance between the standard parallel trends imputation and other plausible imputations. I provide a formal explanation for why such concordance strengthens the credibility of DID conclusions, show that concordance across these imputations implies parallel pretrends (but not vice versa), and offer a decision-theoretic justification for the worst-case strategy. I illustrate the practical implications of the discordance- and pretrends-based approaches with an application in which the two methods produce markedly different assessments of robustness.

In conclusion, it is worth reiterating a point in one of the most prominent sensitivity analysis frameworks — albeit tailored to a different setting:

> "To say that a result is sensitive to small biases is not to assert that bias is present, nor to assert that there is evidence against a treatment effect, but simply to observe that small biases could account for the ostensible effect" (Rosenbaum and Krieger, 1990, p. 497).

A sensitivity analysis does not address whether an assumption is plausible. Rather, a sensitivity analysis speaks to how dependent conclusions are on an assumption (e.g., parallel trends). Judgments about plausibility come into play when one interprets the results of a sensitivity analysis. For example, a sensitivity analysis one mechanically applies to a randomized experiment might show high sensitivity. However, this high sensitivity would not undermine the conclusions from that experiment because violations of randomization are known to be false.

Analogously, the discordance-based sensitivity model developed in this paper is designed to capture dependence on — rather than the plausibility of — the parallel trends assumption, indicating high dependence exactly in the settings where intuition suggests it should (as illustrated in Figure 1). In particular, the model is motivated by scenarios in which the standard parallel trends imputation yields conclusions that differ substantially from those based on pretrends-based evidence. Such discordance indicates that conclusions rely heavily on the parallel trends assumption, since alternative assumptions — justifying different counterfactual imputations — would yield markedly different results.

For practical purposes, because parallel pretrends are a necessary but not sufficient condition for low sensitivity to violations of parallel trends, the discordance-based model will often show low sensitivity when pretrends are parallel, producing sensitivity analyses that agree with pretrends-based approaches. However, when substantial discordance persists — such as when the control group's post-period mean breaks from its pretrend — the two assessments of sensitivity can diverge, sometimes dramatically. Ultimately, the choice between this model and pretrends-based alternatives requires careful judgment about which benchmark — discordance or pretrends — is more appropriate for conceptualizing dependence on parallel trends. That choice can carry important implications for the robustness of the conclusions one draws from DID designs.

Appendix A.

Proof of Proposition 1

Proof. The expected value of $\overline{\tau}$ with respect to π is

$$\begin{aligned}&p_{(0,T),(g,v)}\overline{\tau}_{(0,T),(g,v)} + p_{(0,T),\neg(g,v)}\overline{\tau}_{(0,T),\neg(g,v)}\\&+ p_{\neg(0,T),(g,v)}\overline{\tau}_{\neg(0,T),(g,v)} + p_{\neg(0,T),\neg(g,v)}\overline{\tau}_{\neg(0,T),\neg(g,v)},\end{aligned} \tag{37}$$

where

- $\overline{\tau}_{(0,T),(g,v)}$ denotes the upper bound of the ATT when both set restrictions in (22) and (28) hold;
- $\overline{\tau}_{(0,T),\neg(g,v)}$ and $\overline{\tau}_{\neg(0,T),(g,v)}$ denote the upper bound of the ATT when only one of the two restrictions in (22) and (28) holds; and
- $\overline{\tau}_{\neg(0,T),\neg(g,v)}$ denotes the upper bound of the ATT when neither of the restrictions in (22) and (28) holds.

If both restrictions in (22) and (28) hold, then $\bar{\epsilon}_{1,T}(0)$ must lie in the intersection of

$$[\bar{\epsilon}_{0,T}(0) - M\lambda_{g,v},\, \bar{\epsilon}_{0,T}(0) + M\lambda_{g,v}] \quad \text{and} \quad [\bar{\epsilon}_{g,v}(0) - M\lambda_{g,v},\, \bar{\epsilon}_{g,v}(0) + M\lambda_{g,v}].$$

For any two intervals $[a, b]$ and $[c, d]$, their intersection is given by

$$[a, b] \cap [c, d] = [\max\{a, c\},\, \min\{b, d\}]$$

whenever $\max\{a, c\} \leq \min\{b, d\}$. Applying this characterization yields

$$[\max\{\bar{\epsilon}_{0,T}(0) - M\lambda_{g,v},\, \bar{\epsilon}_{g,v}(0) - M\lambda_{g,v}\},\, \min\{\bar{\epsilon}_{0,T}(0) + M\lambda_{g,v},\, \bar{\epsilon}_{g,v}(0) + M\lambda_{g,v}\}]. \tag{38}$$

Therefore, when both restrictions in (22) and (28) hold, the upper bound of the ATT is attained by minimizing $\bar{\epsilon}_{1,T}(0)$. Given the restriction in (38), the least possible value of $\bar{\epsilon}_{1,T}(0)$ is $\max\{\bar{\epsilon}_{0,T}(0) - M\lambda_{g,v},\, \bar{\epsilon}_{g,v}(0) - M\lambda_{g,v}\}$. Hence, the upper bound of the ATT is

$$\overline{\tau}_{(0,T),(g,v)} = \bar{\epsilon}_{1,T} - \max\{\bar{\epsilon}_{0,T}(0) - M\lambda_{g,v},\, \bar{\epsilon}_{g,v}(0) - M\lambda_{g,v}\},$$

which, because

$$\begin{aligned}\overline{\tau}_{0,T} &= \bar{\epsilon}_{1,T} - (\bar{\epsilon}_{0,T}(0) - M\lambda_{g,v}) \text{ and}\\ \overline{\tau}_{g,v} &= \bar{\epsilon}_{1,T} - (\bar{\epsilon}_{g,v}(0) - M\lambda_{g,v}),\end{aligned}$$

simplifies to

$$\overline{\tau}_{(0,T),(g,v)} = \min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\}.$$

Consequently, the expected value of $\overline{\tau}$ in (37) is

$$\begin{aligned}&p_{(0,T),(g,v)}\min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} + p_{(0,T),\neg(g,v)}\overline{\tau}_{(0,T),\neg(g,v)}\\&+ p_{\neg(0,T),(g,v)}\overline{\tau}_{\neg(0,T),(g,v)} + p_{\neg(0,T),\neg(g,v)}\overline{\tau}_{\neg(0,T),\neg(g,v)}.\end{aligned} \tag{39}$$

Then, again noting that the set restrictions in (22) and (28) imply

$$\begin{aligned}\overline{\tau}_{0,T} &= \bar{\epsilon}_{1,T} - \bar{\epsilon}_{0,T} + M\lambda_{g,v}\\ \overline{\tau}_{g,v} &= \bar{\epsilon}_{1,T} - \bar{\epsilon}_{g,v} + M\lambda_{g,v},\end{aligned}$$

we can plug in $\bar{\epsilon}_{1,T} - \bar{\epsilon}_{0,T} + M\lambda_{g,v}$ for $\overline{\tau}_{(0,T),\neg(g,v)}$ and $\bar{\epsilon}_{1,T} - \bar{\epsilon}_{g,v} + M\lambda_{g,v}$ for $\overline{\tau}_{\neg(0,T),(g,v)}$ into the distribution of the ATT's upper bound in (39):

$$\begin{aligned}&p_{(0,T),(g,v)} \min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} + p_{(0,T),\neg(g,v)} \left(\bar{\epsilon}_{1,T} - \bar{\epsilon}_{0,T} + M\lambda_{g,v}\right)\\ &\quad + p_{\neg(0,T),(g,v)} \left(\bar{\epsilon}_{1,T} - \bar{\epsilon}_{g,v} + M\lambda_{g,v}\right) + p_{\neg(0,T),\neg(g,v)}\overline{\tau}_{\neg(0,T),\neg(g,v)}.\end{aligned} \tag{40}$$

Consequently, the expected squared error of the ATT's upper bound from the parallel trends imputation, $\mathrm{E}_\pi[(\bar{\tau} - \bar{\tau}_{0,T})^2]$, is

$$\begin{aligned}&p_{(0,T),(g,v)} \left(\min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} - \bar{\tau}_{0,T}\right)^2 + p_{(0,T),\neg(g,v)} \left(\bar{\epsilon}_{1,T} - \bar{\epsilon}_{0,T} + M\lambda_{g,v} - \bar{\tau}_{0,T}\right)^2\\ &\quad + p_{\neg(0,T),(g,v)} \left(\bar{\epsilon}_{1,T} - \bar{\epsilon}_{g,v} + M\lambda_{g,v} - \bar{\tau}_{0,T}\right)^2 + p_{\neg(0,T),\neg(g,v)} \left(\overline{\tau}_{\neg(0,T),\neg(g,v)} - \bar{\tau}_{0,T}\right)^2.\end{aligned} \tag{41}$$

Then, noting that

$$\begin{aligned}\bar{\epsilon}_{1,T} - \bar{\epsilon}_{0,T} + M\lambda_{g,v} - \bar{\tau}_{0,T} &= 0 \text{ and}\\ \bar{\epsilon}_{1,T} - \bar{\epsilon}_{g,v} + M\lambda_{g,v} - \bar{\tau}_{0,T} &= \bar{\epsilon}_{1,T} - \bar{\epsilon}_{g,v} + M\lambda_{g,v} - \left(\bar{\epsilon}_{1,T} - \bar{\epsilon}_{0,T} + M\lambda_{g,v}\right)\\ &= \bar{\epsilon}_{0,T} - \bar{\epsilon}_{g,v}\end{aligned}$$

yields

$$\begin{aligned}&p_{(0,T),(g,v)} \left(\min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} - \bar{\tau}_{0,T}\right)^2 + p_{\neg(0,T),(g,v)} \left(\bar{\epsilon}_{0,T} - \bar{\epsilon}_{g,v}\right)^2\\ &\quad + p_{\neg(0,T),\neg(g,v)} \left(\overline{\tau}_{\neg(0,T),\neg(g,v)} - \bar{\tau}_{0,T}\right)^2,\end{aligned}$$

which, since $(\bar{\epsilon}_{0,T} - \bar{\epsilon}_{g,v})^2 = \lambda_{g,v}^2$ by the definition of $\lambda_{g,v} \coloneqq |\bar{\epsilon}_{g,v}(0) - \bar{\epsilon}_{0,T}(0)|$ in (21), implies that $\mathrm{E}_\pi[(\bar{\tau} - \bar{\tau}_{0,T})^2]$ in (41) is

$$\begin{aligned}&p_{(0,T),(g,v)} \left(\min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} - \bar{\tau}_{0,T}\right)^2 + p_{\neg(0,T),(g,v)}\lambda_{g,v}^2\\ &\quad + p_{\neg(0,T),\neg(g,v)} \left(\overline{\tau}_{\neg(0,T),\neg(g,v)} - \bar{\tau}_{0,T}\right)^2.\end{aligned} \tag{42}$$

When $\min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} = \overline{\tau}_{0,T}$, the expected squared error in (42) is

$$p_{\neg(0,T),(g,v)}\lambda_{g,v}^2 + p_{\neg(0,T),\neg(g,v)} \left(\overline{\tau}_{\neg(0,T),\neg(g,v)} - \bar{\tau}_{0,T}\right)^2. \tag{43}$$

Alternatively, when $\min\{\overline{\tau}_{0,T},\, \overline{\tau}_{g,v}\} = \overline{\tau}_{g,v}$, the expected squared error in (42) is

$$\left(p_{(0,T),(g,v)} + p_{\neg(0,T),(g,v)}\right)\lambda_{g,v}^2 + p_{\neg(0,T),\neg(g,v)} \left(\overline{\tau}_{\neg(0,T),\neg(g,v)} - \bar{\tau}_{0,T}\right)^2. \tag{44}$$

Both (43) and (44) are strictly increasing in $\lambda_{g,v}$ whenever $p_{\neg(0,T),(g,v)} > 0$.

Since $|x| = \sqrt{x^2}$ for all $x \in \mathbb{R}$ and the square-root function is strictly increasing on $[0, \infty)$, strict increase of the expected squared error in $\lambda_{g,v}$ implies strict increase of the expected absolute error, $\mathrm{E}_\pi[|\bar{\tau} - \bar{\tau}_{0,T}|]$. The reasoning is analogous for the ATT's lower bound, $\underline{\tau}$, which completes the proof. ■

Proof of Proposition 2

Proof. It follows from the triangle inequality that

$$\begin{aligned}|\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,v}(0)| &= |(\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,T}(0)) + (\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{0,v}(0))| \\ &\leq |\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,T}(0)| + |\bar{\epsilon}_{0,v}(0) - \bar{\epsilon}_{0,T}(0)|.\end{aligned}$$

Then, since $|x| + |y| \leq 2\max\{|x|, |y|\}$ for all $x, y \in \mathbb{R}$, it follows that

$$|\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,T}(0)| + |\bar{\epsilon}_{0,v}(0) - \bar{\epsilon}_{0,T}(0)| \leq 2\max\{|\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,T}(0)|, |\bar{\epsilon}_{0,v}(0) - \bar{\epsilon}_{0,T}(0)|\}$$

and, from the definition of $\lambda_{g,v}$ for all $(g, v) \in \mathcal{G} \times \mathcal{V}$ in (21), that

$$|\bar{\epsilon}_{1,v}(0) - \bar{\epsilon}_{0,T}(0)| + |\bar{\epsilon}_{0,v}(0) - \bar{\epsilon}_{0,T}(0)| \leq 2\max\{\lambda_{1,v}, \lambda_{0,v}\},$$

which completes the proof. ■

Proof of Proposition 3

Proof. From (31), the worst-case regret of weights $\boldsymbol{w}$ is

$$R(\boldsymbol{w}) \coloneqq \sup_{\psi \in \Psi} \left\{ \psi - M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w_{g,v} \lambda_{g,v} \right\},$$

which, since $M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w_{g,v} \lambda_{g,v}$ does not depend on ψ, can be expressed as

$$R(\boldsymbol{w}) = \left(\sup_{\psi \in \Psi} \psi \right) - M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w_{g,v} \lambda_{g,v}.$$

By the definition of Ψ in (30), the supremum of ψ is attained at the largest feasible upper bound implied by the sensitivity model, which is

$$\sup_{\psi \in \Psi} \psi = \sup_{\boldsymbol{w'} \in \Delta_{\mathcal{G} \times \mathcal{V}}} M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w'_{g,v} \lambda_{g,v},$$

thereby implying that the worst-case regret is

$$R(\boldsymbol{w}) = \sup_{\boldsymbol{w'} \in \Delta_{\mathcal{G} \times \mathcal{V}}} M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w'_{g,v} \lambda_{g,v} - M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w_{g,v} \lambda_{g,v}. \tag{45}$$

Then, since the mapping from the weights $\boldsymbol{w'}$ to the value $M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w'_{g,v} \lambda_{g,v}$ is continuous in $\boldsymbol{w'}$ and the feasible set $\Delta_{\mathcal{G} \times \mathcal{V}}$ is a compact simplex, the extreme value theorem implies that the supremum is attained. That is, there exists some $\tilde{\boldsymbol{w}} \in \Delta_{\mathcal{G} \times \mathcal{V}}$ such that

$$\sup_{\boldsymbol{w'} \in \Delta_{\mathcal{G} \times \mathcal{V}}} M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} w'_{g,v} \lambda_{g,v} = M \sum_{(g,v) \in \mathcal{G} \times \mathcal{V}} \tilde{w}_{g,v} \lambda_{g,v}.$$

Because the supremum is attained at a weight vector in the simplex $\Delta_{\mathcal{G}\times\mathcal{V}}$, the first term in (45) can equivalently be written as the following constrained maximization problem:

$$\begin{aligned}\max_{\boldsymbol{w'}\in\mathbb{R}^{|\mathcal{G}\times\mathcal{V}|}} \quad & M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w'_{g,v}\lambda_{g,v} \\ \text{subject to} \quad & w'_{g,v} \geq 0 \quad \text{for all } (g,v)\in\mathcal{G}\times\mathcal{V}, \\ & \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w'_{g,v} = 1.\end{aligned} \tag{46}$$

Since the objective function in (46) is linear in $\boldsymbol{w'}$ and the feasible set is $\Delta_{\mathcal{G}\times\mathcal{V}}$, the optimal solution places all weight on the set $\mathcal{A} := \{(g,v)\in\mathcal{G}\times\mathcal{V} : \lambda_{g,v} = \max_{(g',v')\in\mathcal{G}\times\mathcal{V}} \lambda_{g',v'}\}$, i.e., on the group period pairs that attain the greatest discordance. Consequently, the worst-case regret is

$$R(\boldsymbol{w}) = M\left(\max_{(g,v)\in\mathcal{G}\times\mathcal{V}} \lambda_{g,v} - \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\lambda_{g,v}\right). \tag{47}$$

Because

$$\sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\lambda_{g,v} \leq \max_{(g,v)\in\mathcal{G}\times\mathcal{V}} \lambda_{g,v} \text{ for all } \boldsymbol{w}\in\Delta_{\mathcal{G}\times\mathcal{V}},$$

the worst-case regret in (47) is nonnegative and attains its minimum value of 0 whenever

$$\sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\lambda_{g,v} = \max_{(g,v)\in\mathcal{G}\times\mathcal{V}} \lambda_{g,v},$$

which holds for $\boldsymbol{w} = \boldsymbol{w}^{\max}$ in (26), thereby completing the proof. ■

Proposition 4 and Proof

Proposition 4. *For any $M \geq 0$ and $\boldsymbol{w}$ in the unit $|\mathcal{G}\times\mathcal{V}|$-simplex, denoted by $\Delta_{\mathcal{G}\times\mathcal{V}}$, the bound of the sensitivity model in* (23)*, i.e.,*

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| \leq M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}|\bar{\epsilon}_{g,v}(0) - \bar{\epsilon}_{0,T}(0)|, \tag{48}$$

is a sharp upper bound. That is, the bound in (48) *cannot be tightened further because there exists at least one configuration of $\bar{\epsilon}_{1,T}(0)$, $\bar{\epsilon}_{0,T}(0)$, and $\bar{\epsilon}_{g,v}(0)$ for all $(g,v)\in\mathcal{G}\times\mathcal{V}$ in which the inequality in* (48) *holds with equality.*

Proof. The proof proceeds by construction. First, let

$$\bar{\epsilon}_{1,T}(0) = \bar{\epsilon}_{0,T}(0) + M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\left(\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{g,v}(0)\right).$$

Then, note that

$$\begin{aligned}|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| &= |[\bar{\epsilon}_{0,T}(0) + M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}\left(\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{g,v}(0)\right)] - \bar{\epsilon}_{0,T}(0)| \\ &= |M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}(\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{g,v}(0))|,\end{aligned}$$

which, since $M \geq 0$, implies that

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| = M| \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}(\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{g,v}(0))|.$$

Then let $\bar{\epsilon}_{g,v}(0) \leq \bar{\epsilon}_{0,T}(0)$ for all $(g,v) \in \mathcal{G}\times\mathcal{V}$ so that each $\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{g,v}(0)$ is nonnegative. Consequently, it follows that

$$|\bar{\epsilon}_{1,T}(0) - \bar{\epsilon}_{0,T}(0)| = M \sum_{(g,v)\in\mathcal{G}\times\mathcal{V}} w_{g,v}|\bar{\epsilon}_{0,T}(0) - \bar{\epsilon}_{g,v}(0)|,$$

thereby establishing that the upper bound in (23) is sharp. ■

Acknowledgments

I thank Donald P. Green, Macartan Humphreys, Naoki Egami, Jake Bowers, Fredrik Sävje, Anna Wilke, Georgiy Syunyaev, John Marshall, Sanders Korenman, and Dahlia Remler for helpful comments. I also thank audiences at PolMeth XXXVII, the Columbia University Political Science Department's Seminar in Comparative Politics, and the Marxe School Research Seminar. I am grateful to Omer Rondel for valuable research assistance and to Laurence Wilse-Samson for graciously sharing the data used in this paper's empirical application. Any remaining errors are my own.

References

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, Princeton, NJ, 2008.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

William G Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

William G. Cochran. Observational studies. In Theodore Alfonso Bancroft, editor, *Statistical Papers in Honor of George W. Snedecor*, chapter 6, pages 77–90. Iowa State University Press, Ames, IA, 1972.

Peter L. Cohen, Matt A. Olson, and Colin B. Fogarty. Multivariate one-sided testing in matched observational studies as an adversarial game. *Biometrika*, 107(4):809–825, 2020.

Thomas D. Cook and Donald T. Campbell. *Quasi-experimentation: Design & Analysis Issues for Field Settings.* Houghton Mifflin Company, Boston, MA, 1979.

Jonathan Crush. "The long-averted clash": Farm labour competition in the south african countryside. *Canadian Journal of African Studies/La Revue canadienne des études africaines*, 27(3):404–423, 1993.

Cornelius W. De Kiewiet. *A History of South Africa: Social and Economic.* Clarendon Press, Oxford, UK, 1941.

Matei Demetrescu, Manuel Frondel, Lukas Tomberg, and Colin Vance. Fixed effects, lagged dependent variables, and bracketing: Cautionary remarks. *Political Analysis*, 33(4):378–392, 2025.

Peng Ding and Fan Li. A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, 27(4):605–615, 2019.

Taryn Dinkelman and Martine Mariotti. The long-run effects of labor migration on human capital formation in communities of origin. *American Economic Journal: Applied Economics*, 8(4):1–35, 2016.

Taryn Dinkelman, Grace Kumchulesi, and Martine Mariotti. Labor migration, capital accumulation, and the structure of rural labor markets. *The Review of Economics and Statistics*, 2024.

Saul Dubow. *Apartheid, 1948 – 1994.* Oxford Histories. Oxford University Press, Oxford, UK, 2014.

Naoki Egami and Soichiro Yamauchi. Using multiple pre-treatment periods to improve Difference-in-Differences and Staggered Adoption designs. *Political Analysis*, 31(2):195–212, 2023.

Charles H. Feinstein. *An Economic History of South Africa: Conquest, Discrimination, and Development.* Cambridge University Press, New York, NY, 2005.

Jonathan Guryan. Desegregation and black dropout rates. Technical Report NBER Working Paper No. 8345, National Bureau of Economic Research, Cambridge, MA, June 2001.

Susan Haack. Double-aspect foundherentism: A new theory of empirical justification. *Philosophy and Phenomenological Research*, 53(1):113–128, 1993.

Susan Haack. *Evidence and Inquiry: Towards Reconstruction in Epistemology.* Wiley, Hoboken, NJ, 1995.

Raiden Hasegawa and Dylan Small. Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics*, 73(4):1424–1432, 2017.

Miguel Á. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

Ariella Kahn-Lang and Kevin Lang. The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 38(3):613–620, 2020.

Luke J. Keele, Dylan S. Small, Jesse Y. Hsu, and Colin B. Fogarty. Patterns of effects and sensitivity analysis for differences-in-differences. Working Paper, February 2019.

Thomas Leavitt and Laura A Hatfield. Averaged prediction models (APM): Identifying causal effects in controlled pre-post settings with application to gun policy. *Annals of Applied Statistics*, 19(3):1826–1846, 2025.

Charles F. Manski and John V. Pepper. How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *The Review of Economics and Statistics*, 100(2):232–244, 2018.

Dan O'Meara. *Forty Lost Years: The Apartheid State and the Politics of the National Party, 1948 – 1994*. Ohio University Press, Athens, OH, 1996.

Ashesh Rambachan and Jonathan Roth. A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591, 2023.

Paul R. Rosenbaum. Cochran's causal crossword. *Observational Studies*, 1(1):205–211, 2015.

Paul R. Rosenbaum. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, Cambridge, MA, 2017.

Paul R. Rosenbaum and Abba M. Krieger. Sensitivity of two-sample permutation inferences in observational studies. *Journal of the American Statistical Association*, 85(410):493–498, 1990.

Jonathan Roth and Pedro H. C. Sant'Anna. When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747, 2023.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, Boston, MA, 2002.

Leonard M. Thompson. *A History of South Africa*. Yale University Press, New Haven, CT, 1990.

David Welsh. *The Rise and Fall of Apartheid*. Reconsiderations in Southern African History. University of Virginia Press, Charlottesville, VA, 2009.

Laurence Wilse-Samson. Structural change and democratization: Evidence from rural apartheid. Unpublished working paper, `https://www.columbia.edu/~lhw2110/wilse_samson_apartheid.pdf`, November 2013.

Ting Ye, Luke Keele, Raiden Hasegawa, and Dylan S Small. A negative correlation strategy for bracketing in difference-in-differences. *Journal of the American Statistical Association*, 119(547):2256–2268, 2024.